

Introduction to Statistics

INTRODUCTION TO STATISTICS

An Excel-Based Approach

VALERIE WATTS

Fanshawe College Pressbooks
London Ontario



Introduction to Statistics by Valerie Watts is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

CONTENTS

Acknowledgements	xi
About this Book	xii
Changes From Previous Version	xv

Part I. Sampling and Data

1.1 Introduction to Sampling and Data	3
1.2 Definitions of Statistics, Probability, and Key Terms	5
1.3 Sampling and Data	12
1.4 Frequency, Frequency Tables, and Levels of Measurement	19
1.5 Experimental Design and Ethics	27
1.6 Exercises	36
1.7 Answers to Selected Exercises	55

Part II. Descriptive Statistics

2.1 Introduction to Descriptive Statistics	63
2.2 Histograms, Frequency Polygons, and Time Series Graphs	65
2.3 Measures of Central Tendency	83
2.4 Skewness and the Mean, Median, and Mode	98
2.5 Measures of Location	103
2.6 Measures of Dispersion	120
2.7 Exercises	134

Part III. Probability

3.1 Introduction to Probability	169
3.2 The Terminology of Probability	171
3.3 Contingency Tables	178
3.4 The Complement Rule	185
3.5 The Addition Rule	189
3.6 Conditional Probability	197
3.7 Joint Probabilities	209
3.8 Exercises	225

Part IV. Discrete Random Variables

4.1 Introduction to Discrete Random Variables	239
4.2 Probability Distribution of a Discrete Random Variable	241
4.3 Expected Value and Standard Deviation for a Discrete Probability Distribution	246
4.4 The Binomial Distribution	259
4.5 The Poisson Distribution	274
4.6 Exercises	284

Part V. Continuous Random Variables and the Normal Distribution

5.1 Introduction to Continuous Random Variables	305
5.2 Probability Distribution of a Continuous Random Variable	308
5.3 The Normal Distribution	315
5.4 The Standard Normal Distribution	322
5.5 Calculating Probabilities for a Normal Distribution	332
5.6 Exercises	346

Part VI. The Central Limit Theorem and Sampling Distributions

6.1 Introduction to Sampling Distributions and the Central Limit Theorem	361
6.2 Sampling Distribution of the Sample Mean	363
6.3 Sampling Distribution of the Sample Proportion	375
6.4 Exercises	388

Part VII. Confidence Intervals for Single Population Parameters

7.1 Introduction to Confidence Intervals	395
7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation	398
7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation	416
7.4 Confidence Intervals for a Population Proportion	428
7.5 Calculating the Sample Size for a Confidence Interval	439
7.6 Exercises	450

Part VIII. Hypothesis Tests for Single Population Parameters

8.1 Introduction to Hypothesis Testing	461
8.2 Null and Alternative Hypotheses	463
8.3 Outcomes and the Type I and Type II Errors	470
8.4 Distributions Required for a Hypothesis Test	477
8.5 Rare Events, the Sample, Decision, and Conclusion	480
8.6 Hypothesis Tests for a Population Mean with Known Population Standard Deviation	488
8.7 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation	502
8.8 Hypothesis Tests for a Population Proportion	514

8.9 Exercises	534
---------------	-----

Part IX. Statistical Inference for Two Populations

9.1 Introduction to Statistical Inference with Two Populations	547
9.2 Statistical Inference for Two Population Means with Known Population Standard Deviations	549
9.3 Statistical Inference for Two Population Means with Unknown Population Standard Deviations	568
9.4 Statistical Inference for Matched Samples	590
9.5 Statistical Inference for Two Population Proportions	611
9.6 Exercises	631

Part X. Statistical Inferences Using the Chi-Square Distribution

10.1 Introduction to Statistical Inferences Using the Chi-Square Distribution	645
10.2 The Chi Square Distribution	647
10.3 Statistical Inference for a Single Population Variance	653
10.4 The Goodness-of-Fit Test	667
10.5 The Test of Independence	684
10.6 Exercises	700

Part XI. Statistical Inference Using the F-Distribution

11.1 Introduction to Statistical Inferences Using the F-Distribution	715
11.2 The F-Distribution	717
11.3 Statistical Inference for Two Population Variances	723
11.4 One-Way ANOVA and Hypothesis Tests for Three or More Population Means	738

11.5 Exercises	757
----------------	-----

Part XII. Simple Linear Regression and Correlation

12.1 Introduction to Linear Regression and Correlation	769
12.2 Linear Equations	771
12.3 Scatter Diagrams	776
12.4 Correlation	785
12.5 The Regression Equation	794
12.6 Coefficient of Determination	805
12.7 Standard Error of the Estimate	809
12.8 Exercises	815

Part XIII. Multiple Regression

13.1 Introduction to Multiple Regression	833
13.2 Multiple Regression	834
13.3 Standard Error of the Estimate	847
13.4 Coefficient of Multiple Determination	853
13.5 Testing the Significance of the Overall Model	863
13.6 Testing the Regression Coefficients	873
13.7 Multicollinearity	890
13.8 Exercises	892
13.9 Answers to Select Exercises	899
References	901
Versioning History	916

ACKNOWLEDGEMENTS

This open textbook has been adapted by Dr. Valerie Watts in partnership with the OER Design Studio and the Library Learning Commons at Fanshawe College in London, Ontario

This work is part of the FanshaweOpen learning initiative and is made available through a Creative Commons Attribution-ShareAlike 4.0 International License unless otherwise noted.



We would like to acknowledge and thank the following authors/entities who have graciously made their work available for the remixing, reusing, and adapting of this text:

- Introductory Business Statistics by Alexander Holmes, Barbara Illowsky, Susan Dean, and OpenStax is licensed under a Creative Commons Attribution 4.0 License.
- Introductory Statistics by Barbara Illowsky, Susan Dean, and OpenStax is licensed under a Creative Commons Attribution 4.0 License.

Collaborators

This project was a collaboration between the author and the team in the OER Design Studio at Fanshawe. The following staff and students were involved in the creation of this project:

- Melanie Mitchell Sparkes, *Instructional Designer*
- Shauna Roch, *Project Lead*
- Jenn Ayers, *Project Coordinator*
- Alyssa Giles, *Graphic Design*
- Wilson Poulter, *Copyright*

ABOUT THIS BOOK

Introduction to Statistics: An Excel-Based Approach introduces students to the concepts and applications of statistics, with a focus on using Excel to perform statistical calculations. The book is written at an introductory level, designed for students in fields other than mathematics or engineering, but who require a fundamental understanding of statistics. The text emphasizes understanding and application of statistical tools over theory, but some knowledge of algebra is required.

Although the text focuses on concepts and applications, every effort has been made to provide both information essential to understanding a topic and sound methodological development. Generally accepted terminology and notation for each topic is used throughout without becoming overly focused on technical details.

In place of manual calculations and the use of probability distribution tables, the text utilizes Excel in the application of statistical analysis. Because of the prevalence and use of Excel in a wide-range of fields, it is important for students to understand how to leverage Excel's statistical capabilities. Throughout the text, information is provided on the appropriate Excel function required for a calculation and the solutions to examples illustrate how to use the corresponding Excel function to solve problems.

The text is organized into thirteen chapters, and then divided into subchapters by concept or topic. The chapters are as follows:

Chapter 1: Sampling and Data	Chapter 1 covers key definitions, terms, and terminology used in statistics, as well exploring different sampling methods, types of data, and level of measurement.
Chapter 2: Descriptive Statistics	Chapter 2 examines the different descriptive statistics, both graphical and numerical, required to organize, summarize, and describe data.
Chapter 3: Probability	Chapter 3 introduces the concept of probability, probability terminology, different approaches to probability, and various probability rules.
Chapter 4: Discrete Random Variables	Chapter 4 explores discrete random variables and their probability distributions, the mean and standard deviation of a discrete random variable, and examines the binomial and Poisson distributions.
Chapter 5: Continuous Random Variables and the Normal Distribution	Chapter 5 covers continuous random variables, focusing on the normal distribution and probability problems associated with the normal distribution.
Chapter 6: The Central Limit Theorem and Sampling Distributions	Chapter 6 examines the sampling distributions of sample mean and the sampling distribution of the sample proportion.
Chapter 7: Confidence Intervals for Single Population Parameters	Chapter 7 explores the construction, use and interpretation of confidence intervals to estimate a population mean or a population proportion, as well as determining the sample size necessary for the required accuracy of a confidence interval.
Chapter 8: Hypothesis Tests for Single Population Parameters	Chapter 8 introduces the formal hypothesis testing procedure, focusing on conducting and drawing a conclusion from a hypothesis test on a population mean or a population proportion.
Chapter 9: Statistical Inference for Two Populations	Chapter 9 extends confidence intervals and hypothesis testing to the difference between two population means or the difference between two population proportions.
Chapter 10: Statistical Inference Using the χ^2-Distribution	Chapter 10 covers the use of the χ^2 -distribution in statistical inference, including confidence intervals and hypothesis testing for a population variance, the goodness-of-fit test, and the test of independence.
Chapter 11: Statistical Inference Using the F-Distribution	Chapter 11 examines the use of the F -distribution in statistical inference, including confidence intervals and hypothesis testing for the ratio of two population variances and the one-way ANOVA test on the equality of three or more population means.
Chapter 12: Simple Linear Regression and Correlation	Chapter 12 explores the linear relationship between two variables through the simple linear regression model, including methods to assess the validity of the model.
Chapter 13: Multiple Regression	Chapter 13 extends the linear regression model to include more than one independent variable, including methods to assess the validity of the model.

For the Student

Each sub-chapter in this text begins with a list of relevant learning objectives and concludes with a concept review that highlights the key topics in the sub-chapter. Where appropriate, videos are included to review, enhance, and extend the material covered in the text. At the end of each chapter, a series of exercises are included to check retention and assess understanding.

Accessibility Statement

We are actively committed to increasing the accessibility and usability of the textbooks we produce. Every attempt has been made to make this OER accessible to all learners and is compatible with assistive and adaptive technologies. We have attempted to provide closed captions, alternative text, or multiple formats for on-screen and off-line access.

The web version of this resource has been designed to meet Web Content Accessibility Guidelines 2.0, level AA. In addition, it follows all guidelines in Appendix A: Checklist for Accessibility of the *Accessibility Toolkit – 2nd Edition*.

In addition to the web version, additional files are available in a number of file formats including PDF, EPUB (for eReaders), and MOBI (for Kindles).

If you are having problems accessing this resource, please contact us at oyer@fanshawec.ca.

Please include the following information:

- The location of the problem by providing a web address or page description
- A description of the problem
- The computer, software, browser, and any assistive technology you are using that can help us diagnose and solve your issue (e.g., Windows 10, Google Chrome (Version 65.0.3325.181), NVDA screen reader)

Feedback

To provide feedback on this text please contact oyer@fanshawec.ca.

CHANGES FROM PREVIOUS VERSION

This book is an adaptation of Introductory Statistics by Open Stax, licensed under a Creative Commons Attribution 4.0 license. Additional content was incorporated from Introduction to Business Statistics by Open Stax, licensed under a Create Commons Attribution 4.0 license, and other open materials.

The following is a summary of changes that were made in this version:

<p>Overall</p>	<ul style="list-style-type: none"> • Book reformatted for Pressbooks. • Chapters were moved, combined, or separated. • Sub-chapters were moved, combined, or separated, with some new sub-chapters added. • Latex code updated throughout. • Moved learning objectives from beginning of each chapter to appropriate sub-chapter • Revised and updated learning objectives at the start of every sub-chapter. • Revised and updated chapter and sub-chapter titles throughout. • Revised, rewrote, and updated content throughout. • Added instructions for using Excel throughout. • Replaced all by-hand and graphing calculator solutions with Excel calculations and solutions. • Added additional images and videos. • End matter from each chapter was removed, except for exercises sub-chapter. • Reformatted and renumbered exercises in exercise sub-chapters, with unnecessary exercises removed. • Removed appendices. • Removed collaborative exercises. • Removed statistics lab activities from the end of each chapter. • Added chapter on multiple regression.
<p>Chapter 1: Sampling and Data</p>	<ul style="list-style-type: none"> • Added definition of cumulative frequency and added cumulative frequency to example.
<p>Chapter 2: Descriptive Statistics</p>	<ul style="list-style-type: none"> • Removed sub-chapters on stem-and-leaf plots and box plots. • Changed order of sub-chapters. • Removed some examples and exercises. • Added content on range. • Removed formulas for finding percentiles manually. • Changed the interpretation of percentiles and quartiles to “strictly less than”. • Added an example to calculate multiple modes in Excel. • Removed the content on the law of large numbers and the mean. • Removed content on grouped standard deviation. • Removed content on sampling variability of a statistic. • Removed the explanation of the manual calculation of the standard deviation.

<p>Chapter 3: Probability</p>	<ul style="list-style-type: none"> • Replaced sub-chapters on independent and mutually exclusive events and two basic rules of probability with separate sub-chapters for the complement rule, the addition rule, conditional probability, and joint probabilities. • Changed order of the sub-chapters. • Added examples about sample space, outcomes, events, and probability. • Added content about different approaches to probability. • Moved content about complement, or, conditional, and joint probabilities to separate sub-chapters. • Moved content on mutually exclusive and independent events to sub-chapters on the addition rule and conditional probability, respectively. • Added image illustrating the complement. • Added examples on using the complement. • Added content on repeated trial experiments. • Removed sub-chapter on trees and Venn diagrams.
<p>Chapter 4: Discrete Random Variables</p>	<ul style="list-style-type: none"> • Removed sub-chapters on geometric and hypergeometric distributions. • Removed some examples and exercises. • Updated the definition of binomial experiment. • Removed the probability distribution notation. • Added content about the mean and standard deviation of a Poisson distribution. • Added additional examples and exercises for calculating binomial and Poisson probabilities.
<p>Chapter 5: Continuous Random Variables and the Normal Distribution</p>	<ul style="list-style-type: none"> • Combined chapters on Continuous Random Variables and the Normal Distribution. • Removed sub-chapters on the uniform and exponential distributions. • Removed the probability distribution notation. • Removed references to probability density function. • Moved the content on the Empirical rule to a different sub-chapter. • Removed references to the standard normal distribution table. • Removed some examples and exercises. • Added additional examples and exercises for calculating normal probabilities. • Added images of the normal distribution.
<p>Chapter 6: The Central Limit Theorem and Sampling Distributions</p>	<ul style="list-style-type: none"> • Removed sub-chapter on the central limit theorem for sums. • Added new sub-chapter on the sampling distribution of the sample proportion. • Moved content about distribution of the sample proportion from another chapter. • Added content about binomial distribution for the distribution of the sample proportion. • Updated notation for mean and standard deviation of the sample mean. • Rewrote the description of the sampling distribution of the sample means and the central limit theorem. • Removed some examples and exercises. • Added additional examples and exercises for the sampling distributions. • Removed sub-chapter on using the central limit theorem. • Changed the notation for sample proportion.

<p>Chapter 7: Confidence Intervals for Single Population Parameters</p>	<ul style="list-style-type: none"> • Added images of normal and t-distributions. • Updated some of the notation. • Added separate sub-chapter on finding the sample size and moved relevant content about finding the sample size from other sub-chapters. • Removed content on working backwards to find the error or sample mean. • Removed some examples and exercises. • Added additional examples and exercises. • Removed references to the t-distribution table. • Removed the probability distribution notation. • Moved content about the distribution of the sample proportion to another chapter. • Removed content on “plus four” confidence interval for proportions.
<p>Chapter 8: Hypothesis Tests for Single Population Parameters</p>	<ul style="list-style-type: none"> • Changed the null hypothesis to an :equal to”. • Replaced sub-chapter on additional information and hypothesis test examples with three new sub-chapters on hypothesis tests for population means and population proportions. • Moved content on hypothesis tests for means and hypothesis test for proportions to relevant new sub-chapters. • Removed some examples and exercises. • Added additional examples and exercises. • Added steps to conduct a hypothesis test. • Added images of normal and t-distributions.
<p>Chapter 9: Statistical Inference for Two Population Parameters</p>	<ul style="list-style-type: none"> • Reorganized the content of the chapter. • Added content on the distribution of the difference of the sample means and the distribution of the difference of the sample proportions. • Added steps to conduct the hypothesis tests. • Moved content about independent samples to a different sub-chapter. • Added definition of matched samples. • Added content on confidence intervals for the difference in two population parameters. • Added images of normal and t-distributions. • Removed content on Cohen’s Effect size. • Removed some examples and exercises. • Added additional examples and exercises.

<p>Chapter 10: Statistical Inferences using the χ^2 -Distribution</p>	<ul style="list-style-type: none"> • Reorganized the content of the chapter. • Removed sub-chapters on test of homogeneity and comparison of the χ^2 tests. • Added images of χ^2-distributions. • Removed some examples and exercises. • Added additional examples and exercises. • Added content on the distribution of the sample variance. • Added steps to conduct the hypothesis tests. • Added content on confidence intervals for population variance.
<p>Chapter 11: Statistical Inferences using the F -Distribution</p>	<ul style="list-style-type: none"> • Reorganized the content of the chapter. • Removed sub-chapter on the F-distribution and the F-ratio and moved content to sub-chapter on one-way ANOVA. • Added images of F-distributions. • Removed some examples and exercises. • Added additional examples and exercises. • Added content on the distribution of the ratio of sample variances. • Added content on confidence intervals for the ratio of population variances. • Added steps to conduct the hypothesis tests.
<p>Chapter 12: Simple Linear Regression and Correlation</p>	<ul style="list-style-type: none"> • Removed sub-chapter on testing the significance of the correlation coefficient, prediction, and outliers. • Reorganized the content of the chapter. • Added sub-chapters on correlation, coefficient of determination, and standard error of the estimate. • Removed some examples and exercises. • Added additional examples and exercises. • Added content on dependent and independent variables. • Moved content on correlation into separate sub-chapter. • Simplified discussion about the least squares method. • Moved content on predictions into sub-chapter on the regression equation. • Moved content on coefficient of determination into separate sub-chapter. • Added content on standard error of the estimate.
<p>Chapter 13: Multiple Regression</p>	<ul style="list-style-type: none"> • Created this new chapter.

PART I

SAMPLING AND DATA

Chapter Outline

- 1.1 Introduction to Sampling and Data
- 1.2 Definitions of Statistics, Probability, and Key Terms
- 1.3 Sampling and Data
- 1.4 Frequency, Frequency Tables, and Levels of Measurement
- 1.5 Experimental Design and Ethics
- 1.6 Exercises
- 1.7 Answers to Selected Exercises

1.1 INTRODUCTION TO SAMPLING AND DATA



We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. Photo by David Sim, CC BY 4.0.

You are probably asking yourself the question, “When and where will I use statistics?” If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate, just to mention a few. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or “fact.” Statistical methods can help you make the “best educated guess.”

Because you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what “good” data can be distinguished from “bad.”

Attribution

“Chapter 1 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.2 DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

LEARNING OBJECTIVES

- Recognize and differentiate between key terms used in statistics.

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. In this course, we will learn how to organize and summarize data. The organization and summation of data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After we have studied probability and probability distributions, we will use formal methods for drawing conclusions from “good” data. These formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of the data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (or likelihood) of an event occurring. For example, if we toss a **fair** coin four times, the outcomes may

not be two heads and two tails. However, if we toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is 50%. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads, or 49.8% heads. This is very close to the expected probability of 50%.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in a particular course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide whether or not to buy a lottery ticket. In the study of statistics, we use the power of mathematics through probability calculations to analyze and interpret the data.

Key Terms

In statistics, we generally want to study a population. A **population** is a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll typically sample between 1,000 and 2,000 people. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can actually contains 16 ounces of a carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Because we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The

sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A variable, notated by capital letters such as X and Y , is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. Numerical variables take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. In this case, Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense). **Data** are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE

The words **mean** and **average** are often used interchangeably. The substitution of one word for the other is common practice. The technical term for mean is “arithmetic mean,” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

EXAMPLE

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first-year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution:

- The **population** is all first year students attending ABC College this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).
- The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term (the population mean).
- The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample (the sample mean).
- The **variable** could be the amount of money spent (excluding books) by one first year student. Let X be the amount of money spent (excluding books) by one first year student attending ABC College.
- The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

TRY IT

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy.

We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Click to see Solution

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X be the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

TRY IT

Determine what the key terms refer to in the following study. A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

- | | | |
|---------------------|--------------------|-------------------|
| 1. Population _____ | 3. Parameter _____ | 5. Variable _____ |
| 2. Statistic _____ | 4. Sample _____ | 6. Data _____ |

- a. all students who attended the college last year.
- b. the cumulative GPA of one student who graduated from the college last year.
- c. 3.65, 2.80, 1.50, 3.90.
- d. a group of students who graduated from the college last year, randomly selected.
- e. the average cumulative GPA of students who graduated from the college last year.

- f. all students who graduated from the college last year.
- g. the average cumulative GPA of students in the study who graduated from the college last year.

Click to see Solution

- | | | |
|------|------|------|
| 1. f | 3. e | 5. b |
| 2. g | 4. d | 6. c |

EXAMPLE

Determine what the key terms refer to in the following study. As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which cars crashed	56 kilometers/hour
Location of “drive” (i.e., dummies)	Front seat

Cars with dummies in the front seats were crashed into a wall at a speed of 56 kilometers per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.
- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

- The **data** are either: yes, had head injury, or no, did not.

EXAMPLE

Determine what the key terms refer to in the following study. An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution:

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.
- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Attribution

“1.1 Definitions of Statistics, Probability, and Key Terms“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.3 SAMPLING AND DATA

LEARNING OBJECTIVES

- Identify data as qualitative or quantitative.
- Apply various types of sampling methods to data collection.

Data may come from a population or from a sample. Generally, small letters like x or y are used to represent data values. Most data can be put into the one of two categories: qualitative or quantitative.

Qualitative data are the result of categorizing or describing attributes of a population. Qualitative data are also called **categorical data**. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, the number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are **quantitative continuous data**, assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π , $\frac{3\pi}{4}$ and so on. If you and your friends carry backpacks with books in them to

school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

EXAMPLE

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are quantitative discrete data.

TRY IT

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Click to see Solution

- Quantitative discrete data.

EXAMPLE

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, and 4.3. Weights are quantitative continuous data.

TRY IT

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Click to see Solution

- Quantitative continuous data.

Sampling

Gathering information about an entire population often costs too much, is too time consuming, or is virtually impossible. Instead, we use a sample of the population. In order to get accurate conclusions about the population from the sample, a **sample should have the same characteristics as the population it represents**. Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling,

each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in the following table.

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Ludquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360 0.99832 0.14669 0.51470 0.40581 0.73381 0.04399

Lisa reads two-digit groups from these random numbers until she has chosen three class members (that is, she reads 0.94360 as groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However, the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two

digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.

To choose a **stratified sample**, divide the population into groups called **strata**, and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups), and then randomly select some of the clusters. All the members from the selected clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n -th piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population from 1 to 20,000, and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is **non-random** is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. Such a sample is not random because only those customers in the store on that particular day have the opportunity to be in the sample. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased because they may

favor a certain group. It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**, where a member of the population may only be chosen once. Surveys are typically done without replacement. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Consequently, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose we want to pick a sample of 1,000 randomly for a survey. For any particular sample of 1,000, if we are sampling **with replacement**:

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If we are sampling **without replacement**:

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Comparing the fractions $\frac{999}{10,000}$ and $\frac{999}{9,999}$ to four decimal places, these numbers are equivalent. So we can see that the chance of selecting a small sample from a large population is basically the same, whether or not the sampling is done with replacement.

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and we are sampling **with replacement for any particular sample**, then the chance of picking the first person is 10 out of 25, and the chance of picking a **different** second person is 9 out of 25 (we replace the first person). If we sample without replacement, then the chance of picking the first person is still 10 out of 25 but the chance of picking the second person (who is different) is 9 out of 24. Comparing the fractions $\frac{9}{25} = 0.36$ and $\frac{9}{24} = 0.3750$, these numbers are not equivalent.

When we analyze data, it is important to be aware of **sampling errors** and **non-sampling errors**. The actual process of sampling causes sampling error, which is the difference between the

actual population parameter and the corresponding sample statistic. For example, the sample may not be large enough. Factors not related to the sampling process cause **non-sampling errors**. For example, a defective counting device can cause a non-sampling error. In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=21#oembed-1>

Watch this video: Statistics: Sources of Bias by Mathispower4u [4:43] (transcript available).

Attribution

“1.2 Data, Sampling, and Variation in Data and Sampling“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.4 FREQUENCY, FREQUENCY TABLES, AND LEVELS OF MEASUREMENT

LEARNING OBJECTIVES

- Classify data by level of measurement.
- Create and interpret frequency tables.

Once we have a set of data, we need to organize it so that we can analyze how frequently each datum occurs in the set. However, when calculating the frequency, we may need to round our answers so that they are as precise as possible.

A simple way to round off answers is to carry the final answer to one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth because the data are whole numbers. Most answers will be rounded off in this manner.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be applied to every set of data. In addition to being classified as quantitative or qualitative, data is classified into four levels of measurement. They are (from lowest to highest level):

Qualitative Data		Quantitative Data	
Nominal Scale Level	Ordinal Scale Level	Interval Scale Level	Ratio Scale Level

Data that is measured using a **nominal scale** is data that can be placed into categories. Colors, names, labels, favorite foods, and yes/no survey responses are examples of nominal level data. Nominal scale data are not ordered, which means the categories of the data are not ordered. For example, trying to “order” people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful. Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung, and Apple. This is just a list of different brand names, and there is no agreed upon order for the categories. Some people may prefer Apple but that is a matter of opinion. Because nominal data consists of categories, nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data in that the data can be placed into categories, but there is a big difference. The categories of ordinal scale data can be ordered or ranked. An example of ordinal scale data is a list of the top five national parks in the United States because the parks can be ranked from one to five. Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. In ordinal scale data, the differences between two pieces of data cannot be measured or calculated. Similar to nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using an **interval scale** is similar to ordinal level data because it has a definite ordering. However, the differences between interval scale data can be measured or calculated, but the data does not have a starting point. Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements (Celsius and Fahrenheit), 40° is equal to 100° minus 60° . The differences in temperature can be measured and make sense. But there is no starting point to the temperature scales because 0° is not the absolute lowest temperature. Temperatures like -10°F and -15°C exist, and are colder than 0° . Interval level data can be used in calculations, but ratios do not make sense and cannot be done. For example, 80°C is not four times as hot as 20°C (nor is 80°F four times as hot as 20°F). So there is no meaning to the ratio of 80 to 20 (or four to one) in either temperature scale. In general, ratios have no meaning in interval scale data.

Data that is measured using the **ratio scale** takes care of the ratio problem, and gives us the most information. Ratio scale data is like interval scale data, but it has a starting point to the scale (a 0 point) and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The data can be put in order from lowest to

highest: 20, 68, 80, 92. The differences between the data have meaning: 92 minus 68 is 24. Ratios can be calculated: 80 is four times 20. The smallest possible score is 0.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=23#oembed-3>

Watch this video: Nominal, ordinal, interval and ratio data: How to Remember the differences by NurseKillam [11:03]
(transcript available)

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are recorded in the table below:

5	6	3	3	2	4	7	5	2	3
5	6	5	4	4	3	5	2	5	3

The following table lists the different data values in ascending order and their frequencies.

Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

A **frequency** is the number of times a value of the data occurs. According to the table, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column is 20, which is the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies divide each frequency by the total number of students in the sample—in this case 20. Relative frequencies can be written as fractions, percents, or decimals. The sum of the values in the relative frequency column is 1 or 100%.

Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20} = 0.15$
3	5	$\frac{5}{20} = 0.25$
4	3	$\frac{3}{20} = 0.15$
5	6	$\frac{6}{20} = 0.30$
6	2	$\frac{2}{20} = 0.10$
7	1	$\frac{1}{20} = 0.05$

Cumulative frequency is the accumulation of the previous frequencies. To find the cumulative frequencies, add all of the previous frequencies to the frequency for the current row, as shown in the table below. The last entry of the cumulative frequency column is the number of observations in the data.

Frequency Table of Student Work Hours with Relative and Cumulative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY
2	3	0.15	3
3	5	0.25	$3 + 5 = 8$
4	3	0.15	$8 + 3 = 11$
5	6	0.30	$11 + 6 = 17$
6	2	0.10	$17 + 2 = 19$
7	1	0.05	$19 + 1 = 20$

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the table below. The last entry of the cumulative relative frequency column is 1 or 100%, indicating that 100% of the data has been accumulated.

Frequency Table of Student Work Hours with Relative, Cumulative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	0.15	3	0.15
3	5	0.25	8	$0.15 + 0.25 = 0.40$
4	3	0.15	11	$0.40 + 0.15 = 0.55$
5	6	0.30	17	$0.55 + 0.30 = 0.85$
6	2	0.10	19	$0.85 + 0.10 = 0.95$
7	1	0.05	20	$0.95 + 0.05 = 1.00$

NOTE

Because of rounding of the relative frequencies, the relative frequency column may not always

sum to 1 or 100%, and the last entry in the cumulative relative frequency column may not be 1 or 100%. However, they each should be close to 1 or 100%. If all of the decimals are kept in the calculations, the relative frequency column will sum to 1 or 100% and the last cumulative relative frequency will be 1 or 100%.

CREATING A FREQUENCY DISTRIBUTION IN EXCEL

In order to create a frequency distribution and its corresponding histogram in Excel, we need to use the Analysis ToolPak. Follow these instructions to install the Analysis ToolPak add-in in Excel.

1. Enter the data into an Excel worksheet.
2. Determine the classes for the frequency distribution. Using these classes, create a **Bin** column that contains the **upper limit** for each class.
3. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
4. In the **Data Analysis** window, select **Histogram**. Click **OK**.
5. In the **Input** range, enter the cell range for the data.
6. In the **Bin** range, enter the cell range for the **Bin** column.
7. Select the location where you want the output to appear.
8. Select **Chart Output** to produce the corresponding histogram for the frequency distribution.
9. Click **OK**.

This website provides additional information on using Excel to create a frequency distribution.





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=23#oembed-1>

Watch this video: Frequency Distributions by Joshua Emmanuel [8:40] (transcript available).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=23#oembed-2>

Watch this video: How to Construct a Histogram in Excel using built-in Data Analysis by Joshua Emmanuel [1:58] (transcript available).

Concept Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off final answers to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

There are four levels of measurement for data:

- **Nominal scale level:** the data are categories, but the data cannot be ordered or used in calculations
- **Ordinal scale level:** the data are categories and the data can be ordered, but the differences cannot be measured.
- **Interval scale level:** the data have definite order or rank, but no starting point. The differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** the data have a definite order or rank with a starting point. The

differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on your block own two pets? Frequency, relative frequency, cumulative frequency, and cumulative relative frequency are measures that answer questions like these.

Attribution

“1.3 Frequency, Frequency Tables, and Levels of Measurement“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.5 EXPERIMENTAL DESIGN AND ETHICS

LEARNING OBJECTIVES

- Understand aspects of experimental design.
- Apply ethical behaviour in statistical analysis.

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=25#oembed-1>

Watch this video: Observational Studies and Experiments by ProfessorMcComb [3:05] (transcript available).

Suppose you want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? No, it does not. There are many differences between the two groups compared, in addition to vitamin E consumption. People who take vitamin E often take other steps to improve their health, such as exercise, diet, other vitamin supplements, or choosing not to smoke. Any one of these factors could be influencing a person's health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is the cause of a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design their experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Therefore, different outcomes measured in the response variable must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment (McClung et al., 2007).

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, they do not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

EXAMPLE

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Solution:

- The ***population*** is men aged 50 to 84.
- The ***sample*** is the 400 men who participated.
- The ***experimental units*** are the individual men in the study.
- The ***explanatory variable*** is the oral medication.
- The ***treatments*** are aspirin and a placebo.
- The ***response variable*** is whether a subject had a heart attack.

EXAMPLE

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first

three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.
2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

Solution:

1. The explanatory variable is scent and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. However, researchers timing the mazes can be blinded. The researcher who is observing a subject will not know which mask is being worn.

EXAMPLE

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution:

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

TRY IT

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

1. Describe the explanatory and response variables in the study.
2. What are the treatments?
3. What should you consider when selecting participants?
4. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
5. Identify any lurking variables that could interfere with this study.
6. How can blinding be used in this study?

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world’s top journals including *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Recently, an extensive investigation involving three universities where Stapel worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. “It was a quest for

aesthetics, for beauty—instead of the truth,” he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high (Levelt et al., 2012).

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations;
- altering data in existing datasets;
- changing measuring instruments without reporting the change; and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel’s fraud states that, “statistical flaws frequently revealed a lack of familiarity with elementary statistics”(n.a, 2013). Many of Stapel’s co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don’t want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

EXAMPLE

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

1. She selects a block where she is comfortable walking because she knows many of the people living on the street.
2. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
3. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution:

1. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher

needs to select areas in the community at random.

2. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
3. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

TRY IT

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

1. The survey is commissioned by the seller of a popular brand of apple juice.
2. There are only two types of juice included in the study: apple juice and cranberry juice.
3. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
4. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

Concept Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control

group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule” (Gelman, 2013). Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Attribution

“1.4 Experimental Design and Ethics“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.6 EXERCISES

1. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B: 3 14 11 5 16 17 28 41 31 18 14 14 26 25 121 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Determine what the key terms refer to in the example for Researcher A.

- a. population
 - b. sample
 - c. parameter
 - d. statistic
 - e. variable
2. For each of the following eight exercises, identify: the population, the sample, the parameter, the statistic, the variable, and the data. Give examples where appropriate.
- a. A fitness center is interested in the mean amount of time a client exercises in the center each week.
 - b. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.
 - c. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.
 - d. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.
 - e. A politician is interested in the proportion of voters in his district who think he is doing a good job.
 - f. A marriage counselor is interested in the proportion of clients she counsels who stay married.

- g. Political pollsters may be interested in the proportion of people who will vote for a particular cause.
 - h. A marketing company is interested in the proportion of people who will buy a particular product.
3. Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.
- a. What is the population she is interested in?
 - b. Consider the following: X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of a:
 - c. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:
4. "Number of times per week" is what type of data?
5. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.
- a. The sampling method was
 - b. "Duration (amount of time)" is what type of data?
 - c. The colors of the houses around the park are what kind of data?
 - d. The population is _____
6. The table contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012. Use the table to answer the following questions.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

- What is the proportion of deaths between 2007 and 2012?
 - What percent of deaths occurred before 2001?
 - What is the percent of deaths that occurred in 2003 or after 2010?
 - What is the fraction of deaths that happened before 2012?
 - What kind of data is the number of deaths?
 - Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
 - What contributed to the large number of deaths in 2010? In 2004? Explain.
7. For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).
- A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.
 - A market researcher polls every tenth person who walks into a store.
 - The first 50 people who walk into a sporting event are polled on their television preferences.

d. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

8. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

a. Complete the tables using the data provided:

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0.5–6.5				
6.5–12.5				
12.5–18.5				
18.5–24.5				
24.5–30.5				
30.5–36.5				
36.5–42.5				
42.5–48.5				

Researcher B

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0.5– 6.5				
6.5– 12.5				
12.5– 18.5				
18.5– 24.5				
24.5– 30.5				
30.5– 36.5				
36.5 – 45.5				

- b. Determine what the key term data refers to in the above example for Researcher A.
 - c. List two reasons why the data may differ.
 - d. Can you tell if one researcher is correct and the other one is incorrect? Why?
 - e. Would you expect the data to be identical? Why or why not?
 - f. How might the researchers gather random data?
 - g. Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?
 - h. Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?
9. Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher B

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

- Give a reason why the data may differ.
 - Would the sample size be large enough if the population is the students in the school?
 - Would the sample size be large enough if the population is school-aged children and young adults in the United States?
 - Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?
 - As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?
10. A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed

200 stroke patients recovering over a period of several weeks. The first study collected the data in the table. The second study collected the data in the table.

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

- a. Given what you know, which study is correct?
 - b. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?
 - c. Both groups that performed the study concluded that the software works. Is this accurate?
 - d. The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?
 - e. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from 44?
11. Is a sample size of 1,000 a reliable measure for a population of 5,000?
12. Is a sample of 500 volunteers a reliable measure for a population of 2,500?
13. A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y ?" Is this a fair question?
14. Is a sample size of two representative of a population of five?
15. Is it possible for two experiments to be well run with similar sample sizes to get different data?
16. For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.
- a. number of tickets sold to a concert
 - b. percent of body fat
 - c. favorite baseball team
 - d. time in line to buy groceries
 - e. number of students enrolled at Evergreen Valley College
 - f. most-watched television show

- g. brand of toothpaste
 - h. distance to the closest movie theatre
 - i. age of executives in Fortune 500 companies
 - j. number of competing computer spreadsheet software packages
17. A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.
- a. “Number of times per week” is what type of data?
 - b. “Duration (amount of time)” is what type of data?
18. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
- a. Using complete sentences, list three things wrong with the way the survey was conducted.
 - b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.
19. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
20. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
- 21. List some practical difficulties involved in getting accurate results from a telephone survey.
 - 22. List some practical difficulties involved in getting accurate results from a mailed survey.
 - 23. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.
 - 24. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. What type of sampling was used?
 - 25. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the

park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. What sampling method was used?

26. Name the sampling method used in each of the following situations.

- a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
- b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
- c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
- d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
- e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

27. A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had \$2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

- a. Do you consider the sample size large enough for a study of this type? Why or why not?
- b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."

- c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
- d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

28. The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.⁷⁸ Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

- a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
- b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
- c. In the last seven days, on how many days did you exercise for 30 minutes or more?
- d. Do you have health insurance coverage?

29. In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

- a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
- b. What effect does the low response rate have on the reliability of the sample?
- c. Are these problems examples of sampling error or nonsampling error?
- d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called “quota sampling” to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

30. Crime-related and demographic statistics for 47 US states in 1960 were collected from

government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates. Which of the potential problems with samples discussed below could explain this connection?

31. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks: "Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?" As of April 25, 11 people responded to this question. Each participant answered "NO!" Which of the potential problems with samples could explain this connection?

32. A scholarly article about response rates begins with the following quote: "Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research." The Pew Research Center for People and the Press admits: "The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more."³

- a. What are some reasons for the decline in response rate over the past decade?
- b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

33. Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in the table.

LBCC Distance Learning Survey Results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

- a. What percent of the students surveyed do not have a computer at home?
- b. About how many students in the survey live at least 16 miles from campus?

- c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

34. Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

35. A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers. Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

36. What type of measure scale is being used? Nominal, ordinal, interval or ratio.

- a. High school soccer players classified by their athletic ability: Superior, Average, Above average
- b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
- c. The colors of crayons in a 24-crayon box
- d. Social security numbers
- e. Incomes measured in dollars
- f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
- g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
- h. Time of day on an analog watch
- i. The distance in miles to the closest grocery store
- j. The dates 1066, 1492, 1644, 1947, and 1944
- k. The heights of 21–65 year-old women
- l. Common letter grades: A, B, C, D, and F

37. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

- Fill in the blanks in table
- What percent of students take exactly two courses?
- What percent of students take one or two courses?

38. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown below:

Flossing Frequency for Adults with Gum Disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0	27	0.4500		
1	18			
3				0.9333
6	3	0.0500		
7	1	0.0167		

- Fill in the blanks in the table.
- What percent of adults flossed six times per week?
- What percent flossed at most three times per week?

39. Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10

This table was produced.

Frequency of Immigrant Survey Responses

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

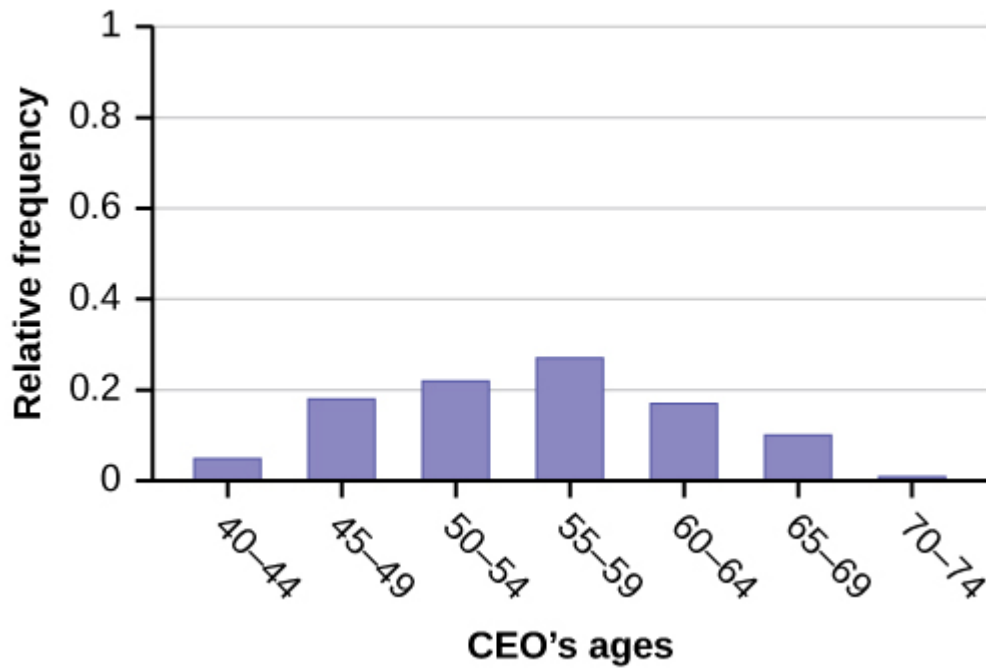
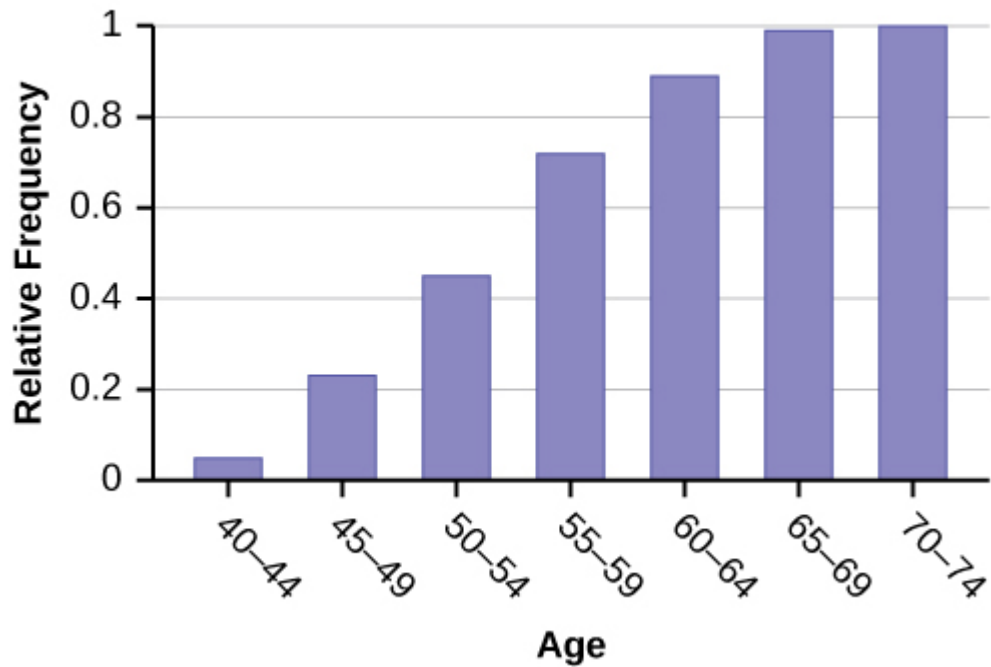
- Fix the errors in the table. Also, explain how someone might have arrived at the incorrect number(s).
 - Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
 - Fix the statement in part (b) to make it correct.
 - What fraction of the people surveyed have lived in the U.S. five or seven years?
 - What fraction of the people surveyed have lived in the U.S. at most 12 years?
 - What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
 - What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?
40. How much time does it take to travel to work? The table below shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

41. *Forbes* magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The table below shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

- What is the frequency for CEO ages between 54 and 65?
- What percentage of CEOs are 65 years or older?
- What is the relative frequency of ages under 50?
- What is the cumulative relative frequency for CEOs younger than 55?
- Which graph shows the relative frequency and which shows the cumulative relative frequency?

Graph A**Graph B**

42. the table below contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Frequency of Hurricane Direct Hits

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

- a. What is the relative frequency of direct hits that were category 4 hurricanes?
- b. What is the relative frequency of direct hits that were AT MOST a category 3 storm?

43. Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

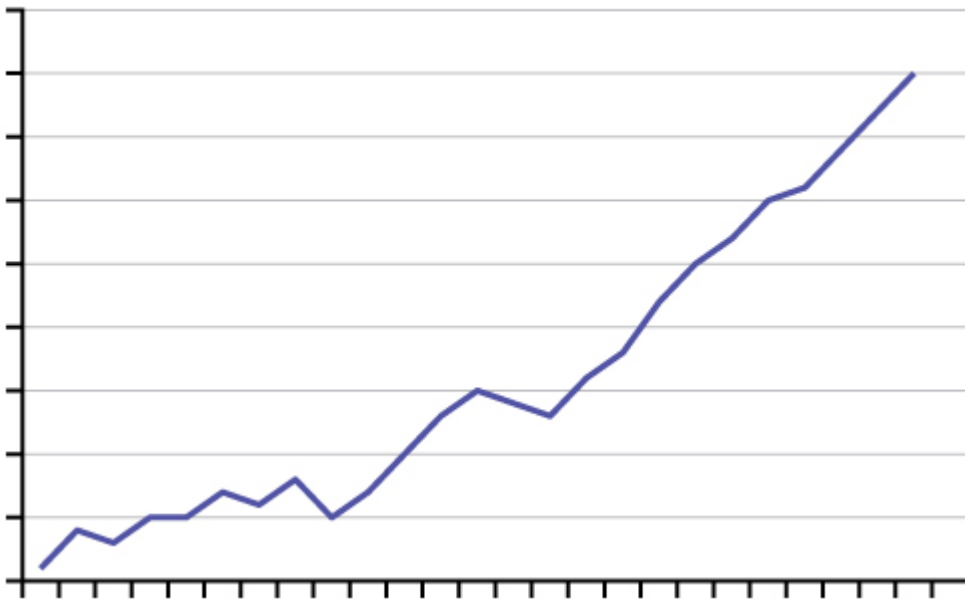
44. Discuss potential violations of the rule requiring informed consent.

- a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
- b. A research study is designed to investigate a new children's allergy medication.
- c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

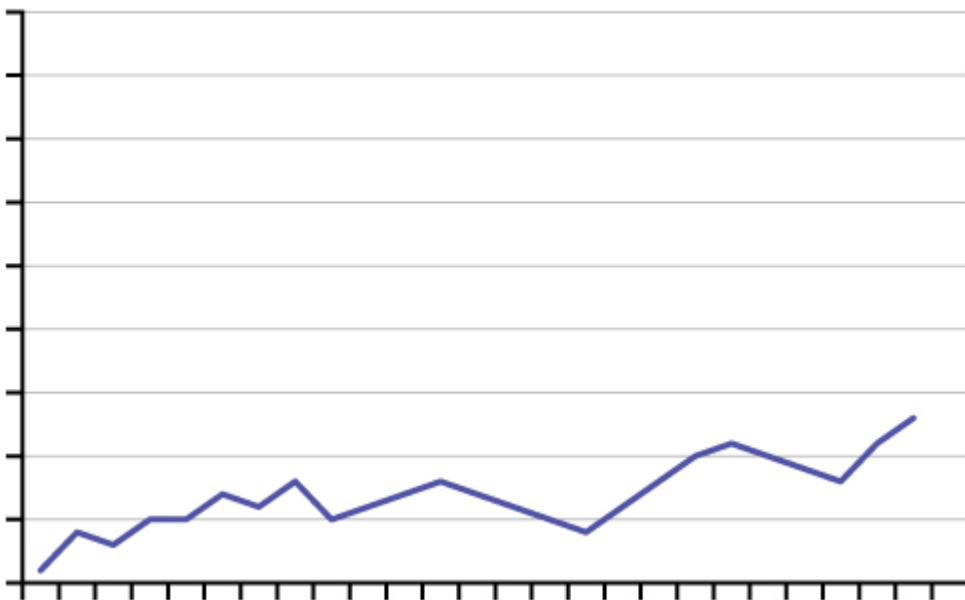
45. How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation. Use key terms from this module to describe the design of this experiment.

46. An advertisement for Acme Investments displays the two graphs in the figure below to show the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

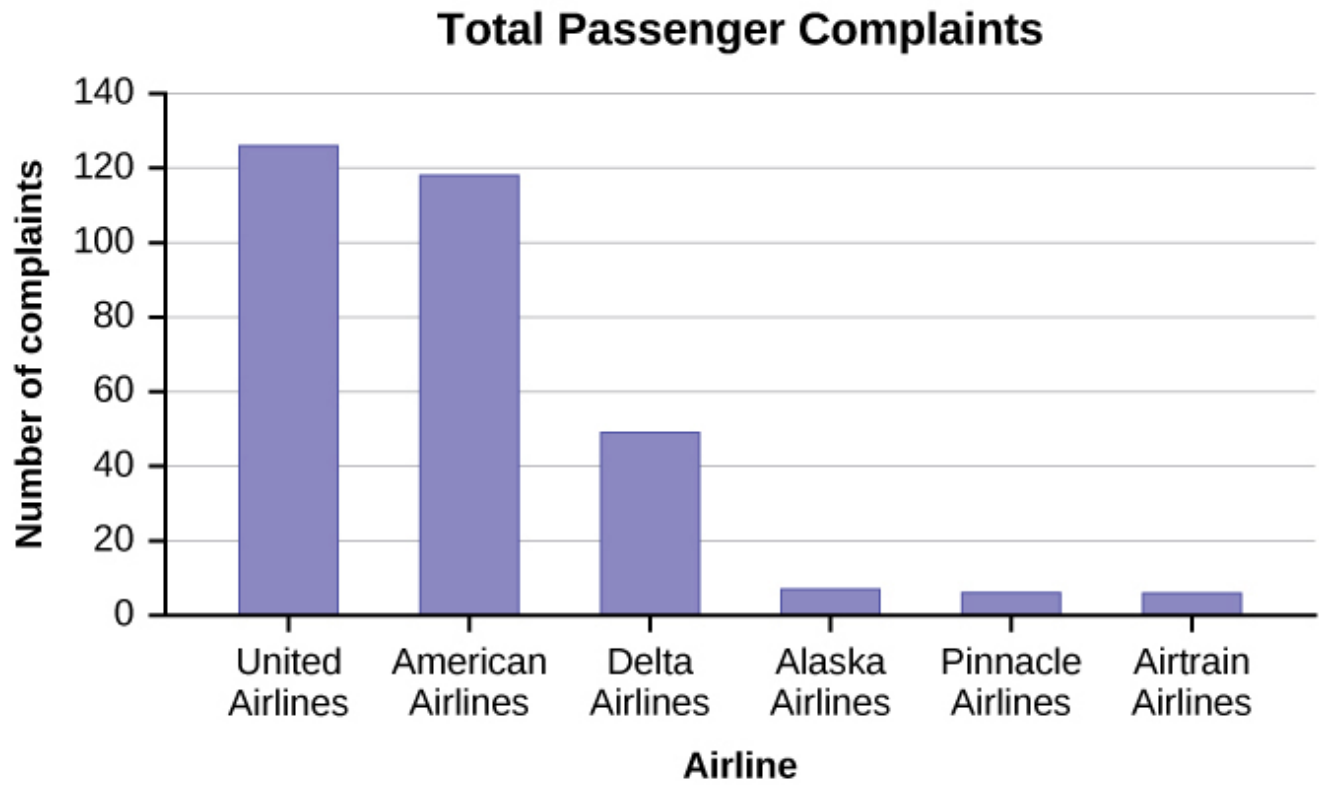
Acme Investments



Other Guy's Investments



47. The graph in the figure below shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?



Attribution

“Chapter 1 Homework” and “Chapter 1 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

1.7 ANSWERS TO SELECTED EXERCISES

1. (a) AIDS patients. (c) The average length of time (in months) AIDS patients live after treatment. (e) X = the length of time (in months) AIDS patients live after treatment

2. (b) all children who take ski or snowboard lessons; a group of these children; the population mean age of children who take their first snowboard lesson; the sample mean age of children who take their first snowboard lesson; X = the age of one child who takes his or her first ski or snowboard lesson; values for X , such as 3, 7, and so on. (d) the clients of the insurance companies; a group of the clients; the mean health costs of the clients; the mean health costs of the sample; X = the health costs of one client; values for X , such as 34, 9, 82, and so on (f) all the clients of this counselor; a group of clients of this marriage counselor; the proportion of all her clients who stay married; the proportion of the sample of the counselor's clients who stay married; X = the number of couples who stay married; yes, no (h) all people (maybe in a certain geographic area, such as the United States); a group of the people; the proportion of all people who will buy the product; the proportion of the sample who will buy the product; X = the number of people who will buy it; buy, not buy.

6.

a. 0.5242

b. 0.03%

c. 6.86%

d. $\frac{823,088}{823,856}$

e. quantitative discrete

f. quantitative continuous

g. In both years, underwater earthquakes produced massive tsunamis.

7. (b) systematic (d) simple random

8. (e) No, we do not have enough information to make such a claim. (f) Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients. (g) This would be convenience sampling and is not random.

9. (b) Yes, the sample size of 150 would be large enough to reflect a population of one school. (d)

Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

10. (a) There is not enough information given to judge if either one is correct or incorrect. (c) The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement. (e) Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

12. No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

14. No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

16. (a) quantitative discrete, 150 (c) qualitative, Oakland A's (e) quantitative discrete, 11, 234 students (g) qualitative, Crest (i) quantitative continuous, 47.3 years

18.

a. The survey was conducted using six similar flights. The survey would not be a true representation of the entire population of air travelers.

Conducting the survey on a holiday weekend will not produce representative results.

b. Conduct the survey during different times of the year. Conduct the survey using flights to and from various locations.

Conduct the survey on different days of the week.

20. Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

22. Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

26. (a) convenience (b) cluster (c) stratified (d) systematic (e) simple random

28.

a. qualitative

b. quantitative discrete

c. quantitative discrete

d. qualitative

30. Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

32.

- a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
- b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

36.

- a. ordinal
- b. interval
- c. nominal
- d. nominal
- e. ratio
- f. ordinal
- g. nominal
- h. interval
- i. ratio
- j. interval
- k. ratio
- l. ordinal

38. (b) 5.00% (c)93.33%

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

40. The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

44.

- Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
- Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
- All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

45.

Explanatory variable: amount of sleep

Response variable: performance measured in assigned tasks

Treatments: normal sleep and 27 hours of total sleep deprivation

Experimental Units: 19 professional drivers

Lurking variables: none – all drivers participated in both treatments

Random assignment: treatments were assigned in random order; this eliminated the effect of any “learning” that may take place during the first experimental session

Control/Placebo: completing the experimental session under normal sleep conditions

Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time⁸⁹. You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most

passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

Attribution

“Chapter 1 Solutions” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART II

DESCRIPTIVE STATISTICS

Chapter Outline

- 2.1 Introduction to Descriptive Statistics
- 2.2 Histograms, Frequency Polygons and Time Series Graphics
- 2.3 Measures of Central Tendency
- 2.4 Skewness and the Mean, Median, and Mode
- 2.5 Measures of Location
- 2.6 Measures of Dispersion
- 2.7 Exercises

2.1 INTRODUCTION TO DESCRIPTIVE STATISTICS



When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. Photo by William Greeson, CC BY 4.0.

Once we have collected data, what do we do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample is often overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that we can summarize and describe data. Your agent might also provide you with a graph of the data.

In this chapter, we will study numerical and graphical ways to describe and display your data.

This area of statistics is called **descriptive statistics**. We will learn how to calculate, and more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps us learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data, and then use more formal tools to analyze the data.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at bar graphs (or histograms), as well as frequency polygons and time-series graphs.

Attribution

“Chapter 2 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.2 HISTOGRAMS, FREQUENCY POLYGONS, AND TIME SERIES GRAPHS

LEARNING OBJECTIVES

- Display data using an appropriate graph: histograms, frequency polygons, and time series graphs.
- Analyze and interpret data presented in a graph.

Histograms

For most of the work we do in this book, we will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** is a visual display of a frequency chart. It consists of contiguous, vertical boxes with both a horizontal axis and a vertical axis. The horizontal axis is labeled with the classes or categories from the frequency chart. The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label on the vertical axis but the scale on the vertical axis will be different. The histogram gives us the shape of the data, the center of the data, and the spread of the data.

Recall that the frequency is the number of times an observation falls into that particular class and the relative frequency is the frequency for the class divided by the total number of data values in the sample. For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then the frequency of the 90% to 100% class is 3 and the relative frequency is $\frac{3}{40} = 0.075$. So, 7.5% of the students received between 90% and 100%.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes,

represent the data. Many histograms consist of 5 to 15 bars or classes for clarity, but the number of bars is determined by the person constructing the histogram. Choose a starting point for the first class to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05. We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495. If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995. If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5. Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=46#oembed-3>

Watch this video: Histograms | Applying mathematical reasoning | Pre-algebra | Khan Academy by Khan Academy [6:07]
(transcript available)

EXAMPLE

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data because height is a measurement.

60	64	64.5	66	66.5	67	67.5	69	70	71
60.5	64	64.5	66	66.5	67	67.5	69	70	71
61	64	64.5	66.5	66.5	67	67.5	69	70	72
61	64	66	66.5	67	67	68	69	70	72
61.5	64	66	66.5	67	67	68	69	70	72
63.5	64.5	66	66.5	67	67.5	69	69.5	70	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	72.5
63.5	64.5	66	66.5	67	67.5	69	69.5	70.5	73
64	64.5	66	66.5	67	67.5	69	69.5	70.5	73.5
64	64.5	66	66.5	67	67.5	69	69.5	71	74

The smallest data value is 60. Because the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Because the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point. Then the starting point is, then, $60 - 0.05 = 59.95$. The largest value is 74, so $74 + 0.05 = 74.05$ is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of classes (you must decide how many classes you want). Suppose we want to have eight classes.

$$\text{Class Width} = \frac{74.05 - 59.95}{8} = 1.76$$

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

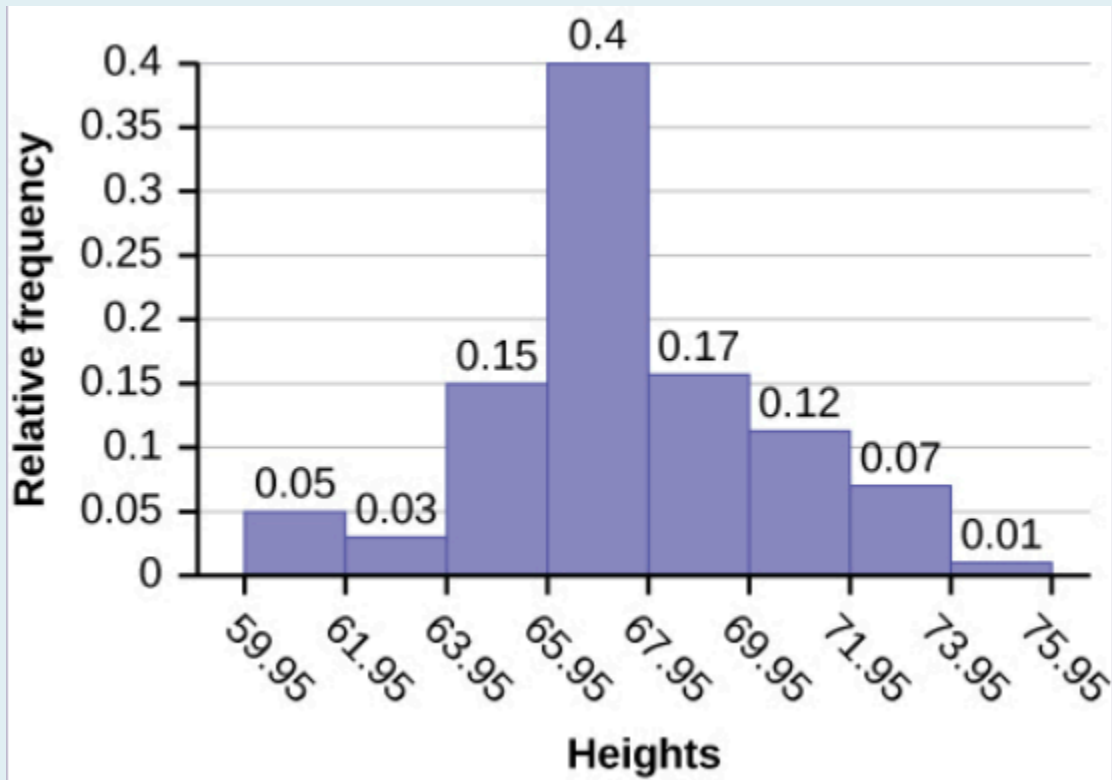
The boundaries for the classes are:

$$\begin{array}{l} 59.95 \\ 59.95 + 2 \\ 61.95 \\ 61.95 + 2 \\ 63.95 \\ 63.95 + 2 \\ 65.95 \\ 65.95 + 2 \\ 67.95 \\ 67.95 + 2 \\ 69.95 \\ 69.95 + 2 \\ 71.95 \\ 71.95 + 2 \\ 73.95 \\ 73.95 + 2 \\ 75.95 \end{array}$$

The heights 60 through 61.5 inches are in the first class 59.95–61.95. The heights that are 63.5 are in the second class 61.95–63.95. The heights that are 64 through 64.5 are in the third class 63.95–65.95.

The heights 66 through 67.5 are in the fourth class 65.95–67.95. The heights 68 through 69.5 are in the fifth class 67.95–69.95. The heights 70 through 71 are in the sixth class 69.95–71.95. The heights 72 through 73.5 are in the seventh class 71.95–73.95. The height 74 is in the last class 73.95–75.95.

The following histogram displays the heights on the x -axis and relative frequency on the y -axis.



NOTE

A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or classes.

TRY IT

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval.

Suppose you choose six bars.

9	9	9.5	9.5	10	10	10	10	10	10
10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	11	11
11	11	11	11	11	11	11	11	11	11
11	11.5	11.5	11.5	11.5	11.5	11.5	11.5	12	12
12	12	12	12	12	12.5	12.5	12.5	12.5	14

Click for the Solution

- Smallest value: 9
- Largest value: 14
- Convenient starting value: $9 - 0.05 = 8.95$
- Convenient ending value: $14 + 0.05 = 14.05$
- Class width: $\frac{14.05 - 8.95}{6} = 0.85$

The calculations suggest using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

EXAMPLE

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data** because books are counted.

1	1	1	1	1	1	1	1	1	1
1	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	5	5	5	5	5	6	6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value to get the starting and ending point. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Because the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____.

Solution:

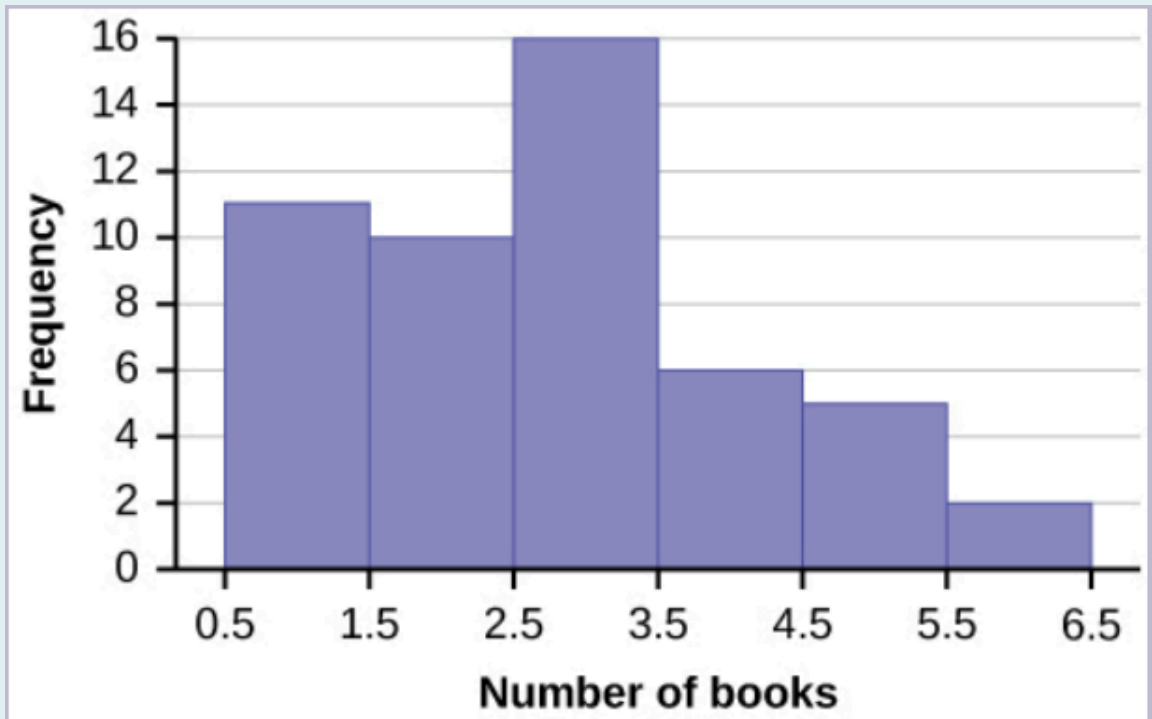
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, the number of bars is 6.

The following histogram displays the number of books on the x -axis and the frequency on the y -axis.



CREATING A FREQUENCY DISTRIBUTION AND HISTOGRAM IN EXCEL

In order to create a frequency distribution and its corresponding histogram in Excel, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter your data into a worksheet.
2. Determine the classes for the frequency distribution. Using these classes, create a **Bin** column that contains the **upper limit** for each class.
3. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
4. In the **Data Analysis** window, select **Histogram**. Click **OK**.
5. In the **Input** range, enter the cell range for the data.
6. In the **Bin** range, enter the cell range for the **Bin** column.
7. Select the location where you want the output to appear.
8. Select **Chart Output** to produce the corresponding histogram for the frequency distribution.
9. Click **OK**.

This website provides additional information on using Excel to create a frequency distribution.

NOTE

The histogram produced by Excel uses the frequency column from the frequency table on the vertical axis, not the relative frequency column.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=46#oembed-1>

Watch this video: Frequency Distributions by Joshua Emmanuel [8:40] (transcript available).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=46#oembed-2>

Watch this video: How to Construct a Histogram in Excel Using Data Analysis by Joshua Emmanuel [1:58] (transcript available).

TRY IT

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data because sports are counted.

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2
2	2	3	3	3	3	3	3	3	3

Fill in the blanks for the following sentence. Because the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Click to see Solution

- 1.5
- 1.5 to 2.5

- 2.5 to 3.5

EXAMPLE

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5



Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Frequency Polygons

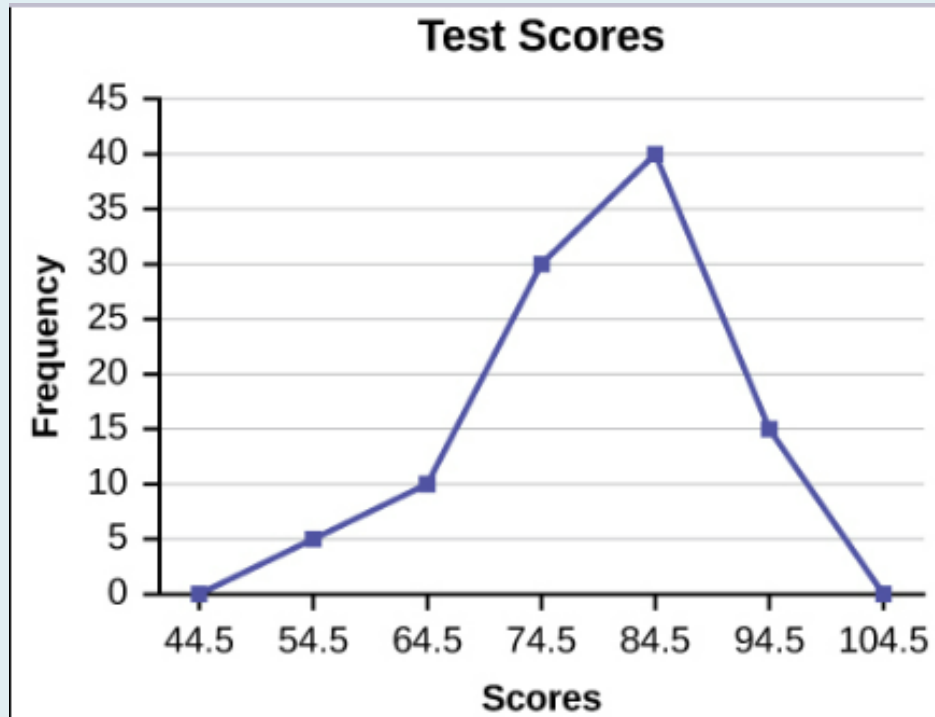
Associated with frequency charts and histograms, frequency polygons are line graphs with the classes on the horizontal axis, frequency on the vertical axis, and the frequencies plotted against the midpoint of the class interval. As with histograms, start by examining the data and decide on the classes, using similar techniques as discussed above. Find the frequency for each class. Plot

the classes on the x -axis and the frequency on y -axis. For each class, add a point on the graph with the x -coordinate equal to the class midpoint and the y -coordinate equal to the frequency of the class. Add points on the horizontal axis at the midpoint of the class before the first class and at the midpoint of the class after the last class. After all the points are plotted, draw line segments to connect them. Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

EXAMPLE

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores		
Lower Bound	Upper Bound	Frequency
49.5	59.5	5
59.5	69.5	10
69.5	79.5	30
79.5	89.5	40
89.5	99.5	15



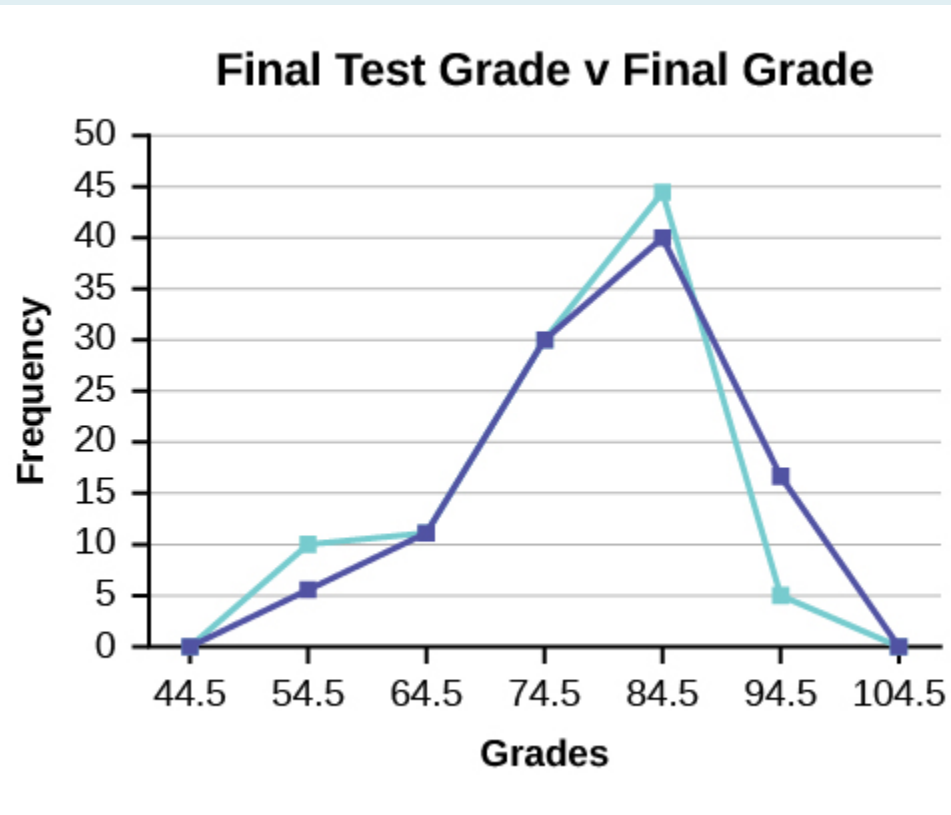
The first label on the x -axis is 44.5. This represents an interval extending from 39.5 to 49.5. Because the lowest test score is 54.5, this interval is used only to allow the graph to touch the x -axis. The point labeled 54.5 represents the next interval, or the first “real” interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the x -axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

EXAMPLE

We will construct an overlay frequency polygon comparing the scores with the students’ final numeric grade.

Frequency Distribution for Calculus Final Test Scores		
Lower Bound	Upper Bound	Frequency
49.5	59.5	5
59.5	69.5	10
69.5	79.5	30
79.5	89.5	40
89.5	99.5	15

Frequency Distribution for Calculus Final Grades		
Lower Bound	Upper Bound	Frequency
49.5	59.5	10
59.5	69.5	10
69.5	79.5	30
79.5	89.5	45
89.5	99.5	5



Time Series Graphs

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Because each date is paired with the temperature reading for the day, we do not have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a **time series graph**.

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments and the vertical axis is used to plot the values of the variable that we are measuring.

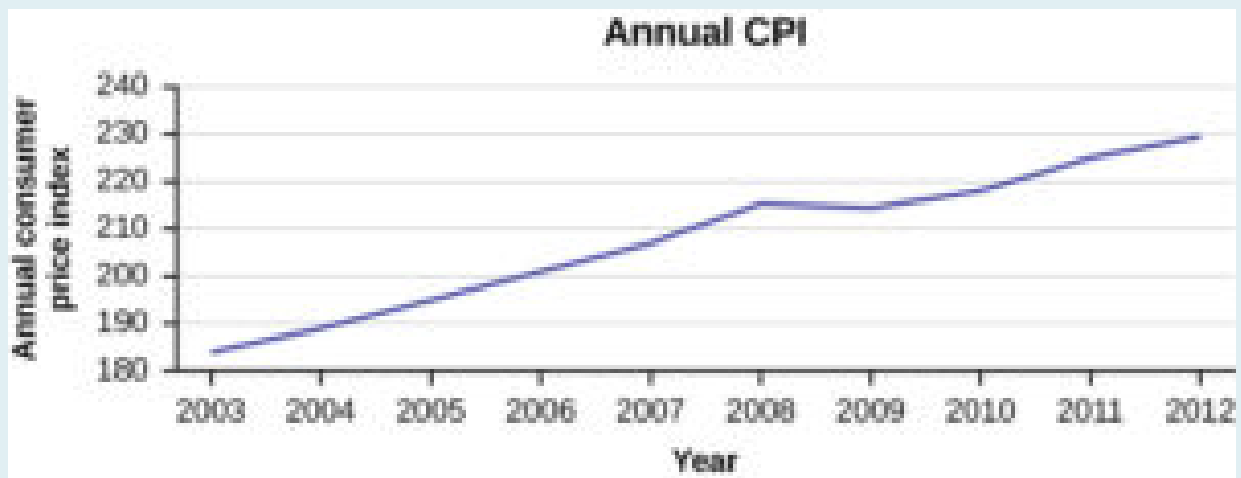
By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

EXAMPLE

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594



Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

Concept Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies or relative frequencies. The heights of the bars correspond to frequency or relative frequency values. Histograms are typically used for large, continuous, quantitative data sets.

A **frequency polygon** can also be used when graphing large data sets with data points that repeat. The data usually goes on the x -axis with the frequency being graphed on the y -axis.

Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

Attribution

“2.2 Histograms, Frequency Polygons, and Time Series Graphs“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.3 MEASURES OF CENTRAL TENDENCY

LEARNING OBJECTIVES

- Recognize, describe, calculate, and analyze the measures of the center of data: mean, median, and mode.

The “center” of a data set is a way of describing location. The two most widely used measures of the “center” of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data, and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term for mean is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

Mean

The **mean** is calculated by adding up all of the values in the data and then dividing the sum by the total number of data values.

The letter used to represent the sample mean is \bar{x} (read x -bar). The Greek letter μ (pronounced “mew”) represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

Consider the sample:

1	1	1	2	2	3	4	4	4	4	4
---	---	---	---	---	---	---	---	---	---	---

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

CALCULATING THE MEAN IN EXCEL

To find the mean in Excel, use the **average(array)** function.

- For **array**, enter the array or cell range containing the data.

The output from the **average** function is the mean of the entered data.

Visit the Microsoft page for more information about the **average** function.

Median

The **median** is the middle value in an **ordered** set of data. You can quickly find the **location** of the median by using the expression $\frac{n+1}{2}$ where n is the total number of data values in the sample.

If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then the median is located in position $\frac{n+1}{2} = \frac{97+1}{2} = 49$ of the ordered list. If the total

number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$ and the median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median.

CALCULATING THE MEDIAN IN EXCEL

To find the median in Excel, use the **median(array)** function.

- For **array**, enter the array or cell range containing the data.

The output from the **median** function is the median of the entered data.

Visit the Microsoft page for more information about the **median** function.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mean and the median.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A40.

For the mean:

Function	average	Answer
Field 1	A1:A40	23.575 months

For the median:

Function	median	Answer
Field 1	A1:A40	24 months

TRY IT

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3	4	5	7	7	7	7	8	8	9
9	10	10	10	10	10	11	12	12	13
14	14	15	15	17	17	18	19	19	19
21	21	22	22	23	24	24	24	24	

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A39.

For the mean:

Function	average	Answer
Field 1	A1:A39	13.949 months

For the median:

Function	median	Answer
Field 1	A1:A39	13 months

EXAMPLE

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the “center”: the mean or the median?

Solution:

$$\mu = \frac{5,000,000 + (49 \times 30,000)}{50} = 129,400$$

$$M = 30,000$$

The median is a better measure of the “center” than the mean because 49 of the values are \$30,000 and one is \$5,000,000. The \$5,000,000 is an outlier. The median of \$30,000 gives us a better sense of the middle of the data.

TRY IT

In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the “center”: the mean or the median?

Click to see Solution

The median is the better measure of the “center” than the mean because 59 of the values are either \$280,000 or \$315,000 and only one is \$2,500,000. The \$2,500,000 is an outlier. Either \$280,000 or \$315,000 gives us a better sense of the middle of the data.

Mode

Another measure of the center of the data is the mode. The **mode** is the most frequently occurring value in the set of data. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A set of data can also have no mode if all of the observations in the data are unique.

Unlike the mean and the median, the mode can be calculated for both qualitative data and quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

CALCULATING THE MODE IN EXCEL

To find the mode in Excel:

- Use the **count** and **mode.mult** function to determine the number of modes in the data. Enter **count(mode.mult(array))** into a cell where **array** is the array or cell range containing the data. This function will output the number of modes present in the data.
- If the output from the **count(mode.mult(array))** function is 1, then the data has a single mode. To find the single mode, use the **mode.sngl(array)** function, where **array** is the array or cell range containing the data. The output from the **mode.sngl** function is the value of single mode in the data.
 - Visit the Microsoft page for more information about the **mode.sngl** function.
- If the output from the **count(mode.mult(array))** function is greater than 1, then the data contains multiple modes. To find the multiple modes:
 - Left click on a cell, hold and drag down to highlight a number of vertical cells equal to the number of modes in the data. For example, if there are 4 modes in the data, highlight 4 cells in the vertical array.
 - In the highlighted cells, enter the **mode.mult(array)** function, where **array** is the array or cell range containing the data.
 - After entering the **mode.mult** function in the vertical array, press **CTRL+SHIFT+ENTER**. Because the output from this function is an array, we must press **CTRL+SHIFT+ENTER** (and not **ENTER**) to produce the array output.
 - The output from the **mode.mult** function are the modes in the data.
 - Visit the Microsoft page for more information about the **mode.mult** function.

EXAMPLE

Statistics exam scores for 20 students are as follows:

50	53	59	59	63	63	72	72	72	72
72	76	78	81	83	84	84	84	90	93

Find the mode.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A20.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))	Answer
Field 1	A1:A20	1

Because the output from the **count(mode.mult(...))** function is 1, there is only 1 mode in the data.

To find the single mode, we use the **mode.sngl** function:

Function	mode.sngl	Answer
Field 1	A1:A20	72

By examining the data, we can see that 72 is the most frequently occurring value (5 times) and that 72 is the only value that occurs 5 times.

TRY IT

The number of books checked out from the library from 25 students are as follows:

0	0	0	1	2
3	3	4	4	5
5	7	7	7	7
8	8	8	9	10
10	11	11	12	12

Find the mode.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A25.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))	Answer
Field 1	A1:A25	1

Because the output from the **count(mode.mult(...))** function is 1, there is only 1 mode in the data. To find the single mode, we use the **mode.sngl** function:

Function	mode.sngl	Answer
Field 1	A1:A25	7

The most frequent number of books is 7, which occurs four times.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mode.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A40.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))	Answer
Field 1	A1:A40	12

Because the output from the **count(mode.mult(...))** function is 12, there are 12 modes in the data. To find the multiple modes, we use the **mode.mult** function. Left click on a cell, hold and drag down to highlight 12 vertical cells. In the highlighted cells, enter the **mode.mult** function:

Function	mode.mult	Answer
Field 1	A1:A40	8, 15, 16, 17, 22, 24, 26, 27, 29, 33, 34, 44

Because the output from the **mode.mult** function is a (vertical) array after entering the function press **CTRL+SHIFT+ENTER** (not **ENTER** by itself).

TRY IT

Ten credit scores are

645	680	700	720	517	630	598	739	720	680
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Calculate the mode.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A10.

Start by using the **count** function to count the number of modes in the data:

Function	count(mode.mult(...))	Answer
Field 1	A1:A10	2

Because the output from the **count(mode.mult(...))** function is 2, there are 2 modes in the data. To find the multiple modes, we use the **mode.mult** function. Left click on a cell, hold and drag down to highlight 2 vertical cells. In the highlighted cells, enter the **mode.mult** function:

Function	mode.mult	Answer
Field 1	A1:A10	680, 720

Because the output from the **mode.mult** function is a (vertical) array after entering the function press **CTRL+SHIFT+ENTER** (not **ENTER** by itself).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=48#oembed-1>

Watch this video: Finding mean, median, and mode | Descriptive statistics | Probability and Statistics | Khan Academy by Khan Academy [3:54] (transcript available).

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, we do not know the individual data values (we only know intervals and interval frequencies). Therefore, we cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$\text{mean} = \frac{\text{data sum}}{\text{number of data values}}$$

We simply need to modify the definition to fit within the restrictions of a frequency table. Because we do not know the individual data values, we use the midpoint of each interval. The midpoint of an interval is

$$\text{midpoint} = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum(f \times m)}{\sum f}$$

where f is the frequency of the interval and m is the midpoint of the interval.

EXAMPLE

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution:

Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

$$\begin{array}{l} \text{Mean} = \frac{53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1)}{19} \\ = \frac{1460.25}{19} \\ = 76.86 \end{array}$$

TRY IT

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Click to the Solution

The midpoints are 1.75, 5.5, 9.5, 13.5, 17.5.

$$\begin{array}{l} \text{Mean} = \frac{(1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9)}{38} \\ = \frac{409.75}{38} \\ = 10.78 \end{array}$$

When to Use Each Measure of Central Tendency

The measures of central tendency tell us about the center of the data, but often give different answers. So how do we know when to use each? Here are some general rules:

1. The mean is the most frequently used measure of central tendency and is generally considered the best measure of central location.
 2. Median is the preferred measure of central tendency when:
 - a. There are a few extreme values or outliers in the distribution of the data. (Note: Remember that a single outlier can have a great effect on the mean).
 - b. There are some missing or undetermined values in the data
 - c. There is an open ended distribution (For example, if you have a data field which measures the number of children and your options are 0, 1, 2, 3, 4, 5 or “6 or more,” then the “6 or more field” is open ended and makes calculating the mean impossible because we do not know the exact values for this field).
 - d. You have data measured on an ordinal scale.
 3. Mode is the preferred measure when data are measured in a nominal or ordinal scale.
-

Concept Review

The mean, the median, and the mode are measures of the “center” of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode tells us the most frequently occurring datum (or data) in our data set. The mean, median, and mode are extremely helpful when we need to analyze our data, but if the data set consists of ranges which lack specific values, the mean may be impossible to calculate. However, the mean of grouped data can be approximated by multiplying the midpoint of each interval with the frequency, adding up these values and then dividing by the total number of values in the data set.

Attribution

“2.5 Measures of the Center of the Data“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.4 SKEWNESS AND THE MEAN, MEDIAN, AND MODE

LEARNING OBJECTIVES

- Identify the shape of a set of data.

Consider the following data set:

4	5	6	6	6	7	7	7
7	7	7	8	8	8	9	10

This data set can be represented by the following histogram. Each interval has width one, and each value is located in the middle of an interval.

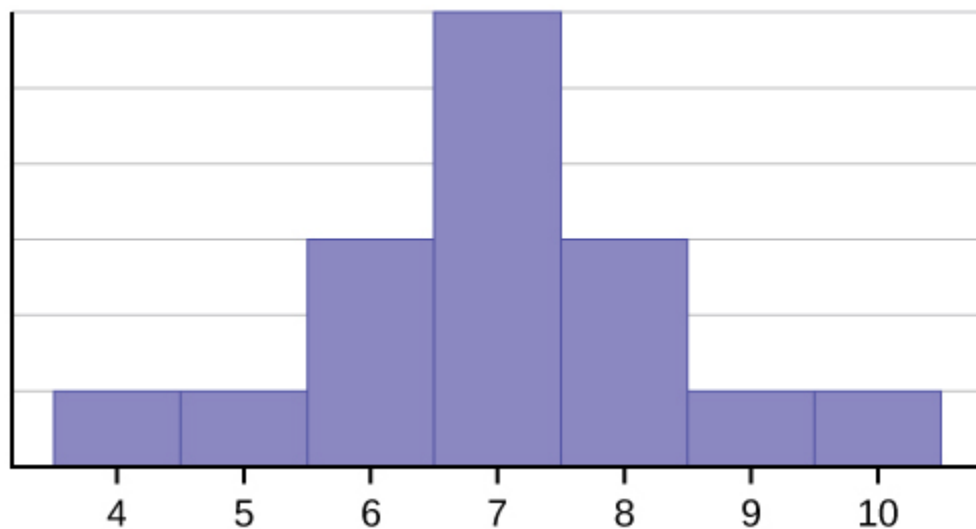


Figure 1

The histogram above displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram so that the shape to the left and the right of the vertical line are mirror images of each other. For the above data set, the mean, the median, and the mode are each seven. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode, and the mode is the same as the mean and median. In a symmetrical distribution that has multiple modes, the modes would be different from the mean and median.

Consider the following data set:

4	5	6	6	6	7	7	7	7	8
---	---	---	---	---	---	---	---	---	---

This data set can be represented by the following histogram. Each interval has width one, and each value is located in the middle of an interval.

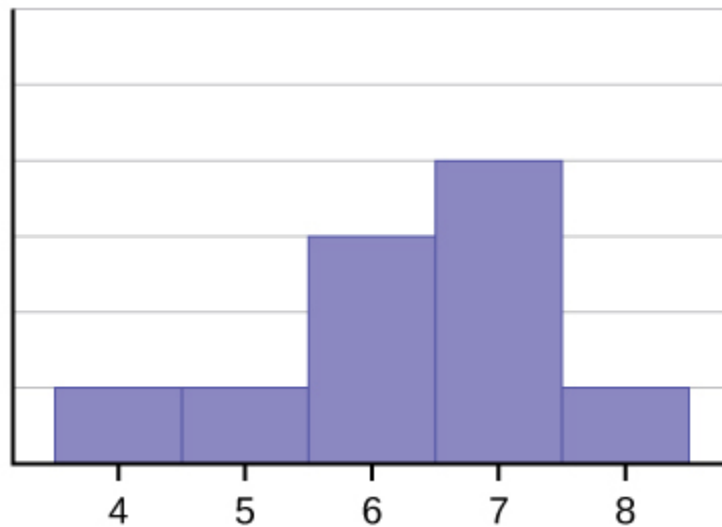


Figure 2

The histogram above is not symmetrical. The right-hand side seems “chopped off” compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left. The mean of this data is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

Consider the following data set:

6	7	7	7	7	8	8	8	9	10
---	---	---	---	---	---	---	---	---	----

This data set can be represented by the following histogram. Each interval has width one, and each value is located in the middle of an interval.

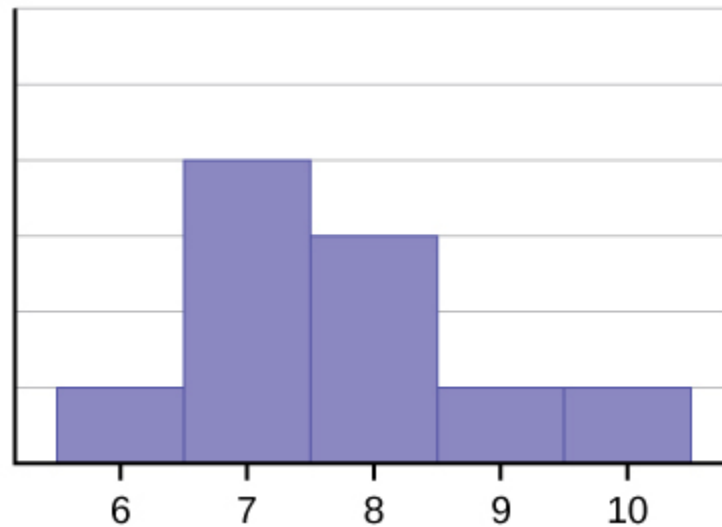


Figure 3

The histogram above is also not symmetrical. In this case, the data is **skewed to the right**. The mean for this data is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize:

- If the distribution of the data is symmetrical, mean = median = mode (assuming there is only one mode). If there are multiple modes in a symmetric distribution, the modes would be different from the mean and the median, but the mean and median would still be equal.
- If the distribution of the data is skewed to the left, mean < median < mode.
- If the distribution of the data is skewed to the right, mean > median > mode.

Skewness and symmetry become important when we discuss probability distributions in later chapters.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=56#oembed-1>

Watch this video: Elementary Business Statistics | Skewness and the Mean, Median, and Mode by Janux [3:57] (transcript available).

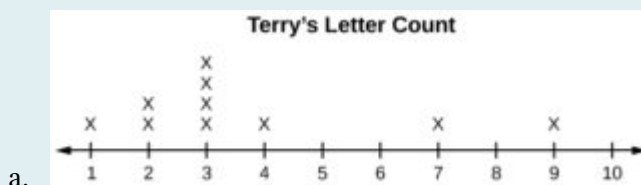
EXAMPLE

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

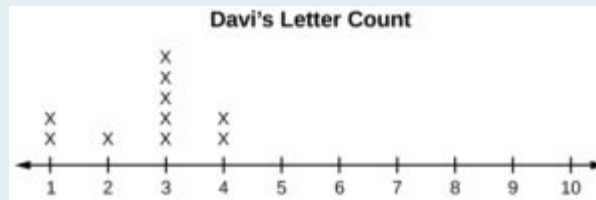
Terry									
7	9	3	3	3	4	1	3	2	2
Davis									
3	3	3	4	1	4	3	2	3	1
Maris									
2	3	4	4	4	6	6	6	8	3

1. Make a dot plot for the three authors and compare the shapes.
2. Calculate the mean for each.
3. Calculate the median for each.
4. Describe any pattern you notice between the shape and the measures of center.

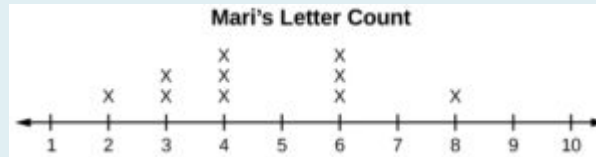
Solution:



Terry's distribution has a right (positive) skew.



Davis' distribution has a left (negative) skew



Maris' distribution is symmetrically shaped.

- b. Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
- c. Terry's median is three, Davis' median is three. Maris' median is four.
- d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Concept Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like Figure 3. A **left (or negative) skewed** distribution has a shape like Figure 2. A **symmetrical** distribution looks like Figure 1.

Attribution

“2.6 Skewness and the Mean, Median, and Mode“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.5 MEASURES OF LOCATION

LEARNING OBJECTIVES

- Recognize, describe, calculate, and interpret the measures of location of data: quartiles and percentiles.

The common measures of location are **quartiles** and **percentiles**. Previously, we learned that the **median** is a number that measures the “center” of the data. But the median can also be thought of as a measure of location because the median is the “middle value” of a set of data. The median is a number that separates ordered data into halves. Half of the values in the data are the same number or smaller than the median and half of the values in the data are the same number or larger.

For example, consider the following data, already ordered from smallest to largest:

1	1	2	2	4	6	6.8
7.2	8	8.3	9	10	10	11.5

Because there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two:

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. We can see that half (or 50%) of the values are less than seven and half (or 50%) of the values are larger than seven.

The median is an example of both a quartile and a percentile. The median is also the second quartile, Q_2 , and the 50th percentile, P_{50} .

Quartiles

Quartiles are numbers that separate the data into quarters (four parts). Like the median, quartiles may or may not be an actual value in the set of data. To find the quartiles, order the data (from smallest to largest) and then find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data and the third quartile, Q_3 , is the middle value of the upper half of the data. To get the idea, consider the same (ordered) data set used above:

1	1	2	2	4	6	6.8
7.2	8	8.3	9	10	10	11.5

The median or **second quartile** is seven. The lower half of the data are:

1	1	2	2	4	6	6.8
---	---	---	---	---	---	-----

The middle value of the lower half of the data is 2. The number 2, which is part of the data, is the **first quartile**, Q_1 . One-fourth (or 25%) of the entire sets of values are the same as or less than 2 and three-fourths (or 75%) of the values are more than two.

The upper half of the data are:

7.2	8	8.3	9	10	10	11.5
-----	---	-----	---	----	----	------

The middle value of the upper half of the data is 9. The **third quartile**, Q_3 , is 9. Three-fourths (or 75%) of the values are less than 9. One-fourth (or 25%) of the values in the data set are greater than or equal to 9.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The *IQR* can help to determine potential **outliers**. A value is suspected to be a potential outlier if it is less than $1.5 \times IQR$ below the first quartile or more than $1.5 \times IQR$ above the third quartile. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors, some kind of abnormality, or they may be a key to understanding the data.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=58#oembed-1>

Watch this video: Median, Quartiles and Interquartile Range by ExamSolutions [12:35] (transcript available).

CALCULATING QUANTILES IN EXCEL

To find quartiles in Excel, use the **quartile.exc(array, quartile number)** function.

- For **array**, enter the array or cell range containing the data.
- For **quartile number**, enter the quartile (1, 2 or 3) being calculated.

The output from the **quartile.exc** function is the value of the corresponding quartile. For example, **quartile.exc(array,1)** returns the value of the first quartile where 25% of the observations in the data are (strictly) less than the value of the first quartile.

Visit the Microsoft page for more information about the **quartile.exc** function.

NOTE

We are using the **quartile.exc** function, and not the **quartile.inc** function, because we want the percent of the observations in the data to be strictly less than the value of the quartile.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=58#oembed-2>

Watch this video: How To Find Quartiles and Construct a Box Plot in Excel by Joshua Emmanuel [4:12] (transcript available).

EXAMPLE

For the following 13 real estate prices, calculate the three quartiles and the *IQR*. Determine if any prices are potential outliers. The prices are in dollars.

389,950	230,500	158,000	479,000	639,000	114,950	5,500,000
387,000	659,000	529,000	575,000	488,800	1,095,000	

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A13.

For the first quartile Q_1 :

Function	quartile.exc	Answer
Field 1	A1:A13	\$308,750
Field 2	1	

For the second quartile Q_2 :

Function	quartile.exc	Answer
Field 1	A1:A13	\$488,800
Field 2	2	

For the third quartile Q_3 :

Function	quartile.exc	Answer
Field 1	A1:A13	\$649,000
Field 2	3	

For the IQR: $IQR = 649,000 - 308,750 = \$340,250$

To determine if there are any outliers:

$$\begin{aligned} 1.5 \times IQR &= 1.5 \times 340,250 = 510,375 \\ Q_1 - 1.5 \times IQR &= 308,750 - 510,375 = -201,625 \\ Q_3 + 1.5 \times IQR &= 649,000 + 510,375 = 1,159,375 \end{aligned}$$

No house price is less than $-\$201,625$. However, $\$5,500,000$ is more than $\$1,159,375$. Therefore, $\$5,500,000$ is a potential **outlier**.

NOTE

Quartiles have the same units as the data. In this case, the data is measured in dollars, so the quartiles are also in dollars.

TRY IT

For the following 11 salaries, calculate the three quartiles and the *IQR*. Are any of the salaries outliers? The salaries are in dollars.

33,000	72,000	54,000
64,500	68,500	120,000
28,000	69,000	40,500
54,000	42,000	

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A11.

For the first quartile Q_1 :

Function	quartile.exc	Answer
Field 1	A1:A11	\$40,500
Field 2	1	

For the second quartile Q_2 :

Function	quartile.exc	Answer
Field 1	A1:A11	\$54,000
Field 2	2	

For the third quartile Q_3 :

Function	quartile.exc	Answer
Field 1	A1:A11	\$69,000
Field 2	3	

For the IQR: $IQR = 69,000 - 40,500 = \$28,500$

To determine if there are any outliers:

$$\begin{aligned} 1.5 \times IQR \text{ \& } &= \text{ \& } 1.5 \times 28,500 = 42,750 \\ Q_1 - 1.5 \times IQR \text{ \& } &= \text{ \& } 40,500 - 42,750 = -2,250 \\ Q_3 + 1.5 \times IQR \text{ \& } &= \text{ \& } 69,000 + 42,750 = 111,750 \end{aligned}$$

No salary is less than $-\$2,250$. However, $\$120,000$ is more than $\$111,750$, so $\$120,000$ is a potential outlier.

TRY IT

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A									
69	96	81	79	65	76	83	99	89	67
90	77	85	98	66	91	77	69	80	94

Test Scores for Class B									
90	72	80	92	90	97	92	75	79	68
70	80	99	95	78	73	71	68	95	100

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data for Class A into column A from cell A1 to A20 and the data for Class B into column B from cell B1 to B20.

Class A

For the first quartile Q_1 :

Function	quartile.exc	Answer
Field 1	A1:A20	70.75
Field 2	1	

For the third quartile Q_3 :

Function	quartile.exc	Answer
Field 1	A1:A20	90.75
Field 2	3	

For the IQR: $IQR = 90.75 - 70.75 = 20$

Class BFor the first quartile Q_1 :

Function	quartile.exc	Answer
Field 1	B1:B20	72.25
Field 2	1	

For the third quartile Q_3 :

Function	quartile.exc	Answer
Field 1	B1:B20	94.25
Field 2	3	

For the IQR: $IQR = 94.25 - 72.25 = 22$

The data for Class B has a larger IQR , so the scores between Q_3 and Q_1 (the middle 50% of the data) for the data for Class B are more spread out and not clustered about the median.

Percentiles

Percentiles are numbers that separate the (ordered) data into hundredths (100 parts). Like quartiles, percentiles may or may not be part of the data. The n th percentile, P_n , is the value where $n\%$ of the observations in the data are less than the value of the n th percentile. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. The 90th percentile means that 90% of test scores are less than your score and 10% of the test scores are the same or greater than your test score.

Quartiles are special percentiles. The first quartile, Q_1 , is the same as the 25th percentile and the third quartile, Q_3 , is the same as the 75th percentile. The median is the 50th percentile.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into an SAT score of at least 1220.

Percentiles are mostly used with very large data sets. Therefore, if you were to say that 90% of

the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

CALCULATING PERCENTILES IN EXCEL

To find the k th percentiles in Excel, use the **percentile.exc(array, percent)** function.

- For **array**, enter the array or cell range containing the data.
- For **percent**, enter the percentile (as a decimal) being calculated. For example, if we are calculating the 60th percentile, we would enter 0.6 for the percent in the **percentile.exc** function.

The output from the **percentile.exc** function is the value of the corresponding percentile. For example, **percentile.exc(array,0.6)** returns the value of the 60th percentile where 60% of the observations in the data are (strictly) less than the value of the 60th percentile.

Visit the Microsoft page for more information about the **percentile.exc** function.

NOTE

We are using the **percentile.exc** function, and not the **percentile.inc** function, because we want the percent of the observations in the data to be strictly less than the value of the percentile.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=58#oembed-3>

Watch this video: Percentiles – How to calculate Percentiles, Quartiles, ... by Joshua Emmanuel [3:43] (transcript available).

EXAMPLE

Listed are twenty-nine ages (in years) for trees found in the Saint Louis Botanical Garden.

18	21	22	25	26	27	29	30	31	33
36	37	41	42	47	52	55	57	58	62
64	67	69	71	72	73	74	76	77	

1. Find the 70th percentile.
2. Find the 83rd percentile.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A29.

For the 70th percentile P_{70} :

Function	percentile.exc	Answer
Field 1	A1:A29	64 years
Field 2	0.7	

For the 83rd percentile P_{83} :

Function	percentile.exc	Answer
Field 1	A1:A29	71.9 years
Field 2	0.83	

NOTE

Percentiles have the same units as the data. In this case, the data is measured in years, so the percentiles are also in years.

TRY IT

Listed are 29 ages (in years) for Academy Award winning best actors.

18	21	22	25	26	27	29	30	31	33
36	37	41	42	47	52	55	57	58	62
64	67	69	71	72	73	74	76	77	

Calculate the 20th percentile and the 55th percentile.

Click to see Solution

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A29.

For the 20th percentile P_{20} :

Function	percentile.exc	Answer
Field 1	A1:A29	27 years
Field 2	0.2	

For the 55th percentile P_{55} :

Function	percentile.exc	Answer
Field 1	A1:A29	53.5 years
Field 2	0.55	

Interpreting Percentiles and Quartiles

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than the value of the n th percentile. For example, 15% of the data values are less than the value of the 15th percentile. Note that low percentiles always correspond to lower data values and high percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.” The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered “good,” but in other contexts a high percentile might be considered “good”. In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles or quartiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text. When writing the interpretation of a percentile or quartile in the context of the given data, the sentence should contain the following information:

- Information about the context of the situation being considered,
- The data value (value of the variable) that represents the percentile/quartile.
- The percent of individuals or items with data values below the percentile/quartile.

EXAMPLE

On a timed math test, the first quartile for the time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution:

- Interpretation: 25% of students finished the exam in less than 35 minutes.
- In this context, a low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

TRY IT

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Click to see Solution

- Interpretation: 75% of runners finished the race in less than 11.5 seconds.
- In this context, a lower percentile is good because finishing a race more quickly is desirable.

EXAMPLE

On a 20 question math test, the 70th percentile for the number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution:

- Interpretation: 70% of students answered less than 16 questions correctly.

TRY IT

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Click to see Solution

- Interpretation: 80% of students earned less than 49 points.

EXAMPLE

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

Solution:

- Interpretation: 30% of students are enrolled in less than 7 credit units.
- In this context, there is no “good” or “bad” value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

TRY IT

During a season, the 40th percentile for points scored per player in a game is 8. Interpret the 40th percentile in the context of this situation.

Click to see Solution

- Interpretation: 40% of players scored fewer than 8 points.

Concept Review

The values that divide an ordered set of data into 100 equal parts are called percentiles. Percentiles

are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50% of the other observations in the set.

Quartiles divide data into quarters. The first quartile, Q_1 , is the 25th percentile, the second quartile, Q_2 , is the 50th percentile, and the third quartile, Q_3 , is the 75th percentile. The interquartile range, IQR , is the range of the middle 50% of the data values. The IQR is found by subtracting Q_1 from Q_3 , and can help determine outliers by using the following two expressions.

Attribution

“2.3 Measures of the Location of the Data“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.6 MEASURES OF DISPERSION

LEARNING OBJECTIVES

- Recognize, describe, calculate, and analyze the measures of the spread of data: variance, standard deviation, and range.

It can be misleading to only use the measures of central tendency (mean, median, mode) to describe a data set. Measures of central tendency describe the center of a distribution. Measures of dispersion or variability are used to describe the spread or dispersion of the data. So far in this chapter, we have already seen a measure of dispersion—the interquartile range. The interquartile range describes the spread of the middle 50% of the data. But there are other measures of dispersion, including range, variance, and standard deviation.

Range

The **range** is the difference between the largest and smallest value in a set of data:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Range is not a very good measure of variability because it is based on only two values in the data set (the largest and smallest values) and is highly influenced by outliers. Also, the range does not help us distinguish between two data sets with the same largest and smallest values because the two data sets will have the same range.

EXAMPLE

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows:

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the range.

Solution:

The largest value is 47 and the smallest value is 3, so

$$\text{Range} = 47 - 3 = 44$$

Variance and Standard Deviation

An important characteristic of any set of data is the variation in the data from the mean. In some data sets, the data values are concentrated close to the mean, but in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures, on average, how far data values are from their mean. The standard deviation provides a numerical measure of the overall amount of variation in a data set, and can be used to determine whether a particular data value is close to or far away from the mean.

The standard deviation provides a measure of the overall variation in a data set. The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean because there is little variation or spread in the data. The standard deviation is larger when the data values are more spread out from the mean because there is a lot of variation in the data. The lower case letter s represents the sample standard deviation and the Greek letter σ represents the population standard deviation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. The mean wait time at both supermarkets is five minutes. At supermarket A the standard deviation for the wait time is two minutes and at supermarket B the standard deviation for the wait time is four minutes. Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the mean and wait times at supermarket A are more concentrated near the mean.

As well, the standard deviation can be used to determine whether a data value is close to or far from the mean. For example, suppose that Rosa and Binh both shop at supermarket A where the mean wait time at the checkout is five minutes and the standard deviation is two minutes. Suppose Rosa's wait time is seven minutes and Binh's wait time is one minute:

- Rosa's wait time of seven minutes is **two minutes longer than the mean** of five minutes. Because two minutes is equal to one standard deviation, Rosa's wait time of seven minutes is **one standard deviation above the mean** of five minutes.
- Binh's wait time of one minute is **four minutes less than the mean** of five minutes. Because four minutes is equal to two standard deviations, Binh's wait time of one minute is **two standard deviations below the mean** of five minutes.

A data value that is two standard deviations from the mean is just on the borderline for what many statisticians would consider to be far from the mean. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations.

Calculating the Standard Deviation

If x is a number, then the difference " $x - \text{mean}$ " is called its **deviation from the mean**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but **not** identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance and the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance and the sample standard deviation s is the square root of the sample variance. The standard deviation can be thought of as a special average of the deviations.

To calculate a population standard deviation σ :

1. Add up the deviations from the mean: $x - \mu$
2. Divide the sum in step 1 by the population size N .
3. The population standard deviation is the square root of the value from step 2.

The formula for the population standard deviation is:
$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

To calculate a sample standard deviation s :

1. Add up the deviations from the mean: $x - \bar{x}$
2. Divide the sum in step 1 by the sample size $n - 1$.
3. The sample standard deviation is the square root of the value from step 2.

The formula for the population standard deviation is:
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=60#oembed-2>

Watch this video: How to calculate Standard Deviation and Variance by statisticsfun [5:04] (transcript available).

CALCULATING VARIANCE IN EXCEL

To find the variance in Excel:

- If the data is population data, use the **var.p(array)** function where **array** is the array or cell range containing the data. The output from the **var.p** function is the population variance.
 - Visit the Microsoft page for more information about the **var.p** function.
- If the data is sample data, use the **var.s(array)** function where **array** is the array or cell range containing the data. The output from the **var.s** function is the sample variance.
 - Visit the Microsoft page for more information about the **var.s** function.

NOTE

There are two different functions to calculate variance in Excel because variance is calculated differently depending on whether the data is from a sample or from a population. When calculating variance, make sure that you are using the correct function based on the type of data you are working with (sample or population).

CALCULATING STANDARD DEVIATION IN EXCEL

To find the standard deviation in Excel:

- If the data is population data, use the **stdev.p(array)** function where **array** is the array or cell range containing the data. The output from the **stdev.p** function is the population standard deviation.
 - Visit the Microsoft page for more information about the **stdev.p** function.
- If the data is sample data, use the **stdev.s(array)** function where **array** is the array or cell range containing the data. The output from the **stdev.s** function is the sample standard deviation.
 - Visit the Microsoft page for more information about the **stdev.s** function.

NOTE

There are two different functions to calculate standard deviation in Excel because standard deviation is calculated differently depending on whether the data is from a sample or from a population. When calculating standard deviation, make sure that you are using the correct function based on the type of data you are working with (sample or population).



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=60#oembed-1>

Watch this video: Range, Variance, Standard Deviation in Excel by Joshua Emmanuel [1:10] (transcript available).

EXAMPLE

In a fifth grade class, the teacher was interested in the standard deviation of the ages of her students. The following data are the ages, in years, for a sample of 20 fifth grade students. The ages are rounded to the nearest half year:

9	9.5	9.5	10	10	10	10	10.5	10.5	10.5
10.5	11	11	11	11	11	11	11.5	11.5	11.5

Calculate the mean, the variance, and the standard deviation of the ages of the students. Interpret the standard deviation.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A20.

For the mean:

Function	average	Answer
Field 1	A1:A20	10.525 years

For the variance:

Function	var.s	Answer
Field 1	A1:A20	0.5125

For the standard deviation:

Function	stdev.s	Answer
Field 1	A1:A20	0.7159 years

Interpreting the standard deviation:

On average, the age of any fifth grader is 0.7159 years away from the mean of 10.525 years.

NOTES

1. We are using the **var.s** (not **var.p**) and **stdev.s** (not **stdev.p**) functions to calculate the variance and standard deviation because the data is from a sample.
2. Standard deviation has the same units as the data. In this case, the data is measured in years, so the standard deviation is also in years.
3. There are no units associated with variance.

TRY IT

On a baseball team, the ages, in years, of each of the players are as follows:

21	21	22	23	24
24	25	25	28	29
28	31	32	33	33
34	35	36	36	36
36	38	38	38	40

Find the mean and standard deviation.

Click to see Solution

$$\mu = 30.64 \text{ years} \quad \sigma = 5.99 \text{ years}$$

NOTE

We are using the **var.p** (not **var.s**) and **stdev.p** (not **stdev.s**) functions to calculate the variance and standard deviation because the baseball team is a population.

NOTE

Your concentration should be on what the standard deviation tells you about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or a positive number. When the standard deviation is zero, there is no dispersion—that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean. Outliers in the data can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better “feel” for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data.**

EXAMPLE

Use the following sample of exam scores from Susan Dean's spring pre-calculus class:

33	42	49	49	53	55	55	61
63	67	68	68	69	69	72	73
74	78	80	83	88	88	88	90
92	94	94	94	94	96	100	

Calculate the following:

- The mean.
- The standard deviation.
- The median.
- The first quartile.
- The third quartile.
- *IQR*.

Solution:

Enter the data into an Excel spreadsheet. For this example, suppose we entered the data in column A from cell A1 to A31.

For the mean:

Function	average	Answer
Field 1	A1:A31	73.5

For the median:

Function	median	Answer
Field 1	A1:A31	73

For the standard deviation:

Function	stdev.s	Answer
Field 1	A1:A31	17.92

For the first quartile:

Function	quartile.exc	Answer
Field 1	A1:A31	61
Field 2	1	

For the third quartile:

Function	quartile.exe	Answer
Field 1	A1:A31	90
Field 2	3	

For the IQR: $IQR = 90 - 61 = 29$

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and different standard deviations, then comparing the data values directly can be misleading. In order to directly compare values in different data sets, we can compare how many standard deviations away from the mean of its data set a value is. This is done by calculating the value's *z*-score:

Sample	$z = \frac{x - \bar{x}}{s}$
Population	$z = \frac{x - \mu}{\sigma}$

The value x is z standard deviations away from the mean.

EXAMPLE

Two students, John and Ali, are from different high schools and wanted to find out who had the highest GPA when compared to their school. Which student had the highest GPA when compared to their school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations, the z -score, their GPA is away from the mean of their school.

$$\text{John: } z = \frac{2.85 - 3.00}{0.7} = -0.21$$

$$\text{Ali: } z = \frac{77 - 80}{10} = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** her school's mean, which means that John's GPA is closer to his school's mean than Ali's GPA is to hers.

NOTE

The sign of a z -score is important. A negative z -score tells us that x is below the mean. A positive z -score tells us that x is above the mean. The absolute value of the z -score tells us how many standard deviations away from the mean the value of x is.

TRY IT

Two swimmers, Angie and Beth are from different teams and wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team's mean time. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Click to see Solution

$$\text{Angie: } z = \frac{26.2 - 27.2}{0.8} = -1.25$$

$$\text{Beth: } z = \frac{27.3 - 30.1}{1.4} = -2$$

Angie's time is 1.25 standard deviations **below** her team's mean time and Beth is 2 standard deviations **below** her team's time. So, Angie had the faster time when compared to her team's mean than Beth's time is to hers.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
 - Approximately 95% of the data is within two standard deviations of the mean.
 - More than 99% of the data is within three standard deviations of the mean.
 - This is known as the Empirical Rule.
 - It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric.
-

Concept Review

The standard deviation measures the average spread of the data about the mean. There are different equations to use if are calculating the standard deviation of a sample or of a population. The standard deviation allows us to compare individual data or to the mean of the data numerically.

- The formula for calculating a sample standard deviation is $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$.
 - The formula for calculating a population standard deviation is $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$.
-

Attribution

“2.7 Measures of the Spread of the Data“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

2.7 EXERCISES

1. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in the table. Construct a line graph.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

2. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in the table. Construct a line graph.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

3. Several children were asked how many TV shows they watch each day. The results of the survey are shown in the table. Construct a line graph.

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

4. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The table shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

5. Using the data from Mrs. Ramirez's math class supplied in the tables, construct a bar graph showing the percentages.

6. David County has six high schools. Each school sent students to participate in a county-wide science competition. The table shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

7. Use the data from the David County science competition supplied in the table above. Construct a bar graph that shows the county-wide population percentage of students at each school.

8. The table contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

- Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
- Construct a bar graph for all the states beginning with the letter “A.”
- Construct a bar graph for all the states beginning with the letter “M.”

9. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency
---------------------	-----------	--------------------	-------------------------------

- a. What does the frequency column sum to? Why?
- b. What does the relative frequency column sum to? Why?
- c. What is the difference between relative frequency and frequency for each data value?
- d. What is the difference between cumulative relative frequency and relative frequency for each data value?
- e. To construct the histogram for the data, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

10. Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10– 13	1
14– 17	0
18– 21	15
22– 25	7
26– 29	2

11. Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230– 259	21
260– 289	13
290– 319	5
320– 349	7
350– 379	1
380– 409	1
410– 439	1

12. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlaid frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49– 55	3
56– 62	3
63– 69	1
70– 76	3
77– 83	8
84– 90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

13. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

14. The following data sets list full-time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

- Construct a double time series graph using a common x -axis for both sets of data.
- Which variable increased the fastest? Explain.
- Did Detroit's increase in police officers have an impact on the murder rate? Explain.

15. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher C

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

- Find the relative frequencies for each survey. Write them in the charts.
 - Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
 - In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
 - Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
 - Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
 - Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.
16. Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions.

Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Singles

Amount(\$)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Couples

Amount(\$)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

- Fill in the relative frequency for each group.
- Construct a histogram for the singles group. Scale the x -axis by \$50 widths. Use relative frequency on the y -axis.
- Construct a histogram for the couples group. Scale the x -axis by \$50 widths. Use relative frequency on the y -axis.
- Compare the two graphs:

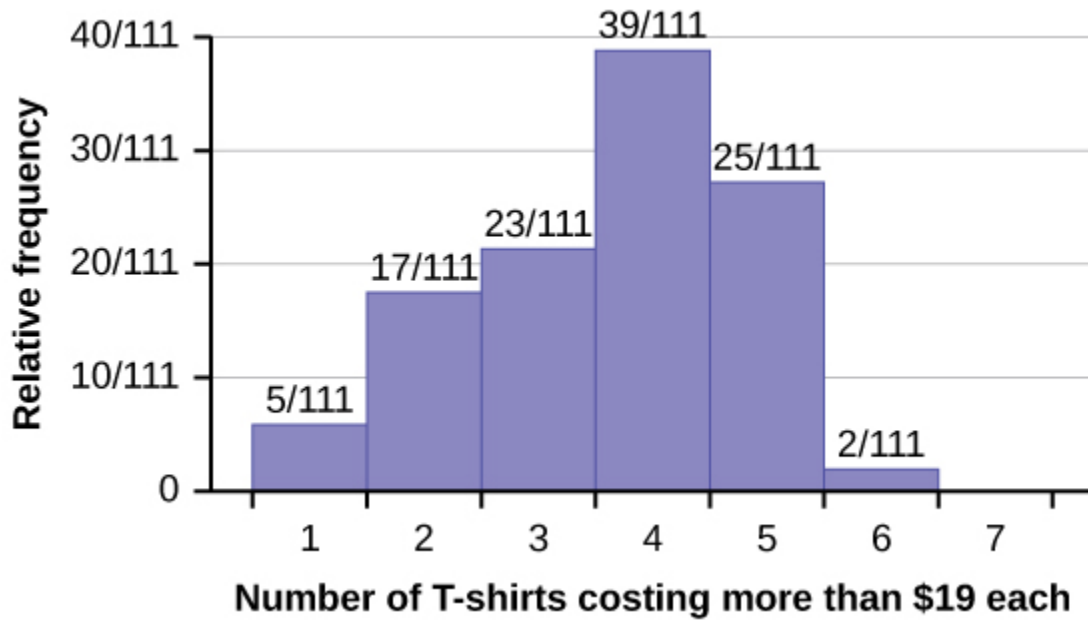
- i. List two similarities between the graphs.
 - ii. List two differences between the graphs.
 - iii. Overall, are the graphs more similar or different?
- e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the x -axis by \$50, scale it by \$100. Use relative frequency on the y -axis.
- f. Compare the graph for the singles with the new graph for the couples:
- i. List two similarities between the graphs.
 - ii. Overall, are the graphs more similar or different?
- g. How did scaling the couples graph differently change the way you compared it to the singles graph?
- h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

17. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

- a. Construct a histogram of the data.
- b. Complete the columns of the chart.

18. Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than \$19 each.



- The percentage of people who own at most three t-shirts costing more than \$19 each is approximately:
- If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

19. Following are the 2010 obesity rates by U.S. states and Washington, DC. Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x -axis with the states.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

20. Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- Find the 40th percentile.
- Find the 78th percentile.

21. Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest*. 18;

18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- a. Find the percentile of 37.
- b. Find the percentile of 72.

22.

- a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- b. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?
- c. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
- d. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.
- e. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- f. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?
- g. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

23. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

24. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

25. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

26. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

27. The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
 - b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?
28. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is \$240,000 in the town you want to move to.
- a. In this town, can you afford 34% of the houses or 66% of the houses?
 - b. Calculate the following:
 - i. First quartile
 - ii. Second quartile
 - iii. Third quartile
 - iv. Interquartile range (*IQR*)
 - v. 10th percentile
 - vi. 70th percentile
29. The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years. Based upon this information, give two reasons why the black median age could be lower than the white median age. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not? How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?
30. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in the table. Also, include left endpoint, but not the right endpoint.

Salary (\$)	Relative Frequency
under 20,000	0.02
20,000 – 25,000	0.09
25,000 – 30,000	0.19
30,000 – 40,000	0.26
40,000 – 50,000	0.18
50,000 – 75,000	0.17
75,000 – 99,999	0.02
100,000 or more	0.01

- What percentage of the survey answered “not sure”?
 - What percentage think that middle-class is from \$25,000 to \$50,000?
 - Construct a histogram of the data.
 - Should all bars have the same width, based on the data? Why or why not?
 - How should the <20,000 and the 100,000+ intervals be handled? Why?
 - Find the 40th and 80th percentiles
 - Construct a bar graph of the data
31. Find the mean for the following frequency tables:

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

32. The following data shows the lengths of boats moored in a marina:

16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35;
37; 39; 40

- Calculate the mean.
- Calculate the median.
- Find the mode.

33. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

- Mean
- Median
- Mode

34. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

- What is the best estimate of the average obesity percentage for these countries?
- The United States has an average obesity rate of 33.9%. Is this rate above average or below?
- How does the United States compare to other countries?

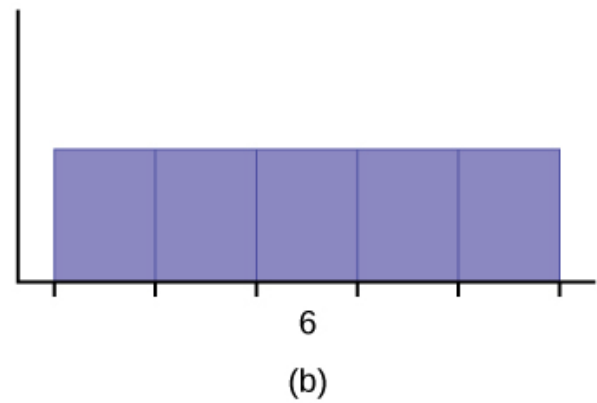
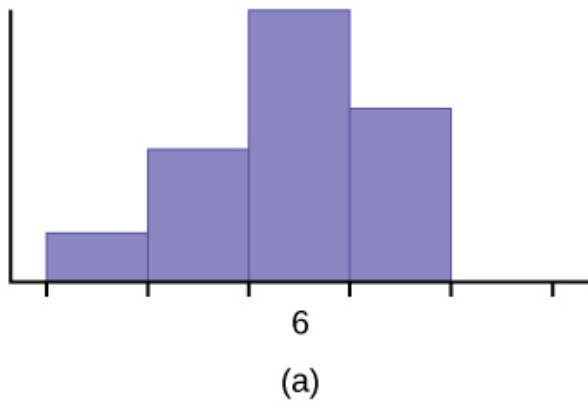
35. The table gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

36. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
\bar{x}	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

- How can you determine which survey was correct?
- Explain what the difference in the results of the surveys implies about the data.
- If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



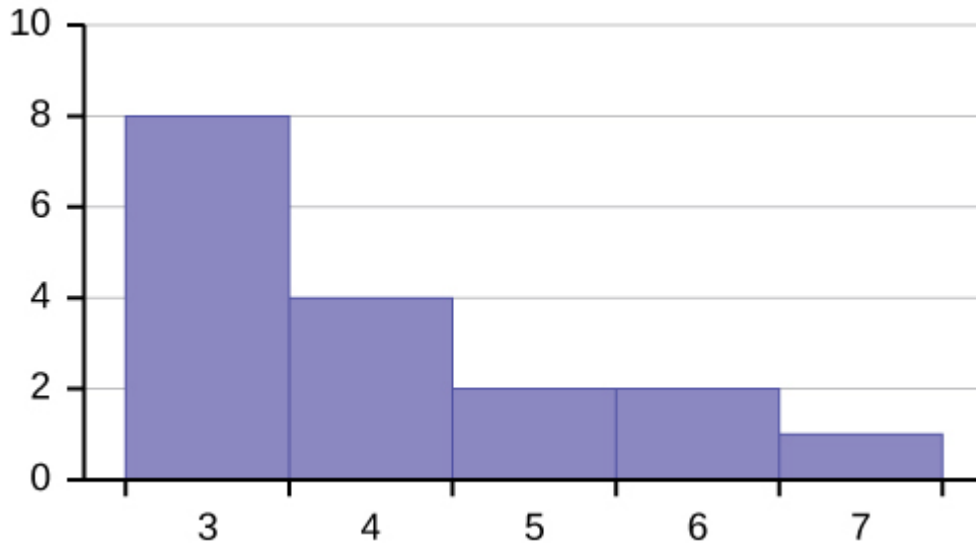
37. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

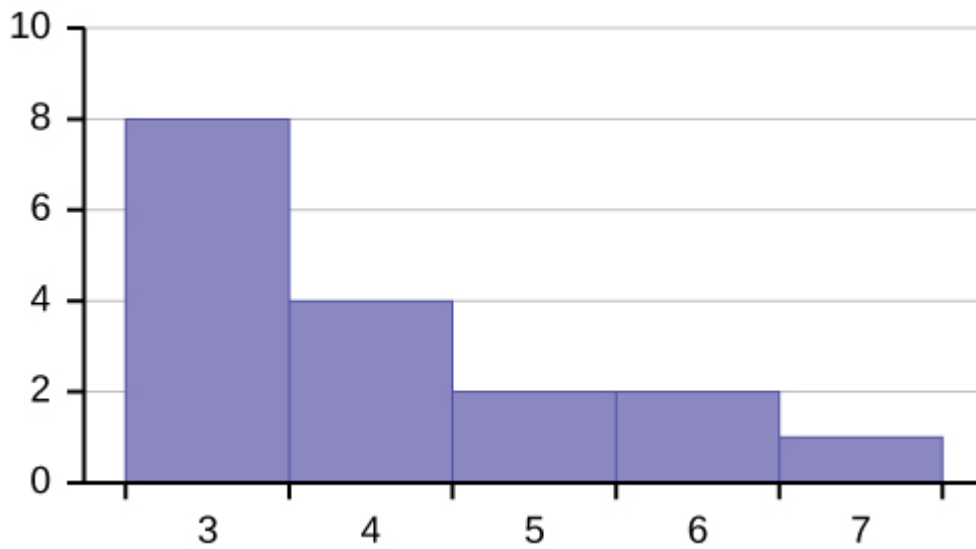
- What is the *IQR*?
 - What is the mode?
 - Is this a sample or the entire population?
38. State whether the data are symmetrical, skewed to the left, or skewed to the right.

- 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5
- 16; 17; 19; 22; 22; 22; 22; 22; 23
- 87; 87; 87; 87; 87; 88; 89; 89; 90; 91

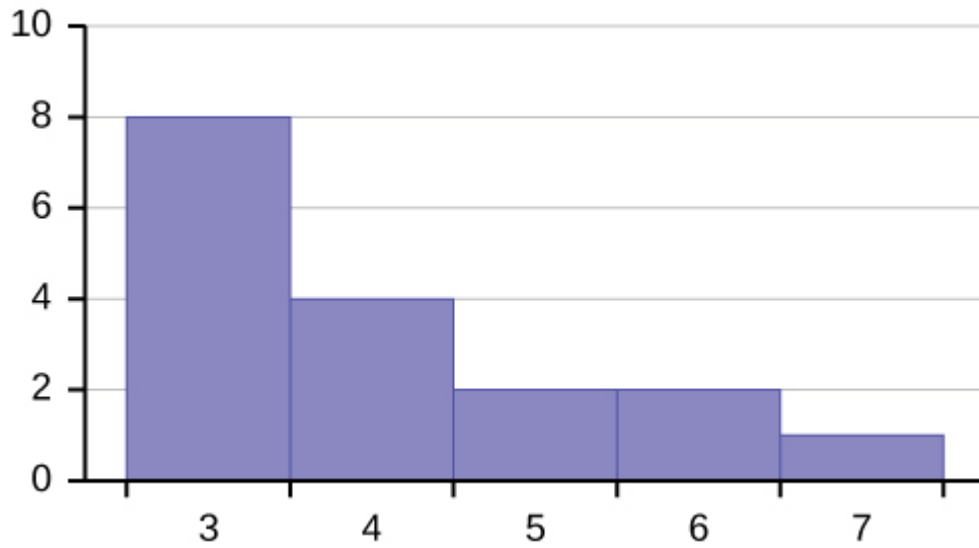
- 39. When the data are skewed left, what is the typical relationship between the mean and median?
- 40. When the data are symmetrical, what is the typical relationship between the mean and median?
- 41. What word describes a distribution that has two modes?
- 42. Describe the shape of this distribution.



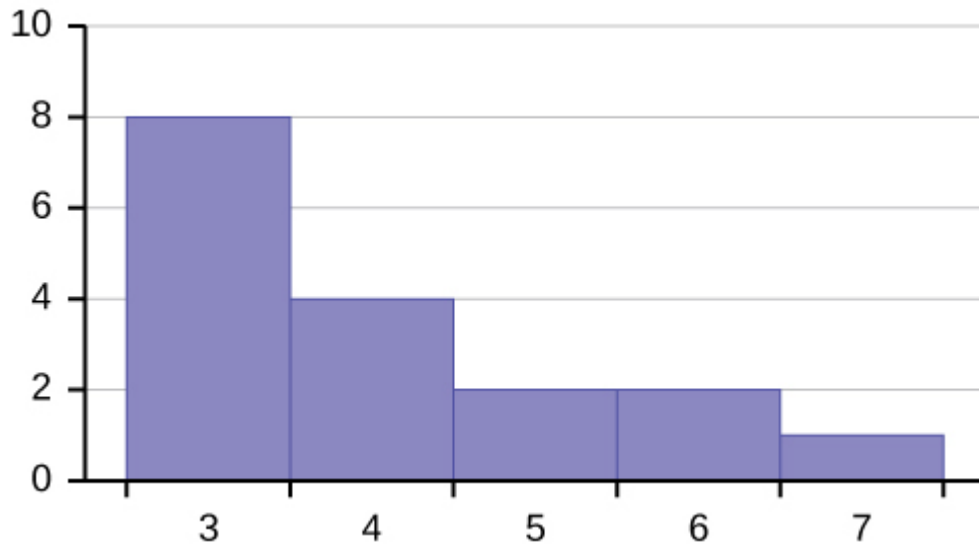
43. Describe the relationship between the mode and the median of this distribution.



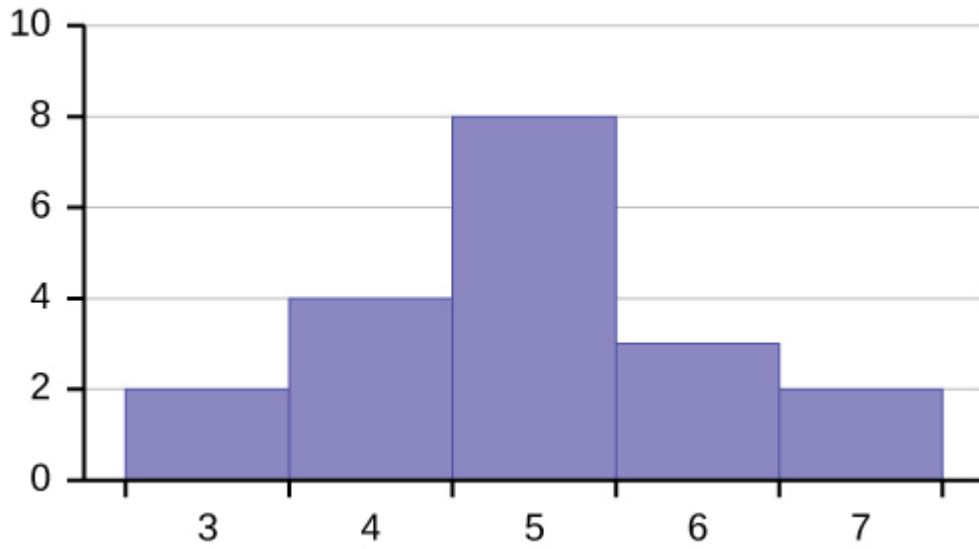
44. Describe the relationship between the mean and the median of this distribution.



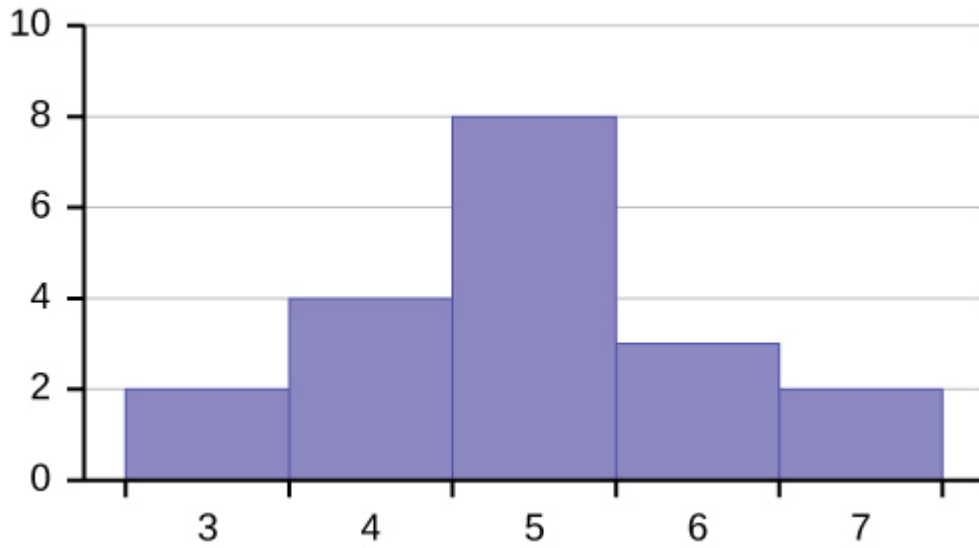
45. Describe the shape of this distribution.



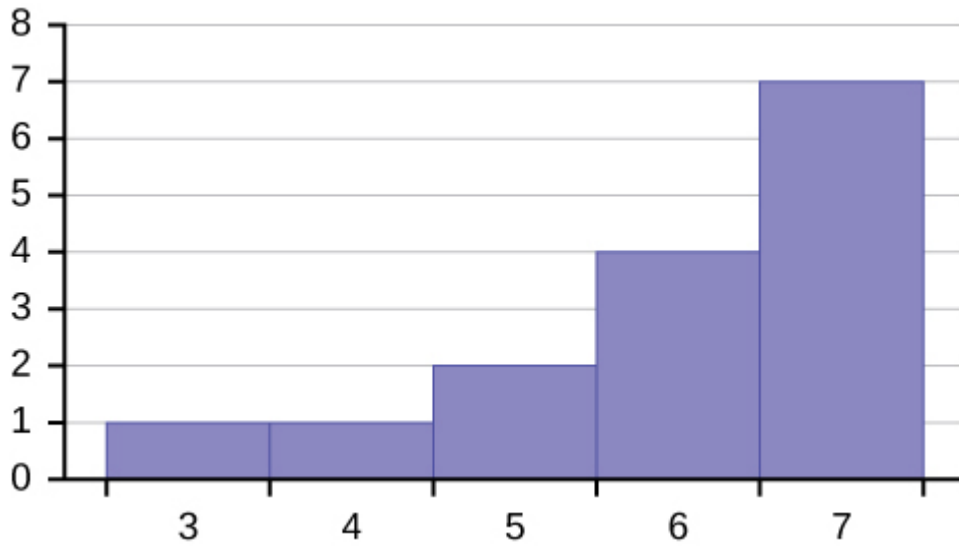
46. Describe the relationship between the mode and the median of this distribution.



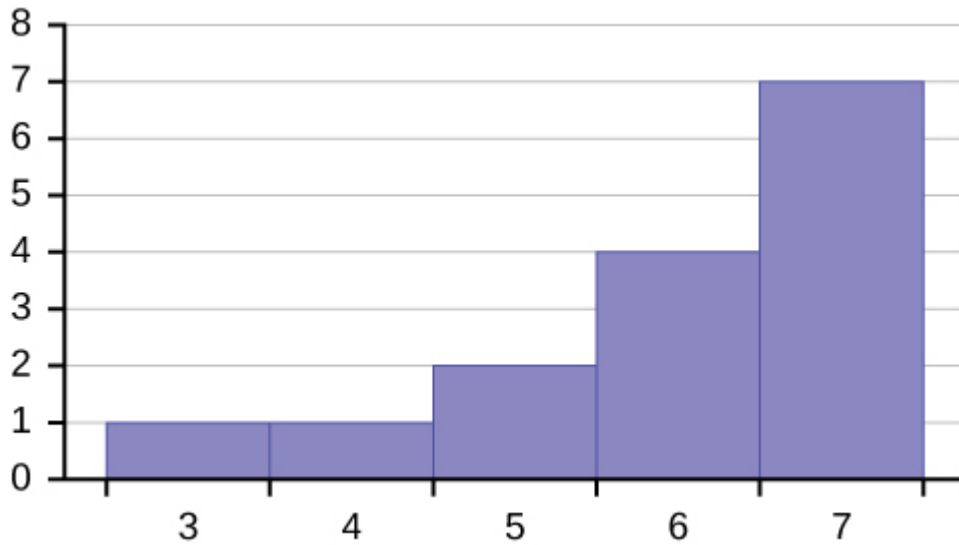
47. Are the mean and the median the exact same in this distribution? Why or why not?



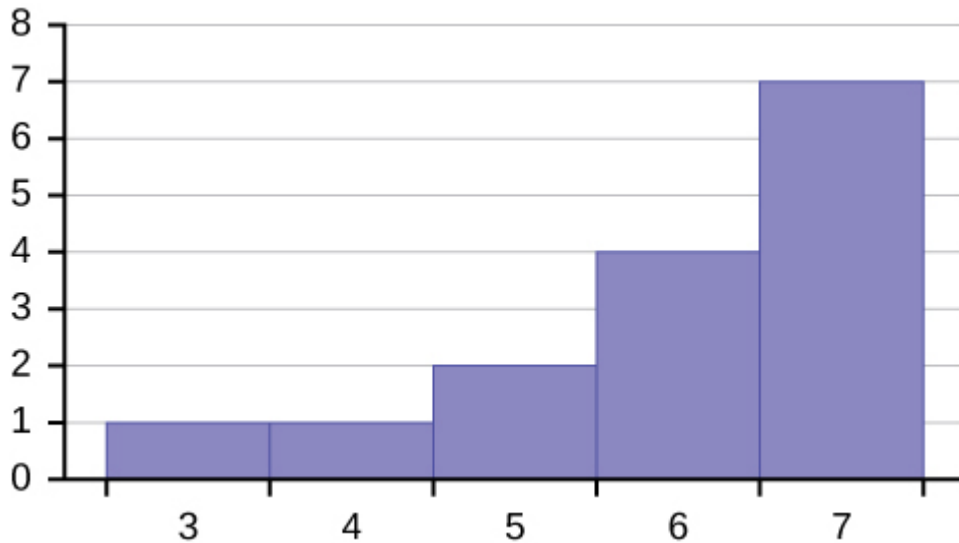
48. Describe the shape of this distribution.



49. Describe the relationship between the mode and the median of this distribution.



50. Describe the relationship between the mean and the median of this distribution.



51. The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

52. Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

53. Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

54. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

55. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

56. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

- What does it mean for the median age to rise?
- Give two reasons why the median age could rise.
- For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

57. The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

- Find the standard deviation and round to the nearest tenth.
- Find the value that is one standard deviation below the mean.

58. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team.

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

- Which baseball player had the higher batting average when compared to his team?
- Use the table above to find the value that is three standard deviations above the mean.
- Use the table below to find the value that is three standard deviations above the mean.

59. Find the standard deviation for the following frequency tables using the formula.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

60. The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median = 1,014 FTES
- $\sigma = 474$ FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- $n = 29$ years

- a. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.
- b. 75% of all years have an FTES:
 - i. at or below what value?
 - ii. at or above what value?
- c. Find the population standard deviation.
- d. What percent of the FTES were from 528.5 to 1447.5? How do you know?
- e. What is the *IQR*? What does the *IQR* represent?
- f. How many standard deviations away from the mean is the median?

The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

- g. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

- h. Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.
- i. Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005–2006 through 2010–2011. Why do you suppose the *IQRs* are so different?

61. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

62. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing \$3,000, a guitar costing \$550, and a drum set costing \$600. The mean cost for a piano is \$4,000 with a standard deviation of \$2,500. The mean cost for a guitar is \$500 with a standard deviation of \$200. The mean cost for drums is \$700 with a standard deviation of \$100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

63. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran one mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

- Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
- Who is the fastest runner with respect to his or her class? Explain why.

64. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4– 20.45	29
20.45– 29.45	13
29.45– 38.45	4
38.45– 47.45	0
47.45– 56.45	2
56.45– 65.45	1
65.45– 74.45	0
74.45– 83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

65. The table gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16– 21.45	23
21.45– 26.9	4
26.9– 32.35	9
32.35– 37.8	7
37.8– 43.25	6
43.25– 48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

66. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

- Find the sample mean \bar{x} .
- Find the approximate sample standard deviation, s .

67. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

- Find the sample mean \bar{x}
- Find the sample standard deviation, s
- Construct a histogram of the data.
- Complete the columns of the chart.
- Find the first quartile.
- Find the median.
- Find the third quartile.
- Construct a box plot of the data.
- What percent of the students owned at least five pairs?
- Find the 40th percentile.
- Find the 90th percentile.

- l. Construct a line graph of the data
- m. Construct a stemplot of the data

68. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212;
215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280;
285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

- a. Organize the data from smallest to largest value.
- b. Find the median.
- c. Find the first quartile.
- d. Find the third quartile.
- e. Construct a box plot of the data.
- f. The middle 50% of the weights are from _____ to _____.
- g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- i. Assume the population was the San Francisco 49ers. Find:
 - i. the population mean, μ .
 - ii. the population standard deviation, σ .
 - iii. the weight that is two standard deviations below the mean.
 - iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
- j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmitt Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

69. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; -1; 2; 0; 5; -3; 1; -1; 6; 5; -2

- a. What is the mean change score?
- b. What is the standard deviation for this population?

- c. What is the median change score?
- d. Find the change score that is 2.2 standard deviations below the mean.

70. In a recent issue of the IEEE SPECTRUM, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- a. Organize the data in a chart.
- b. Find the median, the first quartile, and the third quartile.
- c. Find the 65th percentile.
- d. Find the 10th percentile.
- e. Construct a box plot of the data.
- f. The middle 50% of the conferences last from _____ days to _____ days.
- g. Calculate the sample mean of days of engineering conferences.
- h. Calculate the sample standard deviation of days of engineering conferences.
- i. Find the mode.
- j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

71. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750
; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768;
7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

- a. Organize the data into a chart with five intervals of equal width. Label the two columns “Enrollment” and “Frequency.”
- b. Construct a histogram of the data.
- c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- d. Calculate the sample mean.
- e. Calculate the sample standard deviation.
- f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

72. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

- The 80th percentile is _____
- The number that is 1.5 standard deviations BELOW the mean is approximately _____

73. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

- Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
- If a data value is identified as an outlier, what should be done about it?
- Are any data values further than two standard deviations away from the mean? In some

situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)

- d. Do parts a and c of this problem give the same answer?
 - e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
 - f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?
-

Attribution

“Chapter 2 Homework” and “Chapter 2 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART III

PROBABILITY

Chapter Outline

3.1 Introduction to Probability

3.2 The Terminology of Probability

3.3 Contingency Tables

3.4 The Complement Rule

3.5 The Additional Rule

3.6 Conditional Probability

3.7 Joint Probabilities

3.8 Exercises

3.1 INTRODUCTION TO PROBABILITY



Meteor showers are rare, but the probability of them occurring can be calculated, photo by Ed Sweeney, CC BY 4.0.

It is often necessary to “guess” about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Attribution

“Chapter 3 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.2 THE TERMINOLOGY OF PROBABILITY

LEARNING OBJECTIVES

- Understand and use the terminology of probability.

Every day, decisions are made that involve uncertainty about the outcome. The ability to estimate and understand probability helps us make good decisions. **Probability** is a numerical measure that is associated with how certain we are of outcomes of a particular experiment or activity. Examples of probability used in every day life include the probability that it will rain today and the probability of winning the lottery.

An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance experiment**. An experiment is any activity where the outcome is uncertain. Flipping a coin, rolling a pair of dice, or drawing a card from a deck of cards are all examples of an experiment.

A result of an experiment is called an **outcome**. For example, in the experiment of flipping a coin, a possible outcome is getting heads. The **sample space** of an experiment is the set of all possible outcomes of that experiment. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, in the experiment of flipping a coin, the sample space has two outcomes: heads or tails. In the notation of probability, we would write the sample space of flipping a coin like $S = \{H, T\}$ where H is heads and T is tails.

An **event** is any combination of outcomes. Generally, an event is a collection of outcomes that possess some trait or characteristic. Upper case letters like A and B are used to represent events. For example, if the experiment is to flip a coin two times, event A might be getting at most one head in the two flips. In probability, we are interested in finding the probability of an event. The probability of an event A is written $P(A)$.

EXAMPLE

Suppose a coin is flipped two times.

1. What is the sample space for this experiment?
2. Identify all of the outcomes in the event “exactly one head.”
3. Identify all of the outcomes in the event “at least one tail.”

Solution:

1. $S = \{HH, HT, TH, TT\}$ where H is heads and T is tails. For example, the outcome HT means heads on the first flip and tails on the second flip.
2. The outcomes in the event “exactly one head” are HT and TH . These are the only outcomes in the sample space S where there is exactly one head in the two flips.
3. The outcomes in the event “at least one tail” are HT , TH , and TT . “At least one” means one or more, so we need to include all of the outcomes in the sample space where there is one or more tails.

NOTE

The order in which things happens is important, so the outcomes HT and TH are different outcomes. The outcome HT consists of getting heads on the first flip and tails on the second flip. The outcome TH consists of getting tails on the first flip and heads on the second flip, which is a completely different outcome from HT .

Probability is a numerical measure of the likelihood that an event will occur. The **probability** of an event is the **long-term relative frequency** of that event. Probabilities are numbers between zero and one, inclusive—that is, zero and one and all numbers between these values. Probabilities can be written as fractions, decimals, or percents. $P(A) = 0$ means the event A can never happen—the probability is 0%. $P(A) = 1$ means the event A always happens—the probability is

100%. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur—there is a 50% chance A will happen and a 50% chance A will not happen.

Approaches to Determining Probability

The way that we calculate the probability of an event depends on the situation we are analyzing.

Classical Method Approach to Probability

Most often associated with games of chance, the **classical method approach** requires us to know that the outcomes of an experiment are **equally like to occur**. We have already seen an experiment where the outcomes are equally likely to occur—flipping a coin. **Equally likely** means that each outcome of an experiment occurs with equal probability. In the experiment of tossing a fair coin, you know that you have a 50% chance of getting heads and a 50% chance of getting tails—the outcomes of heads or tails are equally likely to occur. If you roll a fair, six-sided die, you know that you have the same chance $\left(\frac{1}{6}\right)$ of getting any of the six faces—the outcomes of 1, 2, 3, 4, 5, 6 are equally likely to occur.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space.

$$P(A) = \frac{\text{number of outcomes in event } A}{\text{total number of outcomes in the sample space}}$$

EXAMPLE

Suppose a coin is flipped two times.

1. What is the probability of getting “exactly one head?”
2. What is the probability of getting “at least one tail?”

Solution:

Previously, we found the sample space for this experiment: $S = \{HH, HT, TH, TT\}$.

1. The outcomes in the event “exactly one head” are HT and TH . We see that there are 2 outcomes in the event out of the 4 possible outcomes in the sample space. So

$$P(\text{exactly one head}) = \frac{2}{4} = 0.5$$

2. The outcomes in the event “at least one tail” are HT , TH , and TT . We see that there are 3 outcomes in the event out of the 4 possible outcomes in the sample space. So

$$P(\text{at least one tail}) = \frac{3}{4} = 0.75$$

TRY IT

Suppose you roll a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 on the faces.

1. What is the sample space for this experiment?
2. What is the probability of getting at least 5?
3. What is the probability of getting an even number?
4. What is the probability of getting a number less than 4?
5. What is the probability of getting a 7?

Click to see Solution

1. $S = \{1, 2, 3, 4, 5, 6\}$
2. $P(\text{at least } 5) = \frac{2}{6} = 0.3333\dots$

3. $P(\text{even number}) = \frac{3}{6} = 0.5$
4. $P(\text{less than 4}) = \frac{3}{6} = 0.5$
5. $P(7) = \frac{0}{6} = 0$

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair** or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin, but more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home. The spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased because it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces and the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur.

Empirical Method Approach to Probability

The **empirical** or **relative frequency approach** to probability uses results from identical previous experiments that have been performed many times. Probabilities are based on historical or previously recorded data by determining the proportion of times an event occurs within the data. For example, a retail business owner might want to know the probability that a customer spends more than \$50 at their store. To determine this probability, the business owner would look at previous sales, count the number of sales over \$50 and then divide that number by the total number of previous sales.

To calculate an empirical probability, repeat the experiment over a large number of trials and record the result of each trial. To find the probability of event A , count the number of times event A happened and divide by the total number of trials.

$$P(A) = \frac{\text{number of times } A \text{ occurs}}{\text{total number of trials}}$$

To get an accurate probability using this approach, it is important that the experiment is repeated a very large number of times. This important characteristic of probability experiments is known as the **law of large numbers**, which states that as the number of repetitions of an experiment increases, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability.

EXAMPLE

An online retailer wants to know the probability that a transaction will be less than \$30. In 2000 transactions, 650 are less than \$30.

Solution:

$$P(\text{less than } \$30) = \frac{650}{2000} = 0.325$$

Subjective Method Approach to Probability

In the subjective method approach to probability, probabilities are determined by educated guess, personal belief, intuition, or expert reasoning. A subjective probability is essentially a guess, but a guess based on an accumulation of knowledge, understanding, and experience. Estimating the probability the price of a stock goes down over time or the probability a certain sports team will win a championship are examples of subjective probability.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=77#oembed-1>

Watch this video: Probability: Tossing Two Coins by Joshua Emmanuel [5:55] (transcript available).

Concept Review

In this section we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Attribution

“3.1 Terminology“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.3 CONTINGENCY TABLES

LEARNING OBJECTIVES

- Construct and interpret contingency tables.

A **contingency table** provides a way of displaying data that can facilitate calculating probabilities. The table can be used to describe the sample space of an experiment. Contingency tables allow us to break down a sample space when two variables are involved.

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

When reading a contingency table:

- The left-side column lists all of the values for one of the variables. In the table shown above, the left-side column shows the variable about whether or not someone uses a cell phone while driving.
- The top row lists all of the values for the other variable. In the table shown above, the top row shows the variable about whether or not someone had a speeding violation in the last year.
- In the body of the table, the cells contain the number of outcomes that fall into both of the categories corresponding to the intersecting row and column. In the table shown above, the number of 25 at the intersection of the “cell phone user” row and “speeding violation in the last year” column tells us that there are 25 people who have both of these characteristics.

- The bottom row gives the totals in each column. In the table shown above, the number 685 in the bottom of the “no speeding violation in the last year” tells us that there are 685 people who did not have a speeding violation in the last year.
- The right-side column gives the totals in each row. In the table shown above, the number 305 in the right side of the “cell phone user” row tells us that there are 305 people who use cell phones while driving.
- The number in the bottom right corner is the size of the sample space. In the table shown above, the number in the bottom right corner is 755, which tells us that there 755 people in the sample space.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones while driving produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Calculate the following probabilities:

1. What is the probability that a randomly selected person is a cell phone user?
2. What is the probability that a randomly selected person had no speeding violations in the last year?
3. What is the probability that a randomly selected person had a speeding violation in the last year and does not use a cell phone?
4. What is the probability that a randomly selected person uses a cell phone and had no speeding violations in the last year?

Solution:

1. Probability = $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$
2. Probability = $\frac{\text{number of no violations}}{\text{total number in study}} = \frac{685}{755}$
3. Probability = $\frac{\text{number of violations and not cell phone users}}{\text{total number in study}} = \frac{45}{755}$
4. Probability = $\frac{\text{number of cell phone users and no violations}}{\text{total number in study}} = \frac{280}{755}$

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising?
2. What is the probability that a randomly selected athlete had an injury in the last year?
3. What is the probability that a randomly selected athlete does not stretch before exercising and had no injuries in the last year?
4. What is the probability that a randomly selected athlete stretches before exercising and had no injuries in the last year?

Click to see Solution

1. Probability = $\frac{350}{800} = 0.4375$
2. Probability = $\frac{286}{800} = 0.3575$
3. Probability = $\frac{219}{800} = 0.27375$
4. Probability = $\frac{295}{800} = 0.36875$

EXAMPLE

The table below shows a random sample of 100 hikers broken down by gender and the areas of hiking they prefer.

Gender	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16		45
Male			14	55
Total		41		

1. Fill in the missing values in the table
2. What is the probability that a randomly selected hiker is female?
3. What is the probability that a randomly selected hiker prefers to hike on the coast?
4. What is the probability that a randomly selected hiker is male and prefers to hike near lakes and streams?
5. What is the probability that a randomly selected hiker is female and prefers to hike on mountains?

Solution:

1.

Gender	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

2. Probability = $\frac{45}{100} = 0.45$

3. Probability = $\frac{34}{100} = 0.34$

4. Probability = $\frac{25}{100} = 0.25$

5. Probability = $\frac{11}{100} = 0.11$

TRY IT

The table below relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column.
2. Find the probability that a randomly chosen individual from this group is tall.

3. Find the probability that a randomly chosen individual from this group is normal.
4. Find the probability that a randomly chosen individual from this group is obese and short.
5. Find the probability that a randomly chosen individual from this group is underweight and medium.

Click to see Solution

1.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

2. Probability = $\frac{50}{205}$
3. Probability = $\frac{99}{205}$
4. Probability = $\frac{14}{205}$
5. Probability = $\frac{25}{205}$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=79#oembed-1>

Watch this video: Ex: Basic Example of Finding Probability From a Table by Mathispower4u [2:39] (transcript available).

Concept Review

There are several tools we can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have two variables of interest.

Attribution

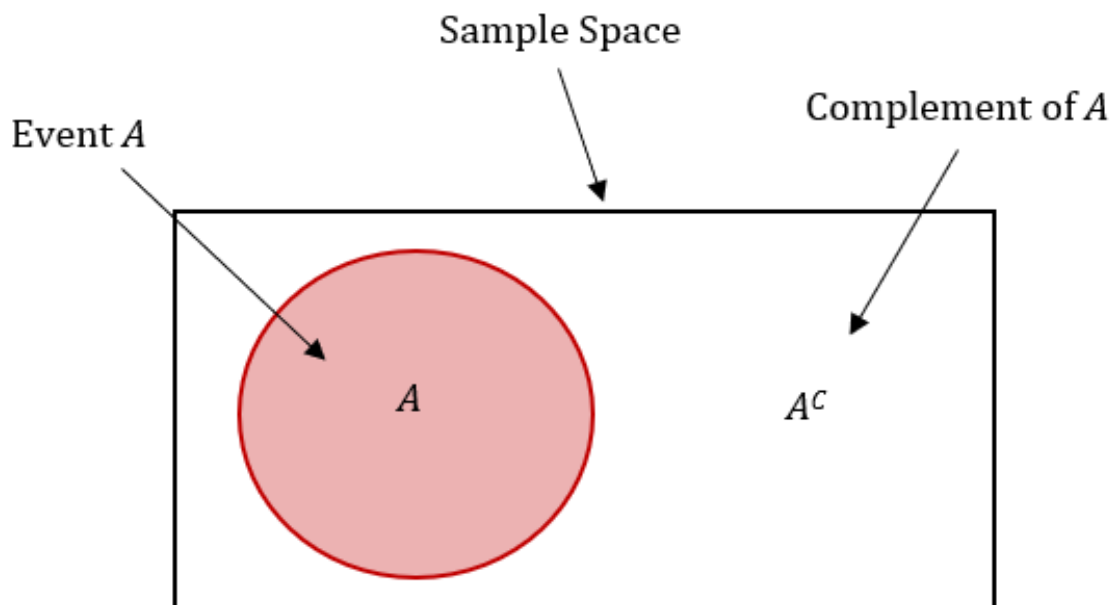
“3.4 Contingency Tables“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.4 THE COMPLEMENT RULE

LEARNING OBJECTIVES

- Calculate probabilities using the complement rule.

The **complement** of an event A is the set of all outcomes in the sample space that are not in A . The complement of A is denoted by A^C and is read “not A .”



EXAMPLE

Suppose a coin is flipped two times. Previously, we found the sample space for this experiment: $S = \{HH, HT, TH, TT\}$ where H is heads and T is tails.

1. What is the complement of the event “exactly one head”?
2. What is the complement of the event “at least one tail.”

Solution:

1. The event “exactly one head” consists of the outcomes HT and TH . The **complement** of “exactly one head” consists of the outcomes HH and TT . These are the outcomes in the sample space S that are NOT in the original event “exactly one head.”
2. The event “at least one tail” consists of the outcomes HT , TH , and TT . The **complement** of “at least one tail” consists of the outcomes HH . These are the outcomes in the sample space S that are NOT in the original event “at least one tail.”

TRY IT

Suppose you roll a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 on the faces. Previously, we found the sample space for this experiment: $S = \{1, 2, 3, 4, 5, 6\}$

1. What is the complement of the event “rolling a 4”?
2. What is the complement of the event “rolling a number greater than or equal to 5”?
3. What is the complement of the event “rolling a even number”?
4. What is the complement of the event “rolling a number less than 4”?

Click to see Solution

1. The complement is $\{1, 2, 3, 5, 6\}$.
2. The complement is $\{1, 2, 3, 4\}$.
3. The complement is $\{1, 3, 5\}$.
4. The complement is $\{1, 2, 3, 4\}$.

The Probability of the Complement

In any experiment, an event A or its complement A^C must occur. This means that $P(A) + P(A^C) = 1$. Rearranging this equation gives us a formula for finding the probability of the complement from the original event:

$$P(A^C) = 1 - P(A)$$

EXAMPLE

An online retailer knows that 30% of customers spend more than \$100 per transaction. What is the probability that a customer spends at most \$100 per transaction?

Solution:

Spending at most \$100 (\$100 or less) per transaction is the complement of spending more than \$100 per transaction.

$$\begin{aligned}
 P(\text{at most } \$100) &= 1 - P(\text{more than } \$100) \\
 &= 1 - 0.3 \\
 &= 0.7
 \end{aligned}$$

TRY IT

At a local college, a statistics professor has a class of 80 students. After polling the students in the class, the professor finds out that 15 of the students play on one of the school's sports team and 60 of the students have part-time jobs.

1. What is the probability that a student in the class does not play on one of the school's sports teams?
2. What is the probability that a student in the class does not have a part-time job?

Click to see Solution

1. $P(\text{no sports team}) = 1 - P(\text{sports team}) = 1 - \frac{15}{80} = 0.8125$
2. $P(\text{no part-time job}) = 1 - P(\text{part-time job}) = 1 - \frac{60}{80} = 0.25$

Concept Review

The complement, A^C , of an event A consists of all of the outcomes in the sample space that are NOT in event A . The probability of the complement can be found from the original event using the formula: $P(A^C) = 1 - P(A)$.

Attribution

“3.1 Terminology” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.5 THE ADDITION RULE

LEARNING OBJECTIVES

- Calculate “or” probabilities using the addition rule.
- Determine if two events are mutually exclusive.

For two events A and B we might want to know the probability that at least one of the two events occurs. For example, we might want to find the probability of rolling a 2 or a 5 in a single roll of a die, or we might want to find the probability that someone has a smartphone or a tablet. In probability terms, we want to find $P(A \text{ or } B)$, the probability that either A or B occurs. In probability, “or” is always an **inclusive** “or,” which means that either A occurs, or B occurs, or both occur.

The Addition Rule for Or Probabilities

To find $P(A \text{ or } B)$, we start by adding the individual probabilities, $P(A)$ and $P(B)$. But this means that the overlap between the two events A and B is counted **twice**: once by $P(A)$ and once by $P(B)$. To correct for this double counting, we need to subtract $P(A \text{ and } B)$, the probability of both events occurring. This gives us the addition rule to find $P(A \text{ or } B)$:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

EXAMPLE

At a local language school, 40% of the students are learning Spanish, 20% of the students are learning German, and 8% of the students are learning both Spanish and German. What is the probability that a randomly selected student is learning Spanish or German?

Solution:

$$\begin{aligned} P(\text{Spanish or German}) &= P(\text{Spanish}) + P(\text{German}) - P(\text{Spanish and German}) \\ &= 0.4 + 0.2 - 0.08 \\ &= 0.52 \end{aligned}$$

EXAMPLE

There are 50 students enrolled in the second year of a business degree program. During this semester, the students have to take some elective courses. 18 students decide to take an elective in psychology, 27 students decide to take an elective in philosophy, and 10 students decide to take an elective in both psychology and philosophy. What is the probability that a student takes an elective in psychology or philosophy?

Solution:

$$\begin{aligned} P(\text{psychology or philosophy}) &= P(\text{psychology}) + P(\text{philosophy}) - P(\text{psychology and philosophy}) \\ &= \frac{18}{50} + \frac{27}{50} - \frac{10}{50} = 0.7 \end{aligned}$$

TRY IT

At a local basketball game, 70% of the fans are cheering for the home team, 25% of the fans are wearing blue, and 12% of the fans are cheering for the home team and wearing blue. What is the probability that a randomly selected fan is cheering for the home team or wearing blue?

Click to see Solution

$$\begin{aligned}
 P(\text{home team or blue}) &= P(\text{home team}) + P(\text{blue}) - P(\text{home team and blue}) \\
 &= 0.7 + 0.25 - 0.12 \\
 &= 0.83
 \end{aligned}$$

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user or has no speeding

violations in the last year?

2. What is the probability that a randomly selected person had a speeding violation in the last year or does not use a cell phone?

Solution:

1.
$$P(\text{cell phone or no violations}) = P(\text{cell phone}) + P(\text{no violations}) - P(\text{cell phone and no violations})$$

$$= \frac{305}{755} + \frac{685}{755} - \frac{280}{755} = \frac{710}{755}$$
2.
$$P(\text{violations or no cell phone}) = P(\text{violations}) + P(\text{no cell phone}) - P(\text{violations and no cell phone})$$

$$= \frac{70}{755} + \frac{450}{755} - \frac{45}{755} = \frac{475}{755}$$

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising or had an injury last year?
2. What is the probability that a randomly selected athlete does not stretch before exercising or had no injuries in the last year?

Click to see Solution

1. Probability = $\frac{350}{800} + \frac{286}{800} - \frac{55}{800} = 0.72625$
2. Probability = $\frac{450}{800} + \frac{514}{800} - \frac{219}{800} = 0.93125$

Mutually Exclusive Events

Two events A and B are **mutually exclusive** if the two events cannot happen at the same time. That is, the events A and B do not share any outcomes and so $P(A \text{ and } B) = 0$. For example, in the experiment of flipping a coin, the events heads and tails are mutually exclusive because it is not possible to have both heads and tails on the top face. In the case of mutually exclusive events, the addition rule is $P(A \text{ or } B) = P(A) + P(B)$.

EXAMPLE

Suppose a bag contains 20 balls. 10 of the balls are white, 7 of the balls are red, and 3 of the balls are blue. Suppose one ball is selected at random from the bag.

1. Are the events “selecting a white ball” and “selecting a red ball” mutually exclusive? Why?
2. What is the probability of selecting a white or red ball?

Solution:

1. The events “selecting a white ball” and “selecting a red ball” are mutually exclusive because the events cannot happen at the same time. It is not possible for the selected ball to be both white and red.
2.
$$P(\text{white or red}) = P(\text{white}) + P(\text{red}) = \frac{10}{20} + \frac{7}{20} = 0.85$$

NOTE

In the calculation of the probability in part 2, there is nothing to subtract. Because the events are mutually exclusive, $P(\text{white and red}) = 0$.

TRY IT

At a local college, 60% of the students are taking a math class, 50% of the students are taking a science class, and 30% of the students are taking both a math and a science class.

1. Are the events “taking a math class” and “taking a science class” mutually exclusive? Explain.
2. What is the probability that a randomly selected student is taking a math class or a science class?

Click to see Solution

1. The events “taking a math class” and “taking a science class” are not mutually exclusive because the events can happen at the same time (i.e. a student can be taking both a math class and a science class). As stated in the question, $P(\text{math and science}) = 0.3 \neq 0$.
2.
$$P(\text{math or science}) = P(\text{math}) + P(\text{science}) - P(\text{math and science}) = 0.6 + 0.5 - 0.3 = 0.8$$

TRY IT

You roll a fair die one time.

1. Are the events “rolling a 4” and “rolling an even number” mutually exclusive?
2. Are the events “rolling a 4” and “rolling an odd number” mutually exclusive?
3. What is the probability of rolling a 4 or rolling an odd number.

Click to see Solution

1. The events “rolling a 4” and “rolling an even number” are not mutually exclusive because the events can happen at the same time (i.e. 4 is an even number).
2. The events “rolling a 4” and “rolling an odd number” are mutually exclusive because the events cannot happen at the same time. It is not possible to roll a die and get a 4 (an even number) and an odd number on the top face at the same time

$$\begin{aligned} \text{3. } & P(\text{4 or odd}) = P(\text{4}) + P(\text{odd}) = \\ & \frac{1}{6} + \frac{3}{6} = \frac{4}{6} \end{aligned}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=84#oembed-1>

Watch this video: Addition Rule for Probability Khan Academy [10:42] (transcript available).

Concept Review

To find the probability of events A or B , use the addition rule:

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. Two events are mutually exclusive if the events cannot happen at the same time.

Attribution

“3.1 Terminology”, “3.2 Independent and Mutually Exclusive Events”, and “3.4 Contingency Tables” in *Introductory Statistics* by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.6 CONDITIONAL PROBABILITY

LEARNING OBJECTIVES

- Calculate conditional probabilities.
- Determine if two events are independent.

A **conditional probability** is the probability of an event A **given** that another event B has already occurred. The idea behind conditional probability is that it reduces the sample space to the part of the sample space that involves just the given event B —except for the event B , everything else in the sample space is thrown away. Once the sample space is reduced to the given event B , we calculate the probability of A occurring within the reduced sample space.

The conditional probability of A given B is written as $P(A|B)$ and is read “the probability of A given B .”

Recognizing a conditional probability and identifying which event is the given event can be challenging. The following sentences are all asking the same conditional probability just in different ways:

- What is the probability a student has a smartphone given that the student has a tablet?
- If a student has a tablet, what is the probability the student has a smartphone?
- What is the probability that a student with a tablet has a smartphone?

The given event is “has a tablet,” so in calculating the conditional probability we would restrict the sample space to just those students that have a tablet and then find the probability a student has a smartphone from among just those students with a tablet.

NOTE

The conditional probability $P(A|B)$ is **NOT** the same as $P(A \text{ and } B)$.

- In the conditional probability $P(A|B)$ we want to find the probability of A occurring **after** B has already happened. In the conditional probability the sample space is restricted to just event B before we calculate the probability of A in the restricted sample space.
- In $P(A \text{ and } B)$ we want to find the probability of events A and B happening **at the same time** in the unrestricted sample space.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user given that they had no speeding violations in the last year?
2. If a randomly selected person does not have a cell phone, what is the probability they had a speeding violation last year?
3. What is the probability that someone with a cell phone did not have a speeding violation last year?

Solution:

1. The given event is “no speeding violations,” so we restrict the table to just the column involving “no speeding violations.” With this restriction, the table would look like this:

	No speeding violation in the last year
Cell phone user	280
Not a cell phone user	405
Total	685

Now, we want to find the probability a person is a cell phone user in this restricted sample space:

$$P(\text{cell phone} | \text{no violations}) = \frac{\text{number of cell phone users in restricted sample space}}{\text{total number in restricted sample space}} = \frac{280}{685}$$

2. The given event is “no cell phone,” so we restrict the table to just the row involving “no cell phone.” With this restriction, the table would look like this:

	Speeding violation in the last year	No speeding violation in the last year	Total
Not a cell phone user	45	405	450

Now, we want to find the probability a person has a speeding violation in the last year in this restricted sample space:

$$P(\text{violation} | \text{no cell phone}) = \frac{\text{number of violations in restricted sample space}}{\text{total number in restricted sample space}} = \frac{45}{450}$$

3. The given event is “cell phone,” so we restrict the table to just the row involving “cell phone.” With this restriction, the table would look like this:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305

Now, we want to find the probability a person does not have a speeding violation in the last year in this restricted sample space:

$$\begin{aligned} P(\text{no violations}|\text{cell phone}) &= \frac{\text{number with no violations in restricted sample space}}{\text{total number in restricted sample space}} \\ &= \frac{280}{305} \end{aligned}$$

NOTE

The conditional probability $P(A|B)$ does not equal the conditional probability $P(B|A)$. In the above example, $P(\text{cell phone}|\text{no violations}) = \frac{280}{685}$ **does not equal**

$$P(\text{no violations}|\text{cell phone}) = \frac{280}{305}.$$

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is the probability that a randomly selected athlete stretches before exercising given that

they had an injury last year?

2. What is the probability that a randomly selected athlete that had no injuries in the last year does not stretch before exercising?
3. If a randomly selected athlete does not stretch before exercising, what is the probability they had an injury in the last year?

Click to see Solution

1. Probability = $\frac{55}{286}$
2. Probability = $\frac{219}{514}$
3. Probability = $\frac{231}{450}$

Calculating Conditional Probabilities Using the Formula

When working with a contingency table as in the above examples, we can simply calculate conditional probabilities by restricting the table to the given event and then finding the required probability in the restricted sample space. Depending on the situation, it might not be possible to workout a conditional probability this way. In these situations we can use the following formula to find a conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

EXAMPLE

At a local language school, 40% of the students are learning Spanish, 20% of the students are learning German and 8% of the students are learning both Spanish and German.

1. What is the probability that a randomly selected student is learning Spanish given that they are learning German?
2. What is the probability that a randomly selected Spanish student is learning German?

Solution:

1.
$$P(\text{Spanish}|\text{German}) = \frac{P(\text{Spanish and German})}{P(\text{German})} = \frac{0.08}{0.2} = 0.4$$
2.
$$P(\text{German}|\text{Spanish}) = \frac{P(\text{Spanish and German})}{P(\text{Spanish})} = \frac{0.08}{0.4} = 0.2$$

EXAMPLE

There are 50 students enrolled in the second year of a business degree program. During this semester, the students have to take some elective courses. 18 students decide to take an elective in psychology, 27 students decide to take an elective in philosophy, and 10 students decide to take an elective in both psychology and philosophy.

1. What is the probability that a student takes an elective in psychology given that they take an elective in philosophy?
2. If a student takes an elective in psychology, what is the probability that they take an elective in philosophy?

Solution:

- $$P(\text{psychology} | \text{philosophy}) = \frac{P(\text{psychology and philosophy})}{P(\text{philosophy})} = \frac{\frac{10}{50}}{\frac{27}{50}} = 0.3704$$
1.
$$P(\text{philosophy} | \text{psychology}) = \frac{P(\text{psychology and philosophy})}{P(\text{psychology})} = \frac{\frac{10}{50}}{\frac{18}{50}} = 0.5556$$

TRY IT

At a local basketball game, 70% of the fans are cheering for the home team, 25% of the fans are wearing blue, and 12% of the fans are cheering for the home team and wearing blue.

1. What is the probability that a randomly selected fan is cheering for the home team given that they are wearing blue?
2. If a randomly selected fan is cheering for the home team, what is the probability they are wearing blue?

Click to see Solution

1.
$$P(\text{home team} | \text{blue}) = \frac{0.12}{0.25} = 0.48$$
2.
$$P(\text{blue} | \text{home team}) = \frac{0.12}{0.7} = 0.1714$$

Independent Events

Two events are **independent** if the probability of the occurrence of one of the events does not affect the probability of the occurrence of the other event. In other words, two events A and B are

independent if the knowledge that one of the events occurred does not affect the chance the other event occurs. For example, the outcomes of two rolls of a fair die are independent events—the outcome of the first roll does not change the probability of the outcome of the second roll. If two events are not independent, then we say the events are **dependent**.

We can test two events A and B for independence by comparing $P(A)$ and $P(A|B)$:

- If $P(A) = P(A|B)$, then the events A and B are independent.
- If $P(A) \neq P(A|B)$, then the events A and B are dependent.

EXAMPLE

At a local language school, 40% of the students are learning Spanish, 20% of the students are learning German and 8% of the students are learning both Spanish and German. Are the events “Spanish” and “German” independent? Explain.

Solution:

To check for independence, we need to check two probabilities: $P(\text{Spanish})$ and $P(\text{Spanish}|\text{German})$. If these probabilities are equal, the events are independent. If the probabilities are not equal, the events are dependent.

From the information provided in the question, $P(\text{Spanish}) = 0.4$. Previously, we calculated $P(\text{Spanish}|\text{German})$:

$$\begin{aligned} P(\text{Spanish}|\text{German}) &= \frac{P(\text{Spanish and German})}{P(\text{German})} \\ &= \frac{0.08}{0.2} = 0.4 \end{aligned}$$

We can see that $P(\text{Spanish}) = P(\text{Spanish}|\text{German})$. Because these two probabilities are equal, the events “Spanish” and “German” are independent. This means that the probability a student is taking Spanish does not affect the probability a student is taking German.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Are the events “cell phone user” and “speeding violation in the last year” independent? Explain.

Solution:

To check for independence, we need to check two probabilities: $P(\text{cell phone})$ and $P(\text{cell phone}|\text{speeding violation})$. If these probabilities are equal, the events are independent. If the probabilities are not equal, the events are dependent.

$$\begin{aligned} P(\text{cell phone}) &= \frac{305}{755} \approx 0.4040 \\ P(\text{cell phone}|\text{speeding violation}) &= \frac{25}{70} \approx 0.3571 \end{aligned}$$

We can see that $P(\text{cell phone}) \neq P(\text{cell phone}|\text{speeding violation})$. Because these probabilities are not equal, the events “cell phone user” and “speeding violation” are dependent. This means that the probability a person is a cell phone user does affect the probability the person had a speeding violation in the last year.

TRY IT

At a local basketball game, 70% of the fans are cheering for the home team, 25% of the fans are wearing blue, and 12% of the fans are cheering for the home team and wearing blue. Are the events “cheering for the home team” and “wearing blue” independent? Explain.

Click to see Solution

Because $P(\text{home team}) = 0.7$ does not equal $P(\text{home team}|\text{blue}) = \frac{0.12}{0.25} = 0.48$, the events “cheering for the home team” and “wearing blue” are dependent.

TRY IT

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

Are the events “does not stretch” and “injury in last year” independent? Explain.

Click to see Solution

Because $P(\text{no stretch}) = \frac{450}{800} = 0.5625$ does not equal

$P(\text{no stretch}|\text{injury}) = \frac{219}{514} = 0.4261$, the events “does not stretch” and “injury in last year” are dependent.

Sampling may be done **with replacement** or **without replacement**, which effects whether or not events are considered independent or dependent.

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent because the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. Depending on the situation, the events are considered to be dependent or not independent.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=86#oembed-1>

Watch this video: Calculating Conditional Probability Khan Academy [6:42] (transcript available).





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=86#oembed-2>

Watch this video: Conditional Probability and Independence Khan Academy [4:06] (transcript available).

Concept Review

A conditional probability is the probability of an event A given that another event B has already occurred. The formula to find a conditional probability is: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. Two events

A and B are independent if the knowledge that one of the events occurred does not affect the chance that the other event occurs. If $P(A) = P(A|B)$, the events A and B are independent. Otherwise the events are dependent.

Attribution

“3.1 Terminology”, “3.2 Independent and Mutually Exclusive Events”, and “3.4 Contingency Tables” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.7 JOINT PROBABILITIES

LEARNING OBJECTIVES

- Calculate joint probabilities.

A **joint probability** is the probability of events A and B happening at the same time. We are interested in both events occurring simultaneously in the unrestricted sample space. We have seen these types of probabilities already when we looked at contingency tables and in the context of “or” probabilities.

EXAMPLE

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

1. What is the probability that a randomly selected person is a cell phone user and had no speeding violations in the last year?
2. What is the probability that a randomly selected person had a speeding violation in the last year and does not use a cell phone?

Solution:

1.
$$\text{Probability} = \frac{\text{number of cell phone users and no violations}}{\text{total number in study}} = \frac{280}{755}$$
2.
$$\text{Probability} = \frac{\text{number of violations and not cell phone users}}{\text{total number in study}} = \frac{45}{755}$$

NOTE

These two probabilities are examples of joint probabilities. For example, in part 1, we want to find the probability that a randomly selected person has both traits: cell phone user and no speeding violations. So, we are interested in both events happening at the same time.

Repeated Trial Experiments

So far, most of the probabilities we have looked at are based on a **single trial experiment** and finding a probability based on that single trial. For example, finding the probability of rolling an even number in a single roll of a die is single trial experiment—we are only rolling the die one time and then we want to find the probability of a particular event happening in that single roll. Even the joint probabilities that we have seen so far, as in the example above, are based on a single trial experiment. We see these types of joint probabilities when we randomly select a single item and then want to find the probability that the item has two different characteristics at the same time.

However, we often want to calculate probabilities associated with **repeated trial experiments**. In a repeated trial experiment, we deal with **identical trials** that are repeated a number of times. For example, flipping a coin three times is an example of a repeated trial experiment—the trial is flipping the coin and then that trial is repeated three identical times.

EXAMPLE

Which of the following are repeated trial experiments? For the repeated trial experiments, identify the trial and the number of repetitions.

1. Finding the probability of rolling an odd number in the roll of die.
2. Finding the probability of drawing five spades from a deck of cards.
3. Finding the probability a randomly selected person has blue eyes and blond hair.
4. Finding the probability that three women from a pool of candidates are selected for a committee.
5. Finding the probability that a student answers ten multiple choice questions correctly.

Solution:

1. Single trial experiment. The die is rolled one time.
2. Repeated trial experiment. The trial is selecting a card from the deck and this trial is repeated five times.
3. Single trial experiment. A single person is selected.
4. Repeated trial experiment. The trial is selecting a women from the candidate pool and this trial is repeated three times.
5. Repeated trial experiment. The trial is answering an individual question and this trial is repeated ten times.

We can think of repeated trial experiments as joint probabilities—event on trial one AND event on trial two AND event on trial three and so on, depending on the number of trials. Suppose in the example of flipping the coin three times we want to find the probability of getting three heads in the three flips. We can think of this as a joint probability—heads on flip one AND heads on flip two AND heads on flip three. We want to calculate probabilities for such repeated trial experiments and, as we will see, the key to such probabilities is to think of the repeated trials as a joint probability.

One thing we must consider in a repeated trial experiment is whether the trials are done with or without replacement because this changes how we calculate the probability as we move from trial to trial.

- **With replacement.** Each member of a population is replaced after it is selected on a trial, and so each member of the population has the possibility of being chosen more than once (on different trials of the experiment). In terms of probability, with replacement means that the probability a member of the population is chosen stays the same from trial to trial. In other words, the trials are independent events because the result of the first trial does not affect the result of the second trial.
- **Without replacement.** Each time a member of a population is selected, it is NOT replaced, and so each member of the population cannot be chosen more than once. In terms of probability, without replacement means that the probability a member of the population is chosen changes from trial to trial. In other words, the trials are dependent events because the result of the first trial does affect the result of the second trial.

When calculating probabilities for repeated trial experiments, it is important that we identify if the experiment is done with or without replacement. Sometimes we will be told directly that the experiment is done with or without replacement. But most of the time we will need to determine if the experiment is done with or without replacement from the context of the question.

EXAMPLE

For each of the following determine if the experiment is done with or without replacement.

1. Flipping a coin three times.
2. Selecting three women for a committee from a pool of candidates.
3. Drawing five cards from deck of cards.
4. Rolling a die six times.
5. Selecting the members of the student executive committee from the student council.

Solution:

1. With replacement. The probability of heads or tails stays the same with each flip.
2. Without replacement. In this case, we want three different women on the committee, so we must select them without replacement. (Selecting with replacement would mean a possibility of the same women being selected three times and then the committee would consist of just a

single person).

3. Depending on the context, this could be with or without replacement. If each card is replaced after it is selected, this would be with replacement. If each card is not replaced after it is selected, this would be without replacement. In this situation, the question would probably include a statement about whether the cards are drawn with or without replacement.
4. With replacement. The probability of rolling any of the numbers stays the same with each roll of the die.
5. Without replacement. In this case, we want the members of the executive committee to be all different, so we must select them without replacement.

The Multiplication Rule for Joint Probabilities

In mathematical terms, “and” means multiply. By thinking of a repeated trial experiment as a joint probability, the basic idea is to multiply the probabilities of the individual trials. Basically, if we think of a repeated trial experiment as a joint probability—event on trial one AND event on trial two AND event on trial three and so on, depending on the number of trials—we can find the probability by multiplying together the probabilities of the trials:

$$\text{Probability} = \text{Prob. on Trial 1} \times \text{Prob. on Trial 2} \times \text{Prob. on Trial 3} \times \dots$$

Unfortunately, it is more complicated than that because we have to work out the probabilities on each trial, and these probabilities are affected by whether the experiment is done with or without replacement.

The multiplication rule to find the probability of A and B in a repeated trial experiment is

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

If we think of A as the first trial and B as the second trial, the probability of A and B is the probability of A (the probability of A on the first trial) times the probability of B given A (the probability of B on the second trial **assuming** that A happened on the first trial).

In the case that the experiment is done with replacement, the events A and B are **independent**, so $P(B|A) = P(B)$ and this rule becomes

$$P(A \text{ and } B) = P(A) \times P(B)$$

We can extend this rule to any number of trials, we just need to keep multiplying as we move from trial to trial.

When finding probabilities associated with repeated trial experiments, remember the following:

- To find the probability we work with the probabilities of the individual trials, multiplying the

probabilities together as we move from trial to trial.

- Identify if the experiment is done with or without replacement and use that information to find the probability on each subsequent trial.

EXAMPLE

A small local high school has 25 students in its graduating class. 18 of the students are going to college next year and the remaining 7 are not going to college next year. Suppose two students are selected at random from the graduating class.

1. What is the probability that both students are going to college next year?
2. What is the probability that exactly one of the students is going to college next year?

Solution:

This is a repeated trial experiment. A trial is selecting a student and there are two trials. The assumption here is that the experiment is done without replacement because we do not want to get the same student twice.

1. We want to get college-bound students on both trials. In other words, college-bound student on trial one AND college-bound student on trial two. On the first trial, the probability of getting a college-bound student is $\frac{18}{25}$. We are selecting without replacement, so after the first trial we assume that we have removed one of the college-bound students. This means that on the second trial, there are only 24 students to pick from (one student was removed on the first trial) and there are only 17 college-bound students left (on the first trial we removed one of the 18 college-bound students). So on the second trial, the probability of getting a college-bound student is $\frac{17}{24}$. The probability of getting two college-bound students is

$$\begin{aligned} \begin{array}{l} \text{\mbox{Probability}} \\ \text{=} \end{array} &= \begin{array}{l} \frac{18}{25} \\ \times \\ \frac{17}{24} \\ = 0.51 \end{array} \end{aligned}$$
2. We want one college-bound student (denoted C) and one non college-bound student (denoted N). In this case, we have to think about the **order** of the selections—there is a difference between college-bound on trial one, non-college bound on trial two (CN) **and** non-college

bound on trial one, college bound on trial two (NC). All possible orders must be accounted for when we calculate the probability. One of the two possible orders must occur: CN OR NC . For each of the individual orders, we multiply the probabilities as we move from trial to trial. The “or” means that we add the probabilities of the different orders. In other words:

$$\text{Probability} = \text{Probability of } CN + \text{Probability of } NC$$

For the CN order (college-bound on trial one, non college-bound on trial two), we want a college-bound student on trial one and the probability of getting a college-bound student is $\frac{18}{25}$. We are selecting without replacement, so after the first trial we assume that we have removed one of the college-bound students. This means that on the second trial, there are only 24 students to pick from (a college-bound student was removed on the first trial) and there are 7 non college-bound students (none of the non college-bound students were removed after the first trial). So on the second trial, the probability of getting a non college-bound student is $\frac{7}{24}$. So the probability of getting the CN order is $\frac{18}{25} \times \frac{7}{24}$.

Similarly for the NC order (non college-bound on trial one, college-bound on trial two), we want a non college-bound student on trial one and the probability of getting a non college-bound student is $\frac{7}{25}$. We are selecting without replacement, so after the first trial we assume that we have removed one of the non college-bound students. This means that on the second trial, there are only 24 students to pick from (a non college-bound student was removed on the first trial) and there are 18 college-bound students (none of the college-bound students were removed after the first trial). So on the second trial, the probability of getting a college-bound student is $\frac{18}{24}$. So the probability of getting the NC order is $\frac{7}{25} \times \frac{18}{24}$.

The probability of getting exactly one college bound student is

$$\begin{aligned} & \text{Probability of } CN + \text{Probability of } NC \\ &= \left(\frac{18}{25} \times \frac{7}{24}\right) + \left(\frac{7}{25} \times \frac{18}{24}\right) \\ &= 0.42 \end{aligned}$$

EXAMPLE

Suppose a fair die is rolled two times.

1. What is the probability of getting two 5's?
2. What is the probability of getting exactly one 2 and one 6 in the two rolls?

Solution:

This is a repeated trial experiment. A trial is rolling a die and there are two trials. The trials are independent (what happens on the first roll does not affect what happens on the second roll).

1. We want to get a 5 on both rolls. In other words, a 5 on roll one AND a 5 on roll two. On the first roll, the probability of getting a 5 is $\frac{1}{6}$. On the second roll, the probability of getting a 5 is $\frac{1}{6}$. Because the rolls are independent, the probability of getting a 5 on the second roll is not affected by what happens on the first roll. The probability of getting two 5's is

$$\begin{aligned} \text{Probability} &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36} \end{aligned}$$
2. We want one 2 and one 6. In this case, we have to think about the **order** of the rolls—there is a difference between 2 on roll one, 6 on roll two **and** 6 on roll one, 2 on roll two. All possible orders must be accounted for when we calculate the probability. One of the two possible orders must occur: 26 OR 62. For of the individual orders, we multiply the probabilities as we move from trial to trial. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = \text{Probability of 26} + \text{Probability of 62}$$

For the 26 order (2 on roll one, 6 on roll two), the probability of getting a 2 on roll one is $\frac{1}{6}$ and the probability of getting a 6 on roll two is $\frac{1}{6}$. So the probability of getting the 26 order is $\frac{1}{6} \times \frac{1}{6}$.

Similarly for the 62 order (6 on roll one, 2 on roll two), the probability of getting a 6 on roll one is $\frac{1}{6}$ and the probability of getting a 2 on roll two is $\frac{1}{6}$. So the probability of getting the 62 order is $\frac{1}{6} \times \frac{1}{6}$.

The probability of getting exactly one 2 and one 6 is

$$\begin{aligned} \text{Probability of } 26 + \text{Probability of } 62 &= \text{Probability of } \\ & \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) \\ &= \frac{1}{18} \end{aligned}$$

TRY IT

A box contains 30 microchips and 9 of those microchips are defective. Suppose two microchips are selected randomly from the box for inspection by the quality control officer.

1. What is the probability that both microchips are defective?
2. What is the probability that exactly one of the microchips is defective?

Click to see Solution

1. Probability = $\frac{9}{30} \times \frac{8}{29} = 0.0828$
2. Probability = $\left(\frac{9}{30} \times \frac{21}{29}\right) + \left(\frac{21}{30} \times \frac{9}{29}\right) = 0.4345$

EXAMPLE

A box contains 5 red cards and 12 white cards. Suppose three cards are drawn at random from the box **without** replacement.

1. What is the probability that all three cards are white?
2. What is the probability that exactly one of the cards is red?
3. What is the probability that at least one card is red?
4. What is the probability that at most one card is white?

Solution:

This is a repeated trial experiment. A trial is selecting a card and there are three trials. There are 17 cards in the box.

1. We want to get a white card on all three draws. In other words, white on draw one AND white on draw two AND white on draw three. On the first draw, the probability of getting a white card is $\frac{12}{17}$. We are selecting without replacement, so after the first draw we assume that we removed a white card from the box. This means that on the second draw, there are only 16 cards left in the box (one card was removed on the first draw) and there are only 11 white cards left (on the first draw we removed one of the 12 white cards). So on the second draw, the probability of getting a white card is $\frac{11}{16}$. After the second draw we assume that we removed white cards from the box on draws one and two. This means that on the third draw, there are only 15 cards left in the box (two cards were removed on the first two draws) and there are only 10 white cards left (white cards were removed on draws one and two). So on the third draw, the probability of getting a white card is $\frac{10}{15}$. The probability of getting three white cards is

$$\begin{aligned} \text{\mbox{Probability}} \ \&= \ \& \ \frac{12}{17} \ \times \ \frac{11}{16} \\ & \ \times \ \frac{10}{15} = 0.3235 \end{aligned}$$

2. We want one red card (R), so the other two cards must be white (W). In this case, we have to think about the **order** of the selection. All possible orders must be accounted for when we

calculate the probability. One of three possible orders must occur: RWW OR WRW OR WWR . For each of the individual orders, we multiply the probabilities as we move from draw to draw. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = P(RWW) + P(WRW) + P(WWR)$$

For the RWW order, the probability of red on draw one is $\frac{5}{17}$. We are selecting without replacement, so after the first trial we assume that we have removed one of the red cards. This means that on the second draw, there are only 16 cards left in the box (one card was removed on the first draw) and all 12 white cards are left (on the first draw we removed a red card). So on the second draw, the probability of getting a white card is $\frac{12}{16}$. After the second draw we assume that we removed a red card on draw one and a white card on draw two. This means that on the third draw, there are only 15 cards left in the box (two cards were removed on the first two draws) and there are only 11 white cards left (one white card was removed on draw two). So on the third draw, the probability of getting a white card is $\frac{11}{15}$. So the probability of

getting the RWW order is $\frac{5}{17} \times \frac{12}{16} \times \frac{11}{15}$.

Using similar logic, the probability of getting the WRW order is $\frac{12}{17} \times \frac{5}{16} \times \frac{11}{15}$ and the probability of getting the WWR order is $\frac{12}{17} \times \frac{11}{16} \times \frac{5}{15}$.

The probability of getting exactly one red card is

$$\begin{aligned} \text{Probability} &= P(RWW) + P(WRW) + P(WWR) \\ &= \left(\frac{5}{17} \times \frac{12}{16} \times \frac{11}{15} \right) + \left(\frac{12}{17} \times \frac{5}{16} \times \frac{11}{15} \right) + \left(\frac{12}{17} \times \frac{11}{16} \times \frac{5}{15} \right) \\ &= 0.4853 \end{aligned}$$

3. We want at least one red card in the three draws. This means we can have exactly one red card or exactly two red cards or exactly three red cards. As before, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. Here, there are seven possible ways of getting at least one red card: RWW OR

WRW OR WWR OR RRW OR RWR OR RRW OR RRR . Of course, we could work out the probabilities of each of these orders and add them all up. But there is a faster way to find this probability—use the complement. The complement of “at least one red card” is “exactly zero red cards.” When we look at the seven possible orders that make up the “at least one red card” event, the complement consists of all of the missing orders. In this case there is only one missing order, WWW , which is the event “exactly zero red cards.” Using the complement, the probability of at least one red card is

$$\begin{aligned} P(\text{at least one red card}) &= 1 - P(\text{exactly zero red card}) \\ &= 1 - P(WWW) \end{aligned}$$

In part 1 of this question, we found the probability of WWW : $\frac{12}{17} \times \frac{11}{16} \times \frac{10}{15}$. So the probability of at least one red card is

$$\begin{aligned} P(\text{at least one red card}) &= 1 - P(WWW) \\ &= 1 - \left(\frac{12}{17} \times \frac{11}{16} \times \frac{10}{15} \right) \\ &= 0.6765 \end{aligned}$$

4. We want at most one white card in the three draws. This means we can have exactly zero white cards or exactly one white card. As before, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. Here, there are four possible ways of getting at most one white card: RRR OR RRW OR RWR OR WRR . Using similar logic to above, the probability of getting the RRR order is

$$\frac{5}{17} \times \frac{4}{16} \times \frac{3}{15}, \text{ the probability of getting the } RRW \text{ order is } \frac{5}{17} \times \frac{4}{16} \times \frac{12}{15},$$

the probability of getting the RWR order is $\frac{5}{17} \times \frac{12}{16} \times \frac{5}{15}$, and the probability of getting

the WRR order is $\frac{12}{17} \times \frac{5}{16} \times \frac{4}{15}$. The probability of getting at most one white card is

$$\begin{aligned} P(\text{Probability}) &= P(RRR) + P(RRW) + P(RWR) + P(WRR) \\ &= \left(\frac{5}{17} \times \frac{4}{16} \times \frac{3}{15} \right) + \left(\frac{5}{17} \times \frac{4}{16} \times \frac{12}{15} \right) \\ &+ \left(\frac{5}{17} \times \frac{12}{16} \times \frac{5}{15} \right) + \left(\frac{12}{17} \times \frac{5}{16} \times \frac{4}{15} \right) \\ &= 0.1912 \end{aligned}$$

TRY IT

A box contains 5 red cards and 12 white cards. Suppose three cards are drawn at random from the box **with** replacement.

1. What is the probability that all three cards are white?
2. What is the probability that exactly one of the cards is red?
3. What is the probability that at least one card is red?
4. What is the probability that at most one card is white?

Click to see Solution

$$1. \text{ Probability} = \frac{12}{17} \times \frac{12}{17} \times \frac{12}{17} = 0.3517$$

$$2. \text{ Probability} = \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{12}{17} \right) = 0.4397$$

$$3. \text{ Probability} = 1 - \left(\frac{12}{17} \times \frac{12}{17} \times \frac{12}{17} \right) = 0.6483$$

$$4. \text{ Probability} = \left(\frac{5}{17} \times \frac{5}{17} \times \frac{5}{17} \right) + \left(\frac{5}{17} \times \frac{5}{17} \times \frac{12}{17} \right) + \left(\frac{5}{17} \times \frac{12}{17} \times \frac{5}{17} \right) + \left(\frac{12}{17} \times \frac{5}{17} \times \frac{5}{17} \right) = 0.2086$$

EXAMPLE

A company produces a popular brand of sports drink. The company is currently running a contest where winning symbols are placed under the bottle caps. 7% of all the bottle caps contain winning symbols. You buy three bottles of the sports drink.

1. What is the probability that all bottles have winning symbols?

2. What is the probability that exactly one of the bottles has a winning symbol?
3. What is the probability that at least one bottle has a winning symbol?

Solution:

This is a repeated trial experiment. A trial is selecting a bottle and there are three trials. This is an experiment without replacement (you do not want to select the same bottle three times). However, because the population of bottles is very, very large, we can treat the experiment as if the selections are made with replace. This means that we can treat the selection of the bottles as independent and so the probability of getting a winning bottle will be 7% on every draw.

1. We want to get a winning symbol on all three bottles. In other words, win on bottle one AND win on bottle two AND win on bottle three. The probability of winning on the first bottle is 7%, the probability of winning on the second bottle is 7%, and the probability of winning on the third bottle is 7%. Because we can treat the selections as independent, the probability of winning does not change from draw to draw. The probability of getting three winning bottles is

$$\begin{array}{l} \text{\mbox{Probability}} \ \&= \ \& 0.07 \ \times \ 0.07 \ \times \ 0.07 = 0.0003 \\ \end{array}$$

2. We want one winning bottle (W), so the other two bottles must be non-winners (N). The probability of winning on any bottle is 7%, so the probability of losing on any bottle is 93%. In this case, we have to think about the **order** of the selection. All possible orders must be accounted for when we calculate the probability. One of the three possible orders must occur: WNN OR NWN OR NNW . For each of the individual orders, we multiply the probabilities as we move from draw to draw. The “or” means that we add the probabilities of the different orders. In other words,

$$\text{Probability} = P(WNN) + P(NWN) + P(NNW)$$

For the WNN order (win on bottle one, non-wins on bottles two and three), the probability of getting a win on bottle one is 0.07 and the probability of getting a non-win on bottle two or bottle three is 0.93. So the probability of getting the WNN order is $0.07 \times 0.93 \times 0.93$. Using similar logic, the probability of getting the NWN order is $0.93 \times 0.07 \times 0.93$ and the probability of getting the NNW order is $0.93 \times 0.93 \times 0.07$.

The probability of getting exactly one winning bottle is

$$\begin{aligned}
 \text{Probability} &= P(WNN) + P(NWN) + P(NNW) \\
 &= (0.07 \times 0.93 \times 0.93) + (0.93 \times 0.07 \times 0.93) + (0.93 \times 0.93 \times 0.07) \\
 &= 0.1816
 \end{aligned}$$

3. We want at least one winning bottle. This means we can have exactly one winning bottle or exactly two winning bottles or exactly three winning bottles. Of course, we could work out the probabilities of each of these orders and add them all up. But the a faster way to find this probability is to use the complement. The complement of “at least one winning bottle” is “exactly zero winning bottles” The “exactly zero winning bottle” is the case NNN (all three bottles are non-winners). Using the complement, the probability of at least one winning bottle is

$$\begin{aligned}
 P(\text{at least one winner}) &= 1 - P(\text{exactly zero winners}) \\
 &= 1 - P(NNN)
 \end{aligned}$$

The probability of zero winning bottles (NNN) is : $0.93 \times 0.93 \times 0.93$. So the probability of at least one winning bottle is

$$\begin{aligned}
 P(\text{at least one winner}) &= 1 - P(NNN) \\
 &= 1 - (0.93 \times 0.93 \times 0.93) \\
 &= 0.1956
 \end{aligned}$$

NOTE

In situations like this example where we are drawing with replacement from a very, very large population, we treat the draws as if they are independent. Because the population is so large, the change in the probability as we go from draw to draw is very, very small, which makes it hardly detectable in the calculation of the answer. In such situations, we can treat the draws as independent. We cannot do this when we are drawing without replacement from a small population (as in the red and white card example above) because there are distinct changes in the probabilities as we move from draw to draw.

Concept Review

In a repeated trial experiment, we deal with identical trials that are repeated a number of times. A repeated trial experiment can be thought of as a joint probability. The multiplication rule for joint probabilities is $P(A \text{ and } B) = P(A) \times P(B|A)$ where A is the event on the first trial and B is the event on the second trial. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be dependent.

Attribution

“3.1 Terminology”, “3.2 Independent and Mutually Exclusive Events”, “3.3 Two Basic Rule of Probability”, and “3.4 Contingency Tables“ in *Introductory Statistics* by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

3.8 EXERCISES

1. In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

- a. The probability that a student does not have long hair.
- b. The probability that a student is male or has short hair.
- c. The probability that a student is a female and has long hair.
- d. The probability that a student is male, given that the student has long hair.
- e. The probability that a student has long hair, given that the student is male.
- f. Of all the female students, the probability that a student has short hair.
- g. Of all students with long hair, the probability that a student is female.
- h. The probability that a student is female or has long hair.
 - i. The probability that a randomly selected student is a male student with short hair.
 - j. The probability that a student is female.

2. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti. Let H be the event of getting a hat. Let N be the event of getting a noisemaker. Let F be the event of getting a finger trap. Let C be the event of getting a bag of confetti.

- a. Find $P(H)$.
- b. Find $P(N)$.
- c. Find $P(F)$.
- d. Find $P(C)$.

3. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

- Let B = the event of getting a blue jelly bean
- Let G = the event of getting a green jelly bean.
- Let O = the event of getting an orange jelly bean.
- Let P = the event of getting a purple jelly bean.
- Let R = the event of getting a red jelly bean.
- Let Y = the event of getting a yellow jelly bean.

- a. Find $P(B)$.
- b. Find $P(G)$.
- c. Find $P(P)$.
- d. Find $P(R)$.
- e. Find $P(Y)$.

4. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

- Let A = the event that a country is in Asia.
- Let E = the event that a country is in Europe.
- Let F = the event that a country is in Africa.
- Let N = the event that a country is in North America.
- Let O = the event that a country is in Oceania.
- Let S = the event that a country is in South America.

- a. Find $P(A)$.
- b. Find $P(E)$.
- c. Find $P(F)$.
- d. Find $P(N)$.
- e. Find $P(O)$.
- f. Find $P(S)$.

5. What is the probability of drawing a red card from a standard deck of 52 cards?

6. What is the probability of drawing a club in a standard deck of 52 cards?

7. What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

8. What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

9. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

- Let I = the event that a player is an infielder.
- Let O = the event that a player is an outfielder.
- Let H = the event that a player is a great hitter.
- Let N = the event that a player is not a great hitter.

- a. Write the symbols for the probability that a player is not an outfielder.
- b. Write the symbols for the probability that a player is an outfielder or is a great hitter.
- c. Write the symbols for the probability that a player is an infielder and is not a great hitter.
- d. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.
- e. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.
- f. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.
- g. Write the symbols for the probability that of all the great hitters, a player is an outfielder.
- h. Write the symbols for the probability that a player is an infielder or is not a great hitter.
- i. Write the symbols for the probability that a player is an outfielder and is a great hitter.
- j. Write the symbols for the probability that a player is an infielder.

10. What is the word for the set of all possible outcomes?

11. What is conditional probability?

12. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

- a. What does $P(E|M)$ mean in words?
- b. What does $P(E \text{ or } M)$ mean in words?

13. Explain what is wrong with the following statements. Use complete sentences.

- a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
- b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

14. E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E | F)$.

15. J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

16. U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

- $P(U \text{ and } V)$
- $P(U|V)$
- $P(U \text{ or } V)$

17. Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ and } R) = 0.1$. Find $P(R)$.

18. A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into the table.

Shirt Number	At most 210	211–250	251–290	More than 290	Total
1–33	21	5	0	0	26
34–66	6	18	7	4	35
66–99	6	12	22	5	45
Total	33	35	29	9	106

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

- What is the probability that the player's shirt number is in the 34-66 category?
- What is the probability that the player weighs at most 210 lbs?
- What is the probability that the player's shirt number is in the 1-33 category and weighs between 211 and 250 lbs?
- What is the probability that the player's shirt number is in the 66-99 category or weighs more than 290 lbs?
- What is the probability that the player's shirt number is in the 34-66 category given that they weigh between 251 and 290 lbs?
- What is the probability that a player weighs at most 210 lbs if their shirt number is in the 1-33 category?
- Are the events "66-99" and more than 290 lbs independent? Explain.

19. At a local college, 20% of the students are studying business, 40% of the students are studying mathematics and 8% of the students are studying both business and mathematics.

- What is the probability that a randomly selected student studies business or mathematics?
- What is the probability that a randomly selected student studies mathematics given that they

study business?

- What is the probability that a randomly selected mathematics student studies business?
- Are the events “business” and “mathematics” independent? Explain.
- Are the events “business” and “mathematics” mutually exclusive? Explain.

20. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers (0,00, 1, 2,...,36), and each number is assigned to a color (green, red or black) and a range. You can place a bet based on number, color, or range.

00	3	6	9	12	15	18	21	24	27	30	33	36	12 to 1
0	2	5	8	11	14	17	20	23	26	29	32	35	12 to 1
	1	4	7	10	13	16	19	22	25	28	31	34	2 to 1
1st Dozen				2nd Dozen				3rd Dozen					
1 to 18		EVEN		♦		♦		ODD		19 to 36			

credit: film8ker/wikibooks

- List the sample space of the 38 possible outcomes in roulette.
- You bet on red. Find the probability of red.
- You bet on “1st Dozen” (meaning the number from 1 to 12). Find the probability of “1st Dozen”.
- You bet on an even number. Find the probability of an even number.
- Is getting an odd number the complement of getting an even number? Why?
- Find two mutually exclusive events.
- Are the events “Even” and “1st Dozen” independent? Explain.
- What is the probability of the event “1 to 18”?

21. Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- What is the probability the card is green?
- What is the probability the card is green given that the card has an even number on it?
- What is the probability the card is green or has an even number on it?
- What is the probability the card is green and has an even number on it?
- Are the events “green” and “even” mutually exclusive? Explain.
- Are the events “green and “even” independent? Explain..

22. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded.

- Suppose three cards are picked without replacement. What is the probability that all three cards are green?
- Suppose three cards are picked without replacement. What is the probability that exactly two of the cards are blue?
- Suppose three cards are picked without replacement. What is the probability that at least one of the cards is red?
- Suppose three cards are picked with replacement. What is the probability that at most one card is green?
- Suppose three cards are picked with replacement. What is the probability that all three cards are green?
- Suppose three cards are picked with replacement. What is the probability that exactly two of the cards are blue?
- Suppose three cards are picked with replacement. What is the probability that at least one of the cards is red?
- Suppose three cards are picked with replacement. What is the probability that at most one card is green?

23. Suppose $P(C) = 0.4$, $P(D) = 0.5$ and $P(C|D) = 0.6$.

- Find $P(C \text{ and } D)$.
- Are C and D mutually exclusive? Why or why not?
- Are C and D independent events? Why or why not?
- Find $P(\text{ or } D)$.

e. Find $P(D|C)$.

24. In 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery.

- What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
- In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement.
- Are "won a green card" and "finalist" independent or dependent events? Justify your answer numerically and also explain why.
- Are "won a green card" and "finalist" mutually exclusive events? Justify your answer numerically and explain why.

25. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

- Are "the hit being made by Hank Aaron" and "the hit being a double" independent events? Explain.
- What is the probability that a hit was made by Babe Ruth?
- What is the probability that a hit was made by Hank Aaron and is a home run?
- What is the probability that a hit was made by Ty Cobb or is a single?
- What is the probability that a hit was a double given that it was by Jackie Robinson?
- What is the probability that a triple was hit by Babe Ruth?

26. United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According

to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

- Find the probability that a person has both type O blood and the Rh- factor.
- Find the probability that a person does NOT have both type O blood and the Rh- factor.

27. At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam.

- Find the probability that a course has a final exam or a research project.
- Find the probability that a course has NEITHER of these two requirements.

28. In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

- Find the probability that a cookie contains chocolate or nuts (he can't eat it).
- Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

29. A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student.

- Find the probability that a student takes a distance learning class and is a part-time student.
- Find the probability that a student is a part-time student given that they take a distance learning class.
- Find the probability that student is a part-time student or takes a distance learning class.
- Are the events “distance learning” and “part-time” independent? Explain.

30. The table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

- Find the probability a musician is female.
- Find the probability that a musician received private instruction.
- Find the probability that a musician is male and is self-taught.
- Find the probability that a musician is female or studied in school.
- Find the probability that a musician is male given that they received private instruction.
- Find the probability that a female musician is self-taught.
- Are the events “female” and “self-taught” independent? Explain.

31. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

- What is the probability that a randomly selected senator has an “Other” affiliation?
- What is the probability that a randomly selected senator is up for reelection in November 2016?
- What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?
- What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?
- Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?
- Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

32. Table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

- a. Complete the table.
 - b. What is the probability that a randomly selected child will have wavy hair?
 - c. What is the probability that a randomly selected child will have either brown or blond hair?
 - d. What is the probability that a randomly selected child will have wavy brown hair?
 - e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
 - f. If B is the event of a child having brown hair, find the probability of the complement of B .
 - g. In words, what does the complement of B represent?
33. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)
- a. Let S be the event that both cookies selected were the same flavor. Find $P(S)$.
 - b. Let T be the event that the cookies selected were different flavors. Find $P(T)$.
 - c. Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.
34. A cup contains three red, four yellow and five blue beads.
- a. Suppose three beads are selected at random without replacement. What is the probability all three beads are blue?
 - b. Suppose three beads are selected at random with replacement. What is the probability all three beads are blue?
 - c. Suppose three beads are selected at random without replacement. What is the probability that exactly one of the beads is red?
 - d. Suppose three beads are selected at random with replacement. What is the probability that exactly one of the beads is red?
 - e. Suppose three beads are selected at random without replacement. What is the probability that at least one bead is yellow?
 - f. Suppose three beads are selected at random with replacement. What is the probability that at least one bead is yellow?
35. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.
- a. Find the probability a driver is female.

- b. Find the probability a driver is 65 or over given that they are female.
- c. Find the probability a driver is 65 or over and female.
- d. In words, explain the difference between the probabilities in part c and part d.
- e. Find the probability a driver is 65 or over.
- f. Are being age 65 or over and being female mutually exclusive events? How do you know?

36. Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

- a. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
- b. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
- c. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

37. When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

- a. Based on the given data, find $P(H)$ and $P(T)$.
- b. Find the probabilities of each possible outcome for the experiment of tossing the coin twice.
- c. Find the probability of obtaining exactly one head in two tosses of the coin.
- d. Find the probability of obtaining at least one head.

Attribution

“Chapter 3 Homework” and “Chapter 3 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART IV

DISCRETE RANDOM VARIABLES

Chapter Outline

- 4.1 Introduction to Discrete Random Variables
- 4.2 Probability Distribution of a Discrete Random Variable
- 4.3 Expected Value and Standard Deviation for a Discrete Probability Distribution
- 4.4 The Binominal Distribution
- 4.5 The Poisson Distribution
- 4.6 Exercises

4.1 INTRODUCTION TO DISCRETE RANDOM VARIABLES



You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. Photo by Leszek Leszczynski, CC BY 2.0.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving **discrete random variables**. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variables

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If X is a random variable, then X is written in words, and x is given as a number.

For example, let X be the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is $TTT, TTH, HTH, HHT, HTT, THT, TTH, HHH$. Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values are 0, 1, 2, 3), X is a discrete random variable.

A **random variable** describes a characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z, \dots and common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters $x, y,$ and z . For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, \dots . Variables in statistics differ from variables in intermediate algebra in the two following ways:

- The domain of the random variable is not necessarily a numerical set. The domain may be expressed in words. For example, if X is hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

Attribution

“Chapter 4 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

4.2 PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

LEARNING OBJECTIVES

- Recognize, understand, and construct discrete probability distributions.

A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=96#oembed-1>

Watch this video: Random Variables and Probability Distributions by Dr Nic's Maths and Stats [4:38]

The **probability distribution** for a random variable lists all the possible values of the random variable and the probability the random variable takes on each value. The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable. A probability distribution can be a table, with a column for the values of the random variable and another column for the corresponding probability, or a graph, like a histogram with the values of the random variable on the horizontal axis and the probabilities on the vertical axis.

In a probability distribution, each probability is between 0 and 1, inclusive. Because all possible values of the random variable are included in the probability distribution, the sum of the probabilities is 1.

EXAMPLE

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X be the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, the values of the random variable are $x = 0, 1, 2, 3, 4, 5$.

In the table, the left column contains all of the possible values of the random variable and the right column, $P(x)$, is the probability that X takes on the corresponding value x . For example, in the first row, the value of the random variable is 0 and the probability the random variable is 0 is $\frac{2}{50}$. In the context of this example, that means that the probability a newborn baby's crying wakes its mother 0 times per week is $\frac{2}{50}$.

x	$P(x)$
0	$\frac{2}{50}$
1	$\frac{11}{50}$
2	$\frac{23}{50}$
3	$\frac{9}{50}$
4	$\frac{4}{50}$
5	$\frac{1}{50}$

Because X can only take on the values 0, 1, 2, 3, 4, and 5, X is a discrete random variable. Note that each probability is between 0 and 1 and the sum of the probabilities is 1:

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

TRY IT

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days a week 15% of the time, one day a week 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

1. Let X be the number of days Nancy _____.
2. X takes on what values?
3. Suppose one week is randomly chosen. Construct a probability distribution table like the one in example above. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Click to see Solution

1. Let X be the number of days Nancy attends class per week.
2. 0, 1, 2, and 3.

3.

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

The $P(x)$ column sums to 1.

EXAMPLE

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Solution:

X is the number of days Jeremiah attends basketball practice per week. X takes on the values 0, 1, and 2.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=96#oembed-2>

Watch this video: Constructing a Probability Distribution for a Random Variable by Khan Academy [6:47]

Concept Review

A probability distribution for a random variable describes how the probabilities are distributed over the random variable—in other words, the probability distribution describes the probability that the random variable takes on a specific value. A probability distribution includes all possible values the random variable can take on and the corresponding probability. Each probability is between 0 and 1, inclusive, and the sum of the probabilities is 1.

Attribution

“4.1 Probability Distribution Function (PDF) for a Discrete Random Variable“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

4.3 EXPECTED VALUE AND STANDARD DEVIATION FOR A DISCRETE PROBABILITY DISTRIBUTION

LEARNING OBJECTIVES

- Calculate and interpret the expected value of a probability distribution.
- Calculate the standard deviation for a probability distribution.

Expected Value of a Probability Distribution

The **expected value** is often referred to as the “**long-term**” **average or mean**. That is, over the long term of repeatedly doing an experiment, you would **expect** this average.

Suppose you toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does the probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment. Probability gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero—the theoretical probability and the relative frequency get closer and closer together. When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This “long-term average” is known as the **mean** or **expected value** of

the experiment and is denoted by μ or $E(x)$. In other words, after conducting many trials of an experiment, you would expect this average value.

The **expected value**, denoted by μ or $E(x)$, is a weighted average where each value of the random variable is weighted by the value's corresponding probability.

$$E(x) = \sum (x \times P(x))$$

EXAMPLE

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value of the number of days per week the men's soccer team plays soccer.

Solution:

First let the random variable X be the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. The table below shows the probability distribution for X , and includes an additional column $x \times P(x)$ that we will use to calculate the expected value. In this new column, we will multiply each x value by its corresponding probability.

x	$P(x)$	$x \times P(x)$
0	0.2	$0 \times 0.2 = 0$
1	0.5	$1 \times 0.5 = 0.5$
2	0.3	$2 \times 0.3 = 0.6$

Add the last column to find the long term average or expected value:

$$\begin{aligned} E(x) &= (0 \times 0.2) + (1 \times 0.5) + (2 \times 0.3) \\ &= 0 + 0.5 + 0.6 \\ &= 1.1 \end{aligned}$$

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week.

NOTE

The expected value does not represent a value that the random variable takes on. The expected value is an average. In this case, the expected value of 1.1 is the average times the team plays per week. To understand what this means, imagine that each week you recorded the number of times the soccer team played that week. You do this repeatedly for many, many, many, weeks. Then you calculate the mean of the numbers you recorded (using the techniques we learned previously)—the mean of these numbers equals 1.1, the expected value. The number of trials must be very, very large in order for the mean of the values recorded from the trials to equal the expected value calculated using the expected value formula.

Standard Deviation of a Probability Distribution

Like data, probability distributions have standard deviations. The **standard deviation**, denoted σ , of a probability distribution for a random variable X describes the spread or variability of the probability distribution. The standard deviation is the standard deviation you expect when doing an experiment over and over.

$$\sigma = \sqrt{\sum (x - \mu)^2 \times P(x)}$$

To calculate the standard deviation of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root.

EXAMPLE

Let X be the number of times per week a newborn baby's crying wakes its mother after midnight. The probability distribution for X is:

x	$P(x)$
0	0.04
1	0.22
2	0.46
3	0.18
4	0.08
5	0.02

Find the expected value and standard deviation of the number of times a newborn baby's crying wakes its mother after midnight.

Solution:

For the expected value:

x	$P(x)$	$x \times P(x)$
0	0.04	0
1	0.22	0.22
2	0.46	0.92
3	0.18	0.54
4	0.08	0.32
5	0.02	0.1

$$\begin{aligned}\mu &= 0 + 0.22 + 0.92 + 0.54 + 0.32 + 0.1 \\ &= 2.1\end{aligned}$$

On average, a newborn wakes its mother after midnight 2.1 times per week.

For the standard deviation: For each value x , multiply the square of its deviation by its probability (each deviation has the format $x - \mu$).

x	$P(x)$	$(x - \mu)^2 \times P(x)$
0	0.04	$(0 - 2.1)^2 \times 0.04 = 0.1764$
1	0.22	$(1 - 2.1)^2 \times 0.22 = 0.2662$
2	0.46	$(2 - 2.1)^2 \times 0.46 = 0.0046$
3	0.18	$(3 - 2.1)^2 \times 0.18 = 0.1458$
4	0.08	$(4 - 2.1)^2 \times 0.08 = 0.2888$
5	0.02	$(5 - 2.1)^2 \times 0.02 = 0.1682$
	Sum	1.05

Add the values in the third column of the table and then take the square root of this sum:

$$\begin{aligned}\sigma &= \sqrt{1.05} \\ &= 1.024\dots\end{aligned}$$

TRY IT

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. Let X be the number of times a post-op patient rings for the nurse. For a random sample of 50 patients, the following information was obtained. What is the expected value? What is the standard deviation?

x	$P(x)$
0	0.08
1	0.16
2	0.32
3	0.28
4	0.12
5	0.04

Click to see Solution

For the expected value:

x	$P(x)$	$x \times P(x)$
0	0.08	0
1	0.16	0.16
2	0.32	0.64
3	0.28	0.84
4	0.12	0.48
5	0.04	0.2

$$\begin{array}{l} \mu = 0 + 0.16 + 0.64 + 0.84 + 0.48 + 0.2 \\ \mu = 2.32 \end{array}$$

For the standard deviation:

x	$P(x)$	$(x - \mu)^2 \times P(x)$
0	0.08	0.430592
1	0.16	0.278784
2	0.32	0.032768
3	0.28	0.129472
4	0.12	0.338688
5	0.04	0.287296

$$\begin{aligned} \sigma &= \sqrt{0.430592 + 0.278784 + 0.032768 + 0.129472 \\ &\quad + 0.338688 + 0.287296} \\ &= 1.22\dots \end{aligned}$$

EXAMPLE

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

Solution:

To do this problem, set up an expected value table for the amount of money you can profit. Let X be the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Because you are interested in your **profit (or loss)**, the values of x are \$100,000 and $-\$2$ dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is

$\frac{1}{10} = 0.1$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = 0.00001$$

Therefore, the probability of winning is 0.00001 and the probability of losing is $1 - 0.00001 = 0.99999$.

The expected value is as follows:

x	$P(x)$	$x \times P(x)$
-2	0.99999	-1.99998
100,000	0.00001	1

$$\begin{aligned} E(x) &= -1.99998 + 1 \\ &= -0.99998 \end{aligned}$$

Because -0.99998 is about -1 , you would, on average, expect to lose approximately \$1 for each game you play. However, each time you play, you either lose \$2 or profit \$100,000. The \$1 is the average (or expected) LOSS per game after playing this game over and over.

TRY IT

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay \$1 to play. If you guess the right suit every time, you get your money back and \$256. What is your expected profit of playing the game over the long term?

Click to see Solution

Let X be the amount of money you profit. The values of x are $-\$1$ (for a loss) and $\$256$ (for a win).

The probability of winning (guessing the correct suit on each draw) is

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = 0.0039$$

The probability of losing is

$$1 - 0.0039 = 0.9961$$

The expected value is as follows:

x	$P(x)$	$x \times P(x)$
-1	0.9961	-0.9961
256	0.0039	0.9984

$$\begin{aligned} E(x) &= -0.9961 + 0.9984 \\ &= 0.0023 \end{aligned}$$

Playing the game over and over again means you would average \$0.0023 in profit per game.

EXAMPLE

Suppose you play a game with a biased coin where the probability of heads is $\frac{2}{3}$. You play each game by tossing the coin once. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead?

1. Define a random variable X .
2. Construct the probability distribution for X .

3. What is the expected value? Do you come out ahead?

Solution:

1. Let X be the amount of profit per game. The values of x are $-\$6$ (for a loss) and $\$10$ (for a win).

2.

x	$P(x)$
10	$\frac{1}{3}$
-6	$\frac{2}{3}$

3.

x	$P(x)$	$x \times P(x)$
10	$\frac{1}{3}$	$\frac{10}{3}$
-6	$\frac{2}{3}$	-4

$$\begin{aligned} E(x) &= \frac{10}{3} + (-4) \\ &= -0.67 \end{aligned}$$

On average, you lose \$0.67 each time you play the game, so you do not come out ahead.

TRY IT

Suppose you play a game with a spinner that has three colours on it: red, green, and blue. The probability of landing on red is 40% and the probability of landing on green is 20%. You play a game by spinning the spinner once. If you land on red, you pay \$10. If you land on blue, you do not pay or win anything. If you land on green, you win \$10. What is the expected value of this game? Do you come out ahead?

Click to see Solution

Let X be the amount won in a game. The values of x are $-\$10$ (for red), 0 (for blue) and $\$10$ (for a green).

x	$P(x)$	$x \times P(x)$
-10	0.4	-4
0	0.4	0
10	0.2	2

$$\begin{aligned} E(x) &= -4 + 0 + 2 \\ &= -2 \end{aligned}$$

On average, you lose \$2 per game. So you do not come out ahead.

TRY IT

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X be the amount of profit from a bet. Find the mean and standard deviation of X .

Click to see Solution

x	$P(x)$	$x \times P(x)$	$(x - \mu)^2 \times P(x)$
100	0.0108	1.08	127.8726
-10	0.9892	-9.892	1.3961

$$\begin{aligned}\mu &= 1.08 + (-9.892) \\ &= -8.812\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{127.7826 + 1.3961} \\ &= 11.3696\dots\end{aligned}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=98#oembed-1>

Watch this video: Mean of a Discrete Random Variable by Khan Academy [4:31]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=98#oembed-2>

Watch this video: Variance and Standard Deviation of a Discrete Random Variable by Khan Academy [6:25]

Concept Review

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

- Mean or Expected Value: $E(x) = \mu = \sum (x \times P(x))$
 - Standard Deviation: $\sigma = \sqrt{\sum ((x - \mu)^2 \times P(x))}$
-

Attribution

“4.2 Mean or Expected Value and Standard Deviation“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

4.4 THE BINOMIAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the binomial probability distribution and apply it appropriately.

There are four characteristics of a **binomial experiment**:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called “success” and “failure,” for each trial. The letter p denotes the probability of a success on any one trial and $1 - p$ denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.
4. For each individual trial, the probability of a success, p , and probability of a failure, $1 - p$, remain the same. Because the n trials are independent, the outcome of one trial does not affect the outcome of another trial.

For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $1 - p = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success $p = 0.6$ and his probability of failure $1 - p = 0.4$ remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X is the number of successes obtained in the n independent trials. The mean of a binomial probability distribution is $\mu = n \times p$ and the standard deviation is $\sigma = \sqrt{n \times p \times (1 - p)}$

Any experiment with the characteristics of a binomial experiment and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A

binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

EXAMPLE

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A “success” could be defined as an individual who withdrew from the course. The random variable X is the number of students who withdraw from the randomly selected elementary physics class.

TRY IT

The state health board is concerned about the amount of fruit available in school lunches. 48% of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a “success” be in this case?

Click to see Solution

- A success would be a school that offers fruit in their lunch every day.

EXAMPLE

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55% and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.

Solution:

If you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $1 - p = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

EXAMPLE

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

1. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
2. If we are interested in the number of students who do their homework on time, then how do we define X ?
3. What values does x take on?
4. What is a “failure,” in words?
5. What is the probability of “failure”?
6. The words “at least” translate as what kind of inequality for the probability question $P(x \dots 40)$.

Solution:

1. failure
2. X is the number of statistics students who do their homework on time.
3. 0, 1, 2, ..., 50
4. Failure is defined as a student who does not complete his or her homework on time.
5. $1 - p = 0.30$
6. "At least" means greater than or equal to (\geq). The probability question is $P(x \geq 40)$.

TRY IT

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Why this is a binomial problem?

Click to see Solution

- There are only two outcomes on any exam (pass or fail).
- There is fixed number of trials ($n = 50$).
- The probability of pass (65%) is the same for each trial.
- The trials are independent. (The fact that any one person passes or fails the exam does not affect whether or not any other person passes or fails.)

EXAMPLE

The following example illustrates a problem that is not binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students?

Solution:

The names of all committee members are put into a box and two names are drawn without replacement. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$ and the probability of a student on the second draw is $\frac{5}{15}$. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

TRY IT

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether or not this is binomial and state why.

Click to see Solution

This is not binomial because the names are not replaced after each draw, which means the probability changes for each time a name is drawn. This violates the condition of independence.

Calculating Binomial Probabilities

CALCULATING BINOMIAL PROBABILITIES IN EXCEL

To calculate probabilities associated with binomial random variables in Excel, use the **binom.dist(x,n,p,logic operator)** function.

- For **x**, enter the number of successes.
- For **n**, enter the number of trials.
- For **p**, enter the probability of success.
- For the logic operator, enter **false** to find the probability of exactly **x** successes and enter **true** to find the probability of at most (less than or equal to) **x** successes.

The output from the **binom.dist** function is:

- the probability of getting exactly **x** success in **n** trials with a probability of success **p** when the logic operator is **false**.
- the probability of at most **x** successes in **n** trials with a probability of success **p** when the logic operator is **true**.

Visit the Microsoft page for more information about the **binom.dist** function.

NOTE

Because we can only enter false or true into the logic operator, the **binom.dist** function can only directly calculate the probability of getting exactly x successes in n trials or getting at most x success in n trials. In order to calculate other binomial probabilities, such as fewer than x successes, more than x successes or at least x successes, we need to manipulate how we use the **binom.dist** function by changing what we enter into the **binom.dist** function, using the complement rule, or both.

EXAMPLE

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. Suppose 20 adult workers are randomly selected.

1. How many adult workers in the sample do you expect to have a high school diploma but do not pursue any further education?
2. What is the probability that exactly 8 of the workers in the sample have a high school diploma but do not pursue further education?
3. What is the probability that at most 12 of the workers in the sample have a high school diploma but do not pursue further education?

Solution:

Let X be the number of workers in the sample who have a high school diploma but do not pursue further education. The number of trials is $n = 20$ and the probability of success is $p = 0.41$.

1. $\mu = n \times p = 20 \times 0.41 = 8.2$. On average, in any sample of 20 workers, 8.2 have a high school diploma but do not pursue further education.
2. We want to find $P(x = 8)$.

Function	binom.dist	Answer
Field 1	8	0.1790
Field 2	20	
Field 3	0.41	
Field 4	false	

The probability that exactly 8 of the workers in the sample have a high school diploma but do not pursue further education is 17.9%.

3. We want to find $P(x \leq 12)$.

Function	binom.dist	Answer
Field 1	12	0.9738
Field 2	20	
Field 3	0.41	
Field 4	true	

The probability that at most 12 of the workers in the sample have a high school diploma but do not pursue further education is 97.38%.

TRY IT

About 32% of students participate in a community volunteer program outside of school. Suppose 30 students are selected at random.

1. What is the expected number of students in the sample that participate in a community volunteer program?
2. What is the probability that exactly 10 of the students in the sample participate in a

community volunteer program?

3. What is the probability that at most 14 of the students in the sample participate in a community volunteer program?

Click to see Solution

1. $\mu = n \times p = 30 \times 0.32 = 9.6$

2.

Function	binom.dist	Answer
Field 1	10	0.1512
Field 2	30	
Field 3	0.32	
Field 4	false	

3.

Function	binom.dist	Answer
Field 1	14	0.9695
Field 2	30	
Field 3	0.32	
Field 4	true	

EXAMPLE

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages and 1.5% of the pages feature signature artists. Suppose 100 pages are randomly selected from the catalog.

1. What is the probability that fewer than 3 of the pages in the sample feature signature artists?
2. What is the probability that more than 5 of the pages in the sample feature signature artists?
3. What is the probability that at least 4 of the pages in the sample feature signature artists?

4. What is the probability that between 2 and 6 of the pages in the sample feature signature artists?

Solution:

1. We want to find $P(x < 3)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate = or \leq probabilities. Because x must be an integer (it is the number of pages), $x < 3$ is the same as $x \leq 2$ (of course, in general, this is not true). So $P(x < 3) = P(x \leq 2)$ and $P(x \leq 2)$ is a probability we can calculate with the binom.dist function.

Function	binom.dist	Answer
Field 1	2	0.8098
Field 2	100	
Field 3	0.015	
Field 4	true	

2. We want to find $P(x > 5)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate = or \leq probabilities. The complement of $>$ is \leq , so $P(x > 5) = 1 - P(x \leq 5)$ and $P(x \leq 5)$ is a probability we can calculate with the binom.dist function.

Function	1-binom.dist	Answer
Field 1	5	0.0177
Field 2	100	
Field 3	0.015	
Field 4	true	

3. We want to find $P(x \geq 4)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate = or \leq probabilities. The complement of \geq is $<$, so
$$P(x \geq 4) = 1 - P(x < 4)$$
. Because x must be an integer (it is the number of pages), $x < 4$ is the same as $x \leq 3$. So
$$P(x \geq 4) = 1 - P(x < 4) = 1 - P(x \leq 3)$$
 and $P(x \leq 3)$ is a probability we can calculate with the binom.dist function.

Function	1-binom.dist	Answer
Field 1	3	0.0642
Field 2	100	
Field 3	0.015	
Field 4	true	

4. We want to find $P(2 \leq x \leq 6)$. We cannot find this probability directly in Excel because the binom.dist function can only calculate = or \leq probabilities. But,
 $P(2 \leq x \leq 6) = P(x \leq 6) - P(x \leq 1)$. So we can calculate $P(2 \leq x \leq 6)$ as the difference of two binom.dist functions.

Function	binom.dist	-binom.dist	Answer
Field 1	6	1	0.4426
Field 2	100	100	
Field 3	0.015	0.015	
Field 4	true	true	

TRY IT

According to a Gallup poll, 60% of American adults prefer saving over spending. Suppose 50 American adults are selected at random.

1. What is the probability that at least 35 adults in the sample prefer saving over spending?
2. What is the probability that fewer than 20 adults in the sample prefer saving over spending?
3. What is the probability between 15 and 25 adults in the sample prefer saving over spending?
4. What is the probability that more than 30 adults prefer saving over spending?

Click to see Solution

1.

Function	1-binom.dist	Answer
Field 1	34	0.0955
Field 2	50	
Field 3	0.6	
Field 4	true	

2.

Function	binom.dist	Answer
Field 1	19	0.0014
Field 2	50	
Field 3	0.6	
Field 4	true	

3.

Function	binom.dist	-binom.dist	Answer
Field 1	25	14	0.0978
Field 2	50	50	
Field 3	0.6	0.6	
Field 4	true	true	

4.

Function	1-binom.dist	Answer
Field 1	30	0.4465
Field 2	50	
Field 3	0.6	
Field 4	true	

TRY IT

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season.

1. What is the expected number shots that scored points in a sample of 80 of DeAndre's shots?
2. What is the probability that DeAndre scored on 60 of the 80 shots?
3. What is the probability that DeAndre scored on more than 50 of the 80 shots?
4. What is the probability that DeAndre scored on between 65 and 75 of the 80 shots?

Click to see Solution

1. $\mu = n \times p = 80 \times 0.613 = 49.04$

2.

Function	binom.dist	Answer
Field 1	60	0.0036
Field 2	80	
Field 3	0.613	
Field 4	false	

3.

Function	1-binom.dist	Answer
Field 1	50	0.3718
Field 2	80	
Field 3	0.613	
Field 4	true	

4.

Function	binom.dist	-binom.dist	Answer
Field 1	75	64	0.0001
Field 2	80	80	
Field 3	0.613	0.613	
Field 4	true	true	



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=100#oembed-1>

Watch this video: Binomial Probability in Excel by Joshua Emmanuel [6:59]

Concept Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n .
2. There are only two possible outcomes, called “success” and, “failure” for each trial. The letter p denotes the probability of a success on one trial and $1 - p$ denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.
4. For each individual trial, the probability of a success, p , and probability of a failure, $1 - p$, remain the same.

The outcomes of a binomial experiment fit a binomial probability distribution. The random

variable X is the number of successes obtained in the n independent trials. The mean of a binomial distribution is $\mu = n \times p$ and the standard deviation is $\sigma = \sqrt{n \times p \times (1 - p)}$.

Attribution

“4.3 Binomial Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

4.5 THE POISSON DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the Poisson probability distribution and apply it appropriately.

There are two main characteristics of a **Poisson experiment**:

1. The Poisson probability distribution gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial distribution if the probability of success is “small” (such as 0.01) and the number of trials is “large” (such as 1,000).

The random variable X associated with a Poisson experiment is the number of occurrences in the interval of interest. In a Poisson distribution, λ is the average number of occurrences in an interval. The mean of a Poisson probability distribution is $\mu = \lambda$ and the standard deviation is $\sigma = \sqrt{\lambda}$.

EXAMPLE

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in five minutes is three?

Solution:

Let X be the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half an hour) is 12, then the average number of loaves put on the shelf in five minutes is $\frac{5}{30} \times 12 = 2$ loaves of bread.

The probability question asks you to find $P(x = 3)$.

Calculating Poisson Probabilities

CALCULATING POISSON PROBABILITIES IN EXCEL

To calculate probabilities associated with a Poisson experiment in Excel, use the **Poisson.dist(x, λ, logic operator)** function.

- For **x**, enter the number of successes over the interval.
- For **λ**, enter the average number of successes over the interval.
- For the logic operator, enter **false** to find the probability of exactly **x** successes and enter **true** to find the probability of at most (less than or equal to) **x** successes.

The output from the **Poisson.dist** function is:

- the probability of getting exactly x successes over the interval when the logic operator is **false**.
- the probability of at most x successes over the interval when the logic operator is **true**.

Visit the Microsoft page for more information about the **Poisson.dist** function.

NOTE

Because we can only enter false or true into the logic operator, the **Poisson.dist** function can only directly calculate the probability of getting exactly x successes or getting at most x success over the interval. In order to calculate other Poisson probabilities, such as fewer than x successes, more than x successes or at least x successes, we need to manipulate how we use the **Poisson.dist** function by changing what we enter into the **Poisson.dist** function, using the complement rule, or both.

EXAMPLE

Leah receives about six telephone calls every two hours.

1. What is the probability that Leah receives exactly 4 calls in the next two hours?
2. What is the probability that Leah receives at most 9 calls in the next two hours?
3. What is the probability that Leah receives at most 2 calls in the next hour?

Solution:

1. The average number of calls in any two hour period is 6, so $\lambda = 6$.

Function	Poisson.dist	Answer
Field 1	4	0.1339
Field 2	6	
Field 3	false	

The probability that Leah receives 4 calls in the next two hours is 13.39%.

2. The average number of calls in any two hour period is 6, so $\lambda = 6$.

Function	Poisson.dist	Answer
Field 1	9	0.9161
Field 2	6	
Field 3	true	

The probability that Leah receives at most 6 calls in the next two hours is 91.61%.

3. The average number of calls in any two hour period is 6. So the average number of calls in one hour is $\frac{6}{2} = 3$.

Function	Poisson.dist	Answer
Field 1	2	0.4232
Field 2	3	
Field 3	true	

The probability that Leah receives at most 6 calls in the next two hours is 42.32%.

TRY IT

The customer service department of a technology company receives an average of 10 phone calls every hour.

1. What is the probability that the customer service department receives exactly 7 phone calls in an hour?
2. What is the probability that the customer service department receives exactly 2 phone calls in a 15 minute period?
3. What is the probability that the customer service department receives at most 4 phone calls in a 30 minute period?
4. What is the probability that the customer service department receives at most 20 phone calls in a three hour period?

Click to see Solution

1.

Function	Poisson.dist	Answer
Field 1	7	0.0901
Field 2	10	
Field 3	false	

2.

Function	Poisson.dist	Answer
Field 1	2	0.2565
Field 2	2.5	
Field 3	false	

3.

Function	Poisson.dist	Answer
Field 1	4	0.4405
Field 2	5	
Field 3	true	

4.

Function	Poisson.dist	Answer
Field 1	20	0.0353
Field 2	30	
Field 3	true	

EXAMPLE

According to Baydin, an email management company, an email user gets, on average, 147 emails over a six hour period.

1. What is the probability that an email user receives fewer than 160 emails over an six hour period?
2. What is the probability that an email user receives more than 40 emails over a two hour period?
3. What is the probability that an email user receives at least 600 emails over a 24 hour period?
4. What is the probability that an email user receives between 150 and 200 emails over a six hour period?

Solution:

1. The average over a six hour period is 147. We want to find $P(x < 160)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate = or \leq probabilities. Because x must be an integer (it is the number of emails), $x < 160$ is the same as $x \leq 159$. So $P(x < 160) = P(x \leq 159)$ and $P(x \leq 159)$ is a probability we can calculate with the Poisson.dist function.

Function	Poisson.dist	Answer
Field 1	159	0.8486
Field 2	147	
Field 3	true	

The probability a user receives fewer than 160 emails over a six hour period is 84.86%.

2. The average over a two hour period is $\frac{147}{3} = 49$. We want to find $P(x > 40)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate = or \leq probabilities. The complement of $>$ is \leq , so $P(x > 40) = 1 - P(x \leq 40)$ and $P(x \leq 40)$ is a probability we can calculate with the Poisson.dist function.

Function	1-Poisson.dist	Answer
Field 1	40	0.8902
Field 2	49	
Field 3	true	

The probability a user receives more than 40 emails over a two hour period is 89.02%.

3. The average over a 24 hour period is $147 \times 4 = 588$. We want to find $P(x \geq 600)$. We cannot find this probability directly in Excel because the Poisson.dist function can only calculate = or \leq probabilities. The complement of \geq is $<$, so $P(x \geq 600) = 1 - P(x < 600)$. Because x must be an integer (it is the number of emails), $x < 600$ is the same as $x \leq 599$. So $P(x \geq 600) = 1 - P(x < 600) = 1 - P(x \leq 599)$ and $P(x \leq 599)$ is a probability we can calculate with the Poisson.dist function.

Function	1-Poisson.dist	Answer
Field 1	599	0.3158
Field 2	588	
Field 3	true	

The probability a user receives at least 600 emails over a 24-hour period is 31.58%.

4. We want to find $P(150 \leq x \leq 200)$. We cannot find this probability directly in Excel

because the Poisson.dist function can only calculate = or \leq probabilities. But, $P(150 \leq x \leq 200) = P(x \leq 200) - P(x \leq 149)$. So we can calculate $P(150 \leq x \leq 200)$ as the difference of two Poisson.dist functions.

Function	Poisson.dist	-Poisson.dist	Answer
Field 1	200	149	0.4132
Field 2	147	147	
Field 3	true	true	

The probability a user receives between 150 and 200 emails over a six hour period is 41.32%.

TRY IT

A car parts manufacturer can produce an average of 25 parts from 100 meters of sheet metal.

1. What is the probability that more than 30 parts can be made from 100 meters of sheet metal?
2. What is the probability that between 10 and 20 parts can be made from 50 meters of sheet metal?
3. What is the probability that fewer than 5 parts can be made from 25 meters of sheet metal?
4. What is the probability that at least 80 parts can be made from 400 meters of sheet metal?

Click to see Solution

1.

Function	1-Poisson.dist	Answer
Field 1	30	0.1367
Field 2	25	
Field 3	true	

2.

Function	Poisson.dist	-Poisson.dist	Answer
Field 1	20	9	0.7813
Field 2	12.5	12.5	
Field 3	true	true	

3.

Function	Poisson.dist	Answer
Field 1	4	0.2530
Field 2	6.25	
Field 3	true	

4.

Function	1-Poisson.dist	Answer
Field 1	79	0.9825
Field 2	100	
Field 3	true	



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=102#oembed-1>

Watch this video: The Poisson Distribution by Dr. Nic's Math and Stats [7:48]

Concept Review

A Poisson probability distribution of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to

approximate the binomial, if the probability of success is “small” (less than or equal to 0.05) and the number of trials is “large” (greater than or equal to 20).

Attribution

“4.6 Poisson Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

4.6 EXERCISES

1. A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X be the number of years a new hire will stay with the company. Let $P(x)$ be the probability that a new hire will stay with the company x years.

a. Complete the table using the data provided.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

b. $P(x = 4) = ?$

c. $P(x \geq 5) = ?$

d. On average, how long would you expect a new hire to stay with the company?

e. What does the column " $P(x)$ " sum to?

2. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

- Define the random variable X .
 - What is the probability the baker will sell more than one batch?
 - What is the probability the baker will sell exactly one batch?
 - On average, how many batches should the baker make?
3. Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.
- Define the random variable X .
 - Construct a probability distribution table for the data.
4. We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?
5. Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.
- Define the random variable X .
 - What values does x take on?
 - Construct a PDF table.
 - Find the probability that Javier volunteers for less than three events each month.
 - Find the probability that Javier volunteers for at least one event each month.
6. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in the following table.

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

- In words, define the random variable X .
- What does it mean that the values zero, one, and two are not included for x in the PDF?

7. Complete the expected value table.

x	$P(x)$	$x \times P(x)$
0	0.2	
1	0.2	
2	0.4	
3	0.2	

8. Find the expected value from the expected value table.

x	$P(x)$	$x \times P(x)$
2	0.1	0.2
4	0.3	1.2
6	0.4	2.4
8	0.2	1.6

9. Find the standard deviation.

x	$P(x)$	$x \times P(x)$	$(x - \mu)^2 \times P(x)$
2	0.1	$2(0.1) = 0.2$	$(2 - 5.4)^2 \times 0.1 = 1.156$
4	0.3	$4(0.3) = 1.2$	$(4 - 5.4)^2 \times 0.3 = 0.588$
6	0.4	$6(0.4) = 2.4$	$(6 - 5.4)^2 \times 0.4 = 0.144$
8	0.2	$8(0.2) = 1.6$	$(8 - 5.4)^2 \times 0.2 = 1.352$

10. Identify the mistake in the probability distribution table.

x	$P(x)$	$x \times P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

11. Identify the mistake in the probability distribution table.

x	$P(x)$	$x \times P(x)$
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

12. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	$P(x)$	$x \times P(x)$
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

- Define the random variable X .
 - Define $P(x)$, or the probability of x .
 - Find the probability that a physics major will do post-graduate research for four years.
 - Find the probability that a physics major will do post-graduate research for at most three years.
 - On average, how many years would you expect a physics major to spend doing post-graduate research?
13. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.
- Let X be the number of years a student will study ballet with the teacher.
 - Let $P(x)$ be the probability that a student will study ballet x years.
- Complete the table using the data provided.

x	$P(x)$	$x \times P(x)$
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

- b. In words, define the random variable X .
- c. $P(x = 4) = ?$
- d. $P(x < 4) = ?$
- e. On average, how many years would you expect a child to study ballet with this teacher?
- f. What does the column $P(x)$ sum to and why?
- g. What does the column " $x \times P(x)$ " sum to and why?
14. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win \$30. If it is not a face card, you pay \$2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?
15. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win \$30. If it is not a face card, you pay \$2. There are 12 face cards in a deck of 52 cards. Should you play the game?
16. A theater group holds a fund-raiser. It sells 100 raffle tickets for \$5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of \$150.
- a. What are you interested in here?
- b. In words, define the random variable X .
- c. List the values that X may take on.
- d. Construct a PDF.
- e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?
17. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win \$6
- If the card is a face card, and the coin lands on Tails, you win \$2
- If the card is not a face card, you lose \$2, no matter what the coin shows.

- a. Find the expected value for this game (expected net gain or loss).
- b. Explain what your calculations indicate about your long-term average profits and losses on this game.
- c. Should you play this game to win money?

18. You buy a lottery ticket to a lottery that costs \$10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one \$500 prize, two \$100 prizes, and four \$25 prizes. Find your expected gain or loss.

19. Complete the PDF and answer the questions.

x	$P(x)$	$x \times P(x)$
0	0.3	
1	0.2	
2		
3	0.4	

- a. Find the probability that $x = 2$.
- b. Find the expected value.

20. Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win \$10. If you roll a four or five, you win \$5. If you roll a one, two, or three, you pay \$6.

- a. What are you ultimately interested in here (the value of the roll or the money you win)?
- b. In words, define the random variable X .
- c. List the values that X may take on.
- d. Construct a PDF.
- e. Over the long run of playing this game, what are your expected average winnings per game?
- f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

21. A venture capitalist, willing to invest \$1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning \$5,000,000 profit, a 30% chance of

returning \$1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning \$3,000,000 profit, a 40% chance of returning \$1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning \$6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

- Construct a PDF for each investment.
- Find the expected value for each investment.
- Which is the safest investment? Why do you think so?
- Which is the riskiest investment? Why do you think so?
- Which investment has the highest expected return, on average?

22. Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X be the number of children married people have.

x	$P(x)$	$x \times P(x)$
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

- Find the probability that a married adult has three children.
- In words, what does the expected value in this example represent?
- Find the expected value.
- Is it more likely that a married adult will have two to three children or four to six children?
How do you know?

23. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in following table. On average, how many years do you expect it to take for an individual to earn a B.S.?

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

24. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.70
5	0.04

- Describe the random variable X in words.
- Find the probability that a customer rents three DVDs.
- Find the probability that a customer rents at least four DVDs.
- Find the probability that a customer rents at most two DVDs.

Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	$P(x)$
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

- e. At which store is the expected number of DVDs rented per customer higher?
- f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
- g. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
- h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

25. A “friend” offers you the following “deal.” For a \$10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth \$6.
- Eighty of the coupons are for a free gift worth \$8.
- Six of the coupons are for a free gift worth \$12.
- Four of the coupons are for a free gift worth \$40.

Based upon the financial gain or loss over the long run, should you play the game?

26. Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

- a. What is the average class size assuming each class is filled to capacity?
- b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student’s class. Define the PDF for X .

- c. Find the mean of X .
- d. Find the standard deviation of X .

27. In a lottery, there are 250 prizes of \$5, 50 prizes of \$25, and ten prizes of \$100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

28. The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

- a. In words, define the random variable X .
- b. What values does the random variable X take on?
- c. Construct the probability distribution function (PDF) table for X .
- d. On average, how many would you expect to answer yes?
- e. What is the standard deviation?
- f. What is the probability that at most five of the freshmen reply “yes”?
- g. What is the probability that at least two of the freshmen reply “yes”?

29. According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

30. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

- a. Define the random variable and list its possible values.
- b. State the distribution of X .
- c. Find the probability that at least four of the 25 patients actually have the flu.
- d. On average, for every 25 patients calling in, how many do you expect to have the flu?

31. A school newspaper reporter decides to randomly survey 12 students to see if they will attend

Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

- In words, define the random variable X .
- List the values that X may take on.
- How many of the 12 students do we expect to attend the festivities?
- Find the probability that at most four students will attend.
- Find the probability that more than two students will attend.

32. The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

- What is the expected number of wins for that upcoming month?
- What is the probability that the San Jose Sharks win six games in that upcoming month?
- What is the probability that the San Jose Sharks win at least five games in that upcoming month?

33. A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

34. A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

35. Six different colored dice are rolled. Of interest is the number of dice that show a one.

- In words, define the random variable X .
- List the values that X may take on.
- On average, how many dice would you expect to show a one?
- Find the probability that all six dice show a one.
- Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

36. More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

- In words, define the random variable X .

- b. List the values that X may take on.
 - c. On average, how many schools would you expect to offer such courses?
 - d. Find the probability that at most ten offer such courses.
 - e. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.
37. Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.
- a. In words, define the random variable X .
 - b. List the values that X may take on.
 - c. How many are expected to attend their graduation?
 - d. Find the probability that 17 or 18 attend.
 - e. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.
38. At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.
- a. In words, define the random variable X .
 - b. List the values that X may take on.
 - c. How many are expected to **not** use the foil as their main weapon?
 - d. Find the probability that six do **not** use the foil as their main weapon.
 - e. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.
39. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.
- a. In words, define the random variable X .
 - b. List the values that X may take on.
 - c. How many seniors are expected to have participated in after-school sports all four years of high school?
 - d. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

- e. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

40. The chance of an IRS audit for a tax return with over \$25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

- In words, define the random variable X .
- List the values that X may take on.
- How many audits are expected in a 20-year period?
- Find the probability that a person is not audited at all.
- Find the probability that a person is audited more than twice.

41. It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

- In words, define the random variable X .
- List the values that X may take on.
- What is the probability that at least eight have adequate earthquake supplies?
- Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
- How many residents do you expect will have adequate earthquake supplies?

42. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being \$1. The player places a bet on a number or object. The “house” rolls three dice. If none of the dice show the number or object that was bet, the house keeps the \$1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her \$1 bet, plus \$1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her \$1 bet, plus \$2 profit. If all three dice show the number or object bet, the player gets back his or her \$1 bet, plus \$3 profit. Let X be the number of matches and Y be the profit per game.

- In words, define the random variable X .
- List the values that X may take on.

- c. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
- d. Calculate the average expected matches over the long run of playing this game for the player.
- e. Calculate the average expected earnings over the long run of playing this game for the player.
- f. Determine who has the advantage, the player or the house.

43. According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X be the number of people who have access to electricity.

- a. What is the probability distribution for X ?
- b. Using the formulas, calculate the mean and standard deviation of X .
- c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
- d. Find the probability that at most ten people in the sample have access to electricity.
- e. Find the probability that more than 25 people in the sample have access to electricity.

44. The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X be the number of people who are literate.

- a. What is the mean of X ?
- b. What is the standard deviation of X ?
- c. Find the probability that more than five people in the sample are literate. Is it more likely that three people or four people are literate.

45. On average, a clothing store gets 120 customers per day.

- a. Assume the event occurs independently in any given day. Define the random variable X .
- b. What values does X take on?
- c. What is the probability of getting 150 customers in one day?
- d. What is the probability of getting 35 customers in the first four hours? Assume the store is open 12 hours each day.
- e. What is the probability that the store will have more than 12 customers in the first hour?
- f. What is the probability that the store will have fewer than 12 customers in the first two hours?

46. On average, eight teens in the U.S. die from motor vehicle injuries per day. As a result, states across the country are debating raising the driving age.

- a. Assume the event occurs independently in any given day. In words, define the random variable X .
- b. What values does X take on?
- c. For the given values of the random variable X , fill in the corresponding probabilities.
- d. Is it likely that there will be no teens killed from motor vehicle injuries on any given day in the U.S? Justify your answer numerically.
- e. Is it likely that there will be more than 20 teens killed from motor vehicle injuries on any given day in the U.S.? Justify your answer numerically.

47. The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let X be the number of calls received at noon.

- a. Find the mean and standard deviation of X .
- b. What is the probability that the office receives at most six calls at noon on Monday?
- c. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
- d. What is the probability that the office receives more than eight calls at noon?

48. The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let X be the number of births in an hour.

- a. Find the mean and standard deviation of X .
- b. What is the probability that the maternity ward will deliver three babies in one hour?
- c. What is the probability that the maternity ward will deliver at most three babies in one hour?
- d. What is the probability that the maternity ward will deliver more than five babies in one hour?

49. A manufacturer of Christmas tree light bulbs knows that 3% of its bulbs are defective. Find the probability that a string of 100 lights contains at most four defective bulbs using both the binomial and Poisson distributions.

50. The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. Find the probability that she has no children.
- d. Find the probability that she has fewer children than the Japanese average.
- e. Find the probability that she has more children than the Japanese average.

51. The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. Find the probability that she has no children.
- d. Find the probability that she has fewer children than the Spanish average.
- e. Find the probability that she has more children than the Spanish average .

52. Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. Find the probability that she has no litters in one year.
- d. Find the probability that she has at least two litters in one year.
- e. Find the probability that she has exactly three litters in one year.

53. The chance of having an extra fortune in a fortune cookie is about 3%. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem, but only use one distribution to solve the problem.

1. In words, define the random variable X .
2. List the values that X may take on.
3. How many cookies do we expect to have an extra fortune?
4. Find the probability that none of the cookies have an extra fortune.
5. Find the probability that more than three have an extra fortune.
6. As n increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

54. According to the South Carolina Department of Mental Health web site, for every 200 U.S.

women, the average number who suffer from anorexia is one. Out of a randomly chosen group of 600 U.S. women determine the following.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. How many are expected to suffer from anorexia?
- d. Find the probability that no one suffers from anorexia.
- e. Find the probability that more than four suffer from anorexia.

55. The chance of an IRS audit for a tax return with over \$25,000 in income is about 2% per year. Suppose that 100 people with tax returns over \$25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to answer the following questions.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. How many are expected to be audited?
- d. Find the probability that no one was audited.
- e. Find the probability that at least three were audited.

56. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. How many seniors are expected to have participated in after-school sports all four years of high school?
- d. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
- e. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

57. On average, Pierre, an amateur chef, drops three pieces of egg shell into every two cake batters he makes. Suppose that you buy one of his cakes.

- a. In words, define the random variable X .
- b. List the values that X may take on.
- c. On average, how many pieces of egg shell do you expect to be in the cake?

- d. What is the probability that there will not be any pieces of egg shell in the cake?
 - e. Let's say that you buy one of Pierre's cakes each week for six weeks. What is the probability that there will not be any egg shell in any of the cakes?
 - f. Based upon the average given for Pierre, is it possible for there to be seven pieces of shell in the cake? Why?
58. The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is ten. We are interested in the number of times her cats wake her up each week.
- a. In words, describe the random variable X .
 - b. Find the probability that her cats will wake her up no more than five times next week.
-

Attribution

“Chapter 4 Homework” and “Chapter 4 Practice” in Introductory Statistics by OpenStax Rice University is licensed under a Creative Commons Attribution 4.0 International License.

PART V

CONTINUOUS RANDOM VARIABLES AND THE NORMAL DISTRIBUTION

Chapter Outline

- 5.1 Introduction to Continuous Random Variables
- 5.2 Probability Distribution of a Continuous Random Variable
- 5.3 The Normal Distribution
- 5.4 The Standard Normal Distribution
- 5.5 Calculating Probabilities for a Normal Distribution
- 5.6 Exercises

5.1 INTRODUCTION TO CONTINUOUS RANDOM VARIABLES



The heights of these radish plants are continuous random variables. “Radishes” by Rev Stan, CC BY 4.0.

A continuous random variable corresponds to data that can be measured. Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few examples of continuous random variables. The field of reliability depends on a variety of continuous random variables.

NOTE

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable because you count the miles. If X is the distance you drive to work, then X is a continuous random variable because you measure the miles. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable because the number of books is a count. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

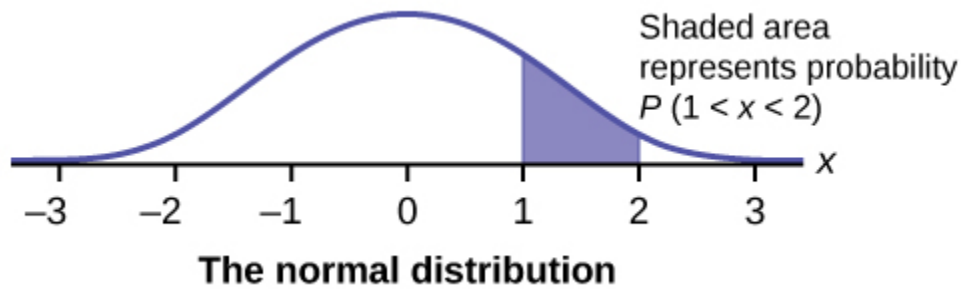
Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. The probability a continuous random variable takes on a value in an interval is the **area** under the curve of the distribution of the continuous random variable. Properties of a continuous random variable X include:

- The outcomes are measured, not counted.
- The entire area under the curve of the distribution of the continuous random variable and above the x -axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < X < d)$ is the probability that the random variable X is in the interval between the values of c and d . The value of $P(c < X < d)$ is the area under the curve, above the x -axis, to the right of c and the left of d .
- The probability that x takes on any single individual value is zero—that is, $P(X = c) = 0$. The area below the curve, above the x -axis, and between $x = c$ and $x = c$ has no width, and therefore no area. Because the probability is equal to the area and the area is 0, the probability is also 0.
- Because probability is equal to area and $P(X = c) = 0$, $P(c < X < d)$ is the same as $P(c \leq X \leq d)$.

Generally, a calculator is needed to find the area under the curve of many continuous probability distributions. However, we will use the built-in functions in Excel to calculate the area under the continuous probability distribution functions. There are many different continuous probability

distributions, including the uniform distribution and the exponential distribution. We will focus on the most important continuous probability distribution—the normal distribution.



The graph shows the Standard Normal Distribution with the area between $x=1$ and $x=2$ shaded to represent the probability that the value of the random variable x is in the interval between one and two.

Attribution

“Chapter 5 Introduction” in *Introductory Statistics* by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

5.2 PROBABILITY DISTRIBUTION OF A CONTINUOUS RANDOM VARIABLE

LEARNING OBJECTIONS

- Recognize and understand continuous probability distributions.

For a continuous random variable, the curve of the probability distribution is denoted by the function $f(x)$. The function $f(x)$ is called a probability density function and $f(x)$ produces the curve of the distribution. The function $f(x)$ is defined so that the **area** between it and the x -axis is equal to a probability.

NOTE

The probability density function $f(x)$ does NOT give us probabilities associated with the continuous random variable. The function $f(x)$ produces the graph of the distribution and the area under this graph corresponds to the probability.

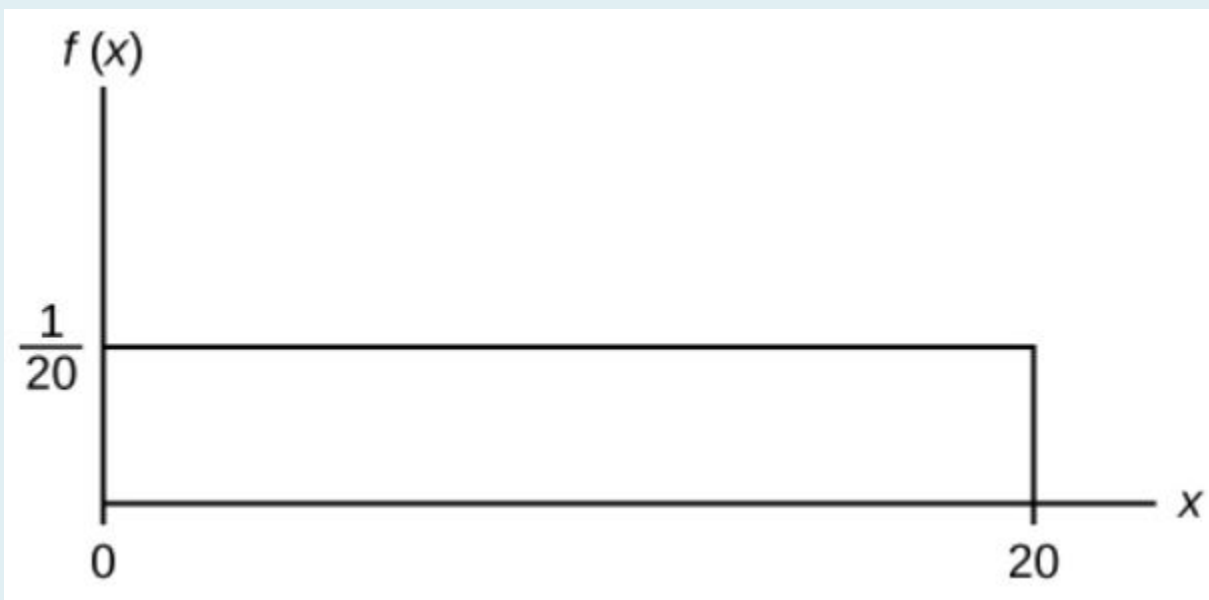
Properties of a continuous probability distribution include:

1. The total area under the curve of the distribution is 1.
2. The probability that the continuous random variable takes on a value in between c and d is the area under the curve of the distribution in between $x = c$ and $x = d$.
3. The probability that the continuous random variable exactly equals a particular number (

$P(x = c)$ is 0.

EXAMPLE

Consider the probability density function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$.



The graph of $f(x) = \frac{1}{20}$ with $0 \leq x \leq 20$ is a horizontal line segment from $x = 0$ to $x = 20$.

Note that the total area under the curve of $f(x)$, above the x -axis, from $x = 0$ to $x = 20$ is

$$\text{Area} = 20 \times \frac{1}{20} = 1$$

Suppose we want to find the area between $f(x)$ and the x -axis for $0 < x < 2$.



In this case, the area equals the area of a rectangle from $x = 0$ to $x = 2$. The area of a rectangle is base \times height, so

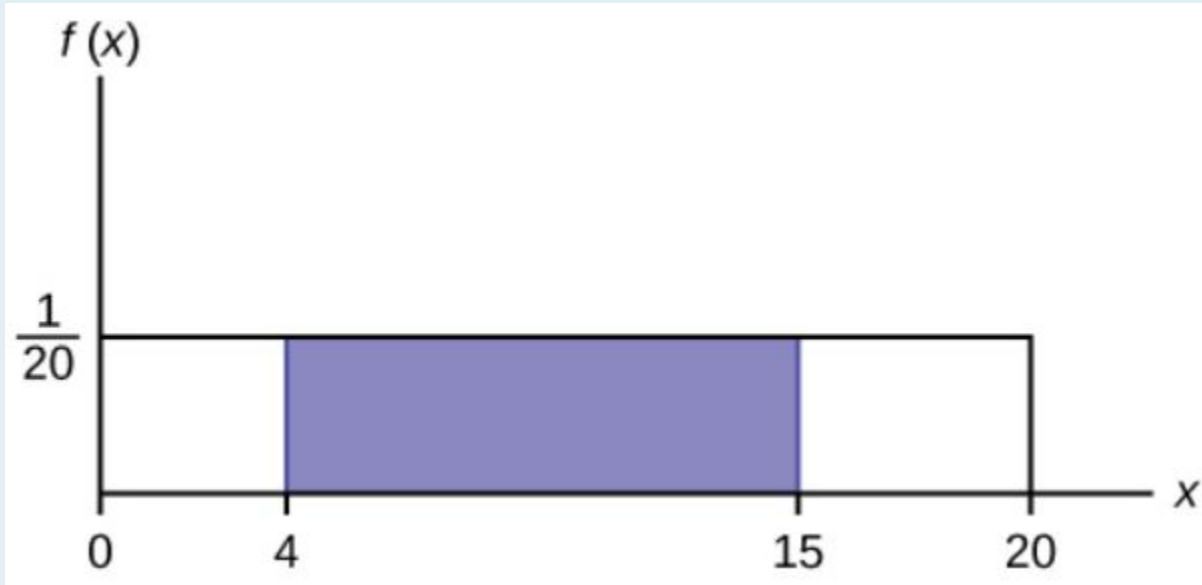
$$\text{Area} = (2 - 0) \times \frac{1}{20} = 0.1$$

The area corresponds to the probability that the associated continuous random variable takes on a value between $x = 0$ and $x = 2$. Because the area is 0.1, the probability that $0 < x < 2$ is 0.1. Mathematically, we can write this as:

$$P(0 < x < 2) = 0.1$$

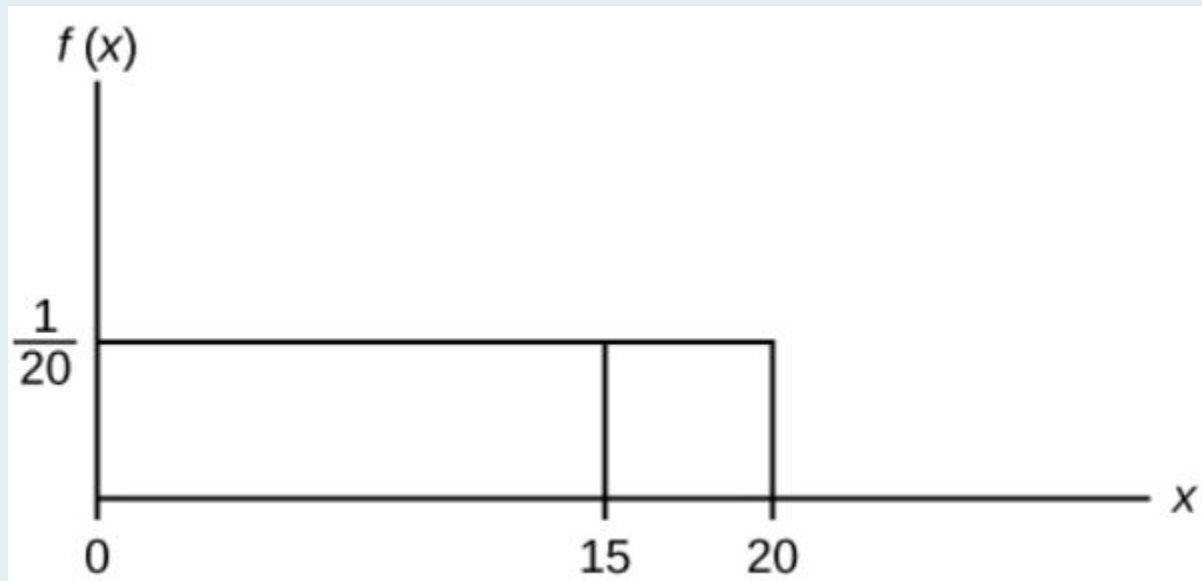
Suppose we want to find the probability that the random variable takes on a value between $x = 4$ and $x = 15$. This corresponds to the area under the curve in between $x = 4$ and $x = 15$.

$$\text{Area} = P(4 < x < 15) = (15 - 4) \times \frac{1}{20} = 0.55$$



Suppose we want to find $P(x = 15)$. This corresponds to the area above $x = 15$, which is just a vertical line. A vertical line has no width (or zero width). So

$$\text{Area} = P(x = 15) = 0 \times \frac{1}{20} = 0$$



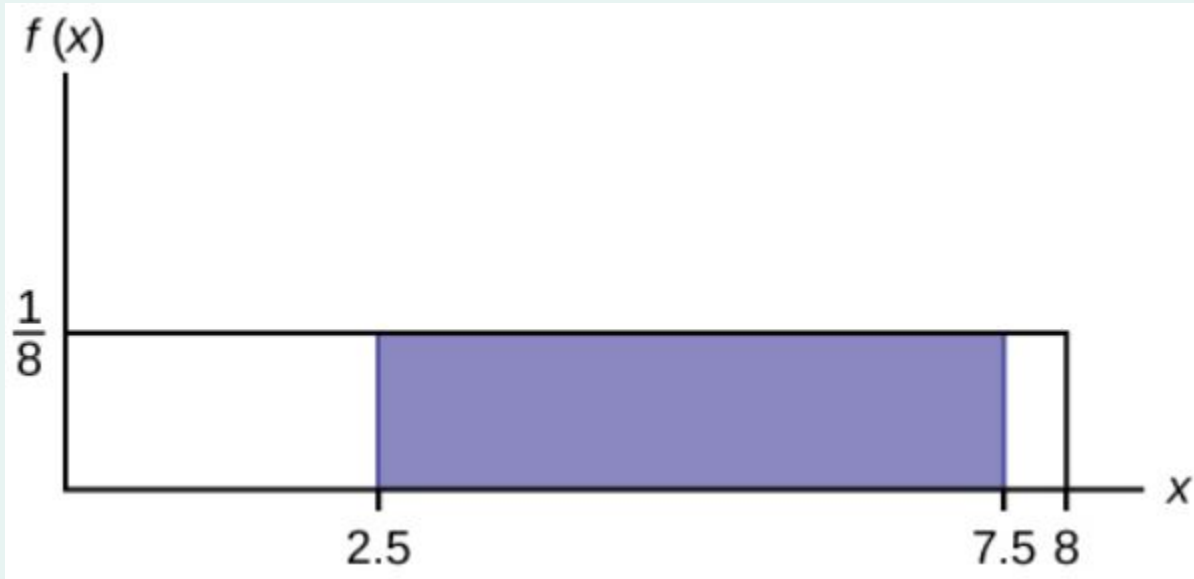
NOTE

The probability density function $f(x) = \frac{1}{20}$ used above is an example of a uniform distribution. The graph of a uniform distribution is always a horizontal line.

TRY IT

Consider the probability density function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Click to see Solution



$$P(2.5 < x < 7.5) = (7.5 - 2.5) \times \frac{1}{8} = 0.625$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=116#oembed-1>

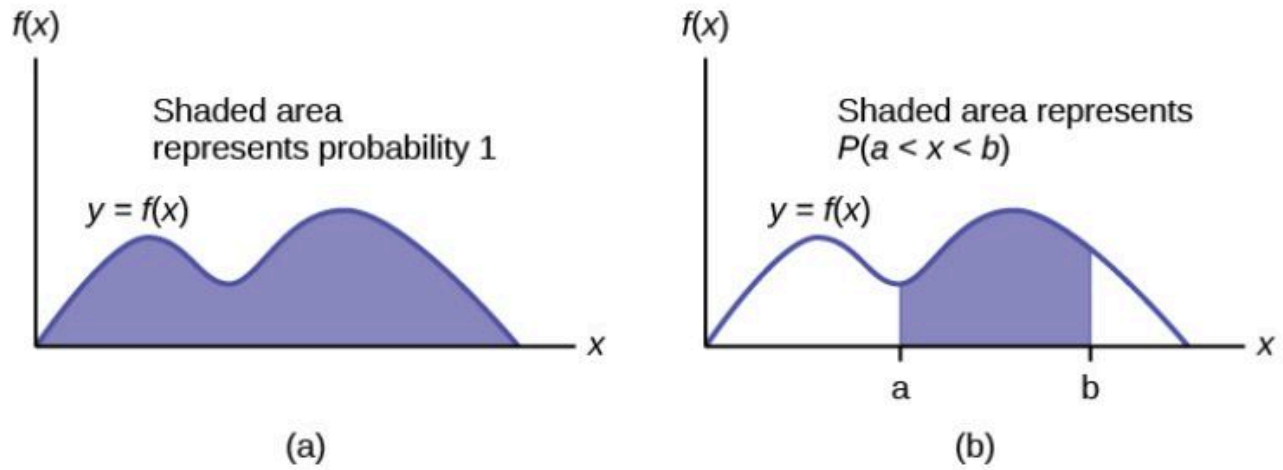
Watch this video: Continuous probability distribution intro by Khan Academy [9:57]

Concept Review

The probability density function describes the curve of a continuous random variables. The area under the probability density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the probability density curve between points a and b is equal to $P(a < x < b)$.

If X is a continuous random variable, the probability density function, $f(x)$, is used to draw the

graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.



Attribution

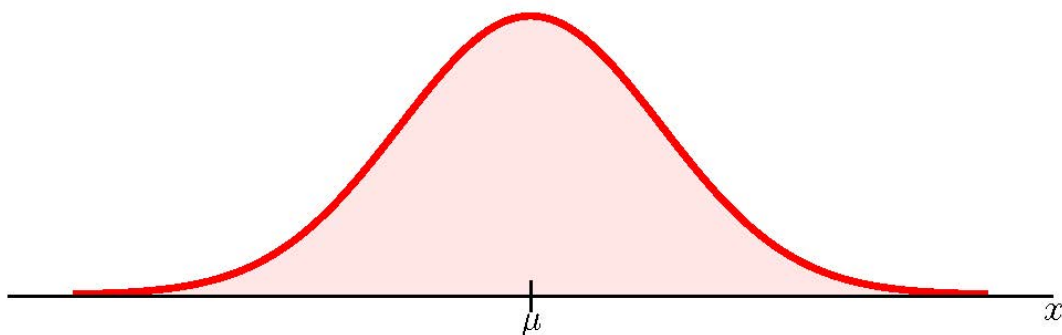
“5.1 Continuous Probability Functions“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

5.3 THE NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Describe properties of the normal distribution.
- Apply the Empirical Rule for normal distributions.

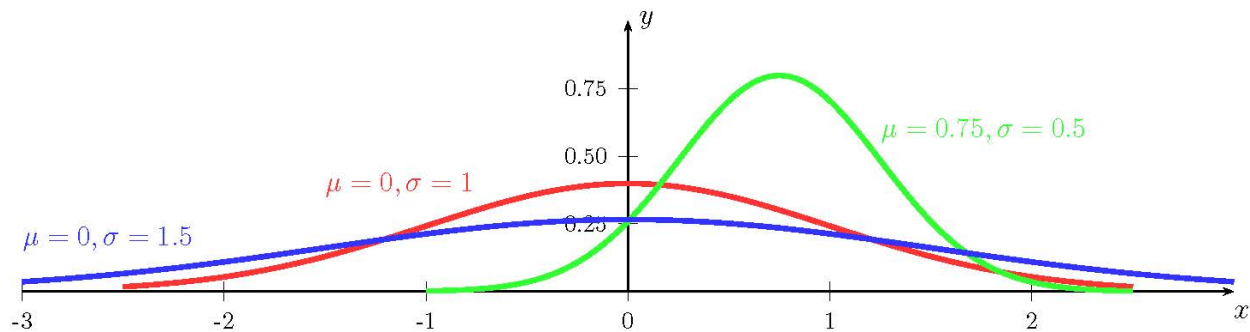
The normal distribution is the most important of all the distributions. It is widely used and even more widely abused. Its graph is a symmetric, bell-shaped curve. You see the bell curve in almost all disciplines, including psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.



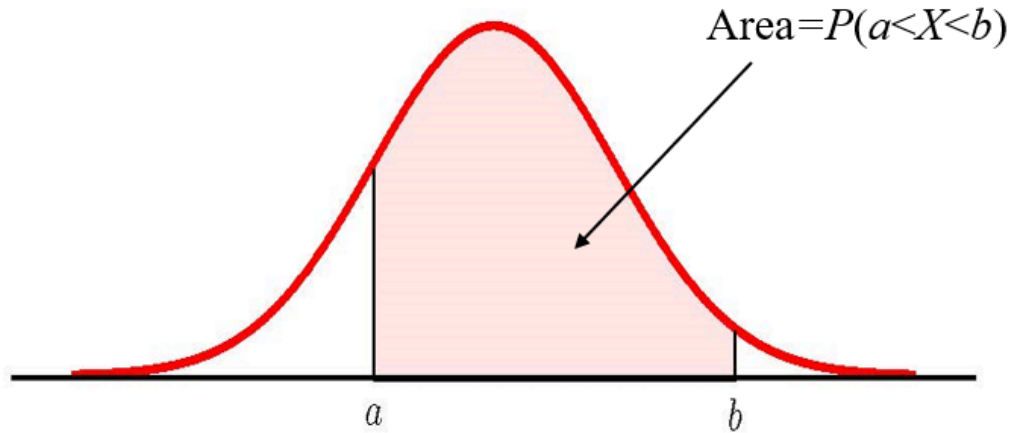
Properties of the normal distribution include:

- The curve of a normal distribution is symmetric and bell-shaped.
- The center of a normal distribution is at the mean μ .
- In a normal distribution, the mean, the median, and the mode are equal.
- The curve is symmetric about a vertical line drawn through the mean.
- The tails of a normal distribution extend to infinity in both directions along the x -axis.
- The standard deviation, σ , of a normal distribution determines how wide or narrow the curve is.
- The total area under the curve of a normal distribution equals 1.

A normal distribution is completely determined by its mean μ and its standard deviation σ , which means there are an infinite number of normal distributions. The mean μ determines the center of the distribution—a change in the value of μ causes the graph to shift to the left or right. The standard deviation σ determines the shape of the bell. Because the area under the curve must equal one, a change in the standard deviation σ causes a change in the shape of the curve—the curve becomes fatter or skinnier depending on the value of σ .



The normal distribution is a continuous probability distribution. As we saw in the previous section, the area under the curve of the normal distribution equals the probability that the corresponding normal random variable takes on a value within a given interval. That is, the probability that the normal random variable is in between a and b equals the area under the normal curve in between $x = a$ and $x = b$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=122#oembed-1>

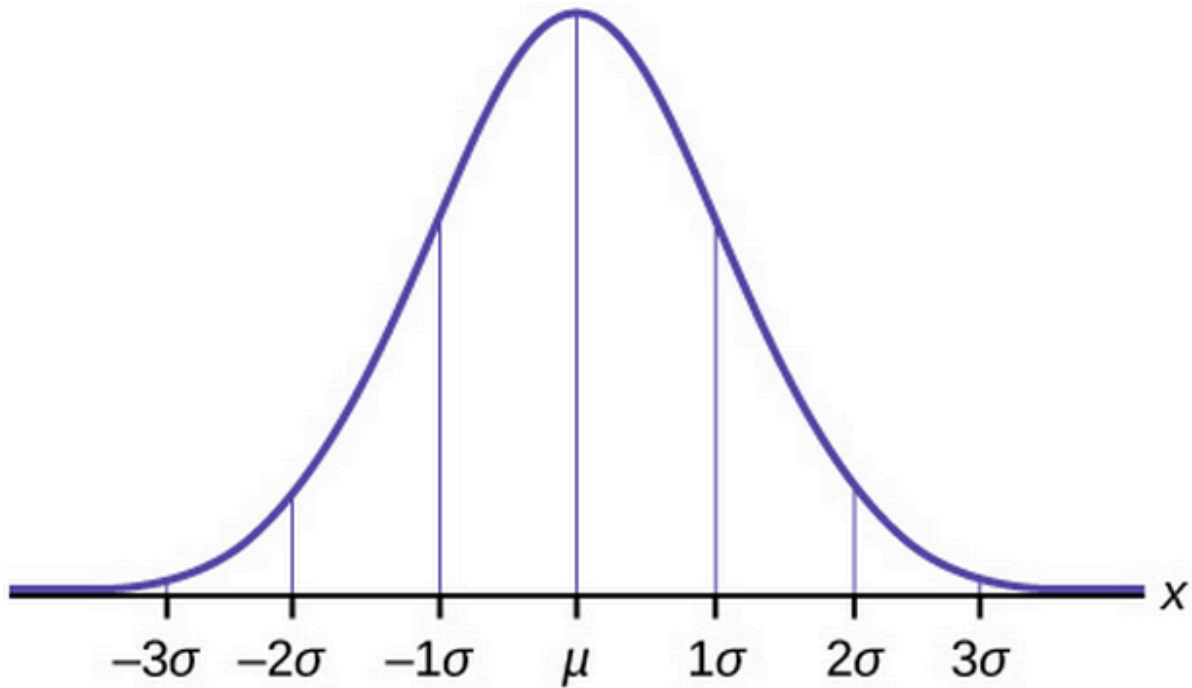
Watch this video: ck12.org normal distribution problems: Quantitative sense of normal distributions | Khan Academy by Khan Academy [10:52]

The Empirical Rule

For a normal distribution with mean μ and standard deviation σ , then the **Empirical Rule** says the following:

- About 68% of the values lie between $-1 \times \sigma$ and $+1 \times \sigma$ of the mean μ .
 - In other words, about 68% of the data fall with one standard deviation of the mean.
- About 95% of the values lie between $-2 \times \sigma$ and $+2 \times \sigma$ of the mean μ .
 - In other words, about 95% of the data fall with two standard deviation of the mean.
- About 99.7% of the values lie between $-3 \times \sigma$ and $+3 \times \sigma$ of the mean μ .
 - In other words, about 99.7% of the data fall with three standard deviation of the mean

The empirical rule is also known as the **68 – 95 – 99.7 rule**.



EXAMPLE

Suppose a normal distribution has a mean 50 and a standard deviation 6.

About 68% of the values lie between $-1 \times \sigma = -1 \times 6 = -6$ and $1 \times \sigma = 1 \times 6 = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50. So 68% of the values in this distribution are between 44 and 56.

About 95% of the values lie between $-2 \times \sigma = -2 \times 6 = -12$ and $2 \times \sigma = 2 \times 6 = 12$ of the mean 50. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations of the mean 50. So 95% of the values in the distribution are between 38 and 62.

About 99.7% of the x values lie between $-3 \times \sigma = -3 \times 6 = -18$ and $3 \times \sigma = 3 \times 6 = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. So 99.7% of the values in the distribution are between 32 and 68.

TRY IT

Suppose a normal distribution has a mean 25 and a standard deviation 5. Between what values of does 68% of the data lie?

Click to see Solution

- Between $25 + (-1) \times 5 = 20$ and $25 + 1 \times 5 = 30$.

EXAMPLE

From 1984 to 1985, the height of 15 to 18-year-old males from Chile follows a normal distribution with mean 172.36cm and standard deviation 6.34cm.

1. About 68% of the heights of 15 to 18-year old males in Chile from 1984 to 1985 lie between what two values?
2. About 95% of the heights of 15 to 18-year old males in Chile from 1984 to 1985 lie between what two values?
3. About 99.7% of the heights of 15 to 18-year old males in Chile from 1984 to 1985 lie between what two values?

Solution:

1. $\mu + (-1) \times \sigma = 172.36 + (-1) \times 6.34 = 166.02$ and
 $\mu + 1 \times \sigma = 172.36 + 1 \times 6.34 = 178.70$
2. $\mu + (-2) \times \sigma = 172.36 + (-2) \times 6.34 = 159.68$ and

$$\mu + 2 \times \sigma = 172.36 + 2 \times 6.34 = 185.04$$

$$3. \mu + (-3) \times \sigma = 172.36 + (-1) \times 6.34 = 153.34 \text{ and}$$

$$\mu + 3 \times \sigma = 172.36 + 2 \times 6.34 = 191.36$$

TRY IT

The scores on a college entrance exam have an approximate normal distribution with a mean 52 points and a standard deviation of 11 points.

1. About 68% of the exam scores lie between what two values?
2. About 95% of the exam scores lie between what two values?
3. About 99.7% of the exam scores lie between what two values?

Click to see Solution

1. About 68% of the scores lie between the values $52 + (-1) \times 11 = 41$ and $52 + 1 \times 11 = 63$.
2. About 95% of the values lie between the values $52 + (-2) \times 11 = 30$ and $52 + 2 \times 11 = 74$.
3. About 99.7% of the values lie between the values $52 + (-3) \times 11 = 19$ and $52 + 3 \times 11 = 85$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=122#oembed-2>

Watch this video: Empirical Rule| Probability and Statistics | Khan Academy by Khan Academy [10:25]

Concept Review

The normal distribution is the most frequently used distribution in statistics. The graph of a normal distribution is a symmetric, bell-shaped curve centered at the mean of the distribution. The probability that a normal random variable takes on a value in inside an interval equals the area under the corresponding normal distribution curve.

For a normal distribution, the empirical rule states that 68% of the data falls within one standard deviation of the mean, 95% of the data falls within two standard deviations, and 99.7% of the data falls within three standard deviations of the mean.

Attribution

“6.1 The Standard Normal Distribution” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

5.4 THE STANDARD NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the standard normal probability distribution and apply it appropriately

The **standard normal distribution** is the normal distribution with $\mu = 0$ and $\sigma = 1$. The normal random variable associated with the standard normal distribution is denoted Z .

For any normal distribution with mean μ and standard deviation σ , a **z-score** is the number of the standard deviations a value x is from the mean. For example, if a normal distribution has $\mu = 5$ and $\sigma = 2$, then for $x = 11$

$$11 = x = \mu + z \times \sigma = 5 + 3 \times 2$$

In this case, $z = 3$. We would say that 11 is three standard deviations above (or to the right of) the mean.

The standard normal distribution is a normal distribution of these **standardized z-scores**. For any normal distribution with mean μ and standard deviation σ , we can transform the normal distribution to the standard normal distribution using the formula

$$z = \frac{x - \mu}{\sigma}$$

where x is a value from the normal distribution. The z -score is the number of standard deviations the value x is above (to the right of) or below (to the left of) the mean μ . Values of x that are larger than the mean have positive z -scores and values of x that are smaller than the mean have negative z -scores. If x equals the mean, then x has a z -score of zero.

EXAMPLE

Suppose a normal distribution has mean $\mu = 5$ and standard deviation $\sigma = 6$.

For $x = 17$, the z -score is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{17 - 5}{6} \\ &= 2 \end{aligned}$$

This tells us that $x = 17$ is **two standard deviations** ($2 \times \sigma$) above or to the right of the mean $\mu = 5$. Notice that $x = 5 + 2 \times 6 = 17$

For $x = 1$, the z -score is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{1 - 5}{6} \\ &= -0.666\dots \end{aligned}$$

This tells us that $x = 1$ is 0.666... standard deviations ($-0.666\dots \times \sigma$) below or to the left of the mean $\mu = 5$. Notice that $x = 5 + (-0.666\dots) \times 6 = 1$

NOTES

- When z is positive, x is above or to the right of the mean μ . In other words, x is greater than μ .
- When z is negative, x is below or to the left of the mean μ . In other words, x is less than μ .

TRY IT

What is the z-score of $x=1$ for a normal distribution with $\mu = 12$ and $\sigma = 3$?

Click to see Solution

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 12}{3} = -3.666\dots$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=124#oembed-1>

Watch this video: Normal Distribution Problems: z-score | Probability and Statistics | Khan Academy by Khan Academy
[7:47]

EXAMPLE

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose the amount of weight (in pounds) a person loses in a month has a normal distribution with $\mu = 5$ and $\sigma = 2$. Fill in the blanks.

1. Suppose a person lost ten pounds in a month. The z -score when $x = 10$ pounds is $z = 2.5$ (verify). This z -score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
2. Suppose a person gained three pounds (a negative weight loss). Then $z =$ _____. This z -score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

1. This z -score tells you that $x = 10$ is 2.5 standard deviations to the **right** of the mean 5.
2. $z = -4$. This z -score tells you that $x = -3$ is 4 standard deviations to the **left** of the mean.

EXAMPLE

Suppose X is a normal random variable with $\mu = 5$ and $\sigma = 6$ and Y is a normal random variable with $\mu = 2$ and $\sigma = 1$.

Suppose $x = 17$:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{17 - 5}{6} \\ &= 2 \end{aligned}$$

The z -score for $x = 17$ is $z = 2$, which means that 17 is 2 standard deviations to the right of the mean $\mu = 5$.

Suppose $y = 4$:

$$\begin{aligned} z &= \frac{y - \mu}{\sigma} \\ &= \frac{4 - 2}{1} \\ &= 2 \end{aligned}$$

The z -score for $y = 4$ is $z = 2$, which means that 4 is 2 standard deviations to the right of the mean $\mu = 2$.

Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of their respective means. In other words, compared to the mean of their corresponding distributions, $x = 17$ and $y = 4$ have the same **relative** position.

NOTE

The z -score allows us to compare data that are scaled differently by considering the data's position relative to its mean. To understand the concept, suppose X represents weight gains for one group of people who are trying to gain weight in a six week period and Y measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Because $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain relative to their means.

TRY IT

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of 4 points. Suppose Jerome scores 10 points in a game. The z -score when $x = 10$ is -1.5 . This score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Click to see Solution

- 1.5, left, 16

EXAMPLE

The height of 15 to 18-year-old males from Chile from 2009 to 2010 follow a normal distribution with mean 170cm and standard deviation 6.28cm.

- Suppose a 15 to 18-year-old male from Chile was 168cm tall from 2009 to 2010. The z -score when $x = 168$ cm is $z =$ _____. This z -score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
- Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = 1.27$. What is the male's height? The z -score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

- $-0.32, 0.32$, left, 170

b. 177.98, 1.27, right

TRY IT

The height of 15 to 18-year-old males from Chile from 2009 to 2010 follow a normal distribution with mean 170cm and standard deviation 6.28cm.

1. Suppose a 15 to 18-year-old male from Chile was 176cm tall from 2009 to 2010. The z -score when $x = 176$ cm is $z =$ _____. This z -score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).
2. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z -score of $z = -2$. What is the male's height? The z -score ($z = -2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Click to see Solution:

1. $z = \frac{x - \mu}{\sigma} = \frac{176 - 170}{6.28} = 0.96$. This z -score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170cm.
2. $x = \mu + z \times \sigma = 170 + (-2) \times 6.28 = 157.44$ cm, The z -score ($z = -2$) tells you that the male's height is 2 standard deviations to the left of the mean.

EXAMPLE

From 2009 to 2010, the height of 15 to 18-year-old males from Chile from 2009 to 2010 follows a normal distribution with mean 170cm and standard deviation 6.28cm. Let X be the height of a 15 to 18-year-old male from Chile in 2009 to 2010.

From 1984 to 1985, the heights of 15 to 18-year-old males from Chile follows a normal distribution with mean 172.36cm and standard deviation 6.34cm. Let Y be the height of a 15 to 18-year-old male from Chile in 1984 to 1985.

Find the z -scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z -score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Solution:

The z -score for $x = 160.58$ is

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{160.58 - 170}{6.28} \\ &= -1.5 \end{aligned}$$

The z -score for $y = 162.85$ is

$$\begin{aligned} z &= \frac{y - \mu}{\sigma} \\ &= \frac{162.85 - 172.36}{6.34} \\ &= -1.5 \end{aligned}$$

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

TRY IT

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT followed a normal distribution with a mean of 496 and a standard deviation of 114.

Find the z -scores for Student 1 with a score of 325 and for Student 2 with a score of 366.21. Interpret each z -score. What can you say about these two students' scores?

Click to see Solution

For Student 1:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{325 - 496}{114} \\ &= -1.5 \end{aligned}$$

For Student 2:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{366.21 - 496}{114} \\ &= -1.138... \end{aligned}$$

Student 2 scored closer to the mean than Student 1 and, because they both had negative z -scores, Student 2 had the better score.

Concept Review

The standard normal distribution is the normal distribution with a mean of 0 and a standard deviation of 1. A z -score is a standardized value that allows us to transform any normal distribution

back to standard normal. The formula for a z -score is $z = \frac{x - \mu}{\sigma}$. The value of the z -score for a value x from a normal distribution with μ and standard deviation σ tells us how many standard deviations x is above (greater than) or below (less than) μ .

Attribution

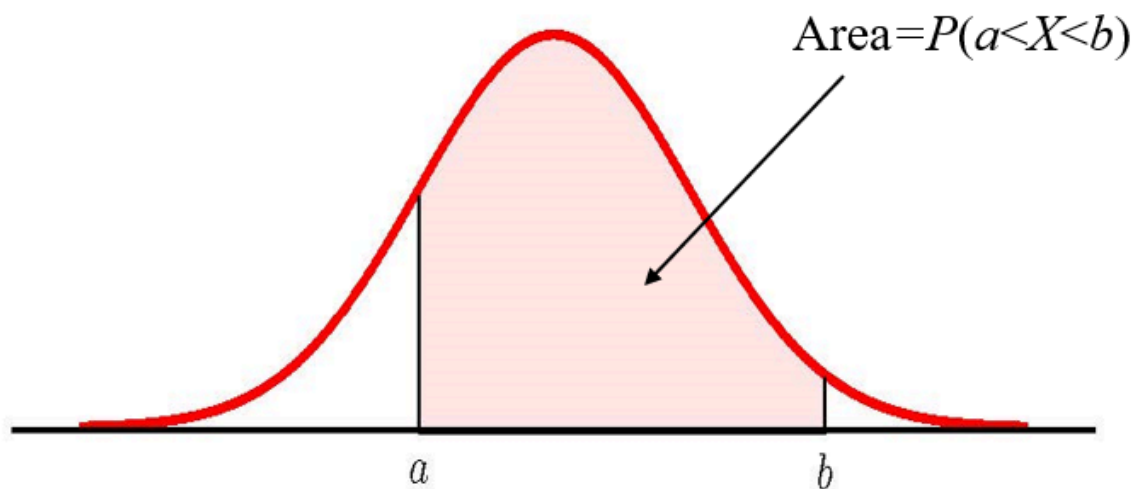
“6.1 The Standard Normal Distribution” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

5.5 CALCULATING PROBABILITIES FOR A NORMAL DISTRIBUTION

LEARNING OBJECTIVES

- Recognize the normal probability distribution and apply it appropriately.
- Calculate probabilities associated with a normal distribution.

Probabilities for a normal random variable X equal the area under the corresponding normal distribution curve. The probability that the value for X falls in between the values $x = a$ and $x = b$ is the area under the normal distribution curve to the right of $x = a$ and to the left of $x = b$.



CALCULATING NORMAL PROBABILITIES IN EXCEL

To calculate probabilities associated with normal random variables in Excel, use the **norm.dist(x,μ,σ,logic operator)** function.

- For **x**, enter the value for x .
- For **μ**, enter the mean of the normal distribution.
- For **σ**, enter the standard deviation of the normal distribution.
- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

The output from the **norm.dist** function is the probability that $X < x$. That is, the output from the **norm.dist** function is the area to the **left** of value of **x**.

Visit the Microsoft page for more information about the **norm.dist** function.

NOTE

The **norm.dist** function always tells us the area to the left of the value entered for **x**.

- To find the area to the right of the value of **x**, we use **1-norm.dist(x,μ,σ,true)**. This corresponds to the probability that $X > x$.
- To find the area in between **x₁** and **x₂** with $x_1 < x_2$, we use **norm.dist(x₂,μ,σ,true)-norm.dist(x₁,μ,σ,true)**. This corresponds to the probability that $x_1 < X < x_2$.

An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the **z**-score for the corresponding value of **x** and the output will be the area to the left **x**.

CALCULATING **Formula does not parse**-VALUES FOR A NORMAL DISTRIBUTION IN EXCEL

Given the area to the left of an (unknown) x -value, use the **norm.inv(probability, μ , σ)** function.

- For **probability**, enter the area to the **left** of x .
- For μ , enter the mean of the normal distribution.
- For σ , enter the standard deviation of the normal distribution.

The output from the **norm.inv** function is the value of x so that the area to left of x equals the given probability. That is, the output from the **norm.inv** function is the value of x so that the $P(X < x) = \text{probability}$.

Visit the Microsoft page for more information about the **norm.inv** function.

NOTE

The **norm.inv** function requires that we enter the area to the **left** of the unknown x -value. If we are given the area to the **right** of the unknown x -value, we enter **1-area to the right** for the probability in the **norm.inv** function. That is, given the area to the **right** of the x -value, we use **norm.inv(1-area, μ , σ)**.

EXAMPLE

The final exam scores in a statistics class are normally distributed with a mean of 63 and a standard deviation of 5.

1. Find the probability that a randomly selected student scored more than 65 on the exam.
2. Find the probability that a randomly selected student scored less than 75.
3. 90% of the students scored less than what value?
4. 30% of the students scored more than what value?

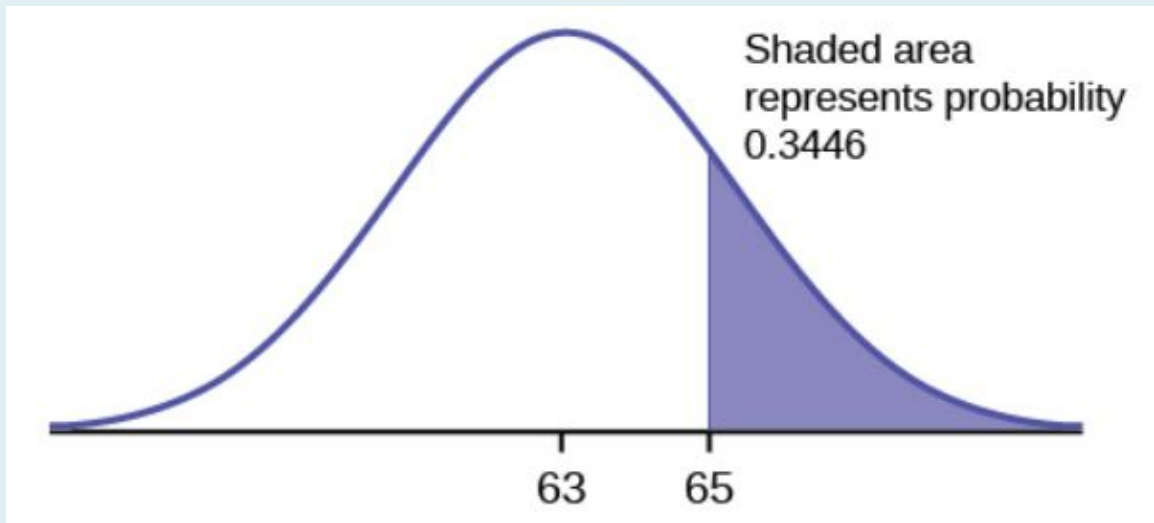
Solution:

Let X be the scores on the final exam.

1. We want to find $P(X > 65)$:

Function	1-norm.dist	Answer
Field 1	65	0.3446
Field 2	63	
Field 3	5	
Field 4	true	

The probability that a student scores more than 65 is 0.3446 (or 34.46%)



2. We want to find $P(X < 75)$:

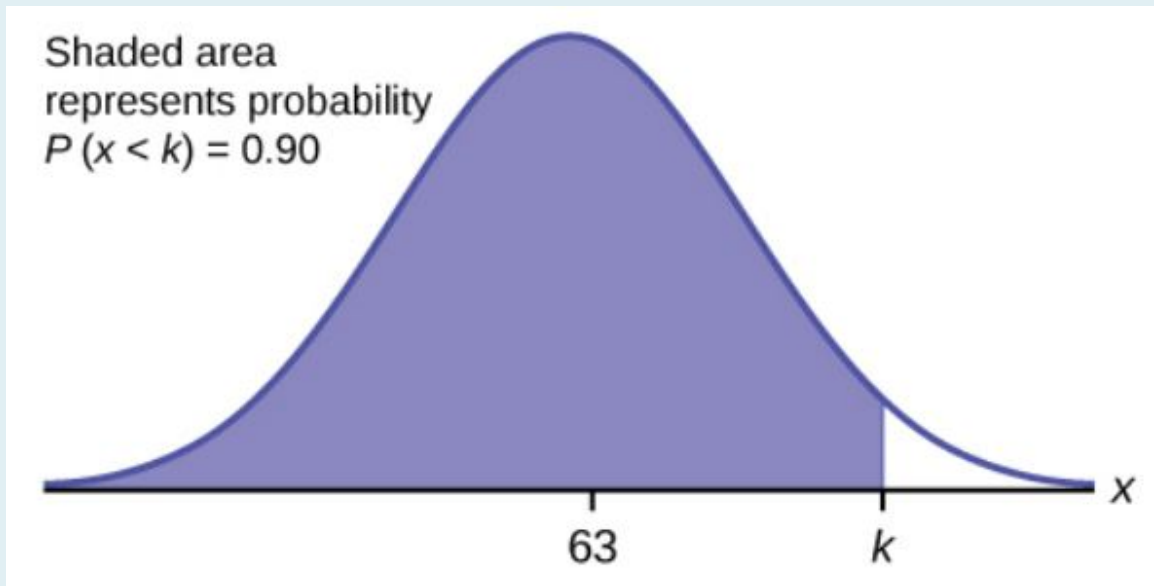
Function	norm.dist	Answer
Field 1	75	0.9918
Field 2	63	
Field 3	5	
Field 4	true	

The probability that a student scores less than 75 is 0.9918 (or 99.18%).

3. We want to find the value of x so that the area to the left of x is 0.9.

Function	norm.inv	Answer
Field 1	0.9	69.41
Field 2	63	
Field 3	5	

90% of the students scored below 69.41 points on the exam.



The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above.

4. We want to find the value of x so that the area to the right of x is 0.3. This is the same as finding the value of x so that the area to left of x is 0.7 (1-0.3).

Function	norm.inv	Answer
Field 1	0.7	65.62
Field 2	63	
Field 3	5	

30% of the students scored more 65.62 points on the exam.

TRY IT

The golf scores for a school team are normally distributed with a mean of 68 and a standard deviation of 3.

1. Find the probability that a randomly selected golfer scored less than 65.
2. Find the probability that a randomly selected golfer scored more than 72.

Click to see Solution

1.

Function	norm.dist	Answer
Field 1	65	0.1587
Field 2	68	
Field 3	3	
Field 4	true	

2.

Function	1-norm.dist	Answer
Field 1	72	0.0912
Field 2	68	
Field 3	3	
Field 4	true	

EXAMPLE

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is 0.5 hour.

1. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
2. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

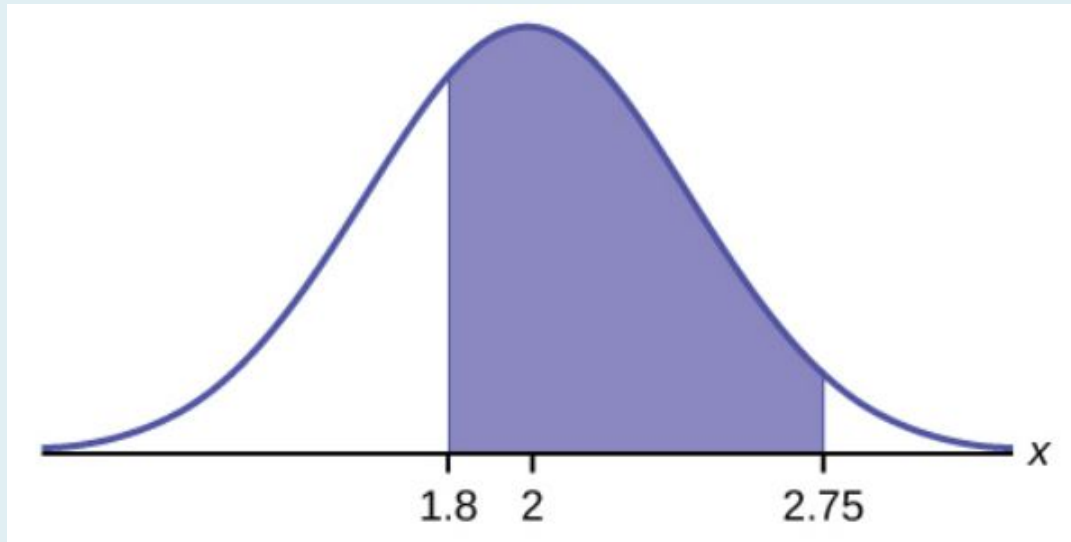
Solution:

Let X be the amount of time (in hours) a household personal computer is used for entertainment.

1. We want to find $P(1.8 < X < 2.75)$.

Function	norm.dist	-norm.dist	Answer
Field 1	2.75	1.8	0.5886
Field 2	2	2	
Field 3	0.5	0.5	
Field 4	true	true	

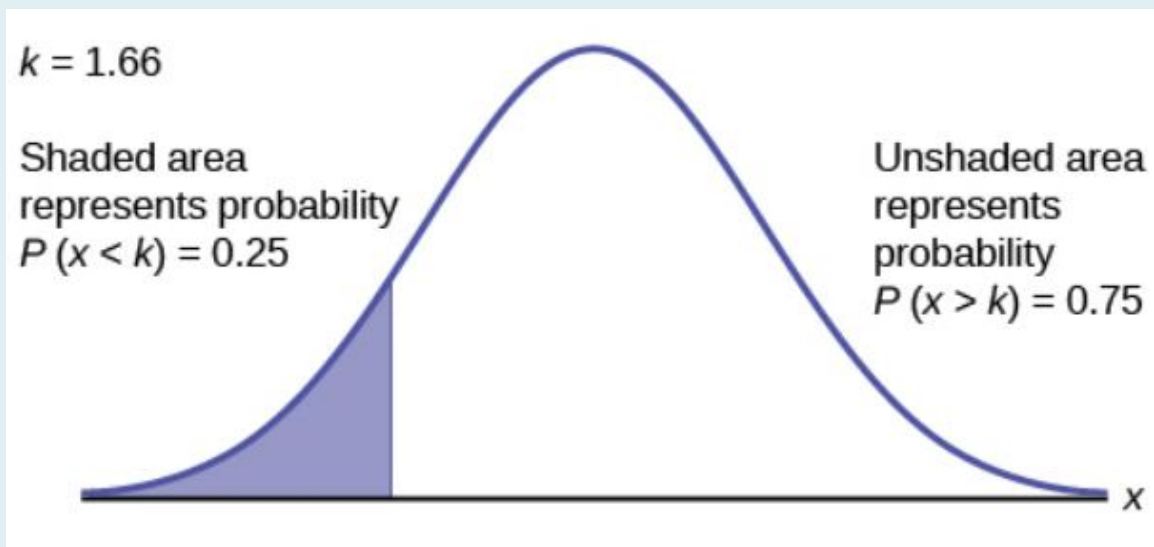
The probability a household computer is used for entertainment between 1.8 and 2.75 hours a day is 0.5886 (or 58.86%).



2. We need to find the value x so that 25% of the number of hours are less than this value.

Function	norm.inv	Answer
Field 1	0.25	1.66
Field 2	2	
Field 3	0.5	

25% of the value are less than 1.66 hours.



TRY IT

The golf scores for a school team are normally distributed with a mean of 68 and a standard deviation of 3. Find the probability that a golfer scored between 66 and 70.

Click to see Solution

Function	norm.dist	-norm.dist	Answer
Field 1	70	66	0.4950
Field 2	68	68	
Field 3	3	3	
Field 4	true	true	

EXAMPLE

There are approximately one billion smartphone users in the world today. In the United States the ages of smartphone users from 13 to 55+ follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

1. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
2. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
3. 80% of the users in the age range 13 to 55+ are less than what age?

4. 40% of the ages that range from 13 to 55+ are at least what age?

Solution:

1.

Function	norm.dist	-norm.dist	Answer
Field 1	64.7	23	0.8186
Field 2	36.9	36.9	
Field 3	13.9	13.9	
Field 4	true	true	

The probability a smartphone user is between 23 and 64.7 years of age is 0.8186 (or 81.86%).

2.

Function	norm.dist	Answer
Field 1	50.8	0.8413
Field 2	36.9	
Field 3	13.9	
Field 4	true	

The probability that a smartphone user is less than 50.8 years of age is 0.8413 (or 84.13%).

3.

Function	norm.inv	Answer
Field 1	0.8	48.6
Field 2	36.9	
Field 3	13.9	

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

4.

Function	norm.inv	Answer
Field 1	0.6	40.42
Field 2	36.9	
Field 3	13.9	

40% of the smartphone users in the age range 13 – 55+ are older than 40.42 years of age.

TRY IT

There are approximately one billion smartphone users in the world today. In the United States the ages of smartphone users from 13 to 55+ follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

- 30% of smartphone users are older than what age?
- What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old.

Click to see Solution

1.

Function	norm.inv	Answer
Field 1	0.7	44.19
Field 2	36.9	
Field 3	13.9	

2.

Function	norm.dist	Answer
Field 1	27	0.2382
Field 2	36.9	
Field 3	13.9	
Field 4	true	

EXAMPLE

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

1. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm.
2. 90% of the diameters of the mandarin oranges are less than what value?
3. 35% of the diameters of the mandarin oranges are greater than what value?

Solution:

1.

Function	1-norm.dist	Answer
Field 1	6	0.2660
Field 2	5.85	
Field 3	0.24	
Field 4	true	

The probability an orange has a diameter greater than 6 cm is 0.2660 (or 26.60%).

2.

Function	norm.inv	Answer
Field 1	0.9	6.16
Field 2	5.85	
Field 3	0.24	

90% of the diameters of the oranges are less than 6.16 cm.

3.

Function	norm.inv	Answer
Field 1	0.65	5.94
Field 2	5.85	
Field 3	0.24	

35% of the diameters of the oranges are greater than 5.94 cm.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=130#oembed-1>

Watch this video: Excel 2013 Statistical Analysis #39: Probabilities for Normal (Bell) Probability Distribution by ExcelIsFun [24:07]

Concept Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. The probability that the value for a normal random variable falls in between the values $x = a$ and $x = b$ is the area under the normal distribution curve to the right of $x = a$ and to the left of $x = b$.

Attribution

“6.2 Using the Normal Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

5.6 EXERCISES

1. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words.
2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?
3. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words.
4. What does a z -score measure?
5. What does standardizing a normal distribution do to the mean?
6. What is the z -score of $x = 12$ if it is two standard deviations to the right of the mean?
7. What is the z -score of $x = 9$ if it is 1.5 standard deviations to the left of the mean?
8. What is the z -score of $x = -2$ if it is 2.78 standard deviations to the right of the mean?
9. What is the z -score of $x = 7$ if it is 0.133 standard deviations to the left of the mean?

10. Suppose X is a normal random variable with a mean of 2 and standard deviation of 6. What value of x has a z -score of three?
11. Suppose X is a normal random variable with a mean of 8 and standard deviation of 1. What value of x has a z -score of -2.25 ?
12. Suppose X is a normal random variable with a mean of 9 and standard deviation of 5. What value of x has a z -score of -0.5 ?
13. Suppose X is a normal random variable with a mean of 2 and standard deviation of 3. What value of x has a z -score of -0.67 ?
14. Suppose X is a normal random variable with a mean of 4 and standard deviation of 2. What value of x is 1.5 standard deviations to the left of the mean?
15. Suppose X is a normal random variable with a mean of 4 and standard deviation of 2. What value of x is 2 standard deviations to the right of the mean?
16. Suppose X is a normal random variable with a mean of 8 and standard deviation of 9. What value of x is 0.67 standard deviations to the left of the mean?
17. Suppose X is a normal random variable with a mean of -1 and standard deviation of 2. What is the z -score of $x = 2$?
18. Suppose X is a normal random variable with a mean of 12 and standard deviation of 6. What is the z -score of $x = 2$?
19. Suppose X is a normal random variable with a mean of 9 and standard deviation of 3. What is the z -score of $x = 9$?

20. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z -score of $x = 5.5$?
21. In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.
22. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.
23. In a normal distribution, $x = -2$ and $z = 6$. This tells you that $x = -2$ is ____ standard deviations to the ____ (right or left) of the mean.
24. In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.
25. In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.
26. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?
27. About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?
28. About what percent of x values lie between the second and third standard deviations (both sides)?
29. Suppose X is a normal random variable with mean 15 and standard deviation 3. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

30. Suppose X is a normal random variable with mean -3 and standard deviation 1 . Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution(i.e., -3).
31. Suppose X is a normal random variable with mean -3 and standard deviation 1 . Between what x values does 34.14% of the data lie?
32. About what percent of x values lie between the mean and three standard deviations?
33. About what percent of x values lie between the mean and one standard deviation?
34. About what percent of x values lie between the first and second standard deviations from the mean (both sides)?
35. About what percent of x values lie between the first and third standard deviations (both sides)?
36. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.
- What is the median recovery time?
 - What is the z -score for a patient who takes ten days to recover?
37. The length of time to find it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?
- The data cannot follow the uniform distribution.
 - The data cannot follow the exponential distribution..
 - The data cannot follow the normal distribution.
38. The heights of the 430 National Basketball Association players were listed on team rosters at

the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean $\mu = 79$ inches and a standard deviation $\sigma = 3.89$ inches. For each of the following heights, calculate the z -score and interpret it using complete sentences.

- a. 77 inches
- b. 85 inches
- c. If an NBA player reported his height had a z -score of 3.5, would you believe him? Explain your answer.

39. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

- a. Calculate the z -scores for the male systolic blood pressures 100 and 150 millimeters.
- b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

40. Kyle's doctor told him that the z -score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$.

- a. Which answer(s) **is/are** correct?
 - i. Kyle's systolic blood pressure is 175.
 - ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
 - iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
 - iv. Kyle's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
- b. Calculate Kyle's blood pressure.

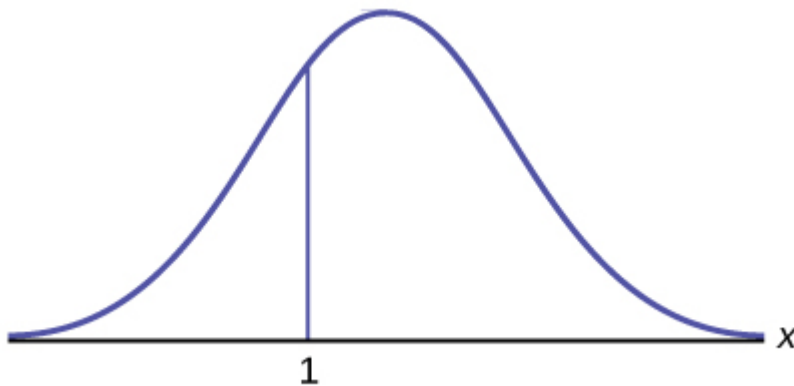
41. Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. Calculate the z -scores that correspond to the following weights and interpret them.

- a. 11 kg
- b. 7.9 kg
- c. 12.2 kg

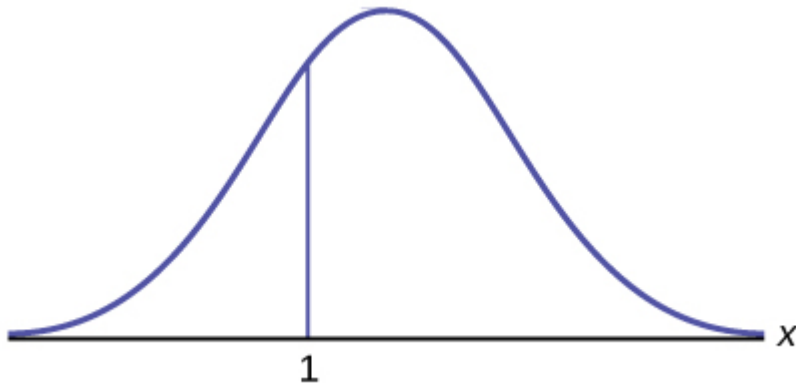
42. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

- a. Calculate the z -score for an SAT score of 720. Interpret it using a complete sentence.
- b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
- c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

43. How would you represent the area to the left of one in a probability statement?

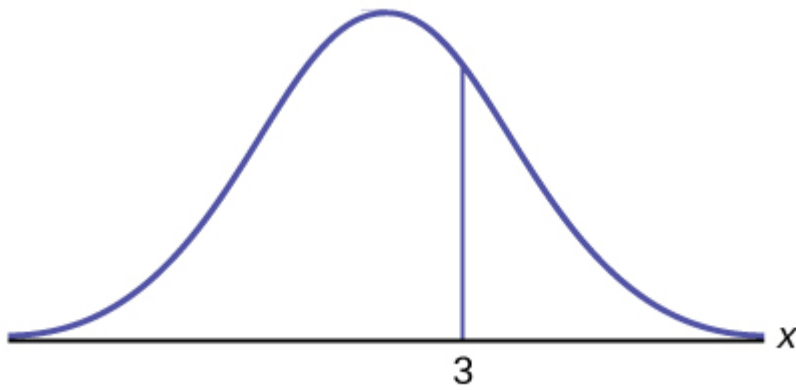


44. What is the area to the right of one?

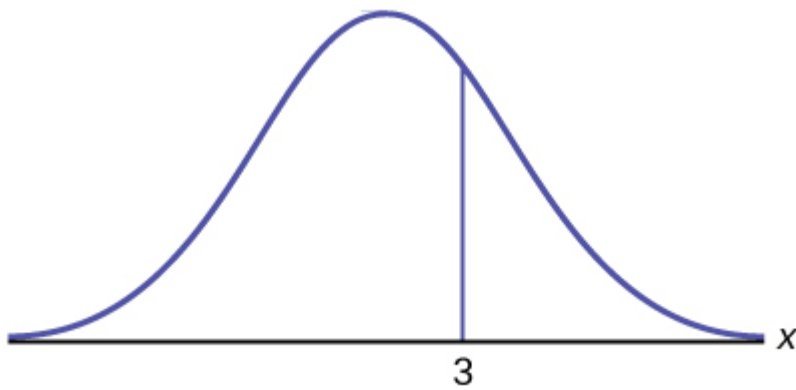


45. Is $P(x < 1)$ equal to $P(x \geq 1)$? Why?

46. How would you represent the area to the left of three in a probability statement?



47. What is the area to the right of three?



48. If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x ?

49. If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x ?
50. Suppose X is a normal random variable with a mean of 54 and 8.
- Find the probability that $x > 56$.
 - Find the probability that $x < 30$.
 - Find the probability that $40 < x < 50$.
 - 80% of the x -values are less than what value?
 - 40% of the x -values are greater than what value?
51. The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years.
- A CD player is guaranteed for three years. Find the probability that a CD player will break down during the guarantee period.
 - Find the probability that a CD player will last between 2.8 and 6 years.
 - 70% of the CD players last how long?
52. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.
- What is the probability of spending more than two days in recovery?
 - 10% of the recovery times are larger than what value?
53. The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.
- Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?
 - Find the probability that it takes at least eight minutes to find a parking space.
 - Seventy percent of the time, it takes more than how many minutes to find a parking space?
54. According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X be the height of the individual.
- Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.

- b. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
- c. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

55. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X be IQ of an individual.

- a. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
- b. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
- c. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

56. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let X be percent of fat calories.

- a. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
- b. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

57. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

- a. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X . Shade the region corresponding to the probability. Find the probability.
- b. 80% of fly balls travel for less than what value?

58. In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

- a. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
- b. What percent of the children spend over ten hours per day unsupervised?
- c. Seventy percent of the children spend at least how long per day unsupervised?

59. In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X be number of votes for President Clinton for an election district.

- a. Is 1,956.8 a population mean or a sample mean? How do you know?
- b. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
- c. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
- d. 25% of the number of votes for President Clinton is higher than what value?

60. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

- a. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
- b. Sixty percent of all trials of this type are completed within how many days?

61. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

- a. Find the percent of her laps that are completed in less than 130 seconds.
- b. The fastest 3% of her laps are under what value.
- c. The middle 80% of her laps are between what values?

62. Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

- a. Ricardo's actual GPA is lower than Anita's actual GPA.
- b. Ricardo is not passing because his z -score is zero.

c. Anita is in the 70th percentile of students at her college.

63. An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the z -scores first, and then use those to calculate the probability.

64. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

65. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

- There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
- There is about a ___ chance that the number of heads will be somewhere between 12 and 28.
- There is about a ___ chance that the number of heads will be somewhere between eight and 32.

66. A \$1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

- somewhere between 34 and 54 prizes.
- somewhere between 54 and 64 prizes.
- more than 64 prizes.

67. Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site. On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

- a. Find the probability that the percent of 18 to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
 - b. 95% of the number of 18 to 34-year-olds who check Facebook before getting out of bed is less than what value?
-

Attribution

“Chapter 6 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART VI

THE CENTRAL LIMIT THEOREM AND SAMPLING DISTRIBUTIONS

Chapter Outline

6.1 Introduction to Sampling Distribution and the Central Limit Theorem

6.2 Sampling Distribution of the Sample Mean

6.3 Sampling Distribution of the Sample Proportion

6.4 Exercises

6.1 INTRODUCTION TO SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM



If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. Photo by John Lodder, CC BY 4.0.

Why are we so concerned with means? Two reasons are that they give us a middle ground for comparison, and they are easy to calculate. In this chapter, we will study means, proportions and their relationship to the **central limit theorem**.

The **central limit theorem** is one of the most powerful and useful ideas in all of statistics. The central limit theorem basically says that if we collect samples of size n from a population with mean μ and standard deviation σ , calculate each sample's mean, and create a histogram of those means, then, under the right conditions, the resulting histogram will tend to have an approximate normal bell shape.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=139#oembed-1>

Watch this video: Central limit theorem | Inferential statistics | Probability and Statistics | Khan Academy by Khan Academy [9:45]

Attribution

“Chapter 7 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

6.2 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

LEARNING OBJECTIVES

- Describe the distribution of the sample mean.
- Solve probability problems involving the distribution of the sample mean.

Suppose all samples of size n are selected from a population with mean μ and standard deviation σ . For each sample, the sample mean \bar{x} is recorded. The probability distribution of these sample means is called **the sampling distribution of the sample means**. The central limit theorem describes the properties of the sampling distribution of the sample means.

THE CENTRAL LIMIT THEOREM

Suppose all samples of size n are taken from a population with mean μ and standard deviation σ . The collection of sample means forms a probability distribution called the **sampling distribution of the sample mean**.

1. The mean of the distribution of the sample means, denoted $\mu_{\bar{x}}$, equals the mean of the population.

$$\mu_{\bar{x}} = \mu$$

2. The standard deviation of the of the sample means (called the standard error of the mean), denoted $\sigma_{\bar{x}}$, equals the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. The distribution of the sample means follows a normal distribution if **one** of the following conditions is met:
- The population the samples are drawn from is normal, regardless of the sample size n .
 - The sample size $n \geq 30$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=141#oembed-2>

Watch this video: Sampling distribution of the sample mean | Probability and Statistics | Khan Academy by Khan Academy

[10:51]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=141#oembed-3>

Watch this video: Standard error of the mean | Inferential statistics | Probability and Statistics | Khan Academy by Khan Academy [15:14]

Because the central limit theorem states that the sampling distribution of the sample means follows a normal distribution (under the right conditions), the normal distribution can be used to answer probability questions about sample means. The z -score for the sampling distribution of the sample means is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where μ is the mean of the population the sample is taken from, σ is the standard deviation of the population the sample is taken from, and n is the sample size.

CALCULATING PROBABILITIES ABOUT SAMPLE MEANS IN EXCEL

Because the distribution the sample means follows a normal distribution (under the right conditions), the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate probabilities associated with a sample mean.

- For x , enter the value for \bar{x} .
- For μ , enter the mean of the sample means μ . Because the mean of the sample means equals the mean of the population the sample is taken from, we enter μ , the mean of the population.
- For σ , enter the standard error of the mean $\frac{\sigma}{\sqrt{n}}$.
- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

NOTE

In this case, we want to calculate probabilities associated with a sample mean. The sample means follow a normal distribution (under the right conditions), which allows us to use the **norm.dist** function to calculate probabilities. Because we are working with sample means, we must enter the **mean** and the **standard distribution** of the **distribution of the sample means** into the **norm.dist** function, and not the mean and standard distribution of the population the samples are taken from. The mean of the sample means equals the mean of the population, so we are entering the value of μ into the second field of the **norm.dist** function. But the standard distribution of the sample means equals $\frac{\sigma}{\sqrt{n}}$, so we must enter this value into third field of the **norm.dist** function.

We use the **norm.dist** function in the same way as we learned previously to calculate the probability a sample mean is less than a given value, a sample mean is greater than a given value, or a sample mean is in between two given values.

An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the **z**-score for the corresponding value of \bar{x} (using the **z**-score for sample means given above).

EXAMPLE

The length of time, in hours, it takes an “over 40” group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. Suppose a sample of size 25 is drawn randomly from the population.

1. Is the distribution of the sample means normal? Explain.
2. What is the mean and the standard distribution of the distribution of the sample means?

3. What is the probability that the mean of the sample is less than 1.7 hours?
4. What is the probability that the mean of the sample is more than 2.2 hours?
5. What is the probability that the sample mean is between 1.8 hours and 2.3 hours?

Solution:

1. Because the population the sample is taken from follows a normal distribution, the distribution of the sample means also follows a normal distribution.
2. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 2$. The standard deviation of the

sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{25}} = 0.1$.

3.

Function	norm.dist	Answer
Field 1	1.7	0.0013
Field 2	2	
Field 3	0.5/sqrt(25)	
Field 4	true	

The probability the sample mean is less than 1.7 hours is 0.0013 (or 0.13%).

Note: Because we are calculating a probability for a sample mean, we enter the standard deviation of the sample means 0.5/sqrt(25) into field 3 (and not the standard deviation of the population).

4.

Function	1-norm.dist	Answer
Field 1	2.2	0.0228
Field 2	2	
Field 3	0.5/sqrt(25)	
Field 4	true	

The probability the sample mean is more than 2.2 hours is 0.0228 (or 2.28%).

5.

Function	norm.dist	-norm.dist	Answer
Field 1	2.3	1.8	0.9759
Field 2	2	2	
Field 3	$0.5/\sqrt{25}$	$0.5/\sqrt{25}$	
Field 4	true	true	

The probability the sample mean is between 1.8 hours and 2.3 hours is 0.9759 (or 97.59%).

TRY IT

The length of time taken on the SAT for a group of students has a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of 60 is drawn randomly from the population.

1. Is the distribution of the sample means normal? Explain.
2. What is the probability that sample mean is between 2.4 hours and 2.8 hours?
3. What is the probability that the sample mean is at least 2.6 hours?
4. What is the probability that the sample mean is at most 2.45 hours?

Click to see Solution

1. The distribution of the sample means is normal because the sample size of 60 is greater than 30.

2.

Function	norm.dist	-norm.dist	Answer
Field 1	2.8	2.4	0.9990
Field 2	2.5	2.5	
Field 3	$0.25/\sqrt{60}$	$0.25/\sqrt{60}$	
Field 4	true	true	

3.

Function	1-norm.dist	Answer
Field 1	2.6	0.0010
Field 2	2.5	
Field 3	0.25/sqrt(60)	
Field 4	true	

4.

Function	norm.dist	Answer
Field 1	2.45	0.0607
Field 2	2.5	
Field 3	0.25/sqrt(60)	
Field 4	true	

EXAMPLE

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years and the standard deviation is 15 years. Suppose a sample of 100 tablet users is taken.

1. What are the mean and standard deviation for the sample mean ages of tablet users?
2. What is the distribution of the sample means? Explain.
3. Find the probability that the sample mean age is more than 30 years.

Solution:

1. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 34$. The standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$.
2. The distribution of the sample means is normal because the sample size of 100 is greater than

30

3.

Function	1-norm.dist	Answer
Field 1	30	0.9962
Field 2	34	
Field 3	15/sqrt(100)	
Field 4	true	

The probability the sample mean is more than 30 years of age is 0.9962 (or 99.62%).

TRY IT

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a start-up game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Click to see Solution

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

Function	norm.dist	-norm.dist	Answer
Field 1	35	29	0.0186
Field 2	28	28	
Field 3	4.8/sqrt(100)	4.8/sqrt(100)	
Field 4	true	true	

There is 1.86% chance that the mean age of men who will play your game is between 29 years and 35 years. Because this is a very low probability, you should not continue your development strategy.

EXAMPLE

The mean number of minutes for app engagement by a tablet user is 8.2 minutes with a standard deviation of 1 minute. Suppose a sample of 60 tablet users is taken.

1. Is the distribution of the sample mean normal? Explain.
2. What are the mean and standard deviation for the sample mean number of minutes for app engagement?
3. Find the probability that the sample mean is between 8 minutes and 8.5 minutes.
4. Find the probability that the sample mean is less than 8.3 minutes.

Solution:

1. Because the sample size of 60 is greater than 30, the distribution of the sample means also follows a normal distribution.
2. The mean of the distribution of the sample means is $\mu_{\bar{x}} = 8.2$. The standard deviation of the sample means is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$.

3.

Function	norm.dist	-norm.dist	Answer
Field 1	8.5	8	0.9293
Field 2	8.2	8.2	
Field 3	$1/\sqrt{60}$	$1/\sqrt{60}$	
Field 4	true	true	

The probability that the sample mean is between 8 and 8.5 minutes is 0.9293 (or 92.93%).

4.

Function	norm.dist	Answer
Field 1	8.3	0.7807
Field 2	8.2	
Field 3	$1/\sqrt{60}$	
Field 4	true	

The probability that the sample mean is less than 8.3 minutes is 0.7807 (or 78.07%).

TRY IT

Cans of a cola beverage claim to contain 16 ounces with a standard deviation of 0.143 ounces. The amounts in a sample of 34 cans are measured and the mean is 16.01 ounces. Find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Click to see Solution

Function	1-norm.dist	Answer
Field 1	16.01	0.3417
Field 2	16	
Field 3	0.143/sqrt(34)	
Field 4	true	

Because there is a 34.17% probability that the average sample volume is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. That is, based on this sample, it is likely that the average volume of the cans is higher than the claimed 16 ounces.

As consumers, we would be glad if the average was higher than 16 ounces because we are likely receiving more cola in the can than what we paid for. As the manufacturer, we would need to inspect our bottling process to determine if the process is working within acceptable limits.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=141#oembed-1>

Watch this video: Excel Statistics 76: Sampling Distribution Of Sample Mean & Central Limit Theorem by ExcellisFun
[24:05]

Concept Review

The distribution of the sample means follows a normal distribution if **one** of the following conditions is met:

- The population the samples are taken from is normal.
- The sample size is greater than or equal to 30.

The mean of the sample means $\mu_{\bar{x}}$ equals the population mean μ . The standard deviation of the sample means $\sigma_{\bar{x}}$ is equal to $\frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation and n is the sample size.

Attribution

“7.1 The Central Limit Theorem for Sample Means (Averages)” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

6.3 SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

LEARNING OBJECTIVES

- Describe the distribution of the sample proportion.
- Solve probability problems involving the distribution of the sample proportion.

The Central Limit Theorem tells us that the distribution of the sample means follow a normal distribution under the right conditions. This allows us to answer probability questions about the sample mean \bar{x} . Now we want to investigate the sampling distribution for another important parameter—the sampling distribution of the sample proportion. Once we know what distribution the sample proportions follow, we can answer probability questions about sample proportions.

A **proportion** is the percent, fraction, or ratio of a sample or population that have a characteristic of interest. The **population proportion** is denoted by p and the **sample proportion** is denoted by \hat{p} .

$$\begin{aligned}\text{Proportion} &= \frac{\text{Number of Items with Characteristic of Interest}}{\text{Total Number of Items}} \\ &= \frac{x}{n}\end{aligned}$$

If the random variable is discrete, such as for categorical data, then the parameter we wish to estimate is the population proportion. This is, of course, the probability of drawing a success in any one random draw. Because we are interested in the number of successes, we are dealing with the binomial distribution. The random variable X is the number of successes and the parameter we wish to know is p , the probability of drawing a success, which is of course the proportion of successes in the population. What is the distribution of the sample proportion \hat{p} ?

THE CENTRAL LIMIT THEOREM FOR SAMPLE PROPORTIONS

Suppose all samples of size n are taken from a population with proportion p . The collection of sample proportions forms a probability distribution called the **sampling distribution of the sample proportion**.

1. The mean of the distribution of the sample proportions, denoted $\mu_{\hat{p}}$, equals the population proportion.

$$\mu_{\hat{p}} = p$$

2. The standard deviation of the of the sample proportions (called the standard error of the proportion), denoted $\sigma_{\hat{p}}$, is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}}$$

3. The distribution of the sample proportion is:
 - Normal if $n \times p \geq 5$ and $n \times (1 - p) \geq 5$.
 - Binomial if one of $n \times p < 5$ and $n \times (1 - p) < 5$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=143#oembed-1>

Watch this video: Sampling Distribution of the Sample Proportion by Khan Academy [9:57]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=143#oembed-2>

Watch this video: Sampling Distribution of the Sample Proportion by Khan Academy [4:34]

When $n \times p \geq 5$ **and** $n \times (1 - p) \geq 5$, the central limit theorem states that the sampling distribution of the sample proportions follows a normal distribution. In this case the normal distribution can be used to answer probability questions about sample proportions and the z-score for the sampling distribution of the sample proportions is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1 - p)}{n}}}$$

where p is the population proportion and n is the sample size.

CALCULATING PROBABILITIES ABOUT SAMPLE PROPORTIONS IN EXCEL (NORMAL)

When the distribution of the sample proportions follows a normal distribution (when $n \times p \geq 5$ and $n \times (1 - p) \geq 5$), the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate probabilities associated with a sample proportion.

- For x , enter the value for \hat{p} .
- For μ , enter the mean of the sample proportions p . Because the mean of the sample proportions equals the proportion of the population the sample is taken from, we enter p , the population proportion.
- For σ , enter the standard error of the proportion $\sqrt{\frac{p \times (1 - p)}{n}}$.

- For the logic operator, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

NOTE

In this case, we want to calculate probabilities associated with a sample proportion. The sample proportions follow a normal distribution (under the right conditions), which allows us to use the **norm.dist** function to calculate probabilities. Because we are working with sample proportions, we must enter the **mean** and the **standard distribution** of the **distribution of the sample proportions** into the **norm.dist** function. The mean of the sample proportions equals the population proportion, so we are entering the value of p into the second field of the **norm.dist**

function. But the standard distribution of the sample proportion equals $\sqrt{\frac{p \times (1 - p)}{n}}$, so we must enter this value into third field of the **norm.dist** function.

We use the **norm.dist** function in the same way as we learned previously to calculate the probability a sample proportion is less than a given value, a sample proportion is greater than a given value, or a sample proportion is in between two given values.

An alternative approach in Excel is to use the **norm.s.dist(z,true)** function. In the **norm.s.dist** function, we enter the z -score for the corresponding value of \hat{p} (using the z -score for sample proportions given above).

EXAMPLE

A recent study asked working adults if they worked most of their time remotely. The study found that 30% of employees spend the majority of their time working remotely. Suppose a sample of 150 working adults is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the mean and standard deviation of the sample proportion?
3. What is the probability that at most 27% of the workers in the sample work remotely most of the time?
4. What is the probability that at least 51 of the workers in the sample work remotely most of the time?
5. What is the probability that between 32% and 35% of the workers in the sample work remotely most of the time?

Solution:

1. $n = 150$ and $p = 0.3$. Checking $n \times p$ and $n \times (1 - p)$:

$$n \times p = 150 \times 0.3 = 45 \geq 5$$

$$n \times (1 - p) = 150 \times (1 - 0.3) = 105 \geq 5$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ the distribution of the sample proportion is normal.

2. The mean of the distribution of the sample proportions is $\mu_{\hat{p}} = 0.3$. The standard deviation of the sample proportions is $\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}} = \sqrt{\frac{0.3 \times (1 - 0.3)}{150}} = 0.0374$.

Function	norm.dist	Answer
Field 1	0.27	0.2113
Field 2	0.3	
Field 3	sqrt(0.3*(1-0.3)/150)	
Field 4	true	

The probability the sample proportion is at most 27% is 0.2113 (or 21.13%).

Note: Because we are calculating a probability for a sample proportion, we enter the mean of the sample proportions 0.3 (which is the population proportion) into field 2 and the standard deviation of the sample proportions sqrt(0.3*(1-0.3)/150) into field 3.

4. In this case, 51 is not a proportion. It is the number of items in the sample that have the

characteristic of interest. We need to convert this 51 out of 150 into a percent: $\frac{51}{150} = 0.34$.

This question is asking us to find the probability that at least 34% of the workers in the sample work remotely most of the time.

Function	1-norm.dist	Answer
Field 1	0.34	0.1425
Field 2	0.3	
Field 3	$\text{sqrt}(0.3*(1-0.3)/150)$	
Field 4	true	

The probability the sample proportion is at least 34% is 0.1425 (or 14.25%).

5.

Function	norm.dist	-norm.dist	Answer
Field 1	0.35	0.32	0.2058
Field 2	0.3	0.3	
Field 3	$\text{sqrt}(0.3*(1-0.3)/150)$	$\text{sqrt}(0.3*(1-0.3)/150)$	
Field 4	true	true	

The probability the sample proportion is between 32% and 35% is 0.2058 (or 20.58%).

TRY IT

According to a recent study, 17.5% of the adult population of Canada are smokers. Suppose a random sample of 200 adult Canadians is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the mean and standard deviation of the sample proportion?

3. What is the probability that less than 32 of the adults in the sample are smokers?
4. What is the probability that more than 20% of the adults in the sample are smokers?
5. What is the probability that between 34 and 44 of the adults in the sample are smokers?

Click to see Solution

1. Because $n \times p = 200 \times 0.175 = 35 \geq 5$ and $n \times (1 - p) = 200 \times (1 - 0.175) = 165 \geq 5$ the distribution of the sample proportions is normal.
2. The mean of the distribution of the sample proportions is $\mu_{\hat{p}} = 0.175$. The standard deviation of the sample proportions is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \times (1 - p)}{n}} = \sqrt{\frac{0.175 \times (1 - 0.175)}{200}} = 0.02687.$$

3.

Function	norm.dist	Answer
Field 1	0.16	0.2883
Field 2	0.175	
Field 3	sqrt(0.175*(1-0.175)/200)	
Field 4	true	

4.

Function	1-norm.dist	Answer
Field 1	0.2	0.1761
Field 2	0.175	
Field 3	sqrt(0.175*(1-0.175)/200)	
Field 4	true	

5.

Function	norm.dist	-norm.dist	Answer
Field 1	0.22	0.17	0.9530
Field 2	0.175	0.175	
Field 3	sqrt(0.175*(1-0.175)/200)	sqrt(0.175*(1-0.175)/200)	
Field 4	true	true	

When one of $n \times p < 5$ or $n \times (1 - p) < 5$, the sampling distribution of the sample proportions

follows a binomial distribution, and so we must use the binomial distribution to answer probability questions about sample proportions. In these cases, we are actually answering probability questions about the number of items with the characteristic of interest, x . In other words, we are answering questions about the number of successes x we get in n trials (the sample size) where the probability of success is the population proportion p . These are exactly the same type of questions we answered previously with the binomial distribution.

CALCULATING PROBABILITIES ABOUT SAMPLE PROPORTIONS IN EXCEL (BINOMIAL)

When the distribution the sample proportions follows a binomial distribution (when one of $n \times p < 5$ or $n \times (1 - p) < 5$), the **binom.dist(x,n,p,logic operator)** function can be used to calculate probabilities associated with a sample proportion.

- For **x**, enter the number of items with the characteristic of interest x .
- For **n**, enter the sample size n . The sample size is the number of trials in the binomial experiment.
- For **p**, enter the population proportion p . The population proportion is the probability of success.
- For the logic operator, enter **true**. **Note:** Because probabilities for sample proportions are generally inequalities ($<$, \leq , $>$, \geq), we enter true for the logic operator. We would only enter false in the case that the probability of the sample proportion exactly equals a given value.

NOTE

We use the **binom.dist** function in the same way as we learned previously to calculate the probability a sample proportion is less than a given value, a sample proportion is at most a given value, a sample proportion is greater than a given value, or a sample proportion is at least a given value.

EXAMPLE

At the local humane society, 3% of the dogs have heartworm disease. Suppose a sample of 60 dogs at the humane society is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the probability that at most 5% of the dogs in the sample have heartworm disease?
3. What is the probability that less than 7 of the dogs in the sample have heartworm disease?
4. What is the probability that more than 8% of the dogs in the sample have heartworm disease?
5. What is the probability that at least 6 of the dogs in the sample have heartworm disease?

Solution:

1. Because $n \times p = 60 \times 0.03 = 1.8 < 5$ the distribution of the sample proportions is binomial.
2. We want to find $P(\hat{p} \leq 0.05)$. Because we are using the binomial distribution, we have to convert 5% into the number of items x in the sample with the required characteristic: $x = 0.05 \times 60 = 3$. In terms of the binomial distribution, we need to find $P(x \leq 3)$.

Function	binom.dist	Answer
Field 1	3	0.8943
Field 2	60	
Field 3	0.03	
Field 4	true	

The probability that at most 5% of the dogs in the sample have heartworm disease is 0.8943 (or 89.43%).

3. We want to find $P(x < 7)$. Because we are using the binomial distribution, this probability is the same as $P(x \leq 6)$.

Function	binom.dist	Answer
Field 1	6	0.9979
Field 2	60	
Field 3	0.03	
Field 4	true	

The probability that less than 7 of the dogs in the sample have heartworm disease is 0.9979 (or 99.79%).

4. We want to find $P(\hat{p} > 0.08)$. Because we are using the binomial distribution, we have to convert 8% into the number of items x in the sample with the required characteristic: $x = 0.08 \times 60 = 4.8$. In terms of the binomial distribution, we need to find $P(x > 4.8)$. This is the same as $1 - P(x \leq 4)$.

Function	1-binom.dist	Answer
Field 1	4	0.0340
Field 2	60	
Field 3	0.03	
Field 4	true	

The probability that more than 8% of the dogs in the sample have heartworm disease is 0.0340 (or 3.4%).

5. We want to find $P(x \geq 6)$. Because we are using the binomial distribution, this probability is the same as $1 - P(x \leq 5)$.

Function	1-binom.dist	Answer
Field 1	5	0.0091
Field 2	60	
Field 3	0.03	
Field 4	true	

The probability that at least 6 of the dogs in the sample have heartworm disease is 0.0091 (or 0.91%).

TRY IT

During the past tax season, 92% of tax returns were filed using an electronic filing system. Suppose a sample of 40 tax returns are selected.

1. What is the distribution of the sample proportions?
2. What is the probability at most 35 of the tax returns in the sample were filed electronically?
3. What is the probability less than 93% of the tax returns in the sample were filed electronically?
4. What is the probability more than 36 of the tax returns in the sample were filed electronically?
5. What is the probability at least 88% of the tax returns in the sample were filed electronically?

Click to see Solution

1. Because $n \times (1 - p) = 40 \times (1 - 0.92) = 3.2 < 5$ the distribution of the sample proportions is binomial.

2.

Function	binom.dist	Answer
Field 1	35	0.2132
Field 2	40	
Field 3	0.92	
Field 4	true	

3.

Function	binom.dist	Answer
Field 1	37	0.6306
Field 2	40	
Field 3	0.92	
Field 4	true	

4.

Function	1-binom.dist	Answer
Field 1	36	0.6007
Field 2	40	
Field 3	0.92	
Field 4	true	

5.

Function	1-binom.dist	Answer
Field 1	33	0.9624
Field 2	40	
Field 3	0.92	
Field 4	true	



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=143#oembed-3>

Watch this video: Excel Statistics 79: Proportions Sampling Distribution by ExcellIsFun [8:54]

Concept Review

The distribution of the sample proportions follows a

- normal distribution if both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$.
- binomial distribution if one of $n \times p < 5$ and $n \times (1 - p) < 5$.

The mean of the sample proportion $\mu_{\hat{p}}$ equals the population proportion p . The standard deviation

of the sample proportions $\sigma_{\hat{p}}$ is equal to $\sqrt{\frac{p \times (1 - p)}{n}}$ where p is the population proportion and n is the sample size.

Attribution

“7.3 The Central Limit Theorem for Proportions“ in Introductory Business Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

6.4 EXERCISES

1. Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Suppose 16 review are selected at random.

Suppose 16 review are selected at random.

- What is the mean, standard deviation, and sample size?
- What is the distribution of the sample means? Explain.
- What is the mean and standard deviation of the sample means?
- Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.
- Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs.
- Why are the probabilities in (d) and (e) different?

2. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

- What is the probability that the 49 balls traveled an average of less than 240 feet?
- What is the probability that the 49 balls traveled an average of 245 feet to 255 feet?
- What is the probability that the 49 balls traveled an average of more than 260 feet?

3. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules) with a standard deviation of two hours. Suppose we randomly sample 36 taxpayers.

- What is the distribution of the sample means? Explain.
- Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
- Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

4. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an

average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.

5. In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

- a. What is the distribution of the sample means? Explain.
- b. What is the mean and standard deviation of the sample means?
- c. What is the probability that the sample mean is less than 170 acres?

6. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen.

- a. What is the distribution of the sample means?
- b. What is the mean and standard deviation of the sample means?
- c. For the group of 16, find the probability that the average percent of fat calories consumed is more than five.

7. The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be \$2,000 per year with a standard deviation of \$8,000. We randomly survey 1,000 residents of that country.

- a. How is it possible for the standard deviation to be greater than the average?
- b. Why is it more likely that the average of the 1,000 residents will be from \$2,000 to \$2,100 than from \$2,100 to \$2,200?

8. NeverReady batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

9. Your company has a contract to perform preventive maintenance on thousands of air-

conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

10. Suppose in a local Kindergarten through 12th grade (K – 12) school district, 53% of the population favor a charter school for grades K through five. A simple random sample of 300 is surveyed.

- a. Find the probability that less than 100 favor a charter school for grades K through 5.
- b. Find the probability that 170 or more favor a charter school for grades K through 5.
- c. Find the probability that no more than 140 favor a charter school for grades K through 5.
- d. Find the probability that there are fewer than 130 that favor a charter school for grades K through 5.
- e. Find the probability that exactly 150 favor a charter school for grades K through 5.

11. Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days.

- a. Find the probability that Janice is the driver at most 20 days.
- b. Find the probability that Roberta is the driver more than 16 days.
- c. Find the probability that Barbara drives between 24 and 30 of those 96 days.

12. A question is asked of a class of 200 freshmen, and 23% of the students know the correct answer. Suppose a sample of 50 students is taken.

- a. What is the mean and standard deviation of the distribution of the sample proportions?
- b. What is the distribution of the sample proportions? Explain.
- c. What is the probability that more than 30% of the students answered correctly?
- d. What is the probability that less than 20% of the students answered correctly?
- e. What is the probability that between 21% and 25% of the students answered correctly?

13. A virus attacks one in three of the people exposed to it. An entire large city is exposed. Suppose a sample of 70 people in the city is taken.

- a. What is the mean and standard deviation of the distribution of the sample proportions?
- b. What is the distribution of the sample proportions? Explain.
- c. What is the probability that between 21 and 40 of the people in the sample were exposed to

the virus?

- d. What is the probability that more than 35% of the people in the sample were exposed to the virus?
 - e. What is the probability that less than 25% of the people in the same were exposed to the virus?
14. A game is played repeatedly. A player wins one-fifth of the time. Suppose a player plays the game 20 times.
- a. What is the mean and standard deviation of the distribution of the sample proportions?
 - b. What is the distribution of the sample proportions? Explain.
 - c. What is the probability that the player wins at most 7 times?
 - d. What is the probability that the player wins at least 30% of the time?
 - e. What is the probability that the player wins less than 15% of the time?
 - f. What is the probability that the player wins more than 10 times?
15. A company inspects products coming through its production process, and rejects defective products. One-tenth of the items are defective. Suppose a sample of 40 items is taken.
- a. What is the mean and standard deviation of the distribution of the sample proportions?
 - b. What is the distribution of the sample proportions? Explain.
 - c. What is the probability that fewer than 7 of the items in the sample are defective?
 - d. What is the probability that more than 15% of the items in the sample are defective?
 - e. What is the probability that at least 3 of the items in the sample are defective?
 - f. What is the probability that at most 20% of the items in the sample are defective?
-

Attribution

“Chapter 7 Practice” in Introductory Business Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART VII

CONFIDENCE INTERVALS FOR SINGLE POPULATION PARAMETERS

Chapter Outline

7.1 Introduction to Confidence Intervals

7.2 Confidence Intervals for a Single Population Mean with Known Population Standard Deviation

7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation

7.4 Confidence Intervals for a Population Proportion

7.5 Calculating the Sample Size for a Confidence Interval

7.6 Exercises

7.1 INTRODUCTION TO CONFIDENCE INTERVALS



Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. Photo by comedy_nose, CC BY 4.0.

Suppose you want to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would obtain a point estimate of the true mean rent of two-bedroom apartments in your town. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would obtain a point estimate for the true proportion of the baskets you make when shooting a basketball.

We use sample data to make generalizations about an unknown population. This part of

statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called **confidence intervals**.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the t -distribution, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=148#oembed-1>

Watch this video: Understanding Confidence Intervals: Statistics Help by Dr Nic's Math and Stats [4:02]

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean \bar{x} and the sample standard deviation s . You would use the sample mean \bar{x} to estimate the population mean and the sample standard deviation s to estimate the population standard deviation. The sample mean \bar{x} is the point estimate for the population mean μ . The sample standard deviation s is the point estimate for the population standard deviation σ . Each of \bar{x} and s is called a **statistic**.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is **likely** to include the unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ , but we do know that the population standard deviation is $\sigma = 1$ and the sample size is $n = 100$. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of

the samples, the sample mean \bar{x} will be within two standard deviations of the population mean μ . For our iTunes example, two standard deviations is $2 \times 0.1 = 0.2$. The sample mean \bar{x} is likely to be within 0.2 units of μ .

Because \bar{x} is within 0.2 units of μ , which is unknown, μ is likely to be within 0.2 units of \bar{x} in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations ($2 \times 0.1 = 0.2$) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples. Suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between $\bar{x} - 0.2 = 2 - 0.2 = 1.8$ and $\bar{x} + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the (unknown) population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is the interval with lower limit 1.8 and upper limit 2.2.**

The 95% confidence interval implies two possibilities. Either the interval 1.8 to 2.2 contains the true mean μ or our sample produced an \bar{x} that is not within 0.2 units of the true mean μ . Because we are 95% confident that the true population mean is inside the interval, the second possibility, that the population mean is not inside the interval, happens for only 5% of all the samples.

Remember that a confidence interval is created for an unknown population parameter like the population mean μ . Confidence intervals for some parameters have the form:

$$\begin{array}{l} \text{\mbox{Lower Limit} \&= \& \text{\mbox{point estimate}} - \text{\mbox{margin} \\ \text{\mbox{Upper Limit} \&= \& \text{\mbox{point estimate}} + \text{\mbox{margin} \\ \text{\mbox{of error}} \end{array}$$

The margin of error depends on the confidence level and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase “margin of error.” Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Attribution

“Chapter 8 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

7.2 CONFIDENCE INTERVALS FOR A SINGLE POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population mean where the population standard deviation is known.

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**, we need \bar{x} , which is the **point estimate** of the unknown population mean μ .

The confidence interval estimate will have the form:

$$\begin{array}{l} \text{\mbox{Lower Limit}} \ \&= \ \& \ \overline{x} - \text{\mbox{margin of error}} \\ \text{\mbox{Upper Limit}} \ \&= \ \& \ \overline{x} + \text{\mbox{margin of error}} \end{array}$$

The margin of error depends on the **confidence level**. The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of their conclusions.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=157#oembed-1>

Watch this video: Confidence Intervals – Introduction by Joshua Emmanuel [3:34]

EXAMPLE

Suppose we have collected data from a sample. The sample mean is 7 and the margin of error is 2.5.

The confidence interval is:

$$\text{Lower Limit} = 7 - 2.5 = 4.5$$

$$\text{Upper Limit} = 7 + 2.5 = 9.5$$

If the confidence level is 95%, then we say that, “We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5.”

TRY IT

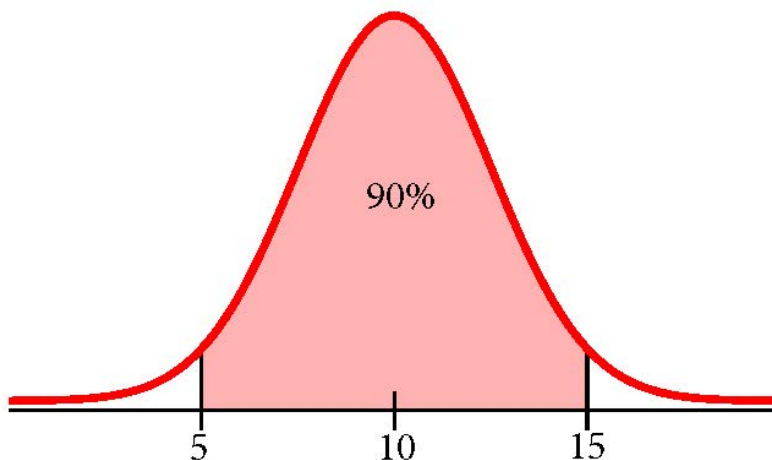
Suppose we have data from a sample. The sample mean is 15 and the margin of error is 3.2. What is the confidence interval estimate for the population mean?

Click to see Solution

$$\begin{array}{l} \text{\mbox{Lower Limit} \&= \& 15 - 3.2 = 11.8 \\ \text{\mbox{Upper Limit} \&= \& 15 + 3.2 = 18.2 \end{array}$$

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval with a lower limit of 5 and an upper limit of 15.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of 10% in both tails, or 5% in each tail, of the normal distribution.



To capture the central 90%, we must go out 1.645 “standard deviations” on either side of the calculated sample mean. The value 1.645 is the z -score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the “standard deviation” used must be appropriate for the parameter we are estimating. So in this section we need to use the standard deviation that applies to sample means,

which is $\frac{\sigma}{\sqrt{n}}$ (the standard deviation of the sample means). The fraction $\frac{\sigma}{\sqrt{n}}$ is commonly called the **standard error of the mean** in order to clearly distinguish the standard deviation for a sample mean from the population standard deviation σ .

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean \bar{x} from the sample data. Remember, in this section we already know the population standard deviation σ .
- Find the z -score that corresponds to the confidence level C .
- Calculate the limits for the confidence interval.
- Write a sentence that interprets the estimate in the context of the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z -score for the Confidence Level

When we know the population standard deviation σ , we use a standard normal distribution to calculate the margin of error and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution. The confidence level C is the area in the middle of the standard normal distribution. The remaining area, $1 - C$, is split equally between the two tails, so each of the tails contains an area equal to $\frac{1 - C}{2}$.

The z -score needed to construct the confidence interval is the z -score so that the **entire** area to the left of z -score equals the area in the middle (the confidence level) plus the area in the left tail $\left(\frac{1 - C}{2}\right)$. That is, the required z -score for the confidence interval is the z -score so that the entire area to the left of the z -score is

$$C + \frac{1 - C}{2}$$

For example, if the confidence level is 95%, then the area in the **center** of the standard normal

distribution is 0.95 and the area in the left tail is $\frac{1 - 0.95}{2} = 0.025$. We would need to find the z -score so that the entire area to the left of the z -score equals $0.95 + 0.025 = 0.975$.

CALCULATING THE **Formula does not parse**-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level (the area in the middle of the distribution) plus the remaining area in the left tail.

Calculating the Margin of Error

The margin of error for a confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is known is

$$\text{Margin of Error} = z \times \frac{\sigma}{\sqrt{n}}$$

where z is the the z -score so the area the left of z is $C + \frac{1 - C}{2}$.

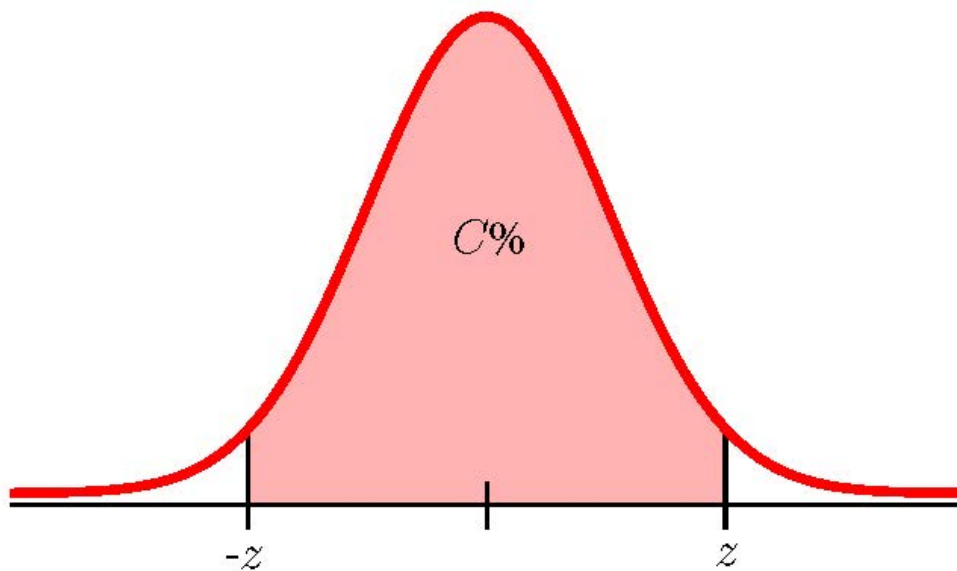
Constructing the Confidence Interval

The limits for the confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is known are

$$\text{Lower Limit} = \bar{x} - z \times \frac{\sigma}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{x} + z \times \frac{\sigma}{\sqrt{n}}$$

where z is the z -score so the area the left of z is $C + \frac{1 - C}{2}$.



Interpreting a Confidence Interval

The interpretation should clearly state the confidence level C , explain what population parameter is being estimated (in this case a **population mean**), and state the confidence interval (both endpoints)—“We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units).”



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=157#oembed-2>

Watch this video: Confidence Interval for a population mean – σ known by Joshua Emmanuel [4:30]

EXAMPLE

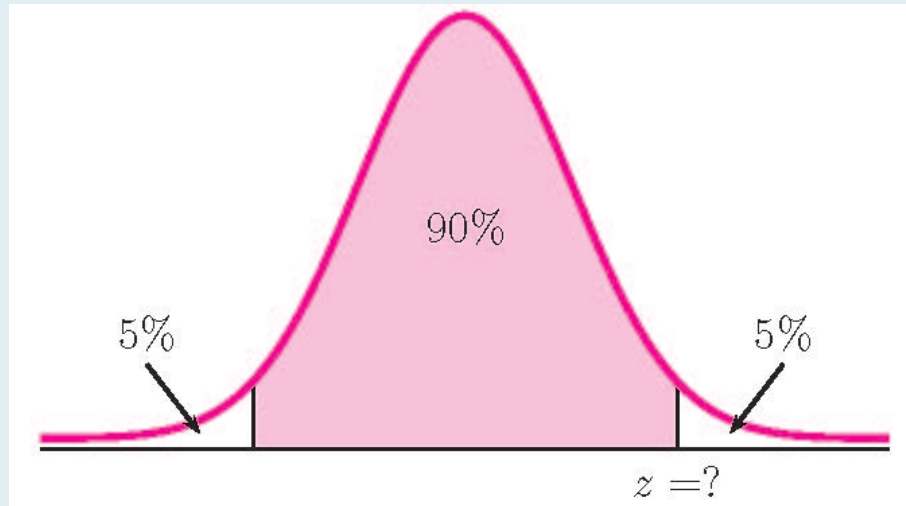
Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and has a sample mean of 68 points.

1. Find a 90% confidence interval for the mean exam score.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that the mean exam score for all the exams is 70? Explain.

Solution:

1. To find the confidence interval, we need to find the z -score for the 90% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is

$$0.9 + \frac{1 - 0.9}{2} = 0.95.$$



Function	norm.s.inv	Answer
Field 1	0.95	1.6448...

So $z = 1.6448\dots$. From the question $\bar{x} = 68$, $\sigma = 3$ and $n = 36$. The 90% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.6448\dots \times \frac{3}{\sqrt{36}} \\ &= 67.18 \\ \text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.6448\dots \times \frac{3}{\sqrt{36}} \\ &= 68.82 \end{aligned}$$

- We are 90% confident that the mean exam score is between 67.18 points and 68.82 points.
- It is not reasonable to conclude that the mean exam score is 70 points because 70 points is outside the confidence interval. (In this case there is a 90% chance that the actual mean exam score is in between 67.18 and 68.82 and only a 10% chance that the mean exam score is outside this interval. So it is unlikely (but not impossible) that the actual mean exam score is a value outside of the confidence interval.)

NOTES

1. When calculating the limits for the confidence interval keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cells containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual population mean captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.
3. 90% of all confidence interval constructed this way contain the true mean exam score. For example, if we constructed 100 of these confidence intervals (using 100 different samples of size 36), we would expect 90 of them to contain the true mean exam score.

TRY IT

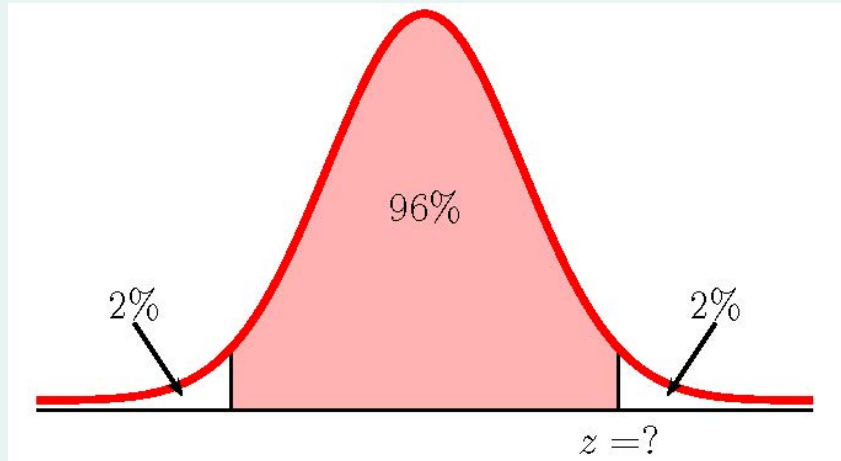
Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

1. Find a 96% confidence interval for the mean delivery time.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to claim that the mean delivery time is 35 minutes? Explain.

Click to see Solution

1.

Function	norm.s.inv	Answer
Field 1	0.98	2.053...



$$\begin{aligned} \text{Lower Limit} &= \overline{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 36 - 2.053... \times \frac{6}{\sqrt{28}} \\ &= 33.67 \\ \text{Upper Limit} &= \overline{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 36 + 2.053... \times \frac{6}{\sqrt{28}} \\ &= 38.05 \end{aligned}$$

- We are 96% confident that the mean delivery time is between 33.67 minutes and 38.05 minutes.
- It is reasonable to conclude that the mean delivery time is 35 minutes because 35 minutes is inside the confidence interval.

EXAMPLE

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy.

Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. This table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

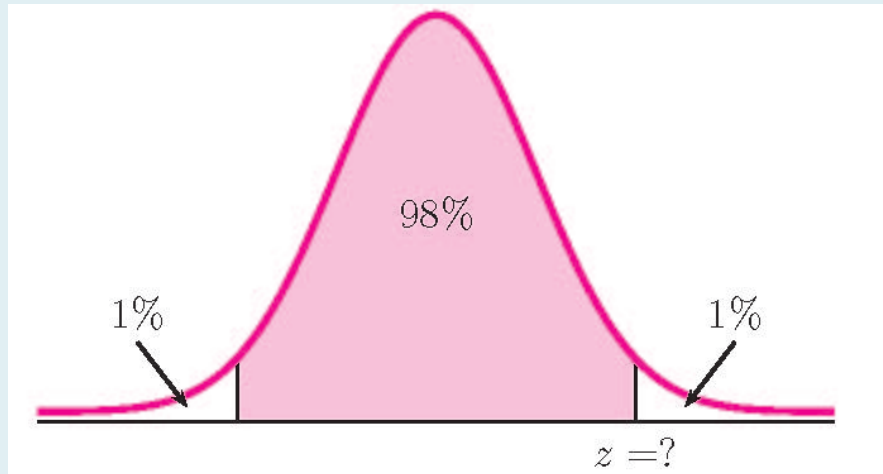
Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

1. Find a 98% confidence interval for the mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.
2. Interpret the confidence interval found in part 1.

Solution:

1. To find the confidence interval, we need to find the z -score for the 98% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is

$$0.98 + \frac{1 - 0.98}{2} = 0.99.$$



Function	norm.s.inv	Answer
Field 1	0.99	2.3263...

So $z = 2.3263\dots$. From the sample data supplied in the question $\bar{x} = 1.0237\dots$ and $n = 30$. The population standard deviation is $\sigma = 0.337$. The 98% confidence interval is

$$\begin{array}{l} \text{Lower Limit} \quad = \quad \overline{x} - z \times \frac{\sigma}{\sqrt{n}} \\ = \quad 1.0237\dots - 2.3263\dots \times \frac{0.377}{\sqrt{30}} \\ = \quad 0.8806 \\ \text{Upper Limit} \quad = \quad \overline{x} + z \times \frac{\sigma}{\sqrt{n}} \\ = \quad 1.0237\dots + 2.3262\dots \times \frac{0.377}{\sqrt{30}} \\ = \quad 1.1839 \end{array}$$

- We are 98% confident that the mean of the Specific Absorption Rates is between 0.8806 watts per kilogram and 1.1839 watts per kilogram.

TRY IT

This table shows a different random sampling of 20 cell phone models. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

1. Construct a 93% confidence interval for the mean SAR for cell phones certified for use in the United States.
2. Interpret the confidence interval found in part 1.

Click to see Solution

1.	Function	norm.s.inv	Answer
	Field 1	0.965	1.8119...

$$\begin{array}{l} \text{Lower Limit} = \bar{x} - z \times \frac{\sigma}{\sqrt{n}} = 0.94... - 1.8119... \times \frac{0.337}{\sqrt{20}} \\ \text{Upper Limit} = \bar{x} + z \times \frac{\sigma}{\sqrt{n}} = 0.94... + 1.8119... \times \frac{0.337}{\sqrt{20}} \\ = 1.0766 \end{array}$$

2. We are 93% confident that the mean of the Specific Absorption Rates is between 0.8035 watts per kilogram and 1.0766 watts per kilogram.

Notice the difference in the confidence intervals calculated in the Example and Try It just completed. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information.

Changing the Confidence Level

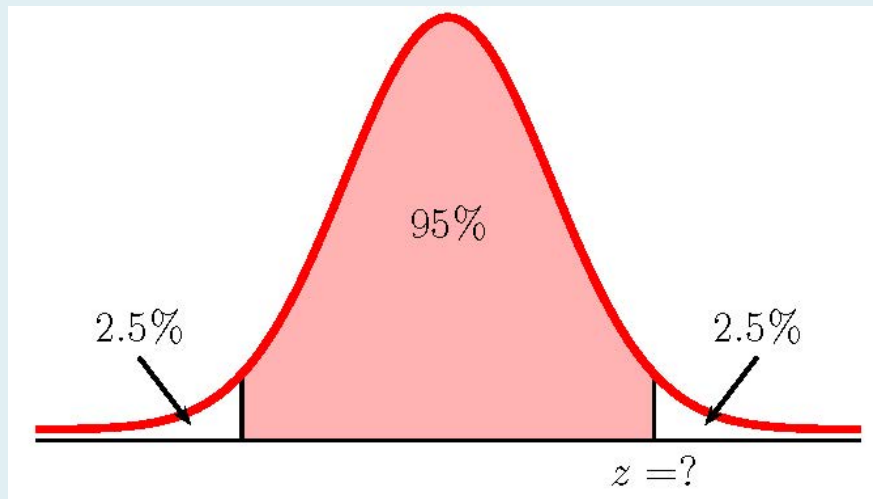
EXAMPLE

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean of 68 points. Previously we found a 90% confidence interval for the mean exam score. Now, find a 95% confidence interval for the mean exam score. Interpret the 95% confidence interval.

Solution:

To find the confidence interval, we need to find the z -score for the 95% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is

$$0.95 + \frac{1 - 0.95}{2} = 0.975.$$



Function	norm.s.inv	Answer
Field 1	0.975	1.9599...

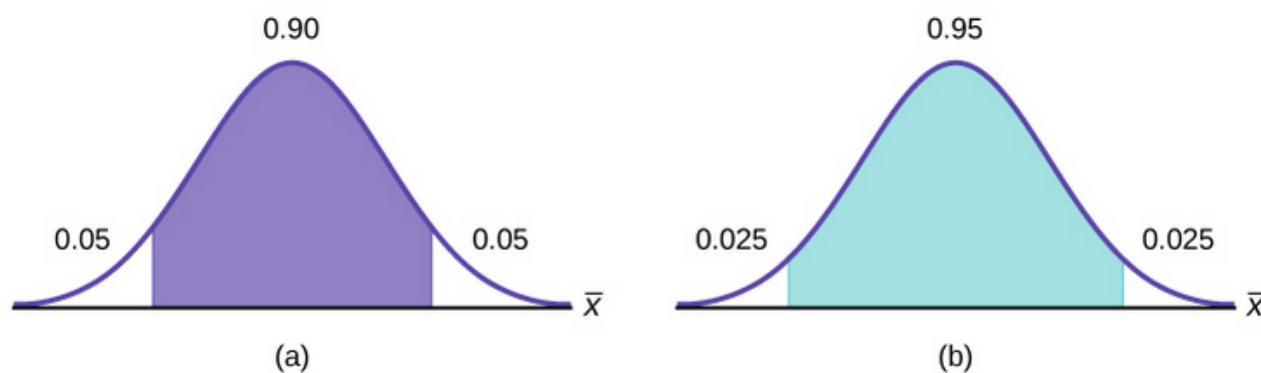
So $z = 1.9599\dots$. From the question $\bar{x} = 68$, $\sigma = 3$ and $n = 36$. The 95% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.9599 \times \frac{3}{\sqrt{36}} \\ &= 67.02 \\ \text{Upper Limit} &= \bar{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.9599 \times \frac{3}{\sqrt{36}} \\ &= 68.98 \end{aligned}$$

We are 95% confident that the mean exam score is between 67.02 points and 68.98 points.

Comparing the Results

For the exam scores examples, the 90% confidence interval has a lower limit of 67.18 and an upper limit of 68.82, and the 95% confidence interval has a lower limit of 67.02 and an upper limit of 68.98. Notice that the 95% confidence interval is wider (the distance between the limits is larger in the 95% confidence interval). If we look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Effect of Changing the Confidence Level

- Increasing the confidence level increases the margin of error, making the confidence interval wider.
- Decreasing the confidence level decreases the margin of error, making the confidence interval narrower.

Changing the Sample Size

EXAMPLE

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. Previously, we found a 90% confidence interval for the mean exam score using a sample of size 36 with a sample mean of 68.

1. Suppose everything is kept the same but the sample size is 100 (instead of 36). Find the 90% confidence interval.
2. Suppose everything is kept the same but the sample size is 25 (instead of 36). Find the 90% confidence interval.

Solution:

1.

Function	norm.s.inv	1.6448...
Field 1	0.95	

$$\begin{aligned} \text{\mbox{Lower Limit}} &= \overline{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.6448... \times \frac{3}{\sqrt{100}} \\ &= 67.51 \\ \text{\mbox{Upper Limit}} &= \overline{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.6448... \times \frac{3}{\sqrt{100}} \\ &= 68.49 \end{aligned}$$

2.

Function	norm.s.inv	1.6448...
Field 1	0.95	

$$\begin{aligned} \text{\mbox{Lower Limit}} &= \overline{x} - z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 - 1.6448... \times \frac{3}{\sqrt{25}} \\ &= 67.01 \\ \text{\mbox{Upper Limit}} &= \overline{x} + z \times \frac{\sigma}{\sqrt{n}} \\ &= 68 + 1.6448... \times \frac{3}{\sqrt{25}} \\ &= 69.27 \end{aligned}$$

Comparing the Results

For the exam scores examples, the 90% confidence interval with a sample size of 36 has a lower limit of 67.18 and an upper limit of 68.82, with a sample size of 100 has a lower limit is 67.51 and an upper limit is 68.49, and with a sample size of 25 has a lower limit is 67.01 and an upper limit is 69.27. When the sample size increased, the confidence interval is narrower. When the sample size decreased, the confidence interval is wider. Generally, the smaller the sample size, the wider the confidence interval needs to be in order to achieve the same level of confidence.

Effect of Changing the Sample Size

- Increasing the sample size causes the margin of error to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the margin of error to increase, making the confidence interval wider.

Concept Review

In this section, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. A confidence interval has the general form:

$$\begin{array}{l} \text{Lower Limit} \text{ \& } = \text{ \& } \overline{x} - \text{margin of error} \\ \text{Upper Limit} \text{ \& } = \text{ \& } \overline{x} + \text{margin of error} \end{array}$$

The general form for a confidence interval for a single population mean, known standard deviation is given by

$$\begin{array}{l} \text{Lower Limit} \text{ \& } = \text{ \& } \overline{x} - z \times \\ \frac{\sigma}{\sqrt{n}} \\ \text{Upper Limit} \text{ \& } = \text{ \& } \overline{x} + z \times \\ \frac{\sigma}{\sqrt{n}} \end{array}$$

where z is the the z -score so the area the left of z is $C + \frac{1 - C}{2}$.

The calculation of the margin of error depends on the size of the sample and the level of confidence required. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding margin of error increases as well. As the sample size increases, the margin of error decreases.

Attribution

“8.1 A Single Population Mean using the Normal Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

7.3 CONFIDENCE INTERVALS FOR A SINGLE POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population mean where the population standard deviation is unknown.

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ , and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation. He found that the actual distribution depends on the sample size. This problem led him to “discover” what is called the **Student’s t -distribution**. The name comes from the fact that Gosset wrote under the pen name “Student.”

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the t -distribution for sample sizes of at most 30. With technology, the practice now is to use the t -distribution whenever s is used as an estimate for σ .

When a simple random sample of size n is taken from a population that has an approximately normal distribution with mean μ , an unknown population standard deviation, and the sample

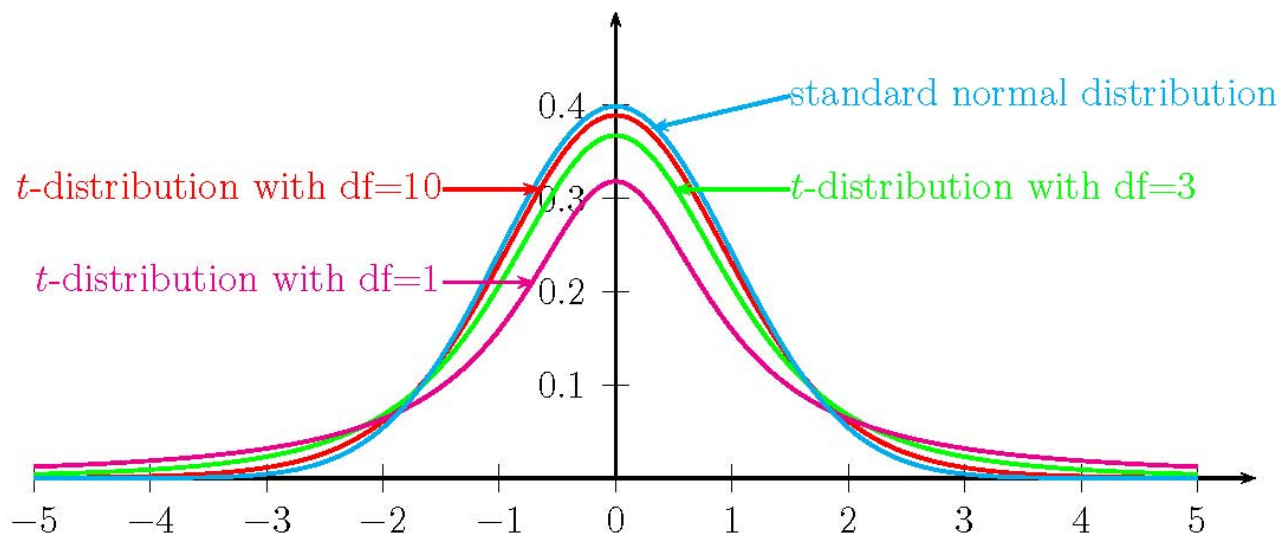
standard deviation s is used as an estimate for the population standard deviation, the distribution of the sample means follows a t -distribution with $n - 1$ degrees of freedom. For each sample size n , there is a different t -distribution. The t -score is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Every t -distribution has a parameter called the **degrees of freedom (df)**. In this case where the t -distribution is used for the distribution of the sample means, the value of the degrees of freedom is $n - 1$. Here the value of $n - 1$ used as the degrees of freedom comes from the calculation of the sample standard deviation s . Because the sum of the deviations is zero, we can find the last deviation once we know the other $n - 1$ deviations. The other $n - 1$ deviations can change or vary freely. Note that the value or formula of the degrees of freedom for the t -distribution will vary depending on the situation in which the t -distribution is used.

Properties of the t -Distribution

- The mean for the t -distribution is 0.
- The graph for the t -distribution is a symmetric, bell-shaped curve, similar to the standard normal curve. The graph is symmetric about the mean 0.
- The t -distribution has more probability in its tails than the standard normal distribution because the spread of the t -distribution is greater than the spread of the standard normal distribution. So the graph of the t -distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the t -distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of t -distribution becomes more like the graph of the standard normal distribution. In fact, the t -distribution with an infinite number of degrees of freedom is the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and does not need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.



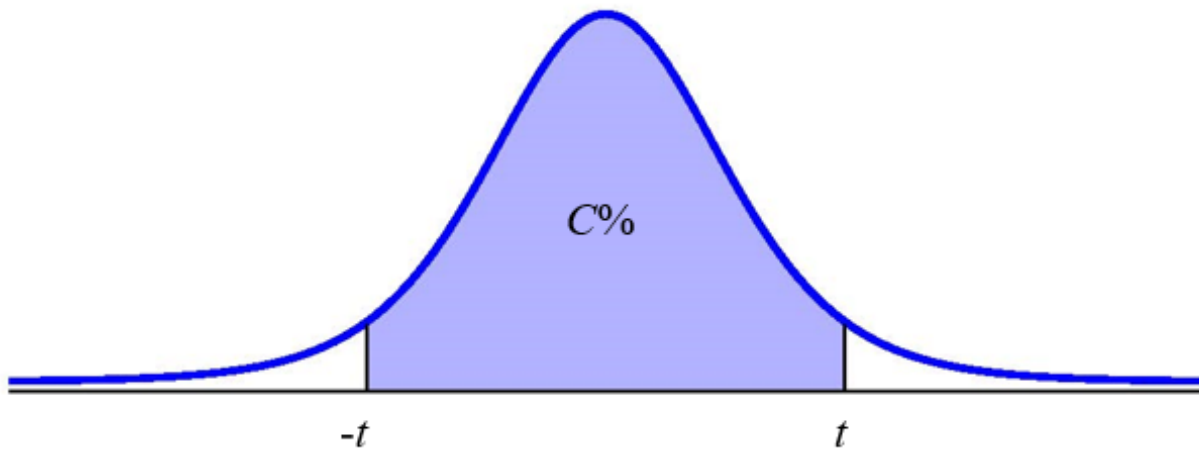
Constructing the Confidence Interval

When finding a confidence interval for an unknown population mean when the population standard deviation is unknown, we use the sample standard deviation s as an estimate for the (unknown) population standard deviation and we use a t -distribution with $n - 1$ degrees of freedom to find the required t -score for the confidence interval. In this case we replace the z -score with a t -score and σ with s in the formulas for the limits of the confidence interval for a population mean.

To construct the confidence interval, take a random sample of size n from the population. Calculate the sample mean \bar{x} and the sample standard deviation s . The limits for the confidence interval with confidence level C for an unknown population mean μ when the population standard deviation σ is **unknown** are

$$\begin{aligned} \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\ \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \end{aligned}$$

where t is the (positive) t -score of the t -distribution with $n - 1$ degrees of freedom so the area under the t -distribution in between $-t$ and t is C .



CALCULATING THE **Formula does not parse**-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the value of the degrees of freedom for the t -distribution. For a confidence interval for a population mean, the degrees of freedom is $n - 1$.

The output from the **t.inv.2t** function is the value of the t -score needed to construct the confidence interval.

Visit the Microsoft page for more information about the **t.inv.2t** function.

NOTE

The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the

middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=163#oembed-1>

Watch this video: Confidence Interval for a population mean – σ unknown by Joshua Emmanuel [7:40]

EXAMPLE

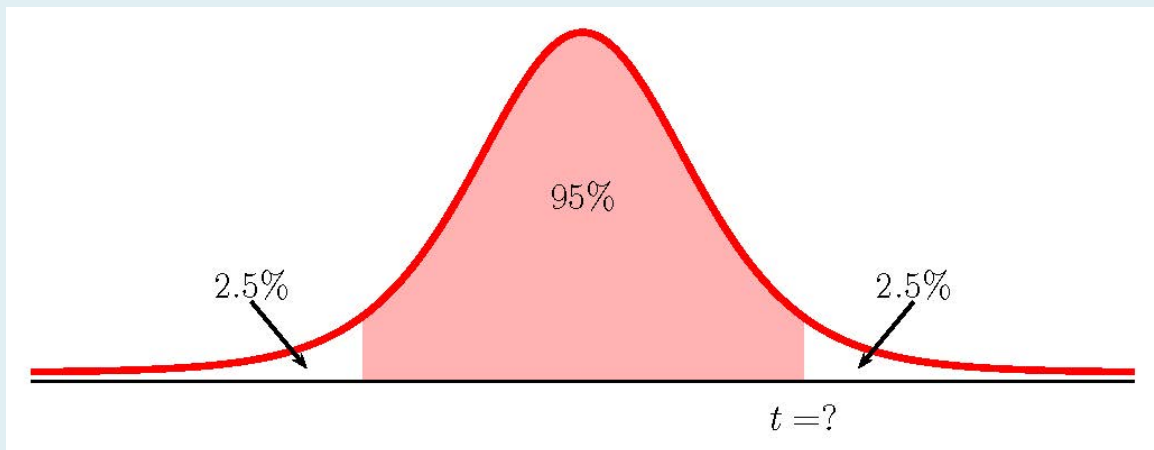
Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below.

8.6	7.3	10.3
9.4	9.2	5.4
7.9	9.6	8.1
6.8	8.7	5.5
8.3	11.4	6.9

1. Construct a 95% confidence interval for the mean sensory rate.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that mean sensory rate is 10? Explain.

Solution:

1. To find the confidence interval, we need to find the t -score for the 95% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.95 = 0.05$. The degrees of freedom for the t -distribution is $n - 1 = 15 - 1 = 14$.



Function	t.inv.2t	Answer
Field 1	0.05	2.1447...
Field 2	14	

So $t = 2.1447\dots$. From the sample data supplied in the question $\bar{x} = 8.226\dots$, $s = 1.672\dots$ and $n = 15$. The 95% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \overline{x} - t \times \frac{s}{\sqrt{n}} \\ &= 8.226... - 2.1447... \times \frac{1.672...}{\sqrt{15}} \\ &= 7.30 \\ \text{Upper Limit} &= \overline{x} + t \times \frac{s}{\sqrt{n}} \\ &= 8.226... + 2.1447... \times \frac{1.672...}{\sqrt{15}} \\ &= 9.15 \end{aligned}$$

2. We are 95% confident that the mean sensory rate is between 7.30 and 9.15.
3. It is not reasonable to conclude that the mean sensory rate is 10 because 10 is outside of the confidence interval.

NOTE

When calculating the limits for the confidence interval keep all of the decimals in the t -score and other values such as \overline{x} and s throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cells containing the t -score, \overline{x} and s , to ensure that all of the decimal places are used in the calculation.

TRY IT

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results.

8.2	8.6	8.9	9.2
9.1	6.9	9.9	7.5
7.7	11.2	10.1	10.5

1. Construct a 97% confidence interval for the mean number of hours slept each night.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to assume that the mean number of hours slept each night is 9 hours? Explain.

Click to see Solution

Function	t.inv.2t	Answer
Field 1	0.03	2.4906...
Field 2	11	

$$\begin{aligned} \text{Lower Limit} &= \overline{x} - t \times \frac{s}{\sqrt{n}} \\ &= 8.9833... - 2.4906... \times \frac{1.2904...}{\sqrt{12}} \\ &= 8.056 \\ \text{Upper Limit} &= \overline{x} + t \times \frac{s}{\sqrt{n}} \\ &= 8.9833... + 2.4906... \times \frac{1.2904...}{\sqrt{12}} \\ &= 9.911 \end{aligned}$$

2. We are 97% confident that the mean number of hours slept each night is between 8.056 hours and 9.911 hours.
3. It is reasonable to assume the mean number of hour slept each night is 9 hours because 9 is inside the confidence interval.

EXAMPLE

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the “In utero/newborn” group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and

nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. This table shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

1. Construct a 90% confidence interval for the mean number of targeted industrial chemicals found in an infant's blood.
2. Interpret the confidence interval found in part 1.

Solution:

1. To find the confidence interval, we need to find the t -score for the 90% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.90 = 0.1$. The degrees of freedom for the t -distribution is $n - 1 = 20 - 1 = 19$

Function	t.inv.2t	Answer
Field 1	0.1	1.7291...
Field 2	19	

So $t = 1.7291\dots$. From the sample data supplied in the question $\bar{x} = 127.45$, $s = 25.9645\dots$ and $n = 20$. The 90% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x} - t \times \frac{s}{\sqrt{n}} \\ &= 127.45 - 1.7291\dots \times \frac{25.9645\dots}{\sqrt{20}} \\ &= 117.41 \\ \text{Upper Limit} &= \bar{x} + t \times \frac{s}{\sqrt{n}} \\ &= 127.45 + 1.7291\dots \times \frac{25.9645\dots}{\sqrt{20}} \\ &= 137.49 \end{aligned}$$

2. We are 90% confident that the mean number of targeted industrial chemicals found in an infant's blood is between 117.41 and 137.49.

TRY IT

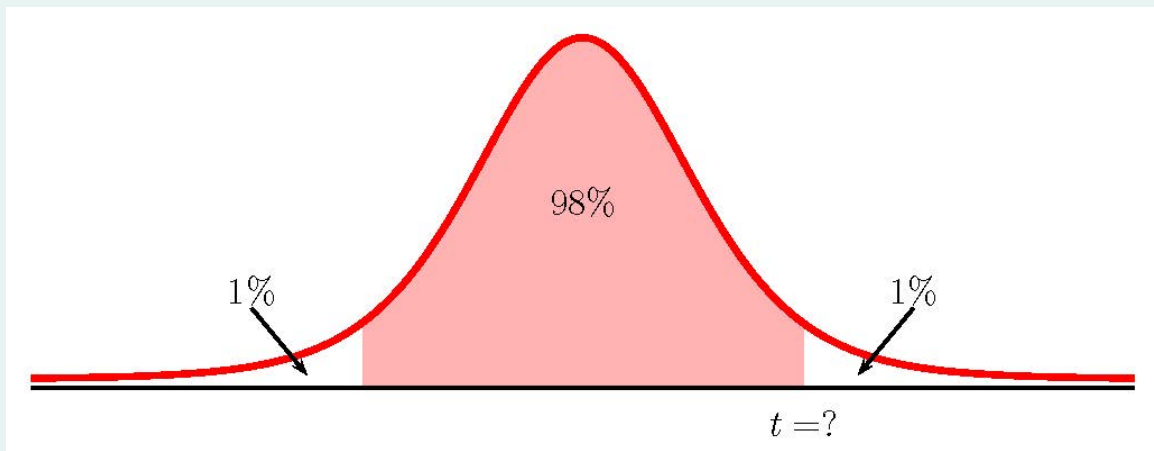
A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in this table.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

1. Construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that the mean number of hours statistics students spend watching television in one week is 5? Explain.

Click to see Solution

1.	Function	t.inv.2t	Answer
	Field 1	0.02	2.6244...
	Field 2	14	



$$\begin{aligned} \text{Lower Limit} &= \overline{x} - t \times \frac{s}{\sqrt{n}} \\ &= 6.133... - 2.6244... \times \frac{5.514...}{\sqrt{15}} \\ &= 2.397 \\ \text{Upper Limit} &= \overline{x} + t \times \frac{s}{\sqrt{n}} \\ &= 6.133... + 2.6244... \times \frac{5.514...}{\sqrt{15}} \\ &= 9.870 \end{aligned}$$

- We are 98% confident that the mean number of hours statistics students will spend watching television in one week is between 2.397 hours and 9.870 hours.
- It is reasonable to assume the mean number of hours statistics students will spend watching television in one week is 5 hours because 5 is inside the confidence interval.

Concept Review

In many cases, the population standard deviation σ for the population being studied is unknown. In these cases, it is common to use the sample standard deviation s as an estimate of σ . The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the t -distribution is much better.

The general form for a confidence interval for a single population mean with unknown population standard deviation is given by

$$\text{Lower Limit} = \bar{x} - t \times \frac{s}{\sqrt{n}}$$

$$\text{Upper Limit} = \bar{x} + t \times \frac{s}{\sqrt{n}}$$

where t is the (positive) t -score of the t -distribution with $n - 1$ degrees of freedom so the area under the t -distribution in between $-t$ and t is C .

Attribution

“8.2 A Single Population Mean using the Student t Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

7.4 CONFIDENCE INTERVALS FOR A POPULATION PROPORTION

LEARNING OBJECTIVES

- Calculate and interpret confidence intervals for estimating a population proportion.

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 37% and 43%.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

A confidence interval for a population proportion is based on the fact that the sample proportions follow an approximately normal distribution when both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. Similar to confidence intervals for population means, a confidence interval for a population proportion is constructed by taking a sample of size n from the population, calculating the sample proportion \hat{p} , and then adding and subtracting the margin of error from \hat{p} to get the limits of the confidence interval.

In order to construct a confidence interval for a population proportion, we must be able to assume the sample proportions follow a normal distribution. As we have seen previously, we can assume the sample proportions follow a normal distribution when both $n \times p \geq 5$ and

$n \times (1 - p) \geq 5$. But in this situation, the population proportion p is unknown so we cannot check the values of $n \times p$ and $n \times (1 - p)$. Because we must take a sample and calculate the sample proportion \hat{p} , we can check the quantities $n \times \hat{p}$ and $n \times (1 - \hat{p})$. For the confidence interval, if both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, we can assume the sample proportions follow a normal distribution.

Calculating the Margin of Error

The margin of error for a confidence interval with confidence level C for an unknown population proportion p is

$$\text{Margin of Error} = z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

where z is the the z -score so the area the left of z is $C + \frac{1 - C}{2}$.

NOTE

In the margin of error formula, the sample proportion \hat{p} is used to estimate the unknown population proportion p . The estimated sample proportion \hat{p} is used because p is the unknown quantity we are trying to estimate with the confidence interval. The sample proportion \hat{p} is calculated from the sample taken to construct the confidence interval where

$$\hat{p} = \frac{\text{number of items in the sample with characteristic of interest}}{n}$$

Constructing the Confidence Interval

The limits for the confidence interval with confidence level C for an unknown population proportion p are

$$\text{Lower Limit} = \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

$$\text{Upper Limit} = \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

where z is the z -score so the area to the left of z is $C + \frac{1 - C}{2}$.

NOTE

The confidence interval can only be used if we can assume the sample proportions follow a normal distribution. This means we must check that $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$ before constructing the confidence interval. If one of $n \times \hat{p}$ or $n \times (1 - \hat{p})$ is less than 5, we cannot construct the confidence interval.

CALCULATING THE Formula does not parse-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z**, enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level (the area in the middle of the distribution) plus the remaining area in the left tail.

EXAMPLE

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones.

1. Construct a 95% confidence interval for the proportion of adult residents of this city who have cell phones.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that 85% of the adult residents of this city have cell phones? Explain.

Solution:

1. The sample proportion is $\hat{p} = \frac{421}{500} = 0.842$. We need to check $n \times \hat{p}$ and $n \times (1 - \hat{p})$:

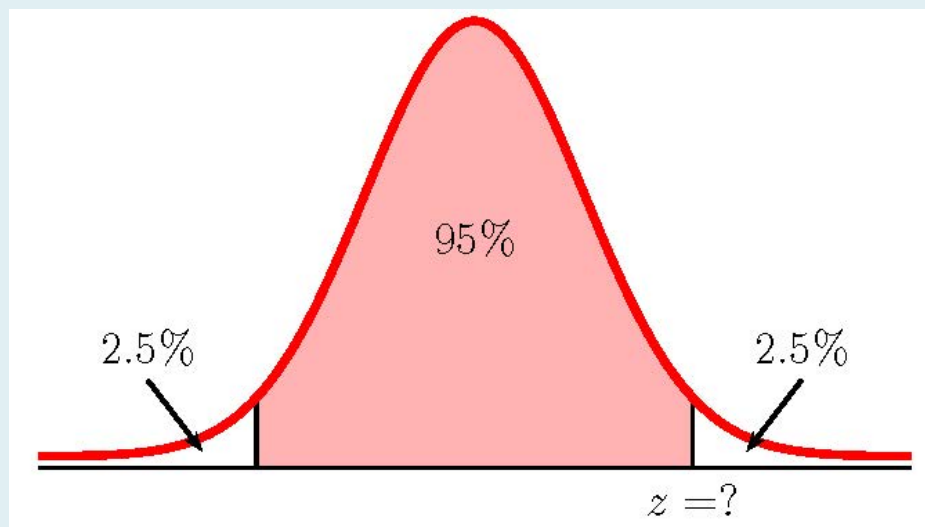
$$n \times \hat{p} = 500 \times 0.842 = 421 \geq 5$$

$$n \times (1 - \hat{p}) = 500 \times (1 - 0.842) = 79 \geq 5$$

Because both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, the sample proportions follow a normal distribution and we can construct the confidence interval.

To find the confidence interval, we need to find the z -score for the 95% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is

$$0.95 + \frac{1 - 0.95}{2} = 0.975.$$



Function	norm.s.inv	Answer
Field 1	0.975	1.9599...

So $z = 1.9599\dots$. The 95% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.842 - 1.9599\dots \times \sqrt{\frac{0.842(1-0.842)}{500}} \\ &= 0.8100 \\ \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.842 + 1.9599\dots \times \sqrt{\frac{0.842(1-0.842)}{500}} \\ &= 0.8740 \end{aligned}$$

2. We are 95% confident that the proportion of adult residents of this city who have cell phones is between 81% and 87.4%.
3. It is reasonable to conclude that 85% of the adult residents of this city have cell phones because 85% is inside the confidence interval.

NOTES

1. When calculating the limits for the confidence interval keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cells containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
2. The limits for the confidence interval are percents. For example, the upper limit of 0.8740 is the decimal form of a percent: 87.4%.
3. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual population proportion captured by the confidence interval (i.e. be specific to the context of the question), and express the limits as percents.
4. 95% of all confidence interval constructed this way contain the proportion of adult residents in this city that have a cell phone. For example, if we constructed 100 of these confidence (using 100 different samples of size 500), we would expect 95 of them to contain the true proportion of adult residents in this city that have a cell phone.

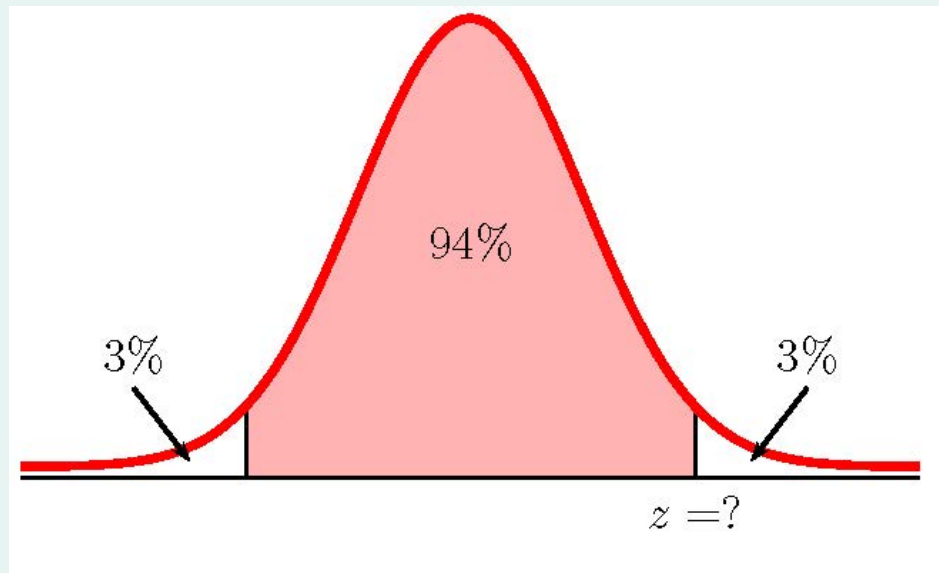
TRY IT

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet.

1. Construct a 94% confidence interval for the proportion of people who own tablets.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to assume that 30% of people own tablets? Explain.

Click to see Solution

1.	Function	norm.s.inv	Answer
	Field 1	0.97	1.8807...



$$\begin{aligned}
 \text{\mbox{Lower Limit}} &= \hat{p} - z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.392 - 1.8807... \times \sqrt{\frac{0.392 \times (1 - 0.392)}{250}} \\
 &= 0.3339 \\
 \text{\mbox{Upper Limit}} &= \hat{p} + z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\
 &= 0.392 + 1.8807... \times \sqrt{\frac{0.392 \times (1 - 0.392)}{250}} \\
 &= 0.4501
 \end{aligned}$$

2. We are 94% confident that the proportion of people who own tablets is between 33.39% and 45.01%.
3. It is not reasonable to claim the proportion of people who own tablets is 30% because 30% is outside the confidence interval.

EXAMPLE

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters.

1. Construct a 90% confidence interval for the percent of students who are registered voters.
2. Interpret the confidence interval found in part 1.

Solution:

1. The sample proportion is $\hat{p} = \frac{300}{500} = 0.6$. We need to check $n \times \hat{p}$ and $n \times (1 - \hat{p})$:

$$n \times \hat{p} = 500 \times 0.6 = 300 \geq 5$$

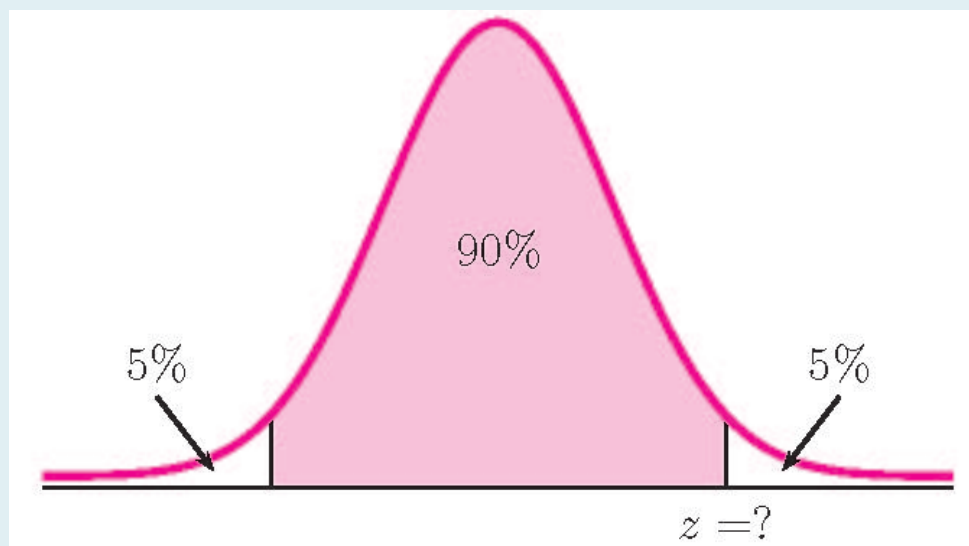
$$n \times (1 - \hat{p}) = 500 \times (1 - 0.6) = 200 \geq 5$$

Because both $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$, the sample proportions follow a normal distribution and we can construct the confidence interval.

To find the confidence interval, we need to find the z -score for the 90% confidence interval.

This means that we need to find the z -score so that the entire area to the left of z is

$$0.90 + \frac{1 - 0.90}{2} = 0.95.$$



Function	norm.s.inv	Answer
Field 1	0.95	1.6448...

So $z = 1.6448\dots$. The 90% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.6 - 1.6448\dots \times \sqrt{\frac{0.6 \times (1-0.6)}{500}} \\ \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.6 + 1.6448\dots \times \sqrt{\frac{0.6 \times (1-0.6)}{500}} \end{aligned}$$

- We are 90% confident that the percent of students who are registered voters is between 56.4% and 63.6%.

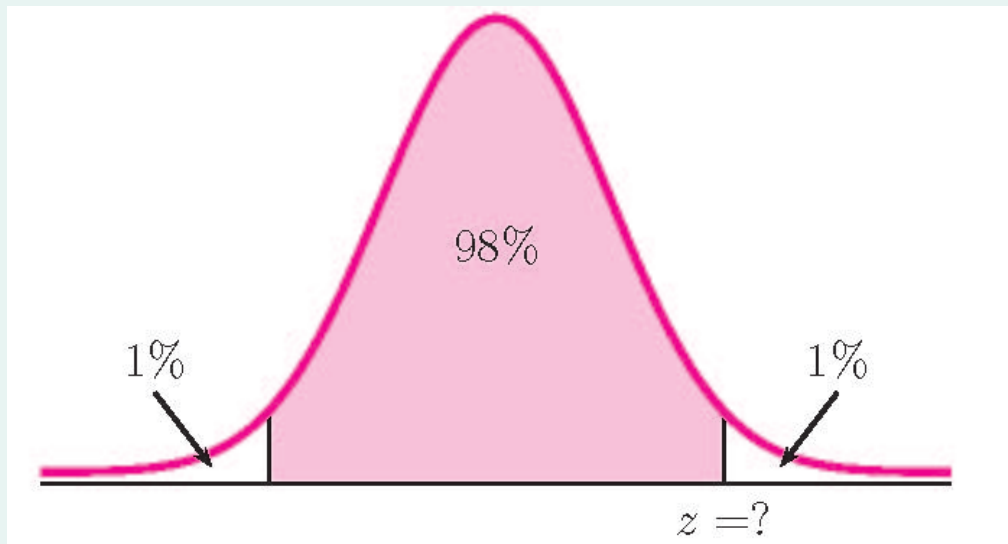
TRY IT

A student polls her school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

- Construct a 98% confidence interval for the proportion of students who are against the new legislation.
- Interpret the confidence interval found in part 1.
- A parents group claims that only 75% of students are against the legislation. Is it reasonable for the group to make this claim? Explain.

Click to see Solution

- | Function | norm.s.inv | Answer |
|----------|------------|-----------|
| Field 1 | 0.99 | 2.3263... |



$$\begin{aligned} \text{Lower Limit} &= \hat{p} - z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.8 - 2.3264 \times \sqrt{\frac{0.8(1-0.8)}{600}} \\ &= 0.7620 \\ \text{Upper Limit} &= \hat{p} + z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.8 + 2.3263 \times \sqrt{\frac{0.8(1-0.8)}{600}} \\ &= 0.8380 \end{aligned}$$

- We are 98% confident that the proportion of students who are against the new legislation is between 76.20% and 83.80%.
- It is not reasonable for the group to claim the proportion is 75% because 75% is outside of the confidence interval.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=166#oembed-1>

Watch this video: Confidence Interval for a population proportion by Excel is Fun [8:34]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=166#oembed-2>

Watch this video: Confidence Interval for a population proportion by Excel is Fun [4:51]

Concept Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

The general form for a confidence interval for a single population proportion is given by

$$\begin{array}{l} \text{Lower Limit} = \hat{p} - z \times \sqrt{\frac{\hat{p}}{n} \times (1 - \hat{p})} \\ \text{Upper Limit} = \hat{p} + z \times \sqrt{\frac{\hat{p}}{n} \times (1 - \hat{p})} \end{array}$$

where z is the the z -score so the area to the left of z is $C + \frac{1 - C}{2}$.

Attribution

“8.3 A Population Proportion“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

7.5 CALCULATING THE SAMPLE SIZE FOR A CONFIDENCE INTERVAL

LEARNING OBJECTIVES

- Calculate the minimum sample size required to estimate a population parameter.

Usually we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly, in both time and product. Simple telephone surveys will cost approximately \$30.00 each, for example, and some sampling requires the destruction of the product. Selecting a sample that is too large is expensive and time consuming. But selecting a sample that is too small can lead to inaccurate conclusions. We want to find the minimum sample size required to achieve the desired level of accuracy in the confidence interval.

Calculating the Sample Size for a Population Mean

The margin of error E for a confidence interval for a population mean is

$$E = \frac{z \times \sigma}{\sqrt{n}}$$

where z is the z -score so that the area under the standard normal distribution in between $-z$ and z is the confidence level C .

Rearranging this formula for n we get a formula for the sample size n :

$$n = \left(\frac{z \times \sigma}{E} \right)^2$$

In order to use this formula, we need values for z , E and σ :

- The value for z is determined by the confidence level of the interval, calculated the same way we calculate the z -score for a confidence interval.
- The value for the margin of error E is set as the predetermined acceptable error, or tolerance, for the difference between the sample mean \bar{x} and the population mean μ . In other words, E is set to the maximum allowable width of the confidence interval.
- An estimate for the population standard deviation σ can be found by one of the following methods:
 - Conduct a small pilot study and use the sample standard deviation from the pilot study.
 - Use the sample standard deviation from previously collected data. Although crude, this method of estimating the standard deviation may help reduce costs significantly.
 - Use $\frac{\text{Range}}{4}$ where Range is the difference between the maximum and minimum values of the population under study.

NOTES

1. Although we do not know the population standard deviation when calculating the sample size, we do not use the t -distribution in the sample size formula. In order to use the t -distribution in this situation, we need the degrees of freedom $n - 1$. But n is the sample size we are trying to estimate. So, we must use the normal distribution to determine the sample size.
2. The value of n determined from the formula is the **minimum** sample size required to achieve the desired level of confidence. The sample size n is a count, and so is an integer. It would be unusual for the value of n generated by the formula to be an integer. Because n is the minimum sample size required, we must **round** the output from the formula **up** to the next integer. If we round the value of n down, the sample size will be below the minimum required sample size.
3. After we have found the sample size n and collected the data for the sample, we use the appropriate confidence interval formula and the sample standard deviation from the actual sample (assuming σ is unknown), and not the estimate of the standard deviation used in the calculation of the sample size.

CALCULATING THE **Formula does not parse**-SCORE FOR SAMPLE SIZE IN EXCEL

To find the z -score to calculate the sample size for a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z** , enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

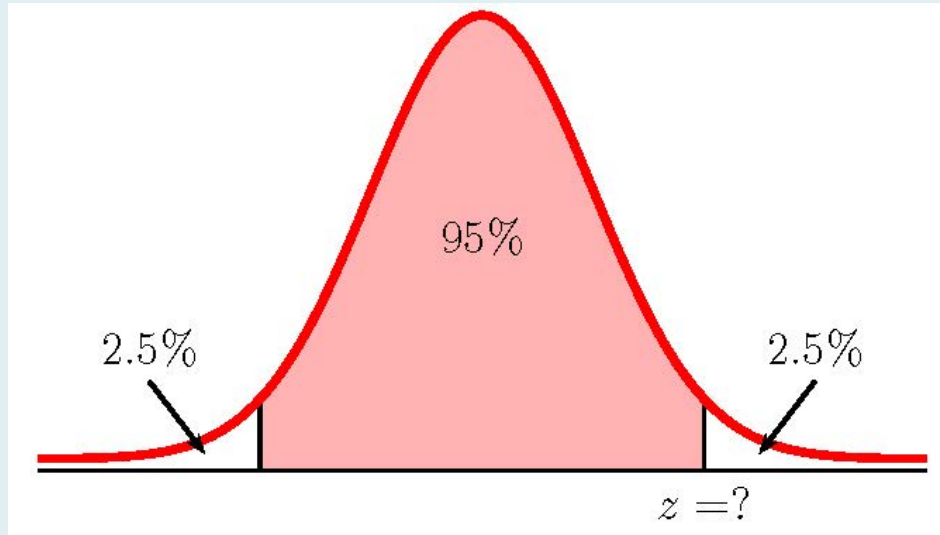
The output from the **norm.s.inv** function is the value of z -score needed to find the sample size.

EXAMPLE

We want to estimate the mean age of Foothill College students. From previous information, an estimate of the standard deviation of the ages of the students is 15 years. We want to be 95% confident that the sample mean age is within two years of the population mean age. How many randomly selected Foothill College students must be surveyed to achieved the desired level of accuracy?

Solution:

To find the sample size, we need to find the z -score for the 95% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is $0.95 + \frac{1 - 0.95}{2} = 0.975$.



Function	norm.s.inv	Answer
Field 1	0.975	1.9599...

So $z = 1.9599\dots$. From the question $\sigma \simeq 15$ and $E = 2$.

$$\begin{aligned} n &= \left(\frac{z \times \sigma}{E} \right)^2 &= & \left(\frac{1.9599\dots \times 15}{2} \right)^2 &= & 216.08\dots & \rightarrow & 217 \text{ \mbox{ students} } \end{aligned}$$

217 students must be surveyed to achieve the desired accuracy.

NOTE

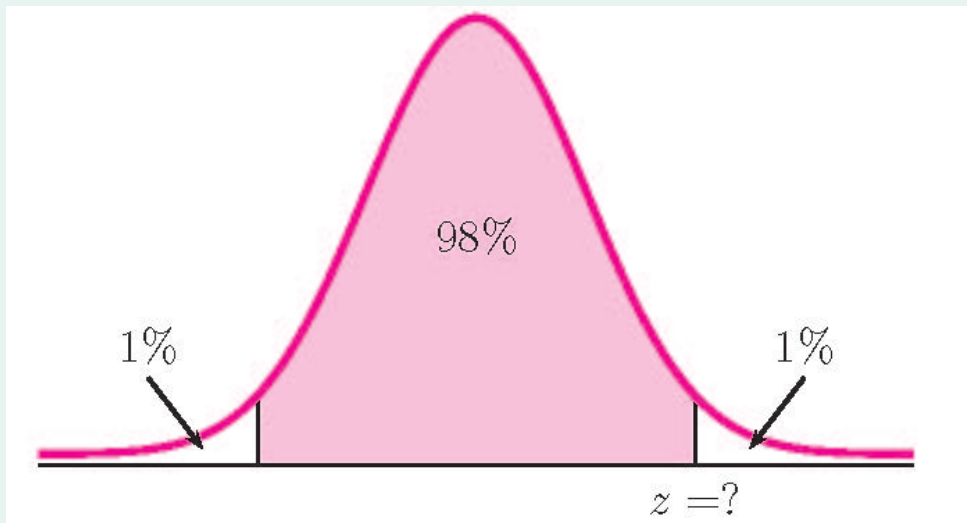
Remember to round the value for the sample size **UP** to the next integer. This ensures that the sample size is an integer and is large enough. Do not forget to include appropriate units with the sample size.

TRY IT

You want to estimate the height of all high school basketball players. You want to be 98% confident with a margin of error of 1.5. From a small pilot study, you estimate the standard deviation to be 3 inches. How large a sample do you need to take to achieve the desired level of accuracy?

Click to see Solution

Function	norm.s.inv	Answer
Field 1	0.99	2.3263...



$$\begin{aligned}
 n &= \left(\frac{z \times \sigma}{E} \right)^2 \\
 &= \left(\frac{2.3263... \times 3}{1.5} \right)^2 \\
 &= 21.6487... \\
 &\Rightarrow 22 \text{ high school basketball players}
 \end{aligned}$$

Calculating the Sample Size for a Population Proportion

The margin of error E for a confidence interval for a population proportion is

$$E = z \times \sqrt{\frac{p \times (1 - p)}{n}}$$

where z is the z -score so that the area under the standard normal distribution in between $-z$ and z is the confidence level C .

Rearranging this formula for n we get a formula for the sample size n :

$$n = p \times (1 - p) \times \left(\frac{z}{E}\right)^2$$

In order to use this formula, we need values for z , E and p :

- The value for z is determined by the confidence level of the interval, calculated the same way we calculate the z -score for a confidence interval.
- The value for the margin of error E is set as the predetermined acceptable error, or tolerance, for the difference between the sample proportion \hat{p} and the population proportion p . In other words, E is set to the maximum allowable width of the confidence interval.
- An estimate for the population proportion p . If no estimate for the population proportion is provided, we use $p = 0.5$.

NOTES

1. The value of n determined from the formula is the **minimum** sample size required to achieve the desired level of confidence. The sample size n is a count, and so is an integer. It would be unusual for the value of n generated by the formula to be an integer. Because n is the minimum sample size required, we must **round** the output from the formula **up** to the next integer. If we round the value of n down, the sample size will be below the minimum required sample size.
2. After we have found the sample size n and collected the data for the sample, we use the appropriate confidence interval formula and the sample proportion from the actual sample.
3. By using 0.5 as an estimate for p in the sample size formula we will get the largest required sample size for the confidence level and margin of error we selected. This is true

because of all combinations of two fractions (the values of p and $1 - p$) that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p , this is the common practice. This may result in oversampling, but certainly not under sampling.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. The table below shows the appropriate sample size at different levels of confidence and different margins of error, assuming $p = 0.5$. Looking at each row, we can see that for the same margin of error, a higher level of confidence requires a larger sample size. Similarly, looking at each column, we can see that for the same confidence level, a smaller margin of error requires a larger sample size.

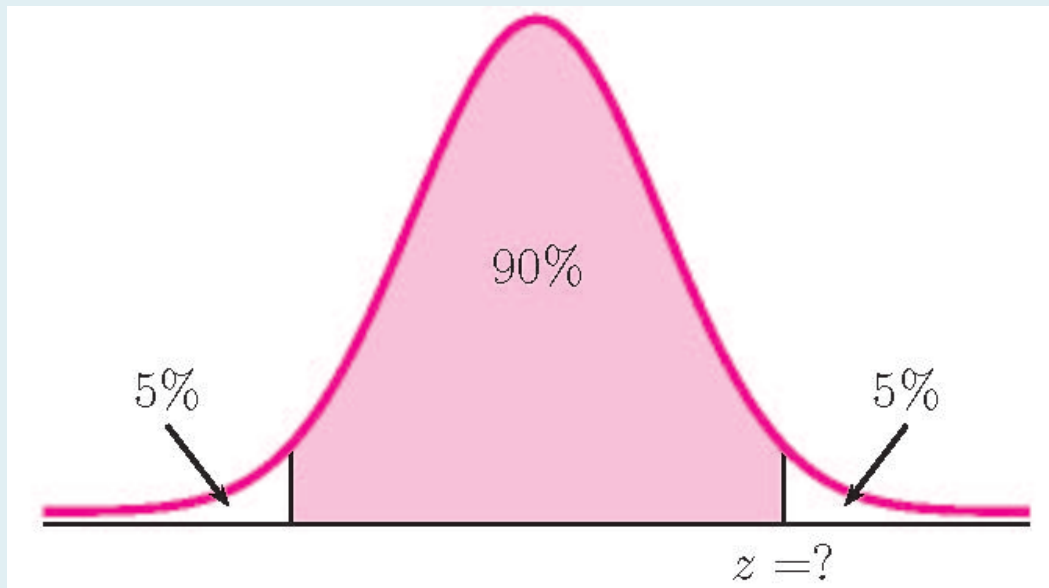
Required Sample Size (90%)	Required Sample Size (95%)	Margin of Error
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

EXAMPLE

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident with a margin of error of 3%?

Solution:

To find the sample size, we need to find the z -score for the 90% confidence interval. This means that we need to find the z -score so that the entire area to the left of z is $0.90 + \frac{1 - 0.90}{2} = 0.95$.



Function	norm.s.inv	Answer
Field 1	0.95	1.6448...

So $z = 1.6,448,....$. From the question $E = 0.03$. Because no estimate of the population proportion is given, $p = 0.5$.

$$\begin{aligned} n &= p \times (1-p) \times \left(\frac{z}{E}\right)^2 \\ &= 0.5 \times (1-0.5) \times \left(\frac{1.6448...}{0.03}\right)^2 \\ &\rightarrow 752 \end{aligned}$$

752 customers aged 50+ must be surveyed to achieve the desired accuracy.

NOTE

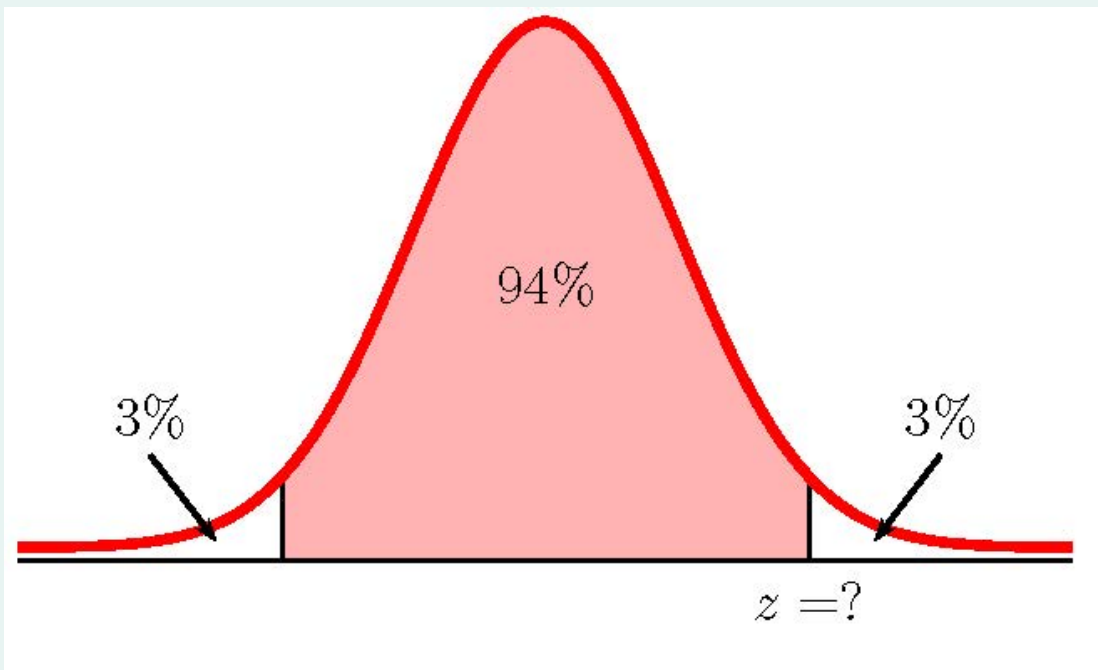
Remember to round the value for the sample size **UP** to the next integer. This ensures that the sample size is large enough. Do not forget to include appropriate units with the sample size.

TRY IT

Suppose an internet marketing company wants to determine the percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 94% confident that the estimated proportion is within 5% of the population proportion of customers who click on ads on their smartphones?

Click to see Solution

Function	norm.s.inv	Answer
Field 1	0.97	1.8807...



$$\begin{aligned}
 n &= p \times (1 - p) \times \left(\frac{z}{E}\right)^2 \\
 &= 0.5 \times (1 - 0.5) \times \left(\frac{1.8807\dots}{0.05}\right)^2 \\
 &= 353.738\dots \\
 &\Rightarrow 354 \text{ customers}
 \end{aligned}$$



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=168#oembed-1>

Watch this video: Sample Size for Confidence Intervals by ExcelIsFun [7:54]

Concept Review

In order to construct a confidence interval, a sample is taken from the population under study. But collecting sample information is time consuming and expensive. The minimum sample size required to achieve the desired level of accuracy is determined before collecting the sample data.

- Sample size for population means: $n = \left(\frac{z \times \sigma}{E}\right)^2$
- Sample size for population proportions: $n = p \times (1 - p) \times \left(\frac{z}{E}\right)^2$

After calculating the value of n from the formula, **round** the value of n **up** to the next integer.

Attribution

“7.2 The Central Limit Theorem for Sums“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

“8.4 Calculating the Sample Size n : Continuous and Binary Random Variables“ in Introductory Business Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

7.6 EXERCISES

1. The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

- Construct a 95% confidence interval for the population mean weight of newborn elephants.
- Interpret the confidence interval found in part (a).
- What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

2. The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

- Construct a 90% confidence interval for the population mean time to complete the forms.
- Interpret the confidence interval found in part (a).
- Is it reasonable to conclude the mean time to complete the forms is 10 minutes? Explain.
- If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
- If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
- Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

3. A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

- Construct a 90% confidence interval for the population mean weight of the heads of lettuce.
- Interpret the confidence interval found in part (a).
- Construct a 95% confidence interval for the population mean weight of the heads of lettuce.

- d. In complete sentences, explain why the confidence interval in part (a) is larger than in part (c).
 - e. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?
 - f. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?
4. The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students.
- a. Construct a 99% confidence interval for the mean age of students at Foothill College.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable for the college to claim that the mean age of its students is 35? Explain.
 - d. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?
 - e. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?
5. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.
- a. Construct a 95% confidence interval for the population mean height of male Swedes.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that the mean height of male Swedes is 75 inches? Explain.
6. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.
- a. Construct a 97% confidence interval for the population mean length of engineering conferences.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that the mean length of the conferences is 3 days? Explain.

7. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

- a. Construct a 90% confidence interval for the population mean time to complete the tax forms.
- b. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
- c. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
- d. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

8. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

- a. Construct a 90% confidence interval for the population mean weight of the candies.
- b. Construct a 98% confidence interval for the population mean weight of the candies.
- c. In complete sentences, explain why the confidence interval in part (b) is larger than the confidence interval in part (a).
- d. In complete sentences, give an interpretation of what the interval in part (b) means.

9. What is meant by the term "90% confident" when constructing a confidence interval for a mean?

- a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
- b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
- c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
- d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

10. The average height of young adult males has a normal distribution with standard deviation of

2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

11. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

- Construct a 99% confidence interval for the population mean time spent waiting.
- Interpret the confidence interval found in part (a).
- Is it reasonable to claim that the mean time spent waiting is 2 hours? Explain.

12. 108 Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

- Construct a 99% confidence interval for the population mean hours spent watching television per month.
- Interpret the confidence interval found in part (a).
- Why would the error bound change if the confidence level were lowered to 95%?

13. In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68.

- Construct a 96% confidence interval for the proportion of Bam-Bam snack pieces per bag.
- Interpret the confidence interval found in part (a).

14. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

- Construct a 95% confidence interval for the mean enrollment at community colleges in the United States.
- Interpret the confidence interval found in part (a).
- Is it reasonable to conclude that the mean enrollment at community colleges in the U.S. is

15,000? Explain.

- d. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

15. Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

- a. Construct a 98% confidence interval for the population mean time wasted.
- b. Explain in a complete sentence what the confidence interval means.

16. A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

- a. Construct a 95% confidence interval for the mean length of time the tranquilizers last.
- b. What does it mean to be “95% confident” in this problem?

17. Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

- a. Construct a 99% confidence interval for the mean length of time children use training wheels.
- b. Interpret the confidence interval found in part (a).
- c. Why would the error bound change if the confidence level were lowered to 90%?

18. The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns. The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

\$46,500.00	\$0	\$40,966.50	\$105,887.20	\$5,175.00
\$29,050.00	\$19,500.00	\$181,557.20	\$31,500.00	\$149,970.80
\$2,555,363.20	\$12,025.00	\$409,000.00	\$60,521.70	\$18,000.00
\$61,810.20	\$76,530.80	\$119,459.20	\$0	\$63,520.00
\$6,500.00	\$502,578.00	\$705,061.10	\$708,258.90	\$135,810.00
\$2,000.00	\$2,000.00	\$0	\$1,287,933.80	\$219,148.30

- Construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle.
- Interpret the confidence interval found in part (a).

19. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

- Construct a 92% confidence interval for the mean number of unoccupied seats per flight.
- Interpret the confidence interval found in part (a).
- Is it reasonable for the airlines to claim that the mean number of unoccupied seats per flight is 20? Explain.

20. In a recent sample of 84 used car sales costs, the sample mean was \$6,425 with a standard deviation of \$3,156. Assume the underlying distribution is approximately normal.

- Construct a 95% confidence interval for the mean cost of a used car.
- Explain what a “95% confidence interval” means for this study.

21. A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; \$1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

- Construct a 95% confidence interval for the mean worth of coupons.
- Interpret the confidence interval found in part (a).
- If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

22. Marketing companies are interested in knowing the percent of women who make the majority of household purchasing decisions.

- a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?
- b. If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

23. Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

- a. Construct a 95% confidence interval for the proportion of households where the women make the majority of the purchasing decisions.
- b. Interpret the confidence interval found in part (a).
- c. Is it reasonable for the marketing company to claim that women make the majority of purchasing decisions in 70% of households? Explain.

24. A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

- a. Construct a 90% confidence interval for the proportion of voters who believe the economy is the most significant issue in the upcoming election.
- b. Interpret the confidence interval found in part (a).
- c. Is it reasonable to claim that 60% of voters believe the economy is the most significant issue in the upcoming election? Explain.
- d. What would happen to the confidence interval if the level of confidence were 95%?

25. The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

- a. Construct a 92% confidence interval for the proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.
 - b. Interpret the confidence interval found in part (a).
26. Insurance companies are interested in knowing the percent of drivers who always buckle up before riding in a car.
- a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
 - b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?
 - c. Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up. Construct a 95% confidence interval for the proportion who claim they always buckle up.
27. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.
- a. Construct a 97% confidence interval for the proportion of people over 50 who ran and died in the same eight-year period.
 - b. Explain what a “97% confidence interval” means for this study.
28. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.
- a. Construct a 93% confidence interval for the proportion of adult Americans who feel that crime is the main problem.
 - b. Interpret the confidence interval found in part (a).
 - c. Is it reasonable to claim that 30% of Americans feel crime is the main problem? Explain.
29. According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed)

feel that “education and our schools” is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

- a. Construct a 90% confidence interval for the proportion of California adults who feel education and schools is one of the top issues facing California.
- b. Interpret the confidence interval found in part (a).
- c. Is it reasonable to claim that 90% of California adults feel education and schools is one of the top issues facing California? Explain.

30. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

- a. Construct a 99% confidence interval for the proportion of American adults who have illegally downloaded music.
- b. Interpret the confidence interval found in part (a).
- c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

31. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

Attribution

“Chapter 3 Homework” and “Chapter 8 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART VIII

HYPOTHESIS TESTS FOR SINGLE POPULATION PARAMETERS

Chapter Outline

8.1 Introduction to Hypothesis Testing

8.2 Null and Alternative Hypothesis

8.3 Outcomes and the Type I and Type II Errors

8.4 Distributions Required for a Hypothesis Test

8.5 Rare Events, the Sample, Decision, and Conclusion

8.6 Hypothesis Tests for a Population Mean with Known Population Standard Deviation

8.7 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation

8.8 Hypothesis Tests for a Population Proportion

8.9 Exercises

8.1 INTRODUCTION TO HYPOTHESIS TESTING



You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. Photo by Robert Neff, CC BY 4.0.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets an average of 35 miles per gallon. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year.

A statistician will make a decision about whether these claims are true or false. This process is

called **hypothesis testing**. A hypothesis test involves collecting data from a sample and evaluating the data. From the evidence provided by the sample data, the statistician makes a decision as to whether or not there is sufficient evidence to reject or not reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single population means and single population proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses. Only one of these hypotheses is true and the hypothesis test will determine which of the hypothesis is **most likely** true.
 2. Collect sample data. (In homework problems, the data or summary statistics will be given to you.)
 3. Determine the correct distribution to perform the hypothesis test.
 4. Analyze the sample data by performing calculations that ultimately will allow you to reject or not reject the null hypothesis.
 5. Make a decision and write a meaningful conclusion.
-

Attribution

“Chapter 9 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.2 NULL AND ALTERNATIVE HYPOTHESES

LEARNING OBJECTIVES

- Describe hypothesis testing in general and in practice.

A hypothesis test begins by considering **two hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints and only one of these hypotheses is true. The hypothesis test determines which hypothesis is **most likely** true.

- The **null hypothesis** is denoted H_0 . It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.
 - The null hypothesis is a claim that a population parameter equals some value. For example, $H_0 : \mu = 5$.
- The **alternative hypothesis** is denoted H_a . It is a claim about the population that is contradictory to the null hypothesis and is what we conclude is true when we reject H_0 .
 - The alternative hypothesis is a claim that a population parameter is greater than, less than, or not equal to some value. For example, $H_a : \mu > 5$, $H_a : \mu < 5$, or $H_a : \mu \neq 5$. The form of the alternative hypothesis depends on the wording of the hypothesis test.
 - An alternative notation for H_a is H_1 .

Because the null and alternative hypotheses are contradictory, we must examine evidence to decide if we have enough evidence to reject the null hypothesis or not reject the null hypothesis. The evidence is in the form of sample data. After we have determined which hypothesis the sample data supports, we make a decision. There are two options for a **decision**. They are “**reject H_0** ”

if the sample information favors the alternative hypothesis or “**do not reject H_0** ” if the sample information is insufficient to reject the null hypothesis.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=175#oembed-1>

Watch this video: Simple hypothesis testing | Probability and Statistics | Khan Academy by Khan Academy [6:24]

EXAMPLE

A candidate in a local election claims that 30% of registered voters voted in a recent election. Information provided by the returning office suggests that the percentage is higher than the 30% claimed.

Solution:

The parameter under study is the proportion of registered voters, so we use p in the statements of the hypotheses. The hypotheses are

$$H_0 : p = 30\%$$

$$H_a : p > 30\%$$

NOTES

1. The null hypothesis H_0 is the claim that the proportion of registered voters that voted equals 30%.
2. The alternative hypothesis H_a is the claim that the proportion of registered voters that voted is greater than (i.e. higher) than 30%.

TRY IT

A medical researcher believes that a new medicine reduces cholesterol by 25%. A medical trial suggests that the percent reduction is different than claimed. State the null and alternative hypotheses.

Click to see Solution

$$H_0 : p = 25\%$$

$$H_a : p \neq 25\%$$

EXAMPLE

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). State the null and alternative hypotheses.

Solution:

$$\begin{array}{l} H_0: \mu = 2 \text{ \mbox{ points}} \\ H_a: \mu \neq 2 \text{ \mbox{ points}} \end{array}$$

EXAMPLE

We want to test whether or not the mean height of eighth graders is 66 inches. State the null and alternative hypotheses.

Solution:

$$\begin{array}{l} H_0: \mu = 66 \text{ \mbox{ inches}} \\ H_a: \mu \neq 66 \text{ \mbox{ inches}} \end{array}$$

EXAMPLE

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

Solution:

$$\begin{array}{l} H_0: \mu = 5 \text{ \mbox{ years}} \\ H_a: \mu < 5 \text{ \mbox{ years}} \end{array}$$

TRY IT

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses.

Click to see Solution

$$\begin{array}{l} H_0: \mu = 45 \text{ \mbox{ minutes}} \\ H_a: \mu < 45 \text{ \mbox{ minutes}} \end{array}$$

EXAMPLE

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

Solution:

$$\begin{array}{l} H_0: p = 6.6\% \\ H_a: p > 6.6\% \end{array}$$

TRY IT

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. State the null and alternative hypotheses.

Click to see Solution

$$\begin{array}{l} H_0: p = 40\% \\ H_a: p > 40\% \end{array}$$

Concept Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we evaluate the **null hypothesis**, typically denoted with H_0 . The null hypothesis is **not** rejected unless the hypothesis test shows otherwise. The null hypothesis always contain an equal sign ($=$). Always write the **alternative hypothesis**, typically denoted with H_a or H_1 , using less than, greater than, or not equals symbols ($<$, $>$, \neq). If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis. But we can never state that a claim is proven true or false. All we can conclude from the hypothesis test is which of the hypothesis is **most likely** true. Because the underlying facts about hypothesis testing is based on probability laws, we can talk only in terms of non-absolute certainties.

Attribution

“9.1 Null and Alternative Hypotheses“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.3 OUTCOMES AND THE TYPE I AND TYPE II ERRORS

LEARNING OBJECTIVES

- Differentiate between Type I and Type II errors in a hypothesis test.

When we perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not the null hypothesis. Ideally, the hypothesis test should tell us to not reject the null hypothesis when the null hypothesis is true and reject the null hypothesis when the null hypothesis is false. However, the outcome of the hypothesis test is based on sample information and probabilities, so there is a chance that the hypothesis test does not correctly identify the truth or falseness of the null hypothesis. The outcomes are summarized in the following table:

	Actual Truth State of the Null Hypothesis	
Outcome of Test	H_0 is True	H_0 is False
Do not reject H_0	Correct Outcome	Type II Error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

- The decision is **not to reject H_0** when **H_0 is true (correct decision)**. That is, the test identifies H_0 is true and in reality H_0 is true, which means the test correctly identified H_0 as true.
- The decision is to **reject H_0** when **H_0 is true** (incorrect decision known as a **Type I error**).

That is, the test identifies H_0 as false but in reality H_0 is true, which means the test did not correctly identify H_0 as true.

- The decision is **not to reject** H_0 when H_0 **is false** (incorrect decision known as a **Type II error**). That is, the test identifies H_0 is true but in reality H_0 is false, which means the test did not correctly identify H_0 as false.
- The decision is to **reject** H_0 when H_0 **is false** (**correct decision** whose probability is called the **Power of the Test**). That is, the test identifies H_0 is false and in reality H_0 is false, which means the test correctly identified H_0 as false.

There are two types of error that can occur in hypothesis testing. Each of the errors occurs with a particular probability.

- A **Type I error** occurs when the null hypothesis is rejected by the test (i.e. the test identifies the null hypothesis as false) but in reality the null hypothesis is true. The probability of a Type I error is denoted by α .
- A **Type II error** occurs when the null hypothesis is not rejected by the test (i.e. the test identifies the null hypothesis as true) but in reality the null hypothesis is false. The probability of a Type II error is denoted by β .

Although the probabilities of a Type I or Type II error should be as small as possible, because they are probabilities of errors, they are rarely zero.

EXAMPLE

Suppose the null hypothesis is

H_0 : Frank's rock climbing equipment is safe.

- **Type I error:** Frank thinks his rock climbing equipment is not safe when in fact the equipment is safe.
 - Frank believes H_0 is false but H_0 is actually true.
- **Type II error:** Frank thinks his rock climbing equipment is safe when in fact the equipment

is not safe.

- Frank believes H_0 is true but H_0 is actually false.

Note that, in this case, the error with the greater consequence is the Type II error. If Frank thinks his rock climbing equipment is safe and it actually is not safe, he will go ahead and use it.

TRY IT

Suppose the null hypothesis is

H_0 : The blood cultures contain no traces of pathogen X .

State the Type I and Type II errors.

Click to see Solution

- **Type I error:** The researcher thinks the blood cultures do contain traces of pathogen X , when in fact, they do not.
- **Type II error:** The researcher thinks the blood cultures do not contain traces of pathogen X , when in fact, they do.

EXAMPLE

Suppose the null hypothesis is

H_0 : The victim of a car accident is alive when they arrive at the ER.

- **Type I error:** The ER staff thinks that the victim is dead when in fact the victim is alive.
- **Type II error:** The ER staff think the victim is alive when in fact the victim is dead.

Note that, in this case, the error with the greater consequence is the Type I error. If the ER staff think the victim is dead, then they will not treat him.

TRY IT

Suppose the null hypothesis is

H_0 : A patient is not sick.

Which type of error has the greater consequence, Type I or Type II? Why?

Click to see Solution

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, they are sick, and so they will not get treatment.

EXAMPLE

A genetics lab claims its product can increase the likelihood a pregnancy will result in a boy being born. Statisticians want to test this claim. Suppose that the null hypothesis is

H_0 : The genetics lab product has no effect on gender outcome.

- **Type I error:** We believe the genetics lab's product can influence gender outcome when in fact the product has no effect.
- **Type II error:** We believe the genetics lab's product cannot influence gender outcome when in fact the product does have an effect.

Note that, in this case, the error with the greater consequence is the Type I error because couples would use the product in hopes of increasing the chances of having a boy.

TRY IT

“Red tide” is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regularly sampling shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Click to see Solution

In this scenario, an appropriate null hypothesis would be

H_0 : The mean level of toxins is at most 800 μg .

- **Type I error:** The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800 μg . The DMF continues the harvesting ban.
- **Type II error:** The DMF believes that toxin levels are within acceptable levels (are at most 800 μg) when, in fact, toxin levels are still too high (more than 800 μg). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

EXAMPLE

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is more serious?

- **Type I:** A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- **Type II:** A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this will most likely influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=177#oembed-1>

Watch this video: Type 1 errors | Inferential statistics | Probability and Statistics | Khan Academy by Khan Academy [3:23]

Concept Review

In every hypothesis test, the outcomes from the test are dependent on sample data and probabilities, which means that the conclusion of the test may not correctly identify the actual truth state of the null hypothesis. Such occurrences are expected. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II** error occurs when a false null hypothesis is not rejected.

Attribution

“9.2 Outcomes and the Type I and Type II Errors“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.4 DISTRIBUTIONS REQUIRED FOR A HYPOTHESIS TEST

LEARNING OBJECTIVES

- Identify the distribution required to conduct a hypothesis test.

Earlier in the course, we discussed sampling distributions: the sampling distribution of the sample mean and the sampling distribution of the sample proportion. These distributions play a role in hypothesis testing.

If the hypothesis test is on a population mean, we use the distribution of the sample means in the hypothesis test. As we learned previously, the distribution of the sample means follows a normal distribution if the population the sample is taken from is normal or if the sample size is large enough ($n \geq 30$). For a hypothesis test on a population mean we use a normal distribution when the population standard deviation is known or a t -distribution when the population standard deviation is unknown.

If the hypothesis test is on a population proportion, we use the distribution of the sample proportions in the hypothesis test. As we learned previously, the distribution of the sample proportions follows a normal distribution if $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ or a binomial distribution if one of $n \times p < 5$ or $n \times (1 - p) < 5$. For a hypothesis test on a population proportion we use either a normal distribution or a binomial distribution, depending on which of the above conditions is met.

Assumptions

When we perform a **hypothesis test of a single population mean** μ and the population standard deviation is **known**, we take a simple random sample from the population. We use a normal

distribution, assuming the population is normal or the sample size is large enough ($n \geq 30$). The z-score we need is the z-score from the distribution of the sample means: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

When we perform a **hypothesis test of a single population mean** μ and the population standard deviation is **unknown**, we take a simple random sample from the population. We use a t -distribution, assuming the population is normal or the sample size is large enough ($n \geq 30$). We use the sample standard deviation to approximate the population standard deviation. The t -score

we need is: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.

When we perform a **hypothesis test of a single population proportion** p , we take a simple random sample from the population. We use a normal distribution when $n \times p \geq 5$ and $n \times (1 - p) \geq 5$. In this case, the z-score we need is the z-score from the distribution of the sample proportions: $z = \sqrt{\frac{p \times (1 - p)}{n}}$. Otherwise, we use a binomial distribution when one of $n \times p < 5$ or $n \times (1 - p) < 5$.

Concept Review

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

Testing a population mean:

- Population standard deviation is known: use a normal distribution, assuming the population is normal or $n \geq 30$.
- Population standard deviation is unknown: use a t -distribution, assuming the population is normal or $n \geq 30$.

Testing a population proportion:

- Use a normal distribution when $n \times p \geq 5$ and $n \times (1 - p) \geq 5$.
- Use a binomial distribution when at least one of $n \times p < 5$ or $n \times (1 - p) < 5$.

Attribution

“9.3 Distribution Needed for Hypothesis Testing“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.5 RARE EVENTS, THE SAMPLE, DECISION, AND CONCLUSION

LEARNING OBJECTIVES

- Define a rare event and identify how a rare event is used in a hypothesis test.
- Define p -value and significance level and identify how they are used in determining the outcome of a hypothesis test.

Establishing the type of distribution, sample size, and known or unknown population standard deviation can help us figure out how to go about a hypothesis test. However, there are several other factors we should consider when working out a hypothesis test.

Rare Events

Suppose we make an assumption about the value of a population parameter (this assumption is the **null hypothesis**). We conduct the hypothesis under the assumption that the null hypothesis is true. Then we randomly select a sample from the population. If the sample has properties that would be very **unlikely** to occur under the assumption the null hypothesis is true, then we would conclude that our assumption about the population is probably **incorrect**. Remember that our assumption is just an **assumption**—it is not a fact, and it may or may not be true. But the sample data we collect is real and the information from that sample is a fact that may or may not support the assumption we make about the null hypothesis.

For example, Didi and Ali are at the birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside of because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a \$100 bill. Didi is the first person to reach into the basket and pull out a bubble.

Her bubble contains a \$100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is such an **unlikely** occurrence, Ali is hoping that what the two of them were told is wrong and there are more \$100 bills in the basket. In this case a “**rare event**” has occurred (Didi getting the \$100 bill) so Ali doubts the original assumption about only one \$100 bill being in the basket.

A **rare event** is something we consider to be **unlikely** to happen (i.e. the probability of that event happening is very small). This is what we are looking for in a hypothesis test. We want to determine if the sample collected for the test is a rare event (unlikely to happen) under the assumption the null hypothesis is true. To determine if the sample is a rare event, we calculate the probability of the sample occurring, assuming that the null hypothesis is true. If the probability of the sample occurring is small, then the sample is a “rare event” and unlikely to occur under the assumption the null hypothesis is true. In such a case we would conclude that the original assumption that the null hypothesis is true must be incorrect, and so we would reject the null hypothesis. If the probability of the sample occurring is not small, then the sample is not a “rare event” and is actually likely to occur under the assumption the null hypothesis is true. In this case we would conclude the original assumption that the null hypothesis is true must be correct, and so we would not reject the null hypothesis.

Remember, a rare event is an event that is unlikely to happen. But unlikely does not mean impossible. The probability of a rare event is very small, which means that the chance of it happening is very small. But as long as the probability is not zero, there is still a possibility the event could happen.

Using the Sample to Test the Null Hypothesis

We use the sample data to calculate the actual probability of getting the selected sample, called the ***p*-value**, under the assumption the null hypothesis is true. The *p*-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample**.

A large *p*-value calculated from the sample data indicates that we should **not reject** the **null hypothesis**. The smaller the *p*-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would **reject** the null hypothesis if the evidence is strongly against it.

EXAMPLE

The customers of a local bakery claim that the height of the bakery's bread is, on average, 15 cm. The baker believes his customers are wrong, and that the average height of the bread is more than 15 cm. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 1 cm and the distribution of the heights is normal. Based on this sample, who is right: the customers or the baker?

Solution:

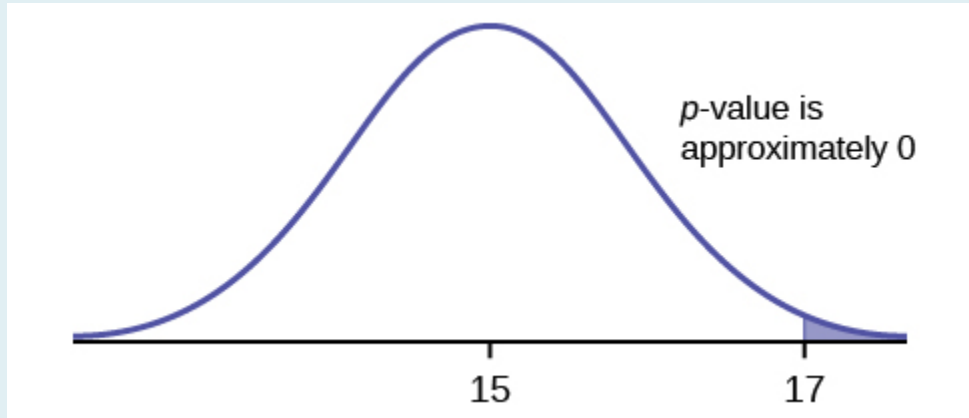
Here, the population under study is the height of the loaves of bread and μ is the average height of the loaves of bread.

The customers' claim is the null hypothesis: $\mu = 15$. The alternative hypothesis is the baker's claim: $\mu > 15$. In mathematical notation, the hypothesis are

$$\begin{array}{l} H_0: \mu = 15 \text{ cm} \\ H_a: \mu > 15 \text{ cm} \end{array}$$

Because the population standard deviation is known ($\sigma = 1$), the distribution we would use is the normal distribution.

Suppose the null hypothesis is true. That is, suppose $\mu = 15$. Under this assumption, we have to ask if the sample mean of 17 is likely or unlikely to occur. The hypothesis test works by asking the question how **unlikely** is this sample mean if the null hypothesis is true. The graph shows how far out the sample mean is on the normal curve. The p -value is the probability that, if we took another sample of size 10, any other sample mean would fall at least as far out as 17 cm.



The p -value is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for sample means. In fact, we are calculating the probability that in a sample of size 10 the sample mean is greater than 17. We learned how to calculate this type of probability when we learned about the sampling distribution of the sample mean:

Function	1-norm.dist	Answer
Field 1	17	0.0000000001
Field 2	15	
Field 3	1/sqrt(10)	
Field 4	true	

So $p\text{-value}=0.0000000001$, which tells us the probability of selecting a sample of size 10 and getting a sample mean greater than 17 is 0.0000000001 under the assumption that the null hypothesis is true ($\mu = 15$). This is a very, very small probability, which tells us that a sample mean of 17 is **unlikely** to happen if the population mean is 15 cm. Because the sample mean of 17 is so unlikely (meaning it is not happening by chance alone), we conclude that the assumption that the mean is 15 cm is wrong. That is, the evidence provided by the sample is **strongly against** the claim of the null hypothesis. So we reject the null hypothesis in favour of the alternative hypothesis. That is, based on the test, we believe the null hypothesis is false and the alternative hypothesis is true. So there is enough evidence to suggest that the average height of the loaves of bread is greater than 15 cm.

TRY IT

A normal distribution has a standard deviation of 1. The original claim is that the mean of the distribution is 12. An alternative claim is that the mean is greater than 12. The hypotheses are:

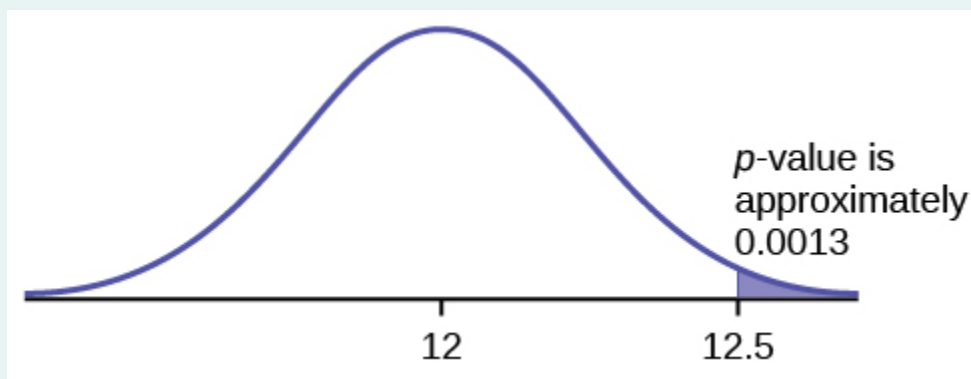
$$\begin{array}{l} H_0: \mu = 12 \\ H_a: \mu > 12 \end{array}$$

In a sample of 36, the sample mean is 12.5. Calculate the p -value.

Click to see Solution

Function	1-norm.dist	Answer
Field 1	12.5	0.0013
Field 2	12	
Field 3	1/sqrt(36)	
Field 4	true	

$$p\text{-value} = 0.0013$$



Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p -value and a preset or preconceived value called the significance level, denoted by α . The significance level is the cut-off value for likely versus unlikely when compared to the p -value. When the p -value is greater than the significance level, the sample is likely to occur under the assumption the null hypothesis is true, and so we would fail to reject the null hypothesis. When the p -value is less than or equal to the significance level, the sample is unlikely to occur under the assumption the null hypothesis is true, and so we would reject the null hypothesis in favour of the alternative hypothesis. A preset value for α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). The significance level may or may not be given at the beginning of the problem.

When we make a **decision** to reject or not reject H_0 , do as follows:

- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

When we “do not reject H_0 ,” it does not mean that we should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

After comparing the p -value and significance level, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

TRY IT

A genetics lab claims its product can increase the likelihood a pregnancy will result in a boy being born. Statisticians want to test this claim. Suppose the hypotheses are

$$\begin{array}{l} H_0: p = 50\% \\ H_a: p > 50\% \end{array}$$

After conducting the hypothesis test, $p\text{-value} = 0.025$. If the significance level is 1%, what is the conclusion of the test?

Click to see Solution

Because the p -value is greater than the significance level ($p\text{-value} = 0.025 > 0.01 = \alpha$), we do not reject the null hypothesis. There is not enough evidence to support the lab's stated claim that their procedures improve the chances of a boy being born.

Concept Review

A rare event is an event that is unlikely to occur. The probability of a rare event happening is very small. In a hypothesis test, we want to determine if the collected sample is a rare event. The p -value is the probability of getting the sample.

In a hypothesis test, the significance level is the cut-off value for likely versus unlikely. The significance level is compared to the p -value for the test.

- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
- If $p\text{-value} > \alpha$, do not reject H_0 .

After determining the outcome of test, we write a conclusion based on the specific context of the question.

Attribution

“9.4 Rare Events, the Sample, Decision and Conclusion“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.6 HYPOTHESIS TESTS FOR A POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population mean with known population standard deviation.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tell us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p -value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p -value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p -value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p -value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p -value.
- **Think about the meaning of the p -value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p -value (for example, 0.001 as opposed to 0.04) even if using a significance level of

0.05. Similarly, for a large p -value such as 0.4, as opposed to a p -value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more confidence that they made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5%.
- An alternative approach for hypothesis testing is to use what is called the **critical value approach**. In this book, we will only use the p -value approach. Some of the videos below may mention the critical value approach, but this approach will not be used in this book.

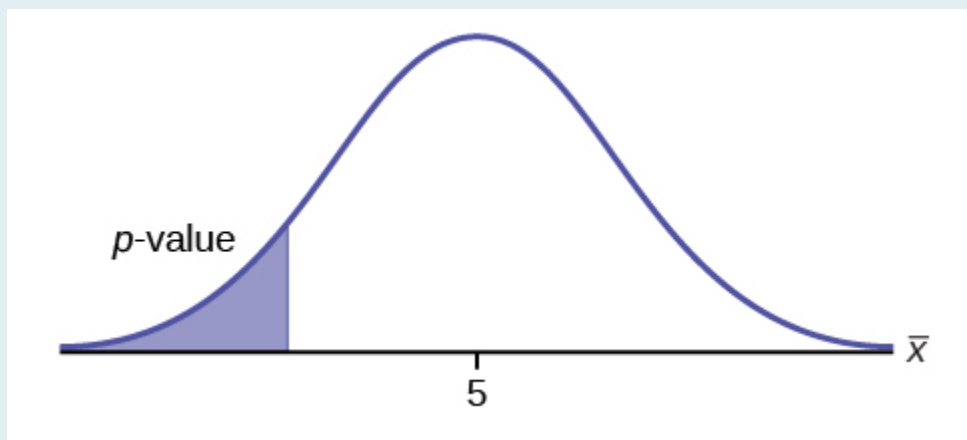
EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$H_0 : \mu = 5$$

$$H_a : \mu < 5$$

Because the alternative hypothesis is a $<$, this is a left-tailed test. The p -value is the area in the left-tail of the distribution.

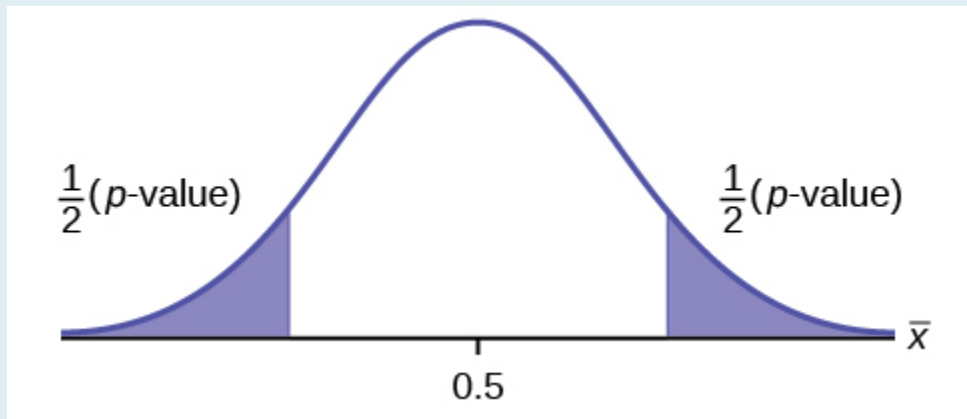


EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: \mu = 0.5 \\ H_a: \mu \neq 0.5 \end{array}$$

Because the alternative hypothesis is a \neq , this is a two-tailed test. The p -value is the sum of the areas in the two tails of the distribution. Each tail contains exactly half of the p -value.

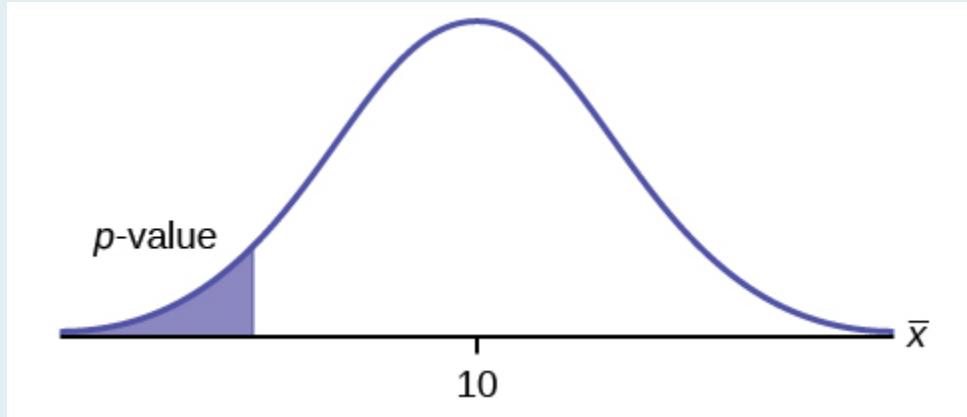


EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: \mu = 10 \\ H_a: \mu < 10 \end{array}$$

Because the alternative hypothesis is a $<$, this is a left-tailed test. The p -value is the area in the left-tail of the distribution.



Steps to Conduct a Hypothesis Test for a Population Mean with Known Population Standard Deviation

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level α .
4. When the population standard deviation is **known**, we use a normal distribution with
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
 to find the p -value. The p -value is the area in the corresponding tail of the normal distribution.
5. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON A POPULATION MEAN WITH KNOWN POPULATION STANDARD DEVIATION

The p -value for a hypothesis test on a population mean is the area in the tail(s) of the distribution of the sample mean. When the population standard deviation is known, use the normal distribution to find the p -value.

The p -value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p -value.

- For x , enter the value for \bar{x} .
- For μ , enter the mean of the sample means μ . Note: Because the test is run assuming the null hypothesis is true, the value for μ is the claim from the null hypothesis.
- For σ , enter the standard error of the mean $\frac{\sigma}{\sqrt{n}}$.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

Use the appropriate technique with the **norm.dist** function to find the area in the left-tail or the area in the right-tail.

EXAMPLE

Jeffrey, as an eight-year old, established a mean time of 16.43 seconds with a standard deviation of 0.8 seconds for swimming the 25-meter freestyle. His dad, Frank, thought that Jeffrey could swim the 25-meter freestyle faster using goggles. Frank bought Jeffrey a new pair of goggles and timed

Jeffrey swimming the 25-meter freestyle 15 different times. In the sample of 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey swim faster than 16.43 seconds. At the 5% significance level, did Jeffrey swim faster wearing the goggles? Assume that the swim times for the 25-meter freestyle are normally distributed.

Solution:

Hypotheses:

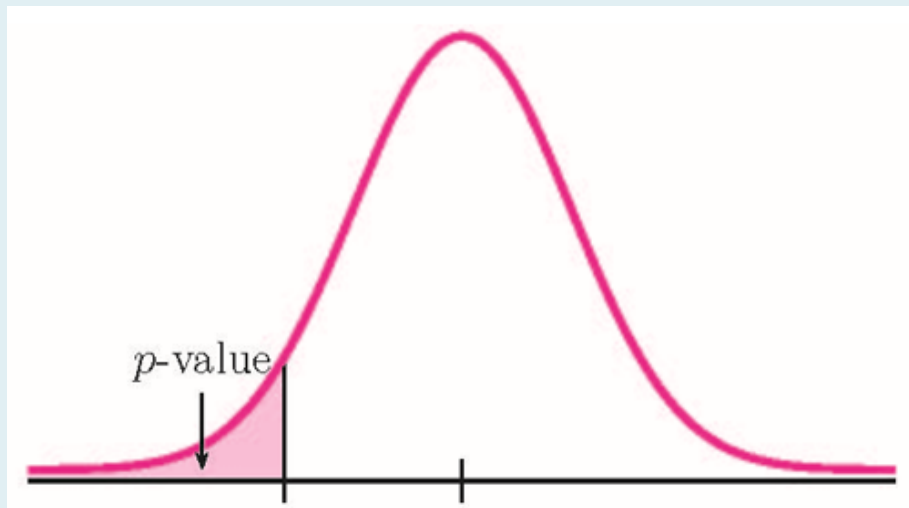
$$H_0 : \mu = 16.43 \text{ seconds}$$

$$H_a : \mu < 16.43 \text{ seconds}$$

***p*-value:**

From the question, we have $n = 15$, $\bar{x} = 16$, $\sigma = 0.8$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 0.8$). So we use a normal distribution to calculate the *p*-value. Because the alternative hypothesis is a $<$, the *p*-value is the area in the left-tail of the distribution.



Function	norm.dist	Answer
Field 1	16	0.0187
Field 2	16.43	
Field 3	0.8/sqrt(15)	
Field 4	true	

So the *p*-value = 0.0187.

Conclusion:

Because $p\text{-value} = 0.0187 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that Jeffrey's mean swim time with the goggles is less than 16.43 seconds.

NOTES

1. The null hypothesis $\mu = 16.43$ is the claim that Jeffrey's mean swim time with the goggles is 16.43 seconds (the same as it is without the goggles).
2. The alternative hypothesis $\mu < 16.43$ is the claim that Jeffrey's swim time with the goggles is less than 16.43 seconds.
3. The p -value is the area in the left tail of the sampling distribution, to the left of $\bar{x} = 16$. In the calculation of the p -value:
 - The function is norm.dist because we are finding the area in the left tail of a normal distribution.
 - Field 1 is the value of \bar{x}
 - Field 2 is the value of μ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu = 16.43$.
 - Field 3 is the standard deviation for the sample means $\frac{\sigma}{\sqrt{n}}$. Note that we are **not** using the standard deviation from the population ($\sigma = 0.8$). This is because the p -value is the area under the curve of the distribution of the sample means, not the distribution of the population.
4. The p -value of 0.0187 tells us that under the assumption that Jeffrey's mean swim time with goggles is 16.43 seconds (the null hypothesis), there is only a 1.87% chance that the mean time for the 15 sample swims is 16 seconds or less. This is a small probability, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis.
5. The Type I error for this problem is to conclude that Jeffrey swims the 25-meter freestyle, on average, in less than 16.43 seconds (the alternative hypothesis) when, in fact, he actually swims the 25-meter freestyle, on average, in 16.43 seconds (the null hypothesis). That is, reject the null hypothesis when the null hypothesis is actually true.
6. The Type II error for this problem is to conclude that Jeffrey swims the 25-meter freestyle, on

average, in 16.43 seconds (the null hypothesis) when, in fact, he actually swims the 25-meter freestyle, on average, in less than 16.43 seconds (the alternative hypothesis). That is, do not reject the null hypothesis when the null hypothesis is actually false.

TRY IT

The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards with a standard deviation of 2 yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws with the new grip. For the 20 throws, Marco's mean distance was 41.5 yards. The coach thought the different grip helped Marco throw farther than 40 yards. At the 5% significance level, is Marco's mean throwing distance higher with the new grip? Assume the throw distances for footballs are normally distributed.

Click to see Solution

Hypotheses:

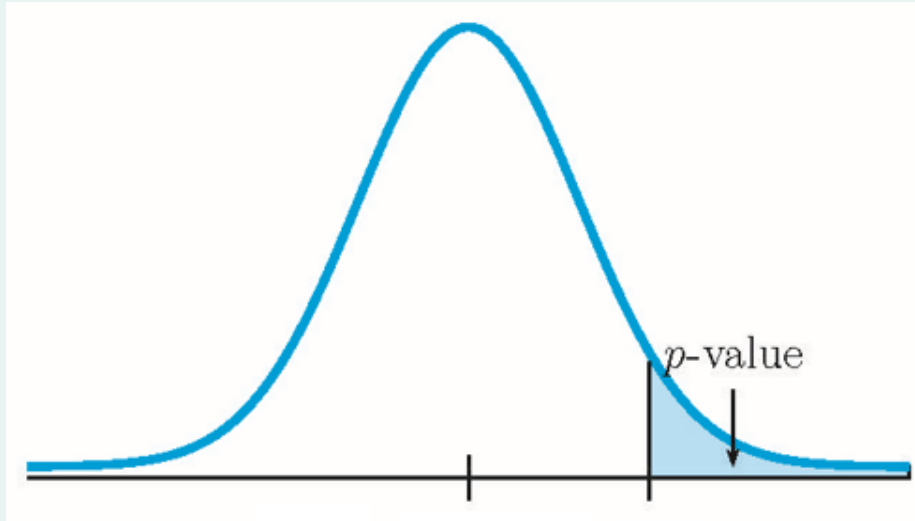
$$H_0 : \mu = 40 \text{ yards}$$

$$H_a : \mu > 40 \text{ yards}$$

p-value:

From the question, we have $n = 20$, $\bar{x} = 41.5$, $\sigma = 2$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 2$). So we use a normal distribution to calculate the p-value. Because the alternative hypothesis is a $>$, the p-value is the area in the right-tail of the distribution.



Function	1-norm.dist	Answer
Field 1	41.5	0.0004
Field 2	40	
Field 3	2/sqrt(20)	
Field 4	true	

So the $p\text{-value} = 0.0004$.

Conclusion:

Because $p\text{-value} = 0.0004 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that Marco's mean throwing distance is greater than 40 yards with the new grip.

NOTES

1. The null hypothesis $\mu = 40$ is the claim that Marco's mean throwing distance with the new grip is 40 yards (the same as it is without the new grip).
2. The alternative hypothesis $\mu > 40$ is the claim that Marco's mean throwing distance with the new grip is greater than 40 yards.
3. The $p\text{-value}$ is the area in the right tail of the normal distribution. To calculate the area in the

right-tail of a normal distribution, we use 1-norm.dist.

- Field 1 is the value of \bar{x}
- Field 2 is the value of μ from the null hypothesis.
- Field 3 is the standard deviation for the sample means $\frac{\sigma}{\sqrt{n}}$.

4. The p -value of 0.0004 tells us that under the assumption that Marco's mean throwing distance with the new grip is 40 yards, there is only a 0.047% chance that the mean throwing distance for the 20 sample throws is more than 40 yards. This is a small probability, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis.

EXAMPLE

A local college states in its marketing materials that the average age of its first-year students is 18.3 years with a standard deviation of 3.4 years. But this information is based on old data and does not take into account that more older adults are returning to college. A researcher at the college believes that the average age of its first-year students has changed. The researcher takes a sample of 50 first-year students and finds the average age is 19.5 years. At the 1% significance level, has the average age of the college's first-year students changed?

Solution:

Hypotheses:

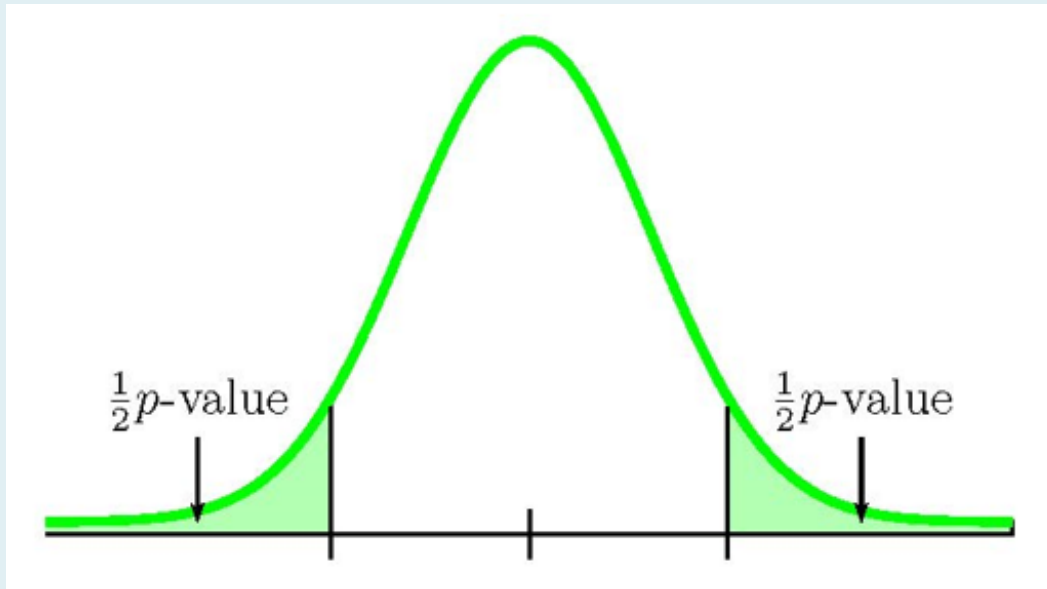
$$H_0 : \mu = 18.3 \text{ years}$$

$$H_a : \mu \neq 18.3 \text{ years}$$

***p*-value:**

From the question, we have $n = 50$, $\bar{x} = 19.5$, $\sigma = 3.4$ and $\alpha = 0.01$.

This is a test on a population mean where the population standard deviation is known ($\sigma = 3.4$). In this case, the sample size is greater than 30. So we use a normal distribution to calculate the *p*-value. Because the alternative hypothesis is a \neq , the *p*-value is the sum of area in the tails of the distribution.



Because there is only one sample, we only have information relating to one of the two tails, either the left tail or the right tail. We need to know if the sample relates to the left tail or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, the sample mean $\bar{x} = 19.5$ is greater than the value of the population mean in the null hypothesis $\mu = 18.3$ ($\bar{x} = 19.5 > 18.3 = \mu$), so the sample information relates to the right-tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the *p*-value is the sum of the area in the two tails and the area in the right-tail is only one half of the *p*-value. The area in the left tail equals the area in the right tail and the *p*-value is the sum of these two areas.

Function	1-norm.dist	Answer
Field 1	19.5	0.0063
Field 2	18.3	
Field 3	3.4/sqrt(50)	
Field 4	true	

So the area in the right tail is 0.0063 and $\frac{1}{2}(p\text{-value}) = 0.0063$. This is also the area in the left tail, so

$$p\text{-value} = 0.0063 + 0.0063 = 0.0126$$

Conclusion:

Because $p\text{-value} = 0.0126 > 0.01 = \alpha$, we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that the average age of the college's first-year students has changed.

NOTES

1. The null hypothesis $\mu = 18.3$ is the claim that the average age of the first-year students is still 18.3 years.
2. The alternative hypothesis $\mu \neq 18.3$ is the claim that the average age of the first-year students has changed from 18.3 years.
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.
 - If the sample mean \bar{x} is less than the population mean μ in the null hypothesis ($\bar{x} < \mu$), then the sample information belongs to the **left tail**.
 - We use **norm.dist($\bar{x}, \mu, \sigma/\text{sqrt}(n), \text{true}$)** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.

- If the sample mean \bar{x} is greater than the population mean μ in the null hypothesis ($\bar{x} > \mu$), then the sample information belongs to the **right tail**.
 - We use **`1-norm.dist($\bar{x}, \mu, \sigma/\text{sqrt}(n), \text{true})$`** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.
- 4. The p -value of 0.0126 is a large probability compared to the 1% significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that the average age of first-year students is 18.3 years is most likely correct.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=191#oembed-1>

Watch this video: Hypothesis Testing: z-test, right tail by ExcelIsFun [33:47]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=191#oembed-2>

Watch this video: Hypothesis Testing: z-test, left tail by ExcelIsFun [10:57]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=191#oembed-3>

Watch this video: Hypothesis Testing: z-test, two tail by ExcelIsFun [9:56]

Concept Review

The hypothesis test for a population mean is a well established process:

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
 2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
 3. Collect the sample information for the test and identify the significance level.
 4. When the population standard deviation is known, find the p -value (the area in the corresponding tail) for the test using the normal distribution.
 5. Compare the p -value to the significance level and state the outcome of the test.
 6. Write down a concluding sentence specific to the context of the question.
-

Attribution

“9.6 Hypothesis Testing of a Single Mean and Single Proportion“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.7 HYPOTHESIS TESTS FOR A POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population mean with unknown population standard deviation.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tell us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p -value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p -value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p -value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p -value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p -value.
- **Think about the meaning of the p -value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p -value (for example, 0.001 as opposed to 0.04) even if using a significance level of

0.05. Similarly, for a large p -value such as 0.4, as opposed to a p -value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more confidence that they made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5%.
- An alternative approach for hypothesis testing is to use what is called the **critical value approach**. In this book, we will only use the p -value approach. Some of the videos below may mention the critical value approach, but this approach will not be used in this book.

Steps to Conduct a Hypothesis Test for a Population Mean with Unknown Population Standard Deviation

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level α .
4. When the population standard deviation is **unknown**, the p -value is the area in the corresponding tail of the t -distribution with:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$df = n - 1$$

5. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON A POPULATION MEAN WITH UNKNOWN POPULATION STANDARD DEVIATION

The p -value for a hypothesis test on a population mean is the area in the tail(s) of the distribution of the sample mean. When the population standard deviation is unknown, use the t -distribution to find the p -value.

If the p -value is the area in the left-tail:

- Use the **t.dist** function to find the p -value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.
 - For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.
- The output from the **t.dist** function is the area under the t -distribution to the left of the entered t -score.
- Visit the Microsoft page for more information about the **t.dist** function.

If the p -value is the area in the right-tail:

- Use the **t.dist.rt** function to find the p -value. In the **t.dist.rt(t-score, degrees of freedom)** function:
 - For **t-score**, enter the value of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$.
 - For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.

- The output from the **t.dist.rt** function is the area under the t -distribution to the right of the entered t -score.
- Visit the Microsoft page for more information about the **t.dist.rt** function.

If the p -value is the sum of area in the tails:

- Use the **t.dist.2t** function to find the p -value. In the **t.dist.2t(t-score, degrees of freedom)** function:
 - For **t-score**, enter the **absolute value** of t calculated from $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. Note: In the **t.dist.2t** function, the value of the t -score must be a **positive** number. If the t -score is negative, enter the absolute value of the t -score into the **t.dist.2t** function.
 - For **degrees of freedom**, enter the degrees of freedom for the t -distribution $n - 1$.
- The output from the **t.dist.2t** function is the sum of areas in the tails under the t -distribution.
- Visit the Microsoft page for more information about the **t.dist.2t** function.

EXAMPLE

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the following scores:

65	67	66	68	72
65	70	63	63	71

The instructor performs a hypothesis test using a 1% level of significance. The test scores are assumed to be from a normal distribution.

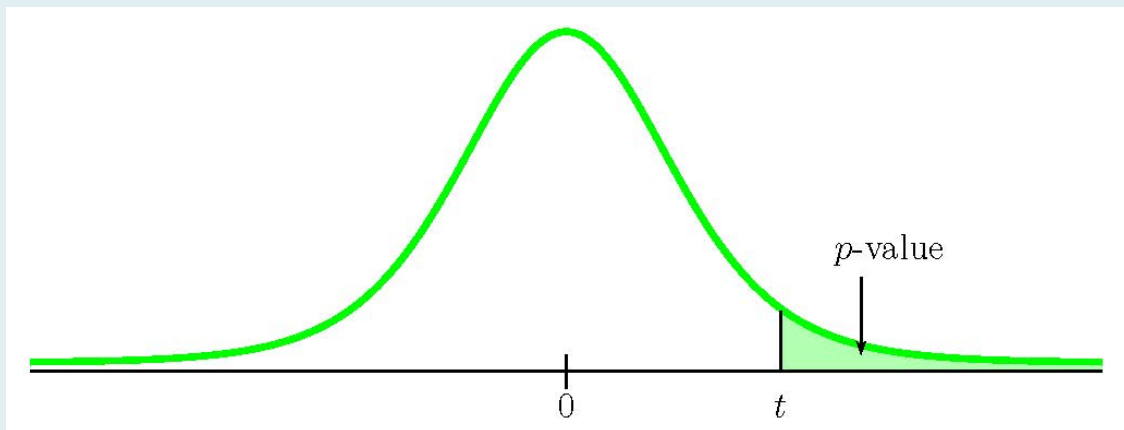
Solution:**Hypotheses:**

$$\begin{array}{l} H_0: \mu = 65 \\ H_a: \mu > 65 \end{array}$$

p-value:

From the question, we have $n = 10$, $\bar{x} = 67$, $s = 3.1972\dots$ and $\alpha = 0.01$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 3.1972\dots$). So we use a t -distribution to calculate the p -value. Because the alternative hypothesis is a $>$, the p -value is the area in the right-tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the t -score:

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{67 - 65}{\frac{3.1972\dots}{\sqrt{10}}} \\ &= 1.9781\dots \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 10 - 1 = 9$.

Function	t.dist.rt	Answer
Field 1	1.9781....	0.0396
Field 2	9	

So the p -value = 0.0396.

Conclusion:

Because p -value = 0.0396 $>$ 0.01 = α , we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that mean score on the test is greater than 65.

NOTES

1. The null hypothesis $\mu = 65$ is the claim that the mean test score is 65.
2. The alternative hypothesis $\mu > 65$ is the claim that the mean test score is greater than 65.
3. Keep all of the decimals throughout the calculation (i.e. in the sample standard deviation, the t -score, etc.) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value.
4. The p -value is the area in the right-tail of the t -distribution, to the right of $t = 1.9781\dots$.
5. The p -value of 0.0396 tells us that under the assumption that the mean test score is 65 (the null hypothesis), there is a 3.96% chance that the mean test score is 65 or more. Compared to the 1% significance level, this is a large probability, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis.

TRY IT

A company claims that the average change in the value of their stock is \$3.50 per week. An investor believes this average is too high. The investor records the changes in the company's stock price over 30 weeks and finds the average change in the stock price is \$2.60 with a standard deviation of \$1.80. At the 5% significance level, is the average change in the company's stock price lower than the company claims?

Click to see Solution

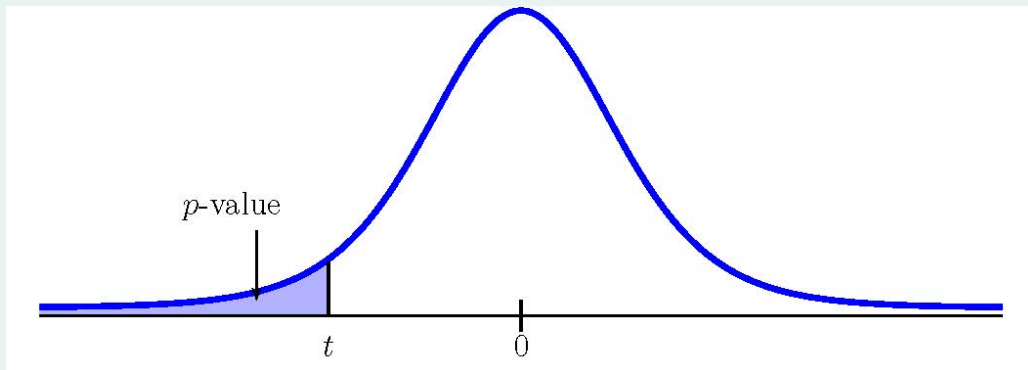
Hypotheses:

$$\begin{array}{l} H_0: \mu = 3.50 \\ H_a: \mu < 3.50 \end{array}$$

***p*-value:**

From the question, we have $n = 30$, $\bar{x} = 2.6$, $s = 1.8$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 1.8$). So we use a t -distribution to calculate the p -value. Because the alternative hypothesis is a $<$, the p -value is the area in the left-tail of the distribution.



To use the **t.dist** function, we need to calculate out the t -score:

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\ &= \frac{2.6 - 3.5}{\frac{1.8}{\sqrt{30}}} \\ &= -1.5699... \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 30 - 1 = 29$.

Function	t.dist	Answer
Field 1	-1.5699....	0.0636
Field 2	29	
Field 3	true	

So the p -value = 0.0636.

Conclusion:

Because p -value = 0.0636 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that average change in the stock price is lower than \$3.50.

NOTES

1. The null hypothesis $\mu = \$3.50$ is the claim that the average change in the company's stock is \$3.50 per week.
2. The alternative hypothesis $\mu < \$3.50$ is the claim that the average change in the company's stock is less than \$3.50 per week.
3. The p -value is the area in the left-tail of the t -distribution, to the left of $t = -1.5699\dots$
4. The p -value of 0.0636 tells us that under the assumption that the average change in the stock is \$3.50 (the null hypothesis), there is a 6.36% chance that the average change is \$3.50 or less. Compared to the 5% significance level, this is a large probability, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the company's claim that the average change in their stock price is \$3.50 per week is most likely correct.

EXAMPLE

A paint manufacturer has their production line set-up so that the average volume of paint in a can is 3.78 liters. The quality control manager at the plant believes that something has happened with the production and the average volume of paint in the cans has changed. The quality control department takes a sample of 100 cans and finds the average volume is 3.62 liters with a standard deviation of 0.7 liters. At the 5% significance level, has the volume of paint in a can changed?

Solution:

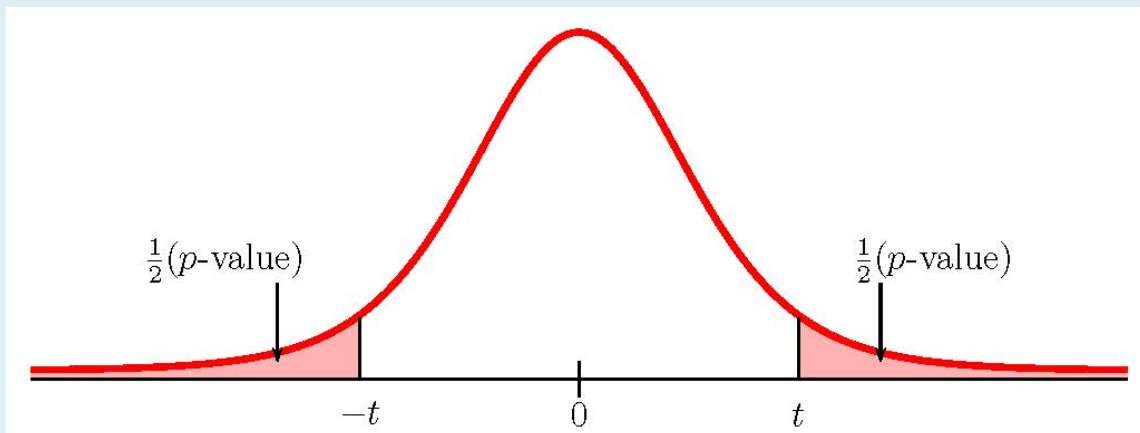
Hypotheses:

$$\begin{array}{l} H_0: \mu = 3.78 \text{ liters} \\ H_a: \mu \neq 3.78 \text{ liters} \end{array}$$

***p*-value:**

From the question, we have $n = 100$, $\bar{x} = 3.62$, $s = 0.7$ and $\alpha = 0.05$.

This is a test on a population mean where the population standard deviation is unknown (we only know the sample standard deviation $s = 0.7$). So we use a t -distribution to calculate the p -value. Because the alternative hypothesis is a \neq , the p -value is the sum of area in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the t -score:

$$\begin{aligned}
 t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\
 &= \frac{3.62 - 3.78}{\frac{0.07}{\sqrt{100}}} \\
 &= -2.2857\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n - 1 = 100 - 1 = 99$.

Function	t.dist.2t	Answer
Field 1	2.2857....	0.0244
Field 2	99	

So the p -value = 0.0244.

Conclusion:

Because p -value = 0.0244 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that average volume of paint in the cans has changed.

NOTES

1. The null hypothesis $\mu = 3.78$ is the claim that the average volume of paint in the cans is 3.78.
2. The alternative hypothesis $\mu \neq 3.78$ is the claim that the average volume of paint in the cans is not 3.78.
3. Keep all of the decimals throughout the calculation (i.e. in the t -score) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value.
4. The p -value is the sum of the area in the two tails. The output from the **t.dist.2t** function is exactly the sum of the area in the two tails, and so is the p -value required for the test. No additional calculations are required.
5. The **t.dist.2t** function requires that the value entered for the t -score is **positive**. A negative t -score entered into the **t.dist.2t** function generates an error in Excel. In this case, the value of

the t -score is negative, so we must enter the absolute value of this t -score into field 1.

6. The p -value of 0.0244 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the average volume of paint in the cans has most likely changed from 3.78 liters.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=196#oembed-1>

Watch this video: Hypothesis Testing: t -test, right tail by ExcellIsFun [11:02]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=196#oembed-2>

Watch this video: Hypothesis Testing: t -test, left tail by ExcellIsFun [7:48]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=196#oembed-3>

Watch this video: Hypothesis Testing: t -test, two tail by ExcellIsFun [8:54]

Concept Review

The hypothesis test for a population mean is a well established process:

1. Write down the null and alternative hypotheses in terms of the population mean μ . Include appropriate units with the values of the mean.
 2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
 3. Collect the sample information for the test and identify the significance level.
 4. When the population standard deviation is unknown, find the p -value (the area in the corresponding tail) for the test using the t -distribution with $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ and $df = n - 1$.
 5. Compare the p -value to the significance level and state the outcome of the test.
 6. Write down a concluding sentence specific to the context of the question.
-

Attribution

“9.6 Hypothesis Testing of a Single Mean and Single Proportion“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.8 HYPOTHESIS TESTS FOR A POPULATION PROPORTION

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for a population proportion.

Some notes about conducting a hypothesis test:

- The null hypothesis H_0 is always an “equal to.” The null hypothesis is the original claim about the population parameter.
- The alternative hypothesis H_a is a “less than,” “greater than,” or “not equal to.” The form of the alternative hypothesis depends on the context of the question.
- The form of the alternative hypothesis tell us if the test is left-tail, right-tail, or two-tail. The alternative hypothesis is the key to conducting the test and finding the correct p -value.
 - If the alternative hypothesis is a “less than”, then the test is left-tail. The p -value is the area in the left-tail of the distribution.
 - If the alternative hypothesis is a “greater than”, then the test is right-tail. The p -value is the area in the right-tail of the distribution.
 - If the alternative hypothesis is a “not equal to”, then the test is two-tail. The p -value is the sum of the area in the two-tails of the distribution. Each tail represents exactly half of the p -value.
- **Think about the meaning of the p -value.** A data analyst (and anyone else) should have more confidence that they made the correct decision to reject the null hypothesis with a smaller p -value (for example, 0.001 as opposed to 0.04) even if using a significance level of 0.05. Similarly, for a large p -value such as 0.4, as opposed to a p -value of 0.056 (a significance level of 0.05 is less than either number), a data analyst should have more confidence that they

made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

- The significance level must be identified before collecting the sample data and conducting the test. Generally, the significance level will be included in the question. If no significance level is given, a common standard is to use a significance level of 5%.

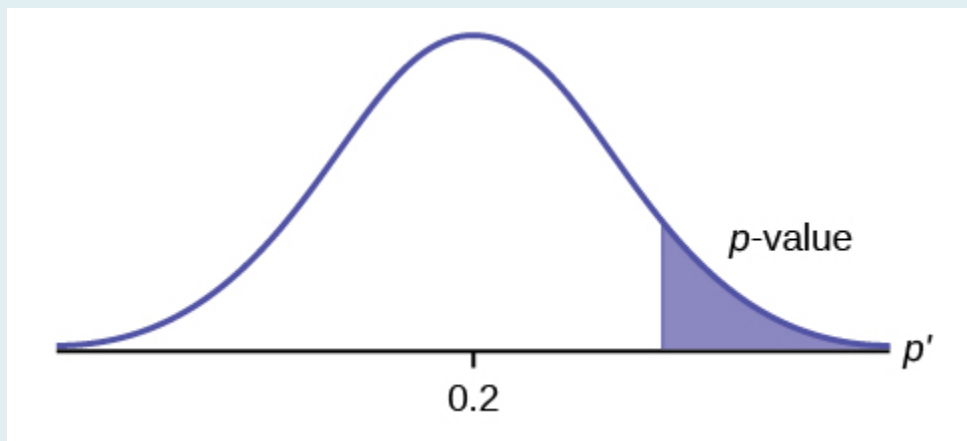
EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$H_0 : p = 20\%$$

$$H_a : p > 20\%$$

Because the alternative hypothesis is a $>$, this is a right-tail test. The p -value is the area in the right-tail of the distribution.

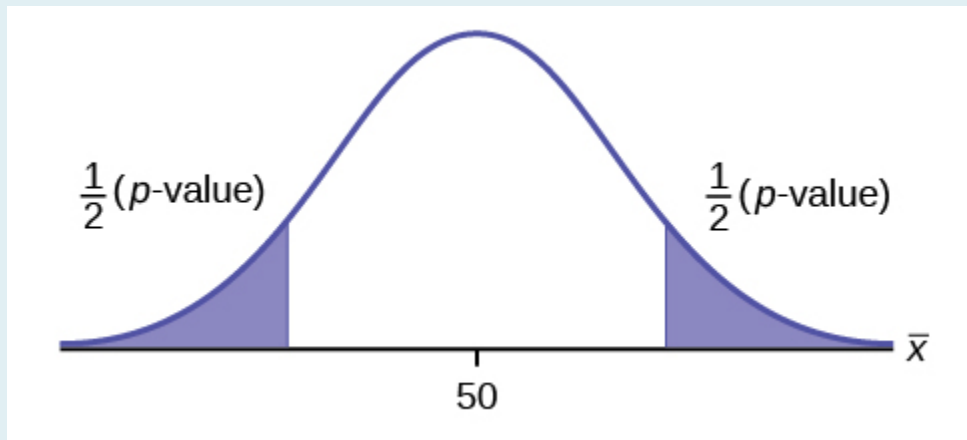


EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: \mu = 50 \\ H_a: \mu \neq 50 \end{array}$$

Because the alternative hypothesis is a \neq , this is a two-tail test. The p -value is the sum of the areas in the two tails of the distribution. Each tail contains exactly half of the p -value.



EXAMPLE

Suppose the hypotheses for a hypothesis test are:

$$\begin{array}{l} H_0: \mu = 10 \\ H_a: \mu < 10 \end{array}$$

Because the alternative hypothesis is a $<$, this is a left-tail test. The p -value is the area in the left-tail of the distribution.

Steps to Conduct a Hypothesis Test for a Population Proportion

1. Write down the null and alternative hypotheses in terms of the population proportion p . Include appropriate units with the values of the proportion.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the appropriate distribution:

- If $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, use the normal distribution with $z = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}}$.
- If one of $n \times p < 5$ or $n \times (1 - p) < 5$, use a binomial distribution.

5. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
6. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON A POPULATION PROPORTION

The p -value for a hypothesis test on a population proportion is the area in the tail(s) of distribution of

the sample proportion. If both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$, use the normal distribution to find the p -value. If at least one of $n \times p < 5$ or $n \times (1 - p) < 5$, use the binomial distribution to find the p -value.

If both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$:

- The p -value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p -value.
 - For **x**, enter the value for \hat{p} .
 - For **μ**, enter the mean of the sample proportions p . Note: Because the test is run assuming the null hypothesis is true, the value for p is the claim from the null hypothesis.
 - For **σ**, enter the standard error of the proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.
- Use the appropriate technique with the **norm.dist** function to find the area in the left-tail or the area in the right-tail.

If at least one of $n \times p < 5$ or $n \times (1 - p) < 5$:

- The p -value is found using the binomial distribution.
- If the alternative hypothesis is a $<$, the p -value is the probability of getting at most x successes in n trials where the probability of success is the claim about the population proportion p in the null hypothesis.
 - The p -value is the output from the **binom.dist(x,n,p,logic operator)** function:
 - For **x**, enter the number of successes.
 - For **n**, enter the sample size.
 - For **p**, enter the value of the population proportion p from the null hypothesis.
 - For the **logic operator**, enter **true**. Note: Because we are calculating an at most probability, the logic operator is always true.
- If the alternative hypothesis is a $>$, the p -value is the probability of getting at least x successes in n trials where the probability of success is the claim about the population proportion p in

the null hypothesis.

- The p -value is the output from the **1-binom.dist(x-1,n,p,logic operator)** function:
 - For **x**, enter the number of successes.
 - For **n**, enter the sample size.
 - For **p**, enter the value of the population proportion p in the null hypothesis.
 - For the **logic operator**, enter **true**. Note: Because we are calculating an at least probability, the logic operator is always true.

EXAMPLE

Marketers believe that 92% of adults own a cell phone. A cell phone manufacturer believes that number is actually lower. In a sample of 200 adults, 87% own a cell phone. At the 1% significance level, determine if the proportion of adults that own a cell phone is lower than the marketers' claim.

Solution:

Hypotheses:

$$H_0 : p = 92\% \text{ of adults own a cell phone}$$

$$H_a : p < 92\% \text{ of adults own a cell phone}$$

p -value:

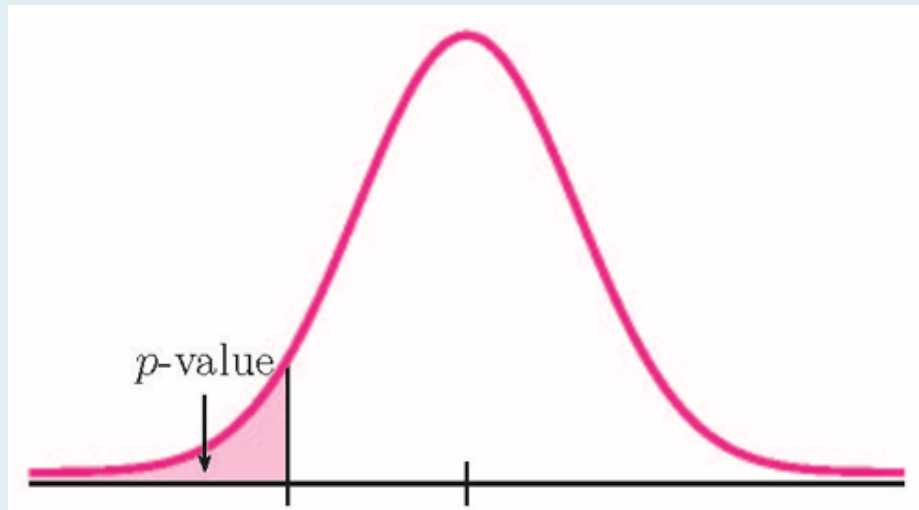
From the question, we have $n = 200$, $\hat{p} = 0.87$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.92$).

$$\begin{aligned} n \times p &= 200 \times 0.92 = 184 \geq 5 \\ n \times (1 - p) &= 200 \times (1 - 0.92) = 16 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ we use a normal distribution to calculate the

p -value. Because the alternative hypothesis is a $<$, the p -value is the area in the left tail of the distribution.



Function	norm.dist	Answer
Field 1	0.87	0.0046
Field 2	0.92	
Field 3	sqrt(0.92*(1-0.92)/200)	
Field 4	true	

So the p -value = 0.0046.

Conclusion:

Because p -value = 0.0046 $<$ 0.01 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that the proportion of adults who own a cell phone is lower than 92%.

NOTES

1. The null hypothesis $p = 92\%$ is the claim that 92% of adults own a cell phone.
2. The alternative hypothesis $p < 92\%$ is the claim that less than 92% of adults own a cell phone.

3. The p -value is the area in the left tail of the sampling distribution, to the left of $\hat{p} = 0.87$.

In the calculation of the p -value:

- The function is `norm.dist` because we are finding the area in the left tail of a normal distribution.
- Field 1 is the value of \hat{p} .
- Field 2 is the value of p from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p = 0.92$.
- Field 3 is the standard deviation for the sample proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.

4. The p -value of 0.0046 tells us that under the assumption that 92% of adults own a cell phone (the null hypothesis), there is only a 0.46% chance that the proportion of adults who own a cell phone in a sample of 200 is 87% or less. This is a small probability, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of adults who own a cell phone is most likely less than 92%.

EXAMPLE

A consumer group claims that the proportion of households that have at least three cell phones is 30%. A cell phone company has reason to believe that the proportion of households with at least three cell phones is much higher. Before they start a big advertising campaign based on the proportion of households that have at least three cell phones, they want to test their claim. Their marketing people survey 150 households with the result that 54 of the households have at least three

cell phones. At the 1% significance level, determine if the proportion of households that have at least three cell phones is less than 30%.

Solution:

Hypotheses:

$H_0 : p = 30\%$ of household have at least 3 cell phones

$H_a : p > 30\%$ of household have at least 3 cell phones

p-value:

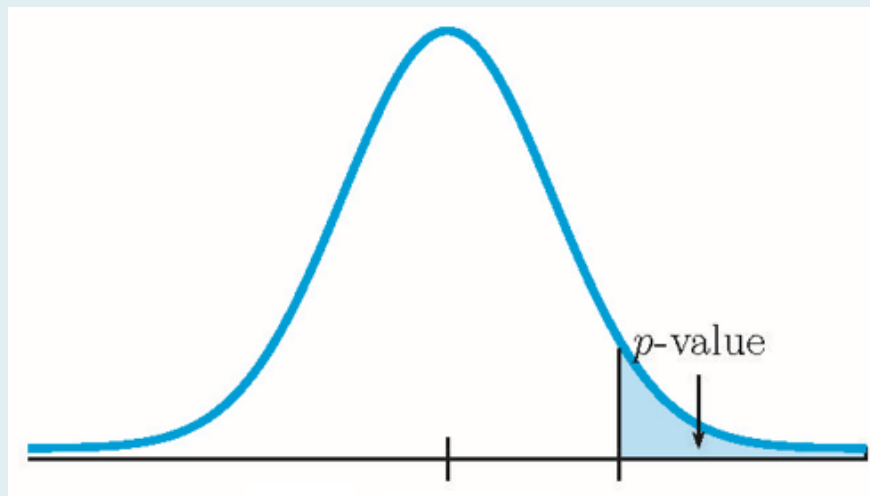
From the question, we have $n = 150$, $\hat{p} = \frac{54}{150} = 0.36$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.3$).

$$n \times p = 150 \times 0.3 = 45 \geq 5$$

$$n \times (1 - p) = 150 \times (1 - 0.3) = 105 \geq 5$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ we use a normal distribution to calculate the p-value. Because the alternative hypothesis is a $>$, the p-value is the area in the right tail of the distribution.



Function	1-norm.dist	Answer
Field 1	0.36	0.0544
Field 2	0.3	
Field 3	$\text{sqrt}(0.3*(1-0.3)/150)$	
Field 4	true	

So the p -value = 0.0544.

Conclusion:

Because p -value = 0.0544 > 0.01 = α , we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that the proportion of households with at least three cell phones is more than 30%.

NOTES

- The null hypothesis $p = 30\%$ is the claim that 30% of households have at least three cell phones.
- The alternative hypothesis $p > 30\%$ is the claim that more than 30% of households have at least three cell phones.
- The p -value is the area in the right tail of the sampling distribution, to the right of $\hat{p} = 0.36$. In the calculation of the p -value:
 - The function is 1-norm.dist because we are finding the area in the right tail of a normal distribution.
 - Field 1 is the value of \hat{p} .
 - Field 2 is the value of p from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p = 0.3$.
 - Field 3 is the standard deviation for the sample proportions $\sqrt{\frac{p \times (1 - p)}{n}}$.
- The p -value of 0.0544 tells us that under the assumption that 30% of households have at least three cell phones (the null hypothesis), there is a 5.44% chance that the proportion of households with at least three cell phones in a sample of 150 is 36% or more. Compared to the 1% significance level, this is a large probability, and so is likely to happen assuming the

null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that 30% of households have at least three cell phones is most likely correct.

TRY IT

A teacher believes that 70% of students in the class will want to go on a field trip to the local zoo. The students in the class believe the proportion is much higher and ask the teacher to verify her claim. The teacher samples 50 students and 39 reply that they would want to go to the zoo. At the 5% significance level, determine if the proportion of students who want to go on the field trip is higher than 70%.

Click to see Solution

Hypotheses:

$$\begin{array}{l} H_0: \text{ } \& \& p = 70\% \text{ \mbox{ of students want to go on the field trip} } \\ H_a: \text{ } \& \& p > 70\% \text{ \mbox{ of students want to go on the field trip} } \end{array}$$

***p*-value:**

From the question, we have $n = 50$, $\hat{p} = \frac{39}{50} = 0.78$, and $\alpha = 0.05$.

$$\begin{aligned} n \times p &= 50 \times 0.7 = 35 \geq 5 \\ n \times (1 - p) &= 50 \times (1 - 0.7) = 15 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1-p) \geq 5$ we use a normal distribution to calculate the p -value. Because the alternative hypothesis is a $>$, the p -value is the area in the right tail of the distribution.

Function	1-norm.dist	Answer
Field 1	0.78	0.1085
Field 2	0.7	
Field 3	$\text{sqrt}(0.7*(1-0.7)/50)$	
Field 4	true	

So the p -value = 0.1085.

Conclusion:

Because p -value = 0.1085 $>$ 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that the proportion of students who want to go on the field trip is higher than 70%.

NOTES

1. The null hypothesis $p = 70\%$ is the claim that 70% of the students want to go on the field trip.
2. The alternative hypothesis $p > 70\%$ is the claim that more than 70% of students want to go on the field trip.
3. The p -value of 0.1085 tells us that under the assumption that 70% of students want to go on the field trip (the null hypothesis), there is a 10.85% chance that the proportion of students who want to go on the field trip in a sample of 50 students is 78% or more. Compared to the 5% significance level, this is a large probability, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the teacher's claim that 70% of students want to go on the field trip is most likely correct.

EXAMPLE

Joan believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joan samples 100 first-time brides and 56 reply that they are younger than their grooms. Use a 5% significance level.

Solution:

Hypotheses:

$H_0 : p = 50\%$ of first-time brides are younger than the groom

$H_a : p \neq 50\%$ of first-time brides are younger than the groom

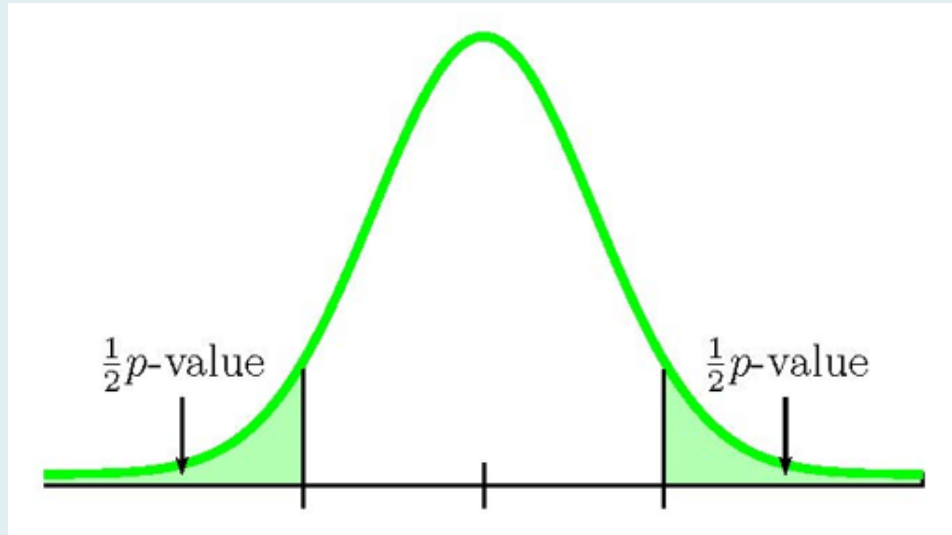
p-value:

From the question, we have $n = 100$, $\hat{p} = \frac{56}{100} = 0.56$, and $\alpha = 0.05$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.5$).

$$\begin{aligned} n \times p &= 100 \times 0.5 = 50 \geq 5 \\ n \times (1 - p) &= 100 \times (1 - 0.5) = 50 \geq 5 \end{aligned}$$

Because both $n \times p \geq 5$ and $n \times (1 - p) \geq 5$ we use a normal distribution to calculate the p-value. Because the alternative hypothesis is a \neq , the p-value is the sum of area in the tails of the distribution.



Because there is only one sample, we only have information relating to one of the two tails, either the left or the right. We need to know if the sample relates to the left or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, the sample proportion $\hat{p} = 0.56$ is greater than the value of the population proportion in the null hypothesis $p = 0.5$ ($\hat{p} = 0.56 > 0.5 = p$), so the sample information relates to the right-tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the p -value is the sum of the area in the two tails and the area in the right-tail is only one half of the p -value. The area in the left tail equals the area in the right tail and the p -value is the sum of these two areas.

Function	1-norm.dist	Answer
Field 1	0.56	0.1151
Field 2	0.5	
Field 3	sqrt(0.5*(1-0.5)/100)	
Field 4	true	

So the area in the right tail is 0.1151 and $\frac{1}{2}(p\text{-value}) = 0.1151$. This is also the area in the left tail,

so

$$p\text{-value} = 0.1151 + 0.1151 = 0.2302$$

Conclusion:

Because $p\text{-value} = 0.2302 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5%

significance level there is not enough evidence to suggest that the proportion of first-time brides that are younger than the groom is different from 50%.

NOTES

1. The null hypothesis $p = 50\%$ is the claim that the proportion of first-time brides that are younger than the groom is 50%.
2. The alternative hypothesis $p \neq 50\%$ is the claim that the proportion of first-time brides that are younger than the groom is different from 50%.
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.
 - If the sample proportion \hat{p} is less than the population proportion p in the null hypothesis ($\hat{p} < p$), the sample information belongs to the **left tail**.
 - We use **norm.dist($\hat{p}, p, \text{sqrt}(p * (1 - p)/n), \text{true}$)** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.
 - If the sample proportion \hat{p} is greater than the population proportion p in the null hypothesis ($\hat{p} > p$), the sample information belongs to the **right tail**.
 - We use **1-norm.dist($\hat{p}, p, \text{sqrt}(p * (1 - p)/n), \text{true}$)** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.
4. The p -value of 0.2302 is a large probability compared to the 5% significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the claim that the proportion of first-time brides who are younger than the groom is most likely correct.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=200#oembed-1>

Watch this video: Hypothesis Testing for Proportions: z-test by ExcelIsFun [7:27]

EXAMPLE

An online retailer believes that 93% of the visitors to its website will make a purchase. A researcher in the marketing department thinks the actual percent is lower than claimed. The researcher examines a sample of 50 visits to the website and finds that 45 of the visits resulted in a purchase. At the 1% significance level, determine if the proportion of visits to the website that result in a purchase is lower than claimed.

Solution:

Hypotheses:

$$H_0 : p = 93\% \text{ of visitors make a purchase}$$

$$H_a : p < 93\% \text{ of visitors make a purchase}$$

p-value:

From the question, we have $n = 50$, $x = 45$, and $\alpha = 0.01$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.93$).

$$\begin{aligned} n \times p &= 50 \times 0.93 = 46.5 \geq 5 \\ n \times (1 - p) &= 50 \times (1 - 0.93) = 3.5 < 5 \end{aligned}$$

Because $n \times (1 - p) < 5$ we use a binomial distribution to calculate the p -value. Because the alternative hypothesis is a $<$, the p -value is the probability of getting at most 45 successes in 50 trials.

Function	binom.dist	Answer
Field 1	45	0.2710
Field 2	50	
Field 3	0.93	
Field 4	true	

So the p -value = 0.2710.

Conclusion:

Because p -value = 0.2710 > 0.01 = α , we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that the proportion of visitors who make a purchase is lower than 93%.

NOTES

1. The null hypothesis $p = 93\%$ is the claim that 93% of visitors to the website make a purchase.
2. The alternative hypothesis $p < 93\%$ is the claim that less than 93% of visitors to the website make a purchase.
3. The p -value is the binomial probability of getting at most 45 successes (the number in the sample with the characteristic of interest) in 50 trials (the sample size) with a probability of success of 93% (the value of p in the null hypothesis). In the calculation of the p -value:
 - The function is binom.dist because we are finding the probability of at most 45 successes.
 - Field 1 is the number of successes x .
 - Field 2 is the sample size n .
 - Field 3 is the probability of success p . This is the claim about the population proportion made in the null hypothesis, so that means we assume $p = 0.93$.
4. The p -value of 0.2710 tells us that under the assumption that 93% of visitors make a purchase (the null hypothesis), there is a 27.10% chance that the number of visitors in a sample of 50 who make a purchase is 45 or less. This is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the

assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the proportion of visitors to the website who make a purchase adults is most likely 93%.

EXAMPLE

A drug company claims that only 4% of people who take their new drug experience any side effects from the drug. A researcher believes that the percent is higher than drug company's claim. The researcher takes a sample of 80 people who take the drug and finds that 10% of the people in the sample experience side effects from the drug. At the 5% significance level, determine if the proportion of people who experience side effects from taking the drug is higher than claimed.

Solution:

Hypotheses:

$H_0 : p = 4\%$ of people experience side effects

$H_a : p > 4\%$ of people experience side effects

p -value:

From the question, we have $n = 80$, $\hat{p} = 0.1$, and $\alpha = 0.05$.

To determine the distribution, we check $n \times p$ and $n \times (1 - p)$. For the value of p , we use the claim from the null hypothesis ($p = 0.04$).

$$n \times p = 80 \times 0.04 = 3.2 < 5$$

Because $n \times p < 5$ we use a binomial distribution to calculate the p -value. Because the alternative hypothesis is a $>$, the p -value is the probability of getting at least 8 successes in 80 trials.

(Note: In the sample of size 80, 10% have the characteristic of interest, so this means that $80 \times 0.1 = 8$ people in the sample have the characteristic of interest.)

Function	1-binom.dist	Answer
Field 1	7	0.0147
Field 2	80	
Field 3	0.04	
Field 4	true	

So the p -value = 0.0147.

Conclusion:

Because p -value = 0.0147 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the proportion of people who experience side effects from taking the drug is higher than 4%.

NOTES

1. The null hypothesis $p = 4\%$ is the claim that 4% of the people experience side effects from taking the drug.
2. The alternative hypothesis $p > 4\%$ is the claim that more than 4% of the people experience side effects from taking the drug.
3. The p -value is the binomial probability of getting at least 8 successes (the number in the sample with the characteristic of interest) in 80 trials (the sample size) with a probability of success of 4% (the value of p in the null hypothesis). In the calculation of the p -value:
 - The function is 1-binom.dist because we are finding the probability of at least 8 successes.
 - Field 1 is $\chi - 1$ where χ is the number of successes. In this case, we are using the compliment rule to change the probability of at least 8 successes into 1 minus the probability of at most 7 successes.
 - Field 2 is the sample size n .
 - Field 3 is the probability of success p . This is the claim about the population proportion made in the null hypothesis, so that means we assume $p = 0.04$.

4. The p -value of 0.0147 tells us that under the assumption that 4% of people experience side effects (the null hypothesis), there is a 1.47% chance that the number of people in a sample of 80 who experience side effects is 8 or more. This is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of people who experience side effects is most likely greater than 4%.

Concept Review

The hypothesis test for a population proportion is a well-established process:

1. Write down the null and alternative hypotheses in terms of the population proportion p . Include appropriate units with the values of the proportion.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the appropriate distribution (normal or binomial).
5. Compare the p -value to the significance level and state the outcome of the test.
6. Write down a concluding sentence specific to the context of the question.

Attribution

“9.6 Hypothesis Testing of a Single Mean and Single Proportion“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

8.9 EXERCISES

1. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.
2. The mean entry level salary of an employee at a company is \$58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.
3. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.
4. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.
5. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.
6. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?
7. The National Institute of Mental Health published an article stating that in any one-year period,

approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

8. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. State the null and alternative hypotheses.

9. The mean price of mid-sized cars in a region is \$32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

10. A sleeping bag is tested to withstand temperatures of -15°F . You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

11. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

12. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis is: the surgical procedure will go well. Which is the error with the greater consequence?

13. A group of divers is exploring an old sunken ship. Suppose the null hypothesis is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

14. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis is: the sample contains E-coli. Which is the error with the greater consequence?

15. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II error?
16. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. What is the Type I error?
17. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. What is the Type I error?
18. Which distributions can you use for hypothesis testing for this chapter?
19. Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.
20. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.
21. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.
22. A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?
23. It is thought that 42% of respondents in a taste test would prefer Brand A. In a particular test of 100 people, 39% preferred Brand A. What distribution should you use to perform a hypothesis test?

24. You are performing a hypothesis test of a single population mean using a Student's t -distribution. What must you assume about the distribution of the data?
25. You are performing a hypothesis test of a single population mean using a Student's t -distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?
26. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of $n \times p$ and $n \times (1 - p)$ in order to use the normal distribution?
27. You are performing a hypothesis test of a single population proportion. You find out that $n \times p$ is less than five. What must you do to be able to perform a valid hypothesis test?
28. When do you reject the null hypothesis?
29. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?
30. The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?
31. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p -value is almost zero. State the null and alternative hypotheses and interpret the p -value.
32. The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p -value is 0.0264. State the null and alternative hypotheses and interpret the p -value.

33. What should you do when $\alpha > p$ -value?

34. What should you do if $\alpha = p$ -value?

35. If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

36. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

- a. Is this a test of means or proportions?
- b. Is the population standard deviation known and, if so, what is it?
- c. Because both σ and s are known, which should be used? Why?
- d. State the distribution to use for the hypothesis test.

37. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

- a. Is this a test of one mean or proportion?
- b. State the null and alternative hypotheses.
- c. Is this a right-tailed, left-tailed, or two-tailed test?
- d. Is the population standard deviation known and, if so, what is it?
- e. State the distribution to use for the hypothesis test.
- f. Find the p -value.
- g. At a significance level of 5%, what is your:
 - i. Decision:
 - ii. Reason for the decision:

iii. Conclusion (write out in a complete sentence):

38. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

- a. Is this a test of one mean or proportion?
- b. State the null and alternative hypotheses.
- c. Is this a right-tailed, left-tailed, or two-tailed test?
- d. State the distribution to use for the hypothesis test.
- e. Find the p -value.
- f. At a significance level of 5%, what is your:
 - i. Decision:
 - ii. Reason for the decision:
 - iii. Conclusion (write out in a complete sentence):

39. Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

40. Assume $H_0 : \mu = 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

41. Assume $H_0 : p = 25\%$ and $H_a : p \neq 25\%$. Is this a left-tailed, right-tailed, or two-tailed test?

42. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

43. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

44. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

45. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?
46. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?
47. A particular brand of tires claims that its deluxe tire averages 50,000 miles before it needs to be replaced. A group of owners believe this number is too high. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. At the 5% significance level, is the data highly inconsistent with the claim?
48. From generation to generation, the mean age when smokers first start to smoke is 19 years. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Does the data support the claim at the 5% level?
49. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is \$1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Does the data support the claim at the 1% level?
50. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Does the data support your claim at the 1% level?
51. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The

number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. At the 5% significance level, should the personnel team believe that the mean number is ten?

52. In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

53. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think? Use a 5% significance level.

54. A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief. Use a 5% significance level.

55. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours? Use a 5% significance level.

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

56. Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a

hypothesis test to determine if the percent at her school is less than 40%. Use a 1% significance level.

57. According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased. Use a 1% significance level.

58. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test. Use a 5% significance level.

59. Registered nurses earned an average annual salary of \$69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than \$69,110 for California nurses. The sample average was \$71,121 with a sample standard deviation of \$7,489. Conduct a hypothesis test. Use a 5% significance level.

60. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. Use a 5% significance level.

61. According to the Center for Disease Control website, in 2011 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized—approximately 1,200 students—small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use

a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

62. A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

63. Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

64. For Americans using library services, the American Library Association claims that 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

65. The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

66. A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Attribution

“Chapter 9 Homework” and “Chapter 9 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART IX

STATISTICAL INFERENCE FOR TWO POPULATIONS

Chapter Outline

- 9.1 Introduction to Statistical Inference with Two Populations
- 9.2 Statistical Inference for Two Population Means with Known Population Standard Deviation
- 9.3 Statistical Inference for Two Population Means with Unknown Population Standard Deviation
- 9.4 Statistical Inference for Matched Samples
- 9.5 Statistical Inference for Two Population Proportions
- 9.6 Exercises

9.1 INTRODUCTION TO STATISTICAL INFERENCE WITH TWO POPULATIONS



If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. Photo by Chloe Lim, CC BY 4.0.

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their test scores.

Previously, we learned to conduct confidence intervals and hypothesis tests on single means and single proportions. We will extend these ideas in this chapter so that we can compare two means or two proportions to each other. The general procedures are similar to any confidence or hypothesis test, following the same basic steps we have already learned, just expanded to include the cases of studying two population parameters.

To compare two means or two proportions, we work with two populations. The groups are classified either as **independent** or **matched** pairs. Independent groups consist of two samples that are independent. That is, one population is independent of the other if the sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. That is, there is some relationship between the samples selected from the two populations. In this book, independent groups are used for either two population means or two population proportions and matched pairs are for two population means.

Attribution

“Chapter 10 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

9.2 STATISTICAL INFERENCE FOR TWO POPULATION MEANS WITH KNOWN POPULATION STANDARD DEVIATIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population means with known population standard deviations.
- Conduct and interpret hypothesis tests for two population means with known population standard deviations.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference in the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**. In the case where the samples are independent, we also have to contend with whether or not we know the population standard deviations.

Two populations are **independent** if the sample taken from population 1 is not related in anyway to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the two populations:

Symbol for:	Population 1	Population 2
Population Mean	μ_1	μ_2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Mean	\bar{x}_1	\bar{x}_2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two population means ($\mu_1 - \mu_2$), we need to use the distribution of the difference in the sample means $\bar{x}_1 - \bar{x}_2$:

- The mean of the distribution of the difference in the sample means is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The standard deviation of the distribution of the difference in the sample means is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- The distribution of the difference in the sample means is normal if **one** of the following is true:
 - Both populations are normally distributed.
 - The sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- Assuming the distribution of the difference of the sample means is normal, the z -score is

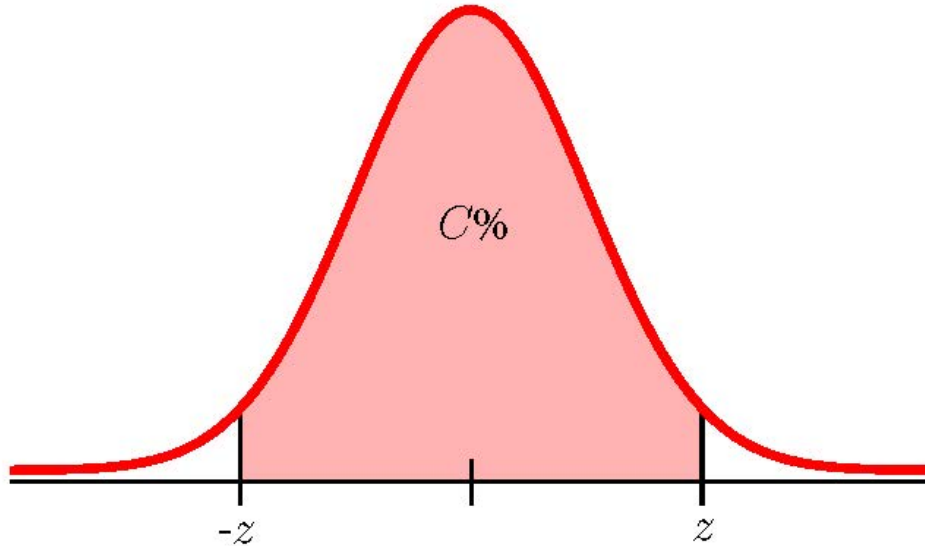
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Constructing a Confidence Interval for the Difference in Two Population Means with Known Population Standard Deviation

Suppose a sample of size n_1 with sample mean \bar{x}_1 is taken from population 1 and a sample of size n_2 with sample mean \bar{x}_2 is taken from population 2 where the populations are independent and the population standard deviations, σ_1 and σ_2 , are **known**. The limits for the confidence interval with confidence level C for the difference in the population means $\mu_1 - \mu_2$ are:

$$\begin{aligned} \text{Lower Limit} &= \overline{x}_1 - \overline{x}_2 - z \times \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \\ \text{Upper Limit} &= \overline{x}_1 - \overline{x}_2 + z \times \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \end{aligned}$$

where z is the positive z -score of the standard normal distribution so that the area under the curve in between $-z$ and z is $C\%$.



NOTE

In order to construct the confidence interval for the difference in two population means with independent samples, we need to check that the distribution of the difference in the sample means follows a normal distribution. This means that we need to check that either the populations are normal or that the sample sizes are large enough (greater than or equal to 30).

CALCULATING THE **Formula does not parse**-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z** , enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of the z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level (the area in the middle of the distribution) plus the remaining area in the left tail.

EXAMPLE

A consumer advocacy group wants to study consumer satisfaction with their shopping experience at the country's two biggest retailers. The group surveyed consumers and asked them to rate one of the

retailers in a number of different categories. An overall satisfaction score out of 100 summarized the responses for each consumer sampled. In a sample of 35 consumers for retailer A, the average overall satisfaction score was 79. In a sample of 30 consumers for retailer B, the average overall satisfaction score was 71. Based on prior experience with the satisfaction rating scale, the population standard deviation for retailer A is assumed to be 10 and the population standard deviation for retailer B is assumed to be 12.

1. Construct a 94% confidence interval for the difference in the mean satisfaction score for the two retailers.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean satisfaction score for retailer A is greater than the mean satisfaction score for retailer B? Explain.

Solution:

1. Let retailer A be population 1 and retailer B be population 2. These populations are independent because there is no relationship between the consumers sampled for each retailer. From the question, we have the following information:

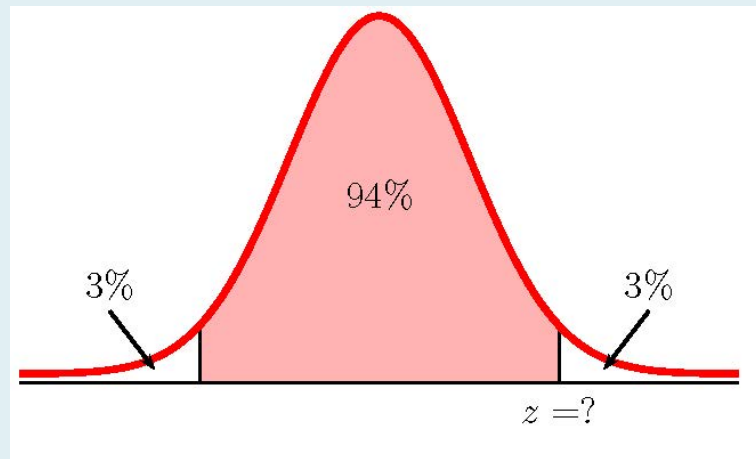
Retailer A	Retailer B
$n_1 = 35$	$n_2 = 30$
$\bar{x}_1 = 79$	$\bar{x}_2 = 71$
$\sigma_1 = 10$	$\sigma_2 = 12$

The normal distribution applies because the sample sizes are both greater than or equal to 30.

To find the confidence interval, we need to find the z -score for the 94% confidence interval.

This means that we need to find the z -score so that the entire area to the left of z is

$$0.94 + \frac{1 - 0.94}{2} = 0.97.$$



Function	norm.s.inv	Answer
Field 1	0.97	1.8807...

So $z = 1.8807\dots$. The 94% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x}_1 - \bar{x}_2 - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 79 - 71 - 1.8807\dots \sqrt{\frac{10^2}{35} + \frac{12^2}{30}} \\ &= 2.796 \\ \text{Upper Limit} &= \bar{x}_1 - \bar{x}_2 + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 79 - 71 + 1.8807\dots \sqrt{\frac{10^2}{35} + \frac{12^2}{30}} \\ &= 13.204 \end{aligned}$$

- We are 94% confident that the difference in the mean satisfaction score for the two retailers is between 2.796 and 13.204.
- Because 0 is outside the confidence interval and both limits are positive, it suggests that the difference in the means $\mu_1 - \mu_2$ is greater than 0. That is, $\mu_1 - \mu_2 > 0$ ($\mu_1 > \mu_2$). This suggests that the mean for population 1 (retailer A) is greater than the mean for population 2 (retailer B). So the mean satisfaction score for retailer A is greater than the mean satisfaction score for retailer B.

NOTES

1. When calculating the limits for the confidence interval keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cells containing the z -score or any other values, to ensure that all of the decimal places are used in the calculation.
2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population means captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.

Steps to Conduct a Hypothesis Test for the Difference in Two Independent Population Means with Known Population Standard Deviations

1. Write down the null hypothesis that there is no difference in the population means:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \end{array}$$

The null hypothesis is always the claim that the two population means are equal ($\mu_1 = \mu_2$).

2. Write down the alternative hypotheses in terms of the difference in the population means.

The alternative hypothesis will be one of the following:

$$H_a : \mu_1 - \mu_2 < 0 \quad (\mu_1 < \mu_2)$$

$$H_a : \mu_1 - \mu_2 > 0 \quad (\mu_1 > \mu_2)$$

$$H_a : \mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.

5. Assuming the population standard deviations are known, use the normal distribution to find the p -value (the area in the corresponding tail) for the test. The z -score is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6. Compare the p -value to the significance level and state the outcome of the test:
- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION MEANS WITH KNOWN POPULATION STANDARD DEVIATIONS

Assuming that the population standard deviations are known, the p -value for a hypothesis test on the difference in two independent population means is the area in the tail(s) of the normal distribution.

The p -value is the area in the tail(s) of a normal distribution, so the **norm.dist(x,μ,σ,logic operator)** function can be used to calculate the p -value.

- For x , enter the value for $\bar{x}_1 - \bar{x}_2$.
- For μ , enter the 0, the value of $\mu_1 - \mu_2$ from the null hypothesis. This is the mean of the

distribution of the differences in the sample means.

- For σ , enter the value of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, the standard deviation of the distribution of the differences in the sample mean.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

As with the previous chapter, use the appropriate technique with the **norm.dist** function to find the area in the left-tail, the area in the right-tail or the sum of the area in tails.

EXAMPLE

A floor cleaning company has been using Wax 1 to wax floors for a long time. A new floor wax, Wax 2, has recently come on the market with the claim that it is longer lasting than Wax 1. The company wants to investigate this claim. The company waxed a sample of 20 floors with Wax 1 and found the average number of months the wax lasted was 2.7 months. The company waxed a sample of 20 floors with Wax 2 and found the average number of months the wax lasted was 2.9 months. Based on previous information, the standard deviation for the length of time Wax 1 lasts is 0.33 months and the standard deviation for the length of time Wax 2 lasts is 0.36 months. Both populations have normal distributions. At the 5% significance level, test if Wax 2 lasts longer, on average, than Wax 1.

Solution:

Let Wax 1 be population 1 and Wax 2 be population 2. These populations are independent because there is no relationship between the length of time each type of wax lasts. From the question, we have the following information:

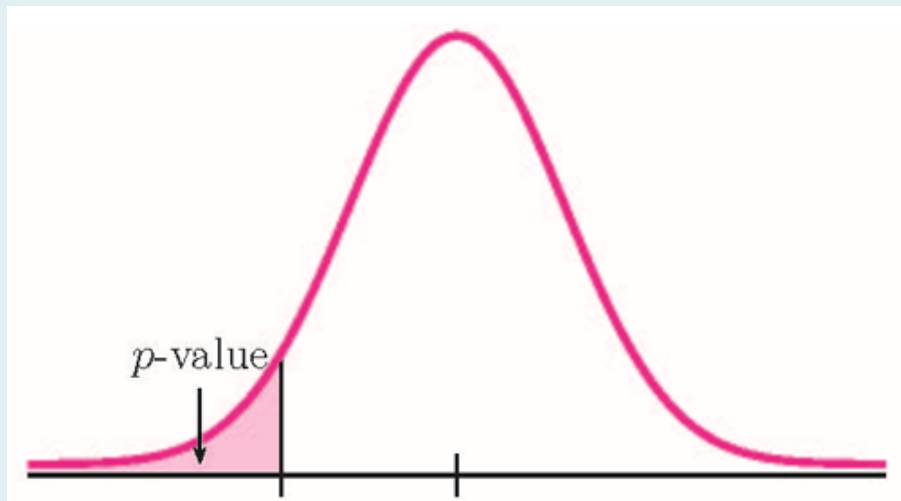
Wax 1	Wax 2
$n_1 = 20$	$n_2 = 20$
$\bar{x}_1 = 2.7$	$\bar{x}_2 = 2.9$
$\sigma_1 = 0.33$	$\sigma_2 = 0.36$

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 < 0 \end{array}$$

p-value:

This is a test on a the difference in two population means where the population standard deviation are known. So we use a normal distribution to calculate the p-value. Because the alternative hypothesis is a $<$, the p-value is the area in the left-tail of the distribution.



Function	norm.dist	Answer
Field 1	2.7-2.9	0.0335
Field 2	0	
Field 3	$\text{sqrt}(0.33^2/20+0.36^2/20)$	
Field 4	true	

So the p-value= 0.0335.

Conclusion:

Because $p\text{-value} = 0.0335 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that Wax 2 lasts longer than Wax 1.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean number of months for Wax 1 equals the mean number of months for Wax 2. That is, the two types of waxes have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 < 0$ is the claim that the mean for Wax 1 is less than the mean for Wax 2 ($\mu_1 < \mu_2$). This is the same as saying that the mean for Wax 2 is larger than the mean for Wax 1.
3. The p -value is the area in the left tail of the normal distribution. In the calculation of the p -value:

- The function is norm.dist because we are finding the area in the left tail of a normal distribution.
- Field 1 is the value of $\bar{x}_1 - \bar{x}_2 = 2.7 - 2.9$
- Field 2 is 0, the value of $\mu_1 - \mu_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu_1 - \mu_2 = 0$.
- Field 3 is the standard deviation for the difference in the sample means

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}$$

4. The p -value of 0.0335 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the mean number of months for Wax 1 is less than the mean number of months for Wax 2. For the company this suggests that they should switch to Wax 2 because of it is longer lasting than Wax 1.

EXAMPLE

A consumer advocacy group wants to compare the revolutions per minute (RPM) for two different engines. The group believes that Engine A has a higher average RPM than Engine B. In a sample of 40 Engine A's, the sample mean number of RPMs was 1550. In a sample of 30 Engine B's, the sample mean number of RPMs was 1500. Based on previous information, the standard deviation for the RPMs for Engine A is 75 and the standard deviation for Engine B is 65. At the 1% significance level, is the average RPM for Engine A higher than for Engine B?

Solution:

Let Engine A be population 1 and Engine B be population 2. These populations are independent because there is no relationship between the RPMs for the two engines. From the questions, we have the following information:

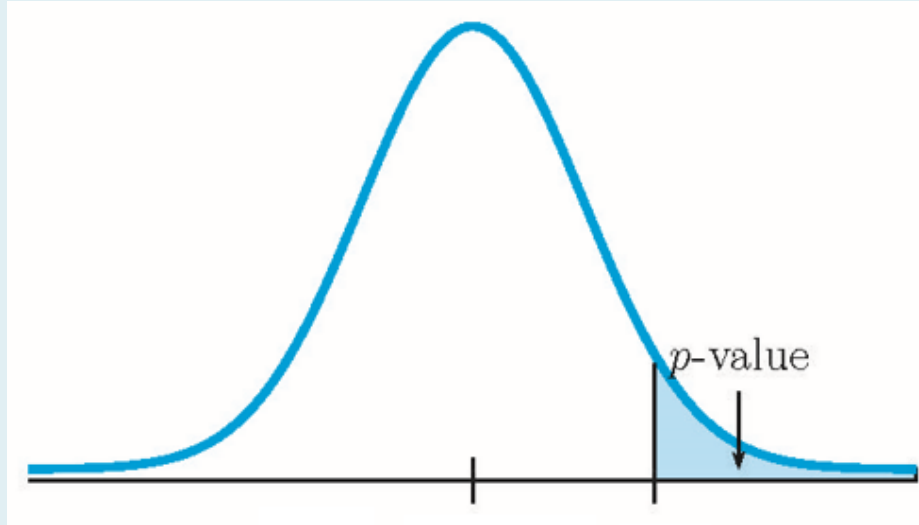
Engine A	Engine B
$n_1 = 40$	$n_2 = 30$
$\bar{x}_1 = 1550$	$\bar{x}_2 = 1500$
$\sigma_1 = 75$	$\sigma_2 = 65$

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 > 0 \end{array}$$

p-value:

This is a test on the difference in two population means where the population standard deviation are known. So we use a normal distribution to calculate the p -value. Because the alternative hypothesis is a $>$, the p -value is the area in the right tail of the distribution.



Function	1-norm.dist	Answer
Field 1	1550-1500	0.0014
Field 2	0	
Field 3	$\text{sqrt}(75^2/40+65^2/30)$	
Field 4	true	

So the $p\text{-value} = 0.0014$.

Conclusion:

Because $p\text{-value} = 0.0014 < 0.01 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that the average RPM for Engine A is higher than for Engine B.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean RPM for Engine A equals the mean RPM for Engine B. That is, the two engines have the same average RPM.
2. The alternative hypothesis $\mu_1 - \mu_2 > 0$ is the claim that the mean RPM for Engine A is greater than the mean RPM for Engine B ($\mu_1 > \mu_2$).
3. The $p\text{-value}$ is the area in the right tail of the normal distribution. In the calculation of the

p -value:

- The function is 1-norm.dist because we are finding the area in the right tail of a normal distribution.
- Field 1 is the value of $\bar{x}_1 - \bar{x}_2 = 1550 - 1500$
- Field 2 is 0, the value of $\mu_1 - \mu_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $\mu_1 - \mu_2 = 0$.
- Field 3 is the standard deviation for the difference in the sample means

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{75^2}{40} + \frac{65^2}{30}}$$

4. The p -value of 0.0014 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the mean RPM for Engine A is greater than the mean RPM for Engine B, just as the consumer advocacy group claimed.

EXAMPLE

The student union at a local college owns two coffee shops on campus: The Study Cafe and Coffee&Books. The student union wants to find out if there is a difference the average amount students spend per transaction at each of the coffee shops. In a sample of 65 transactions at the Study Cafe, the average amount spent was \$9.40. In a sample of 50 transactions at Coffee&Books, the average amount spent was \$10.15. Based on previous information, the standard deviation for the amount spent at the Study Cafe is \$1.35 and the

standard deviation for Coffee&Books B is \$2.70. At the 5% significance level, is there a difference in the average amount spent per transaction at the two coffee shops?

Solution:

Let the Study Cafe be population 1 and Coffee&Books be population 2. These populations are independent because there is no relationship between the amount spent at each coffee shop. From the question, we have the following information:

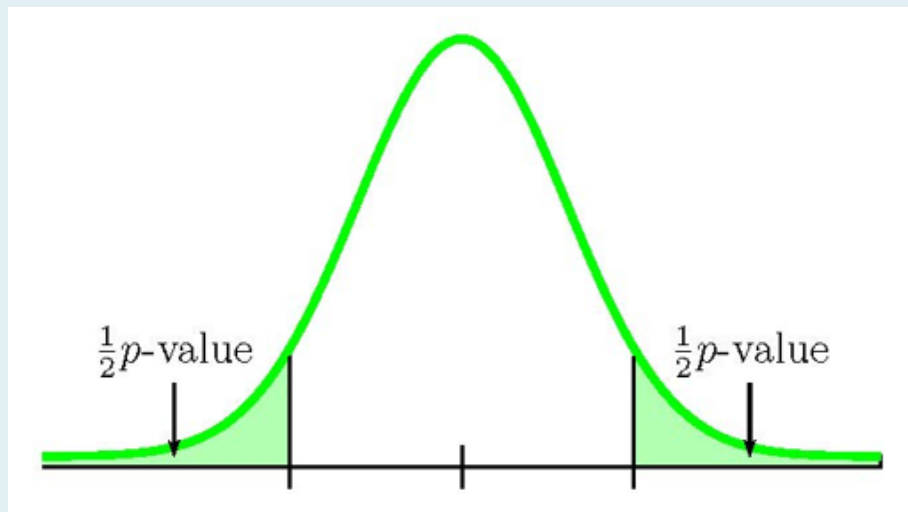
The Study Cafe	Coffee&Books
$n_1 = 65$	$n_2 = 50$
$\bar{x}_1 = 9.40$	$\bar{x}_2 = 10.15$
$\sigma_1 = 1.35$	$\sigma_2 = 2.70$

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 \neq 0 \end{array}$$

p-value:

This is a test on a the difference in two population means where the population standard deviation are known. So we use a normal distribution to calculate the p-value. Because the alternative hypothesis is a \neq , the p-value is the sum of the area in the two tails of the distribution.



We need to know if the sample information relates to the left or right tail because that will determine

how we calculate out the area of that tail using the normal distribution. In this case, the $\bar{x}_1 < \bar{x}_2$ ($9.4 < 10.15$), so the sample information relates to the left tail of the normal distribution. This means that we will calculate out the area in the left tail using **norm.dist**. However, this is a two-tailed test where the p -value is the sum of the area in the two tails and the area in the left tail is only one half of the p -value. The area in the left tail equals the area in the right tail and the p -value is the sum of these two areas.

Function	norm.dist	Answer
Field 1	9.40-10.15	0.0360
Field 2	0	
Field 3	sqrt(1.35^2/65+2.7^2/50)	
Field 4	true	

So the area in the left tail is 0.0360, which means $\frac{1}{2}(p\text{-value}) = 0.0360$. This is also the area in the right tail, so

$$p\text{-value} = 0.0360 + 0.0360 = 0.0720$$

Conclusion:

Because $p\text{-value} = 0.0720 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that there is a difference in the average amount spent at the two coffee shops.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the mean amount spent at the Study Cafe equals the mean amount spent at Coffee&Books. That is, the average amount spent is the same at both coffee shops.
2. The alternative hypothesis $\mu_1 - \mu_2 \neq 0$ is the claim that the mean amount spent at the Study Cafe is different than the mean amount spent at Coffee&Books ($\mu_1 \neq \mu_2$).
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail

represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.

- If the sample mean \bar{x}_1 is **less than** the sample mean \bar{x}_2 ($\bar{x}_1 < \bar{x}_2$), the sample information belongs to the **left tail**.

- We use **norm.dist** $(\bar{x}_1 - \bar{x}_2, 0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{true})$ to find the area in the

left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.

- If the sample mean \bar{x}_1 is **greater than** the sample mean \bar{x}_2 ($\bar{x}_1 > \bar{x}_2$), the sample information belongs to the **right tail**.

- We use **1-norm.dist** $(\bar{x}_1 - \bar{x}_2, 0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{true})$ to find the area in the

right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.

4. The p -value of 0.0720 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the mean amount spent at the Study Cafe equals the mean amount spent at Coffee&Books.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=207#oembed-1>

Watch this video: Confidence Intervals for Two Population Means, Sigma Known by ExcelIsFun [9:52]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=207#oembed-2>

Watch this video: Hypothesis Testing for Two Population Means, Sigma Known by ExcellIsFun [16:47]

Concept Review

The general form of a confidence interval for the difference in two independent population means with known population standard deviations is

$$\text{Lower Limit} = \bar{x}_1 - \bar{x}_2 - z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{Upper Limit} = \bar{x}_1 - \bar{x}_2 + z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where z is the positive z -score of the standard normal distribution so the area under the normal distribution in between $-z$ and z is C .

The hypothesis test for the difference in two independent population means with known population standard deviations is a well established process:

1. Write down the null and alternative hypotheses in terms of the differences in the population means $\mu_1 - \mu_2$.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the normal distribution. Because the population standard deviations are known, we use the normal distribution to find the p -value.
5. Compare the p -value to the significance level and state the outcome of the test.
6. Write down a concluding sentence specific to the context of the question.

Attribution

“10.2 Two Population Means with Known Standard Deviations“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

9.3 STATISTICAL INFERENCE FOR TWO POPULATION MEANS WITH UNKNOWN POPULATION STANDARD DEVIATIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population means with unknown population standard deviations.
- Conduct and interpret hypothesis tests for two population means with unknown population standard deviations.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference in the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**. In the case the samples are independent, we also have to contend with whether or not we know the population standard deviations.

Two populations are **independent** if the sample taken from population 1 is not related in anyway to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the two populations:

Symbol for:	Population 1	Population 2
Population Mean	μ_1	μ_2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Mean	\bar{x}_1	\bar{x}_2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two population means ($\mu_1 - \mu_2$), we need to use the distribution of the difference in the sample means $\bar{x}_1 - \bar{x}_2$:

- The mean of the distribution of the difference in the sample means is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The standard deviation of the distribution of the difference in the sample means is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- The distribution of the difference in the sample means is normal if **one** of the following is true:
 - Both populations are normally distributed.
 - The sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- Assuming the distribution of the difference of the sample means is normal, the z -score is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

As we have seen previously when working with confidence intervals and hypothesis testing for a single population, when the population standard deviation is unknown and we must use the sample standard deviation as an estimate for the population standard deviation, we use a t -distribution. We do the same thing when working with the two population means. When the population standard deviations are unknown, we use the sample standard deviations as estimates for the population standard deviations σ_1 and σ_2 . In this situation, we use a t -distribution for the distribution of the difference in the sample means. So, when the population standard deviations are unknown for a confidence interval or hypothesis test on the difference in two population means, we will use a t -distribution. The t -score and the degrees of freedom are:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

Obviously, the degrees of freedom formula is somewhat complicated. But a computer makes the calculation a bit more manageable. The output from the degrees of freedom formula is rarely a whole number. After calculating the value of df using the above formula, round the output from this formula **down** to the next whole number to get the degrees of freedom for the t -distribution.

Constructing a Confidence Interval for the Difference in Two Population Means with Unknown Population Standard Deviations

Suppose a sample of size n_1 with sample mean \bar{x}_1 and standard deviation s_1 is taken from population 1 and a sample of size n_2 with sample mean \bar{x}_2 and standard deviation s_2 is taken from population 2 where the populations are independent and the population standard deviations are **unknown**. The limits for the confidence interval with confidence level C for the difference in the population means $\mu_1 - \mu_2$ are:

$$\begin{aligned} \text{Lower Limit} &= \bar{x}_1 - \bar{x}_2 - t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ \text{Upper Limit} &= \bar{x}_1 - \bar{x}_2 + t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \end{aligned}$$

where t is the positive t -score of the t -distribution with $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$ so that the area under the curve in between $-t$ and t is $C\%$.

NOTES

1. In order to construct the confidence interval for the difference in two population means with independent samples, we need to check that the distribution of the difference in the

sample means follows a normal distribution. This means that we need to check that either the populations are normal or that the sample sizes are large enough (greater than or equal to 30).

2. When the population standard deviations are unknown, we must use a t -distribution in the construction of the confidence interval.
3. The value of degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.

CALCULATING THE Formula does not parse-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

The output from the **t.inv.2t** function is the value of t -score needed to construct the confidence interval.

NOTE

1. The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.
2. The degrees of freedom for a t -distribution **must** be a **whole number**. The output from the degrees of freedom formula

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

is almost never a whole number. After calculating the value of df using the formula, **round the value down to the next whole number**. Remember to enter the rounded down value of df for the degrees of freedom in the **t.inv.2t** function.

EXAMPLE

A company that manufactures and services photocopiers wants to study the difference in the average repair time for the two different models of photocopiers they make. In a sample of 60 repairs of photocopier A, the mean repair time was 84.2 minutes with a standard deviation of 19.4 minutes. In a sample of 70 repairs of photocopier B, the mean repair time was 91.6 minutes with a standard deviation of 18.8 minutes.

1. Construct a 95% confidence interval for the difference in the mean repair time for the two photocopiers.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean repair times for the photocopiers is the same?

Explain.

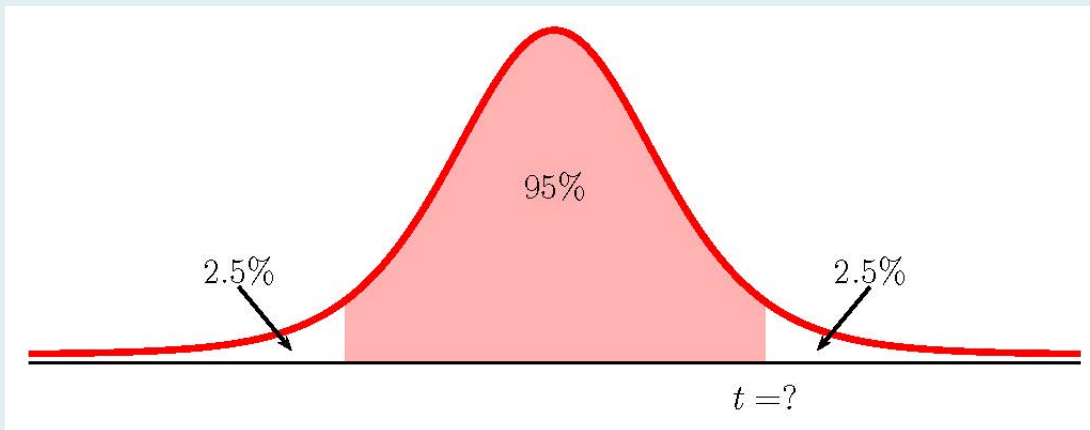
Solution:

- Let photocopier A be population 1 and photocopier B be population 2. These populations are independent because there is no relationship between the repair times for the two photocopiers. From the question we have the following information:

Photocopier A	Photocopier B
$n_1 = 60$	$n_2 = 70$
$\bar{x}_1 = 84.2$	$\bar{x}_2 = 91.6$
$s_1 = 19.4$	$s_2 = 18.8$

To find the confidence interval, we need to find the t -score for the 95% confidence interval. This means that we need to find the t -score so that the area in the tails is $1 - 0.95 = 0.05$.

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = t_{\alpha/2, df} = t_{0.025, 123} = 1.978$$



Function	t.inv.2t	Answer
Field 1	0.05	1.9794...
Field 2	123	

So $t = 1.9794\dots$. The 95% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x}_1 - \bar{x}_2 - t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= 84.2 - 91.6 - 1.9794\dots \sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{70}} \\ &= -14.06 \\ \text{Upper Limit} &= \bar{x}_1 - \bar{x}_2 + t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= 84.2 - 91.6 + 1.9794\dots \sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{70}} \\ &= -0.74 \end{aligned}$$

- We are 95% confident that the difference in the mean repair time for the two photocopiers is between -14.06 minutes and -0.74 minutes.
- Because 0 is outside the confidence interval and both limits are negative, it suggests that the difference in the means $\mu_1 - \mu_2$ is less than 0. That is, $\mu_1 - \mu_2 < 0$ ($\mu_1 < \mu_2$). This suggests that the mean for population 1 (photocopier A) is less than the mean for population 2 (photocopier B). So the mean repair time for photocopier A is less than the mean repair time for photocopier B.

NOTES

- When calculating the limits for the confidence interval keep all of the decimals in the t -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cells containing the t -score and any other values, to ensure that all of the decimal places are used in the calculation.
- When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population means captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.
- The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.

Steps to Conduct a Hypothesis Test for the Difference in Two Independent Population Means with Unknown Population Standard Deviations

1. Write down the null hypothesis that there is no difference in the population means:

$$H_0 : \mu_1 - \mu_2 = 0$$

The null hypothesis is always the claim that the two population means are equal ($\mu_1 = \mu_2$).

2. Write down the alternative hypotheses in terms of the difference in the population means. The alternative hypothesis will be one of the following:

$$\begin{array}{l} H_a: \mu_1 - \mu_2 < 0 \quad \& \quad (\mu_1 < \mu_2) \\ H_a: \mu_1 - \mu_2 > 0 \quad \& \quad (\mu_1 > \mu_2) \\ H_a: \mu_1 - \mu_2 \neq 0 \quad \& \quad (\mu_1 \neq \mu_2) \end{array}$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Assuming the population standard deviations are unknown, use a t -distribution to find the p -value (the area in the corresponding tail) for the test. The t -score and degrees of freedom are

$$t = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

6. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

- Write down a concluding sentence specific to the context of the question.

USING EXCEL TO CALCULATE THE *P*-VALUE FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION MEANS WITH UNKNOWN POPULATION STANDARD DEVIATIONS

Assuming that the population standard deviations are unknown, the *p*-value for a hypothesis test on the difference in two independent population means is the area in the tail(s) of the *t*-distribution.

If the *p*-value is the area in the left tail:

- Use the **t.dist** function to find the *p*-value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:
 - For **t-score**, enter the value of *t* calculated from

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
 - For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right) + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)}$$
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

If the *p*-value is the area in the right tail:

- Use the **t.dist.rt** function to find the *p*-value. In the **t.dist.rt(t-score, degrees of freedom)** function:
 - For **t-score**, enter the value of *t* calculated from

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

If the p -value is the sum of the area in the two tails:

- Use the **t.dist.2t** function to find the p -value. In the **t.dist.2t(t-score, degrees of freedom)** function:

- For **t-score**, enter the **absolute value** of t calculated from

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

. Note: In the **t.dist.2t** function, the value of the t -score must be a **positive** number. If the t -score is negative, enter the absolute value of the t -score into the **t.dist.2t** function.

- For **degrees of freedom**, enter the degrees of freedom calculated using

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

NOTE

The degrees of freedom for a t -distribution **must** be a **whole number**. The output from the degrees of freedom formula

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \times \left(\frac{s_2^2}{n_2}\right)^2}$$

is almost never a whole number. After calculating the value of df using the formula, **round the value down to the next whole number**. Remember to entered the rounded down value of df for the degrees of freedom in the **t.dist** functions.

EXAMPLE

A researcher wants to study the difference between the average amount of time boys and girls aged seven to eleven spend playing sports each day. In a sample of 9 girls, the average number of hours spent playing sports per day is 2 hours with a standard deviation of 0.866 hours. In a sample of 16 boys, the average number of hours spent playing sports per day is 3.2 hours with a standard deviation of 1 hour. Both populations have a normal distribution. At the 5% significance level, is there a difference in the mean amount of time boys and girls aged seven to eleven play sports each day?

Solution:

Let girls be population 1 and boys be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

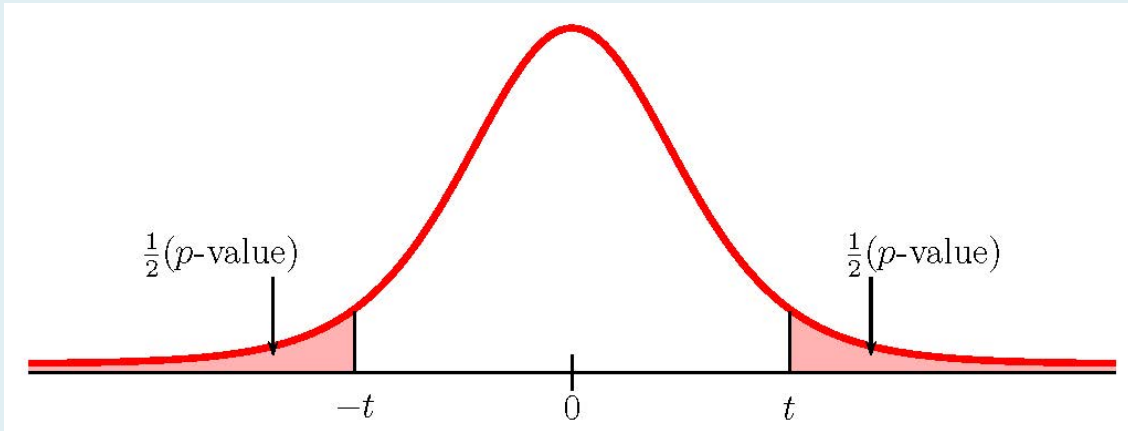
Girls	Boys
$n_1 = 9$	$n_2 = 16$
$\bar{x}_1 = 2$	$\bar{x}_2 = 3.2$
$s = 0.866$	$s_2 = 1$

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 \neq 0 \end{array}$$

p-value:

This is a test on the difference in two population means where the population standard deviation are unknown. So we use a t -distribution to calculate the p -value. Because the alternative hypothesis is a \neq , the p -value is the sum of areas in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the *t*-score and the degrees of freedom:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(2 - 3.2) - 0}{\sqrt{\frac{0.866^2}{9} + \frac{1^2}{16}}} = -3.1423... \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \times \left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{0.866^2}{9} + \frac{1^2}{16}\right)^2}{\frac{1}{9 - 1} \times \left(\frac{0.866^2}{9}\right)^2 + \frac{1}{16 - 1} \times \left(\frac{1^2}{16}\right)^2} = 18.846... \rightarrow 18$$

Function	t.dist.2t	Answer
Field 1	3.1423...	0.0056
Field 2	18	

So the *p*-value = 0.0056.

Conclusion:

Because *p*-value = 0.0056 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that there is a difference in the mean amount of time boys and girls aged seven to eleven play sports each day.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that there is no difference in the mean amount of time boys and girls spend playing sports each day. That is, the two populations have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 \neq 0$ is the claim that there is a difference in the mean amount of time boys and girls spend playing sports each day ($\mu_1 \neq \mu_2$). That is, the two populations have different means.
3. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The **t.dist.2t** function requires that the value entered for the t -score is **positive**. A negative t -score entered into the **t.dist.2t** function generates an error in Excel. In this case, the value of the t -score is negative, so we must enter the absolute value of this t -score into field 1.
6. The p -value of 0.0056 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, there is a difference in the mean amount of time boys and girls spend playing sports each day.

EXAMPLE

A town has two colleges. A local community group believes that students who graduate from

College A have taken more math classes than the students who graduate from College B. In a sample of 11 graduates from College A, the average is 4 math classes per graduate with a standard deviation of 1.5 math classes. In a sample of 9 graduates from College B, the average is 3.5 math classes per graduate with a standard deviation of 1 math class. Both populations have a normal distribution. At the 1% significance level, test the community groups claim that graduates from College A have taken more math classes than graduates from College B.

Solution:

Let College A be population 1 and College B be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

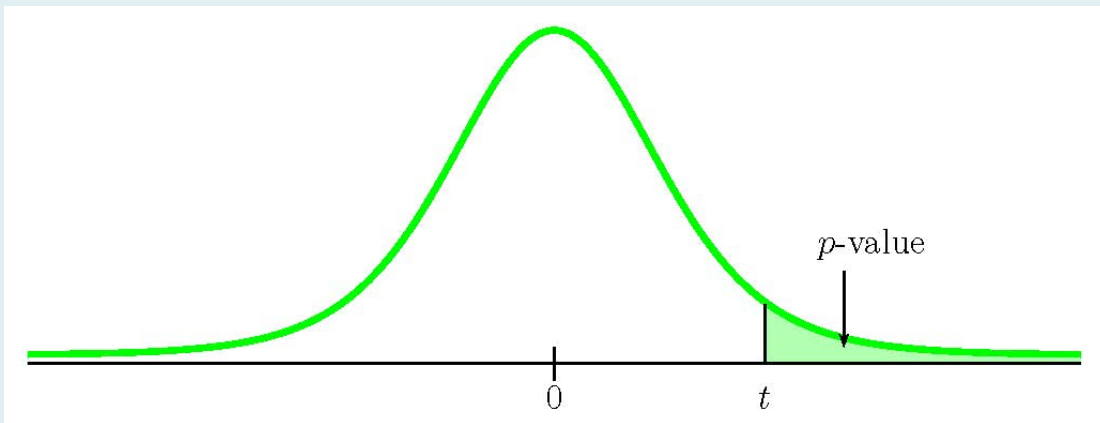
College A	College B
$n_1 = 11$	$n_2 = 9$
$\bar{x}_1 = 4$	$\bar{x}_2 = 3.5$
$s_1 = 1.5$	$s_2 = 1$

Hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \quad H_a: \mu_1 - \mu_2 > 0$$

p-value:

This is a test on a the difference in two population means where the population standard deviation are unknown. So we use a *t*-distribution to calculate the *p*-value. Because the alternative hypothesis is a *>*, the *p*-value is the area in the right tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the *t*-score and the degrees of freedom:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(4 - 3.5) - 0}{\sqrt{\frac{1.5^2}{11} + \frac{1^2}{9}}} = 0.8899... \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}} = \frac{\left(\frac{1.5^2}{11} + \frac{1^2}{9}\right)^2}{\frac{1}{11 - 1} + \frac{1}{9 - 1}} = 17.397... \rightarrow 17$$

Function	t.dist.rt	Answer
Field 1	0.8899...	0.1930
Field 2	17	

So the *p*-value = 0.1930.

Conclusion:

Because *p*-value = 0.1930 > 0.01 = α, we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that, on average, graduates of College A take more math classes than graduates of College B.

NOTES

- The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the average number of math classes taken by graduates of College A equals the average number of math classes taken by graduates of College B. That is, the two populations have the same mean.
- The alternative hypothesis $\mu_1 - \mu_2 > 0$ is the claim that, on average, graduates of College A taken more math classes than graduates of College B ($\mu_1 > \mu_2$).
- Keep all of the decimals throughout the calculation (i.e. in the *t*-score, etc.) to avoid any round-off error in the calculation of the *p*-value. This ensures that we get the most accurate value for the *p*-value. Use Excel to do the calculations, and then click on the cells in

subsequent calculations.

4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The p -value of 0.1930 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, graduates from the two colleges take, on average, the same number of math classes.

EXAMPLE

A professor at a large community college taught both an online section and a face-to-face section of his statistics course. The professor wants to study the difference in the average score on the final exam, believing that the mean score for the online section would be lower than the face-to-face section. The professor randomly selected 30 final exam scores from each section and recorded the scores in the tables below.

Online Section:

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Face-to-Face Section:

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

At the 5% significance level, is the mean of the final exam score for the online section lower than the mean of the final exam score for the face-to-face section?

Solution:

Let the online section be population 1 and the face-to-face section be population 2. These populations are independent because there is no relationship between the two groups. From the questions, we have the following information:

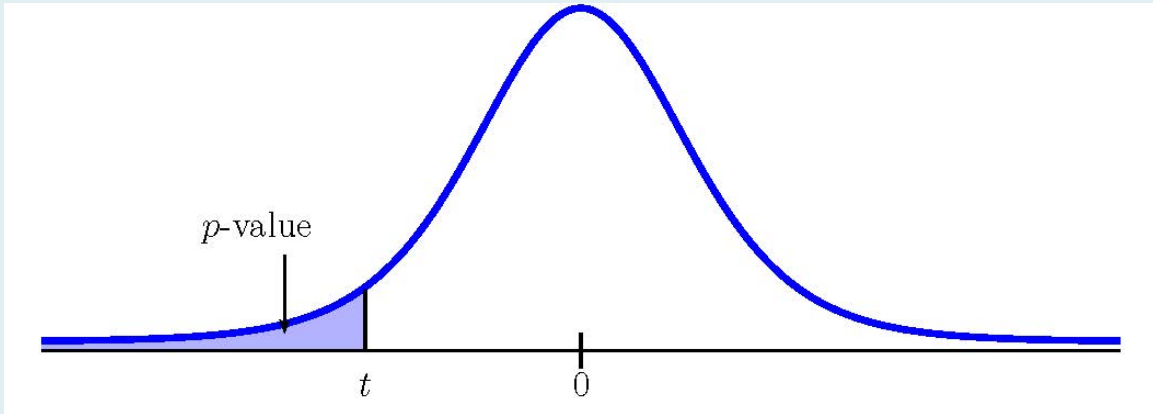
Online	Face-to-Face
$n_1 = 30$	$n_2 = 30$
$\bar{x}_1 = 72.85$	$\bar{x}_2 = 84.98$
$s_1 = 16.918...$	$s_2 = 11.714...$

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 < 0 \end{array}$$

p-value:

This is a test on a the difference in two population means where the population standard deviation are unknown. So we use a *t*-distribution to calculate the *p*-value. Because the alternative hypothesis is a *<*, the *p*-value is the area in the left tail of the distribution.



To use the **t.dist** function, we need to calculate out the *t*-score and the degrees of freedom:

$$\begin{aligned} t &= \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(72.85 - 84.98) - 0}{\sqrt{\frac{16.918...^2}{30} + \frac{11.714...^2}{30}}} \\ &= -3.228... \\ df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2} \\ &= \frac{\left(\frac{16.918...^2}{30} + \frac{11.714...^2}{30}\right)^2}{\frac{1}{30 - 1} \left(\frac{16.918...^2}{30}\right)^2 + \frac{1}{30 - 1} \left(\frac{11.714...^2}{30}\right)^2} \\ &= 51.608... \end{aligned}$$

Function	t.dist	Answer
Field 1	-3.228...	0.0011
Field 2	51	
Field 3	true	

So the $p\text{-value} = 0.0011$.

Conclusion:

Because $p\text{-value} = 0.0011 < 0.05 = \alpha$, we do reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the mean final exam score for the online section is lower than the face-to-face section.

NOTES

1. The null hypothesis $\mu_1 - \mu_2 = 0$ is the claim that the average final exam score is the same for both sections. That is, the two populations have the same mean.
2. The alternative hypothesis $\mu_1 - \mu_2 < 0$ is the claim that average final exam score for the online section is lower than the face-to-face section ($\mu_1 < \mu_2$).
3. Keep all of the decimals throughout the calculation (i.e. in the sample means, sample standard deviations, in the t -score, etc.) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
4. The value of the degrees of freedom must be a whole number. After using the formula, remember to round the value **down** to the next whole number to get the required degrees of freedom for the t -distribution.
5. The p -value of 0.0011 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the average final exam score for the online section is lower than for the face-to-face section.

TRY IT

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is 5 years with a standard deviation of 1.2 years. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8 years. The populations are normally distributed. At the 5% significance level, on average, do workers at Company A stay longer than workers at Company B?

Click to see Solution

Let Company A be population 1 and Company B be population 2.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 > 0 \end{array}$$

p-value:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(5 - 4.5) - 0}{\sqrt{\frac{1.2^2}{15} + \frac{0.8^2}{20}}} = 1.3975... \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}} = \frac{\left(\frac{1.2^2}{15} + \frac{0.8^2}{20}\right)^2}{\frac{1}{15 - 1} + \frac{1}{20 - 1}} = 23.005... \rightarrow 23$$

Function	t.dist.rt	Answer
Field 1	1.3975...	0.0878
Field 2	23	

Conclusion:

Because $p\text{-value} = 0.0878 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that, on average, workers at Company A stay longer than workers at Company B.





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=209#oembed-1>

Watch this video: Confidence Intervals for Two Population Means, Sigma Unknown by ExcelIsFun [16:11]



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=209#oembed-2>

Watch this video: Hypothesis Testing for Two Population Means, Sigma Unknown by ExcelIsFun [17:29]

Concept Review

The general form of a confidence interval for the difference in two independent population means with unknown population standard deviations is

$$\text{Lower Limit} = \bar{x}_1 - \bar{x}_2 - t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper Limit} = \bar{x}_1 - \bar{x}_2 + t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t is the positive t -score of the t -distribution with $\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \times \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \times \left(\frac{s_2^2}{n_2}\right)^2}$ so that the area under the t -distribution in between $-t$ and t is C .

The hypothesis test for the difference in two independent population means with unknown population standard deviations is a well established process:

1. Write down the null and alternative hypotheses in terms of the differences in the population means $\mu_1 - \mu_2$.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the t -distribution with

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Because the population standard deviations are unknown, we use the t -distribution to find the p -value.

5. Compare the p -value to the significance level and state the outcome of the test.
6. Write down a concluding sentence specific to the context of the question.

Attribution

“10.1 Two Population Means with Unknown Standard Deviations“ and “10.2 Two Population Means with Known Standard Deviations“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

9.4 STATISTICAL INFERENCE FOR MATCHED SAMPLES

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for the mean difference for matched samples.
- Conduct and interpret hypothesis tests for matched samples.

The comparison of two population means is very common. Often, we want to find out if the two populations under study have the same mean or if there is some difference in the two population means. The approach we take when studying two population means depends on whether the samples are **independent** or **matched**.

In a **matched sample** experiment, there is some relationship between **pairs of data** in the samples. Inferences on matched samples are typically more accurate than inferences on independent samples because matched samples reduce the variability measures to only the ones within the pairs.

EXAMPLE

In a clinical trial for a new drug, patients are tested before the drug is administered and then the same group of patients are tested after being given the drug. This is a matched sample experiment because the same group of patients is measured before and after the administration of the drug. In

this way, there are a pair of observations (a before measurement and an after measurement) for each patient.

EXAMPLE

A manufacturing company wants to know which of two different production methods allow employees to perform a task the fastest. The table below illustrates the difference in an independent sample design and a matched sample design to test the difference in the average time it takes to perform the task using the two different methods.

Independent Sample Design	Matched Sample Design
<ul style="list-style-type: none"> • The company randomly selects two different groups of employees. • The employees in Group 1 perform the task using Method 1 and their times are recorded. • The employees in Group 2 perform the task using Method 2 and their times are recorded. 	<ul style="list-style-type: none"> • The company randomly selects one group of employees. • Each of the employees in the group perform the task using both methods and their times using each method are recorded.

In the independent sample design, there is no relationship between the two groups of employees. In the matched sample design, there is one group of employees with a pair of observations (a time from Method 1 and a time from Method 2) for each employee.

In matched sample designs, we work with the **differences** in the paired observations. We combine the two samples into a single sample by calculating out the difference between each of the paired observations. Throughout this section, we will use the following notation for the sample size, mean, and standard deviation of the differences in the paired observations:

Symbol for:	Symbol
Population Mean of the Differences in the Paired Data	μ_D
Population Standard Deviation of the Differences in the Paired Data	σ_D
Sample Size of the Differences in the Paired Data	n_D
Sample Mean of the Differences in the Paired Data	\bar{x}_D
Sample Standard Deviation of the Differences in the Paired Data	s_D

In order to construct a confidence interval or conduct a hypothesis test on the mean of the differences in the paired data (μ_D), we need to use the distribution of the differences in the paired data. In such cases, we need the distribution of the differences in the paired data to be normal, either because the differences are assumed to be normal or because the sample size n_D is large enough ($n_D \geq 30$).

By calculating out the differences in the paired data, we combine the two samples into a single sample consisting of the differences in the paired data. We use the differences to construct the confidence interval and run the hypothesis test. The confidence interval on the mean difference μ_D is a confidence interval for a single population mean. Similarly, the hypothesis test on the mean difference μ_D is actually a hypothesis test on a single population mean. In this case, we will follow the exact same procedures as we learned previously for a single population mean confidence interval and hypothesis test, only now the single population consists of the differences in the paired data.

When working with a matched sample design and the differences in the paired data, the population standard deviation will be unknown. So we will need to estimate the population standard deviation with the sample standard deviation. As we have seen previously, this means we must use a t -distribution in the confidence intervals and hypothesis test on the mean of the differences in the paired data.

Constructing a Confidence Interval for the Difference in Two Population Means with Matched Samples

Suppose matched samples, each of size n , are taken from two related populations. The sample mean \bar{x}_D and sample standard deviation s_D for the differences in the matched pairs are calculated. The limits for the confidence interval with confidence level C for the mean difference μ_D are:

$$\text{Lower Limit} = \bar{x}_D - t \times \frac{s_D}{\sqrt{n_D}}$$

$$\text{Upper Limit} = \bar{x}_D + t \times \frac{s_D}{\sqrt{n_D}}$$

where t is the positive t -score of the t -distribution with $\text{df} = n_D - 1$ so that the area under the curve in between $-t$ and t is $C\%$.

NOTES

1. In order to construct the confidence interval for the mean difference, we need to check that the distribution of the differences in the paired data follows a normal distribution. This means that we need to check that either the differences follow a normal distribution or that the sample size is large enough (greater than or equal to 30).
2. When the population standard deviations are unknown, we must use a t -distribution in the construction of the confidence interval.

CALCULATING THE **Formula does not parse**-SCORE FOR A CONFIDENCE INTERVAL IN EXCEL

To find the t -score to construct a confidence interval with confidence level C , use the **t.inv.2t(area in the tails, degrees of freedom)** function.

- For **area in the tails**, enter the **sum** of the area in the tails of the t -distribution. For a confidence interval, the area in the tails is $1 - C$.
- For **degrees of freedom**, enter the degrees of freedom $\text{df} = n_D - 1$.

The output from the **t.inv.2t** function is the value of the t -score needed to construct the confidence interval.

NOTE

The **t.inv.2t** function requires that we enter the **sum** of the area in **both** tails. The area in the middle of the distribution is the confidence level C , so the sum of the area in both tails is the leftover area $1 - C$.

EXAMPLE

A company has two different methods that employees can use to complete a manufacturing task. A sample of workers is taken and the time, in minutes, that each worker takes to complete the task using each method is recorded. The data is shown in the table below. Assume the differences in the paired times have a normal distribution.

Worker	Method 1	Method 2
1	5.5	6.8
2	6.9	6.6
3	6.1	5.1
4	6	6.8
5	7	6.7
6	6.7	6.5
7	6.4	5.8
8	7	6.8
9	6.6	5.3
10	5.7	5.8
11	5.9	6.9
12	7	6.7
13	5.4	6.5
14	5.4	6.3
15	5.3	5

1. Construct a 98% confidence interval for the mean difference in the time it takes the workers to complete the task.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the mean completion time for the two methods is the same? Explain.

Solution:

1. We start by calculating out the differences in the paired data. We will calculate the differences as **Method 1-Method 2**.

Worker	Method 1	Method 2	Difference
1	5.5	6.8	-1.3
2	6.9	6.6	0.3
3	6.1	5.1	1
4	6	6.8	-0.8
5	7	6.7	0.3
6	6.7	6.5	0.2
7	6.4	5.8	0.6
8	7	6.8	0.2
9	6.6	5.3	1.3
10	5.7	5.8	-0.1
11	5.9	6.9	-1
12	7	6.7	0.3
13	5.4	6.5	-1.1
14	5.4	6.3	-0.9
15	5.3	5	0.3

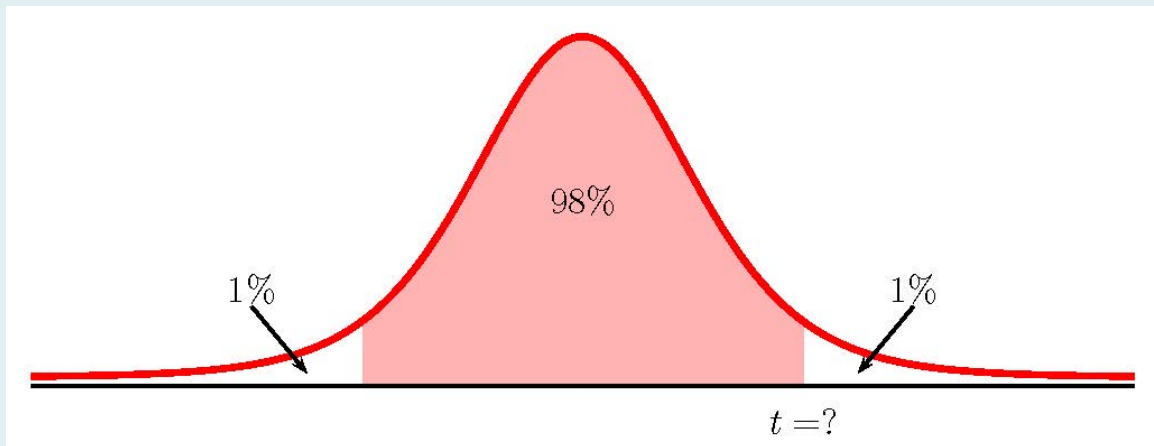
From the difference column, we have $n_D = 15$, $\bar{x}_D = -0.0466\dots$, and $s_D = 0.7936\dots$

To find the confidence interval, we need to find the t -score for the 98% confidence interval.

This means that we need to find the t -score so that the sum of the area in the tails is

$1 - 0.98 = 0.02$. The degrees of freedom for the t -distribution is

$df = n_D - 1 = 15 - 1 = 14$.



Function	t.inv.2t	Answer
Field 1	0.02	2.6244...
Field 2	14	

So $t = 2.6244\dots$. The 98% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \bar{x}_D - t \times \frac{s_D}{\sqrt{n_D}} \\ &= -0.0466\dots - 2.6244\dots \times \frac{0.7936\dots}{\sqrt{15}} \\ &= -0.584 \\ \text{Upper Limit} &= \bar{x}_D + t \times \frac{s_D}{\sqrt{n_D}} \\ &= -0.0466\dots + 2.6244\dots \times \frac{0.7936\dots}{\sqrt{15}} \\ &= 0.491 \end{aligned}$$

- We are 98% confident that the mean difference in the completion times using the two methods is between -0.584 minutes and 0.491 minutes.
- Because 0 is inside the confidence interval, it suggests that the mean difference μ_D is 0. That is, $\mu_D = 0$. This suggests that the mean completion times for the two methods are the same.

NOTES

- When calculating the limits for the confidence interval keep all of the decimals in the t -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the differences, sample mean,

sample standard deviation, and the limits, clicking on the corresponding cells to ensure that all of the decimal places are used in the calculation.

2. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual mean difference captured by the confidence interval (i.e. be specific to the context of the question), and appropriate units for the limits.

Steps to Conduct a Hypothesis Test for the Difference in Two Population Means with Matched Samples

1. Write down the null hypothesis that the mean difference is 0:

$$\begin{array}{l} H_0: \mu_D = 0 \end{array}$$

The null hypothesis is always the claim that there is no difference in the two population means.

2. Write down the alternative hypotheses in terms of the mean difference. The alternative hypothesis will be **one** of the following:

$$\begin{array}{l} H_a: \mu_D \leq 0 \\ H_a: \mu_D > 0 \\ H_a: \mu_D \neq 0 \end{array}$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level.
5. Use a t -distribution to find the p -value (the area in the corresponding tail) for the test. The t -score and degrees of freedom are

$$t = \frac{\overline{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \quad \text{and} \quad df = n_D - 1$$

6. Compare the p -value to the significance level and state the outcome of the test:

- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .

- The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON MATCHED SAMPLES

The p -value for a hypothesis test on the mean difference in matched samples is the area in the tail(s) of the t -distribution.

If the p -value is the area in the left tail:

- Use the **t.dist** function to find the p -value. In the **t.dist(t-score, degrees of freedom, logic operator)** function:
 - For **t-score**, enter the value of t calculated from
$$t = \frac{\overline{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$$
.
 - For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$.
 - For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

If the p -value is the area in the right tail:

- Use the **t.dist.rt** function to find the p -value. In the **t.dist.rt(t-score, degrees of freedom)** function:
 - For **t-score**, enter the value of t calculated from
$$t = \frac{\overline{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$$
.
 - For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$.

If the p -value is the sum of the area in the two tails:

- Use the **t.dist.2t** function to find the p -value. In the **t.dist.2t(t-score, degrees of freedom)** function:
 - For **t-score**, enter the **absolute value** of t calculated from
$$t = \frac{\overline{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}}$$
. Note: In the **t.dist.2t** function, the value of the t -score must be a **positive** number. If the t -score is negative, enter the absolute value of the t -score into the **t.dist.2t** function.
 - For **degrees of freedom**, enter the degrees of freedom calculated using $df = n_D - 1$.

EXAMPLE

A study was conducted to investigate the effectiveness of hypnosis on reducing pain. Eight subjects are randomly selected. Each subject's pain is measured before and after being hypnotized. A lower score indicates less pain. Assume the differences in the before and after scores have a normal distribution. At the 5% significance level, are the pain sensory measurements, on average, lower after hypnosis?

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Solution:

We start by calculating out the differences in the paired data. We will calculate the differences as **before-after**.

Subject	Before	After	Difference
A	6.6	6.8	-0.2
B	6.5	2.4	4.1
C	9	7.4	1.6
D	10.3	8.5	1.8
E	11.3	8.1	3.2
F	8.1	6.1	2
G	6.3	3.4	2.9
H	11.6	2	9.6

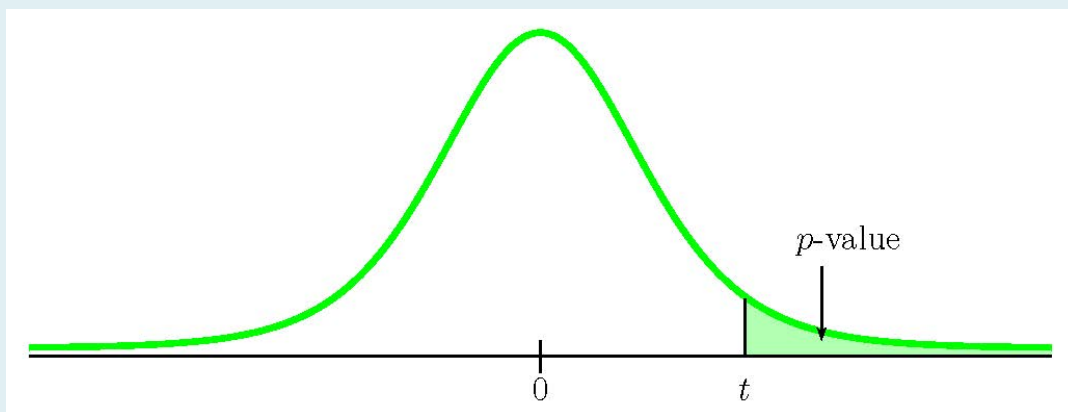
From the difference column, we have $n_D = 8$, $\bar{x}_D = 3.125$, and $s_D = 2.911\dots$

Hypotheses:

$$\begin{array}{l} H_0: \mu_D = 0 \\ H_a: \mu_D > 0 \end{array}$$

***p*-value:**

This is a test on the mean difference in matched samples, so we use a *t*-distribution to calculate the *p*-value. Because the alternative hypothesis is a $>$, the *p*-value is the area in the right tail of the distribution.



To use the **t.dist.rt** function, we need to calculate out the *t*-score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{3.125 - 0}{\frac{2.911\dots}{\sqrt{8}}} \\
 &= 3.0359\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 8 - 1 = 7$.

Function	t.dist.rt	Answer
Field 1	3.0359....	0.0095
Field 2	7	

So the p -value = 0.0095.

Conclusion:

Because p -value = 0.0095 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that, on average, the pain sensory measurements are lower after hypnosis.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on the order of subtraction, so we must decide on the order of subtraction before writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the pain sensory measurements after hypnosis. That is, the average pain sensory measurement is the same before and after hypnosis.
3. For the alternative hypothesis, we are testing that the **after** score is lower than the **before** score. In other words, **before** > **after**. Because we calculated the differences as **before**-**after**, **before** > **after** means **before**-**after** > 0. So the alternative hypothesis is $\mu_D > 0$, the claim that the **before** score is larger than the **after** score (or the **after** score is lower than the **before** score).

4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.
5. The p -value of 0.0095 is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the after score is, on average, lower than the before score.

EXAMPLE

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Nine patients were selected for the new diet and their cholesterol was measured before and after starting the new diet. The results are recorded in the table below. Assume the differences have a normal distribution. At the 5% significance level, was the new diet, on average, successful in lowering patients' cholesterol?

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Solution:

We start by calculating out the differences in the paired data. We will calculate the differences as **after-before**.

Subject	Before	After	Difference
A	209	199	-10
B	210	207	-3
C	205	189	-16
D	198	209	11
E	216	217	1
F	217	202	-15
G	238	211	-27
H	240	223	-17
I	222	201	-21

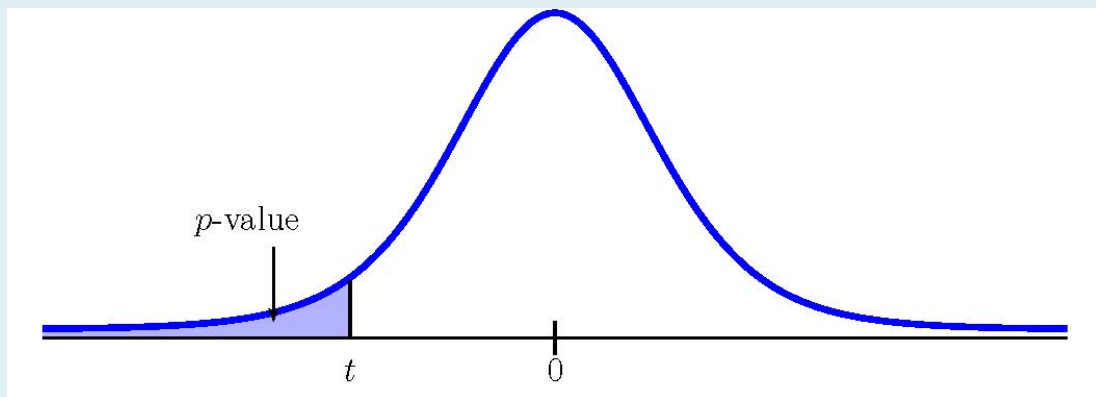
From the difference column, we have $n_D = 9$, $\bar{x}_D = -10.777\dots$, and $s_D = 11.861\dots$

Hypotheses:

$$H_0: \mu_D = 0 \quad H_a: \mu_D < 0$$

***p*-value:**

This is a test on the mean difference in matched samples, so we use a *t*-distribution to calculate the *p*-value. Because the alternative hypothesis is a $<$, the *p*-value is the area in the left tail of the distribution.



To use the **t.dist** function, we need to calculate out the *t*-score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{-10.777\dots - 0}{\frac{11.861\dots}{\sqrt{9}}} \\
 &= -2.725\dots
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 9 - 1 = 8$.

Function	t.dist	Answer
Field 1	-2.725...	0.0130
Field 2	8	
Field 3	true	

So the p -value = 0.0130.

Conclusion:

Because p -value = 0.0130 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that, on average, the new diet lowered the patients' cholesterol levels.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on the order of subtraction, so we must decide on the order of subtraction before writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the patients' cholesterol level. That is, the average cholesterol level is the same before and after the diet.
3. For the alternative hypothesis, we are testing that the **after** score is lower than the **before** score. In other words, **after < before**. Because we calculated the differences as **after-before**, **after < before** means **after-before < 0**. So, the alternative hypothesis is $\mu_D < 0$, the claim that the **after** score is lower than the **before** score.
4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any

round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. Use Excel to do the calculations, and then click on the cells in subsequent calculations.

- The p -value of 0.0224 is a small probability compared to the significance level, and so is unlikely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the after score is, on average, lower than the before score.

EXAMPLE

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The results from their throws are recorded in the table below. Assume the differences are normally distributed. At the 5% significance level, is there a difference in the average distance for the dominant versus weaker hand?

Distance (in feet)	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Solution:

We start by calculating out the differences in the paired data. We will calculate the differences as **dominant-weaker**.

Student	Dominant	Weaker	Difference
1	30	28	2
2	26	14	12
3	34	27	7
4	17	18	-1
5	19	17	2
6	26	26	0
7	20	16	4

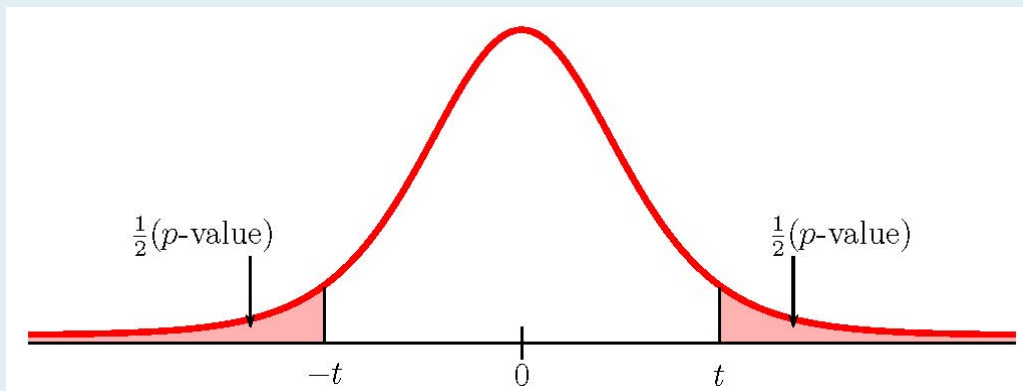
From the difference column, we have $n_D = 7$, $\bar{x}_D = 3.714\dots$, and $s_D = 4.498\dots$

Hypotheses:

$$\begin{array}{l} H_0: \mu_D = 0 \\ H_a: \mu_D \neq 0 \end{array}$$

p-value:

This is a test on the mean difference in matched samples, so we use a *t*-distribution to calculate the *p*-value. Because the alternative hypothesis is a \neq , the *p*-value is the sum of area in the tails of the distribution.



To use the **t.dist.2t** function, we need to calculate out the *t*-score:

$$\begin{aligned}
 t &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \\
 &= \frac{3.714... - 0}{\frac{4.498...}{\sqrt{7}}} \\
 &= 2.184...
 \end{aligned}$$

The degrees of freedom for the t -distribution is $n_D - 1 = 7 - 1 = 6$.

Function	t.dist.2t	Answer
Field 1	2.184....	0.0716
Field 2	6	

So the p -value = 0.0716.

Conclusion:

Because p -value = 0.0716 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that there a difference in the average distance for the dominant versus weaker hand.

NOTES

1. Before writing down the hypotheses, decide on the order of subtraction for calculating the differences. In a matched sample experiment, the form of the alternative hypothesis depends on order of subtraction, so we must decide on the order of subtraction before writing down the hypotheses.
2. The null hypothesis $\mu_D = 0$ is the claim that there is no difference in the average distance. That is, the average distance is the same for both hands.
3. For the alternative hypothesis, we are testing that there is a difference in the dominant hand and weaker hand distances. In other words, **dominant \neq weaker**. So, the alternative hypothesis is $\mu_D \neq 0$, the claim that there is a difference in the distances.
4. Keep all of the decimals throughout the calculation (i.e. in the t -score, etc.) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. Use Excel to do the calculations, and then click on the cells in

subsequent calculations.

5. The p -value of 0.0716 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, on average, the distances are the same for both hands.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=211#oembed-1>

Watch this video: Hypothesis Testing for Matched/Paired Samples by ExcellIsFun [20:48]

Concept Review

The general form of a confidence interval for the mean difference of matched samples is

$$\text{Lower Limit} = \bar{x}_D - t \times \frac{s_D}{\sqrt{n_D}}$$

$$\text{Upper Limit} = \bar{x}_D + t \times \frac{s_D}{\sqrt{n_D}}$$

where t is the positive t -score of the t -distribution with $n_D - 1$ degrees of freedom so the area under the t -distribution in between $-t$ and t is C .

The hypothesis test for matched samples is a well established process:

1. Write down the null and alternative hypotheses in terms of the mean difference μ_D .
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the t -distribution.

Because the population standard deviation is unknown, we use the t -distribution to find the

$$p\text{-value with } t = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} \text{ and } df = n_D - 1.$$

5. Compare the p -value to the significance level and state the outcome of the test.
 6. Write down a concluding sentence specific to the context of the question.
-

Attribution

“10.4 Matched or Paired Samples“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

9.5 STATISTICAL INFERENCE FOR TWO POPULATION PROPORTIONS

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population proportions.
- Conduct and interpret hypothesis tests for two population proportions.

Similar to comparing two population means, the comparison of two population proportions is very common. Often, we want to find out if the two populations under study have the same proportion or if there is some difference in the two population proportions. Unlike two population means, we can only approach the comparison of two population proportions using independent samples. Recall that two populations are **independent** if the sample taken from population 1 is not related in anyway to the sample taken from population 2. In this situation, any relationship between the samples or populations is entirely coincidental.

Throughout this section, we will use subscripts to identify the values for the proportions and sample sizes for the two populations:

Symbol for:	Population 1	Population 2
Population Proportion	p_1	p_2
Sample Size	n_1	n_2
Sample Proportion	\hat{p}_1	\hat{p}_2
Number of Items in Sample with Characteristic of Interest	x_1	x_2

In order to construct a confidence interval or conduct a hypothesis test on the difference in two

population proportions ($p_1 - p_2$), we need to use the distribution of the difference in the sample proportions $\hat{p}_1 - \hat{p}_2$:

- The mean of the distribution of the difference in the sample proportions is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$.

- The standard deviation of the distribution of the difference in the sample proportions is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}.$$

- The distribution of the difference in the sample proportions is normal if $n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$.

- Assuming the distribution of the difference of the sample proportions is normal, the z -score

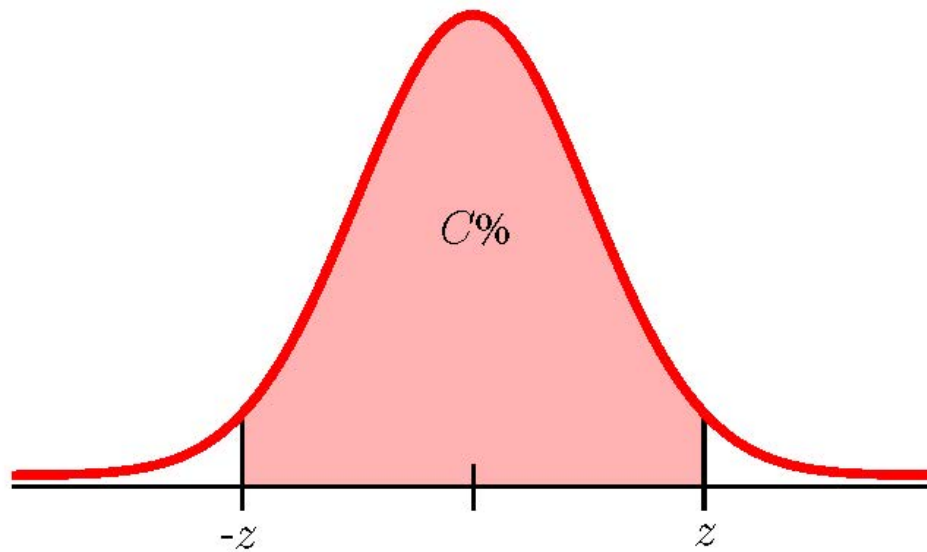
$$\text{is } z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}}.$$

Constructing a Confidence Interval for the Difference in Two Population Proportions

Suppose a sample of size n_1 with sample proportion \hat{p}_1 is taken from population 1 and a sample of size n_2 with sample proportion \hat{p}_2 is taken from population 2. The limits for the confidence interval with confidence level C for the difference in the population proportions $p_1 - p_2$ are:

$$\begin{aligned} \text{\mbox{Lower Limit}} &= \hat{p}_1 - \hat{p}_2 - z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}} \\ \text{\mbox{Upper Limit}} &= \hat{p}_1 - \hat{p}_2 + z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}} \end{aligned}$$

where z is the positive z -score of the standard normal distribution so that the area under the curve in between $-z$ and z is $C\%$.



NOTES

1. In order to construct the confidence interval for the difference in two population proportions, we need to check that the normal distribution applies. This means that we need to check that $n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$.
2. Because the population proportions p_1 and p_2 are often unknown, we replace the values of the population proportions with the sample proportions \hat{p}_1 and \hat{p}_2 in the normal distribution check. That is, when the population proportions are unknown, we check $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies.

CALCULATING THE **Formula does not parse**-SCORE FOR A

CONFIDENCE INTERVAL IN EXCEL

To find the z -score to construct a confidence interval with confidence level C , use the **norm.s.inv(area to the left of z)** function.

- For **area to the left of z** , enter the **entire** area to the left of the z -score you are trying to find.

For a confidence interval, the area to the left of z is $C + \frac{1 - C}{2}$.

The output from the **norm.s.inv** function is the value of the z -score needed to construct the confidence interval.

NOTE

The **norm.s.inv** function requires that we enter the **entire** area to the **left** of the unknown z -score. This area includes the confidence level (the area in the middle of the distribution) plus the remaining area in the left tail.

EXAMPLE

A marketing company places an advertisement for a new brand of deodorant on two different platforms: television and social media. The company wants to study the proportion of people who remembered seeing the advertisement two hours later. In a sample of 200 people who saw the

advertisement on television, 74 remembered seeing it two hours later. In a sample of 300 people who saw the advertisement on social media, 129 remembered seeing it two hours later.

1. Construct a 98% confidence interval for the difference in the proportion of people from the two different platforms that remember seeing the advertisement two hours later.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that the proportion of people from social media who remember seeing the advertisement two hours later is greater than the proportion of people from television? Explain.

Solution:

1. Let television be population 1 and social media be population 2. From the question we have the following information:

Television	Social Media
$n_1 = 200$	$n_2 = 300$
$\hat{p}_1 = \frac{74}{200} = 0.37$	$\hat{p}_2 = \frac{129}{300} = 0.43$

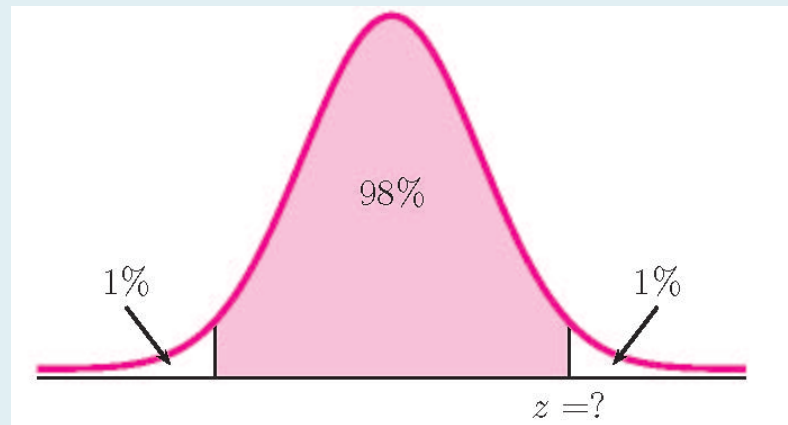
Before constructing the confidence interval, we check that the normal distribution applies:

$$\begin{aligned}
 n_1 \times \hat{p}_1 &= 200 \times 0.37 = 74 \geq 5 \\
 n_1 \times (1 - \hat{p}_1) &= 200 \times (1 - 0.37) = 126 \geq 5 \\
 n_2 \times \hat{p}_2 &= 300 \times 0.43 = 129 \geq 5 \\
 n_2 \times (1 - \hat{p}_2) &= 300 \times (1 - 0.43) = 171 \geq 5
 \end{aligned}$$

To find the confidence interval, we need to find the z -score for the 98% confidence interval.

This means that we need to find the z -score so that the entire area to the left of z is

$$0.98 + \frac{1 - 0.98}{2} = 0.99.$$



Function	norm.s.inv	Answer
Field 1	0.99	2.3263...

So $z = 2.3263\dots$. The 98% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \hat{p}_1 - \hat{p}_2 - z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.37 - 0.43 - 2.3263\dots \times \sqrt{\frac{0.37(1-0.37)}{200} + \frac{0.43(1-0.43)}{300}} \\ &= -0.1636 \\ \text{Upper Limit} &= \hat{p}_1 - \hat{p}_2 + z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.37 - 0.43 + 2.3263\dots \times \sqrt{\frac{0.37(1-0.37)}{200} + \frac{0.43(1-0.43)}{300}} \\ &= 0.0436 \end{aligned}$$

- We are 98% confident that the difference in the proportion of people from the two platforms that remember seeing the advertisement two hours later is between -16.36% and 4.36%.
- Because 0 is inside the confidence interval, it suggests that the difference in the proportions $p_1 - p_2$ is 0. That is, $p_1 - p_2 = 0$. This suggests that the two proportions are equal. So the proportion of people from social media who remember seeing the advertisement two hours is not greater than the proportion of people from television.

NOTES

1. Because the population proportions are unknown, we use the sample proportions in the check for normality.
2. When calculating the limits for the confidence interval keep all of the decimals in the z -score and other values throughout the calculation. This will ensure that there is no round-off error in the answers. You can use Excel to do the calculation of the limits, clicking on the cell containing the z -score and any other values, to ensure that all of the decimal places are used in the calculation.
3. The limits for the confidence interval are percents. For example, the upper limit of 0.0436 is the decimal form of a percent: 4.36%.
4. When writing down the interpretation of the confidence interval, make sure to include the confidence level, the actual difference in the population proportions captured by the confidence interval (i.e. be specific to the context of the question), and express the limits as percents.

Steps to Conduct a Hypothesis Test for the Difference in Two Population Proportions

1. Write down the null hypothesis that there is no difference in the population proportions:

$$\begin{array}{l} H_0: p_1 - p_2 = 0 \end{array}$$

The null hypothesis is always the claim that the two population proportions are equal ($p_1 = p_2$).

2. Write down the alternative hypotheses in terms of the difference in the population proportions. The alternative hypothesis will be **one** of the following:

$$\begin{array}{l} H_a: p_1 - p_2 < 0 \quad \text{and} \quad (p_1 < p_2) \\ H_a: p_1 - p_2 > 0 \quad \text{and} \quad (p_1 > p_2) \\ H_a: p_1 - p_2 \neq 0 \quad \text{and} \quad (p_1 \neq p_2) \end{array}$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.

4. Collect the sample information for the test and identify the significance level.
5. Check the conditions $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies. Use the normal distribution to find the p -value (the area in the corresponding tail) for the test. The z -score is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\overline{p} \times (1 - \overline{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

6. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

NOTES

1. Because the population proportions p_1 and p_2 are often unknown, we replace the values of the population proportions with the sample proportions \hat{p}_1 and \hat{p}_2 in the normal distribution check. That is, when the population proportions are unknown, we check $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies to the calculation of the p -value.
2. Because we are testing the equality of the two population proportions, the z -score for the hypothesis test uses a pooled sample proportion \bar{p} . The pooled sample proportion \bar{p} combines the sample data to create an estimate of the overall proportion of success.

USING EXCEL TO CALCULATE THE P -VALUE FOR A HYPOTHESIS TEST ON TWO INDEPENDENT POPULATION PROPORTIONS

The p -value for a hypothesis test on the difference in two population proportions is the area in the tail(s) of the normal distribution, assuming that the conditions for using a normal distribution are met ($n_1 \times p_1 \geq 5$, $n_1 \times (1 - p_1) \geq 5$, $n_2 \times p_2 \geq 5$ and $n_2 \times (1 - p_2) \geq 5$).

The p -value is the area in the tail(s) of a normal distribution, so the **norm.dist(x, μ , σ , logic operator)** function can be used to calculate the p -value.

- For x , enter the value for $\hat{p}_1 - \hat{p}_2$.
- For μ , enter 0, the value of $p_1 - p_2$ from the null hypothesis. This is the mean of the distribution of the differences in the sample proportions.
- For σ , enter the value of $\sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$. The value for σ is the bottom part of the z -score used in the hypothesis test.
- For the **logic operator**, enter **true**. Note: Because we are calculating the area under the curve, we always enter true for the logic operator.

As with the previous chapter, use the appropriate technique with the **norm.dist** function to find the area in the left-tail, the area in the right-tail or the sum of the area in tails.

EXAMPLE

A cell phone company claimed that iPhones are more popular with adults 30 years old or younger

than with adults over 30 years old. A consumer advocacy group wants to test this claim. In a sample of 1340 adults 30 years old or younger, 134 own an iPhone. In a sample of 250 adults over the age of 30, 15 own an iPhone. At the 5% significance level, is the proportion of adults 30 years old or younger who own an iPhone greater than the proportion of adults over the age of 30 who own an iPhone?

Solution:

Let adults 30 years old or younger be population 1 and adults over 30 years old be population 2. From the question, we have the following information:

30 Years or Younger	Over 30 Years
$n_1 = 1340$	$n_2 = 250$
$x_1 = 134$	$x_2 = 15$
$\hat{p}_1 = \frac{134}{1340} = 0.1$	$\hat{p}_2 = \frac{15}{250} = 0.05$

Hypotheses:

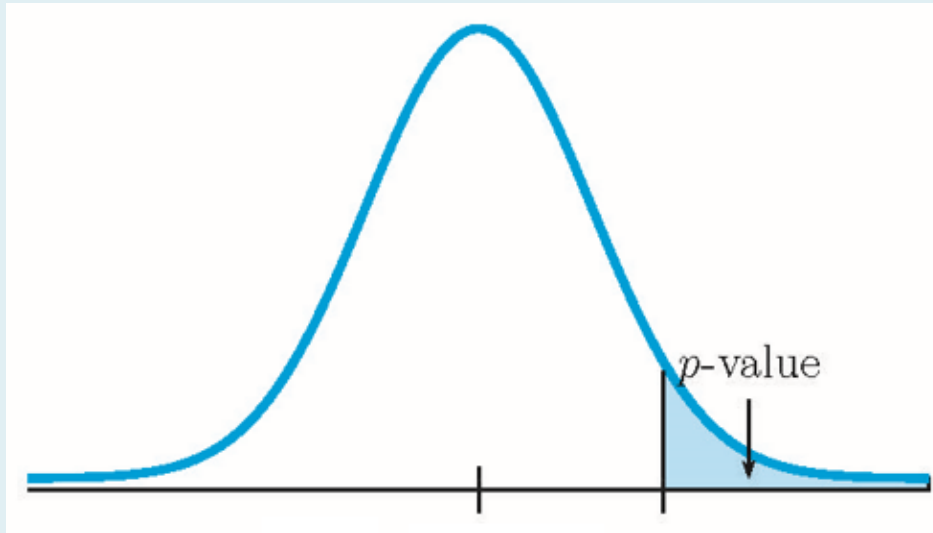
$$\begin{array}{l} H_0: p_1 - p_2 = 0 \\ H_a: p_1 - p_2 > 0 \end{array}$$

p-value:

Before calculating the p -value, we check that the normal distribution applies:

$$\begin{aligned} n_1 \times \hat{p}_1 &= 1340 \times 0.1 = 134 \geq 5 \\ n_1 \times (1 - \hat{p}_1) &= 1340 \times (1 - 0.1) = 1206 \geq 5 \\ n_2 \times \hat{p}_2 &= 250 \times 0.05 = 15 \geq 5 \\ n_2 \times (1 - \hat{p}_2) &= 250 \times (1 - 0.05) = 235 \geq 5 \end{aligned}$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies and so we use a normal distribution to calculate the p -value. Because the alternative hypothesis is a $>$, the p -value is the area in the right tail of the distribution.



The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{134 + 15}{1340 + 250} \\ &= \frac{149}{1590} \\ &= 0.09371\dots\end{aligned}$$

Function	1-norm.dist	Answer
Field 1	0.1-0.05	0.0232
Field 2	0	
Field 3	sqrt(0.09371... *(1-0.09371...)*(1/1340+1/250))	
Field 4	true	

So the p -value = 0.0232.

Conclusion:

Because p -value = 0.0232 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the proportion of adults 30 years old or younger who own an iPhone is greater than the proportion of adults over the age of 30 who own an iPhone.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that the proportion of adults 30 or younger with an iPhone equals the proportion of adults over 30 with an iPhone. That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 > 0$ is the claim that the proportion of adults 30 or younger with an iPhone is greater than the proportion of adults over 30 with an iPhone ($p_1 > p_2$).
3. Make sure to keep all of the decimal places throughout the calculation to avoid any round-off error in the p -value. Perform the calculations of the sample proportions and the pooled sample proportion \bar{p} in Excel and then click on the corresponding cells when completing the fields in the norm.dist function.
4. The p -value is the area in the right tail of the normal distribution. In the calculation of the p -value:

- The function is 1-norm.dist because we are finding the area in the right tail of a normal distribution.
- Field 1 is the value of $\hat{p}_1 - \hat{p}_2 = 0.1 - 0.05$.
- Field 2 is 0, the value of $p_1 - p_2$ from the null hypothesis. Remember, we run the test assuming the null hypothesis is true, so that means we assume $p_1 - p_2 = 0$.
- Field 3 is the value of

$$\sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{0.09371... \times (1 - 0.09371...) \times \left(\frac{1}{1340} + \frac{1}{250} \right)}$$

5. The p -value of 0.0232 is a small probability compared to the significance level, and so is unlikely to happen that assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the proportion of adults 30 years old or younger who own an iPhone is greater than the proportion of adults over the age of 30 who own an iPhone.

EXAMPLE

Two types of medication for hives are tested to determine if there is a difference in the proportions of adult patient reactions. In a sample of 200 adults given medication A, 20 still had hives 30 minutes after taking the medication. In a sample of 200 adults given medication B, 12 still had hives 30 minutes after taking the medication. At the 1% significance level, is there a difference in the proportion of adults who still have hives 30 minutes after taking medications?

Solution:

Let medication A be population 1 and medication B be population 2. From the question, we have the following information:

Medication A	Medication B
$n_1 = 200$	$n_2 = 200$
$x_1 = 20$	$x_2 = 12$
$\hat{p}_1 = \frac{20}{200} = 0.1$	$\hat{p}_2 = \frac{12}{200} = 0.06$

Hypotheses:

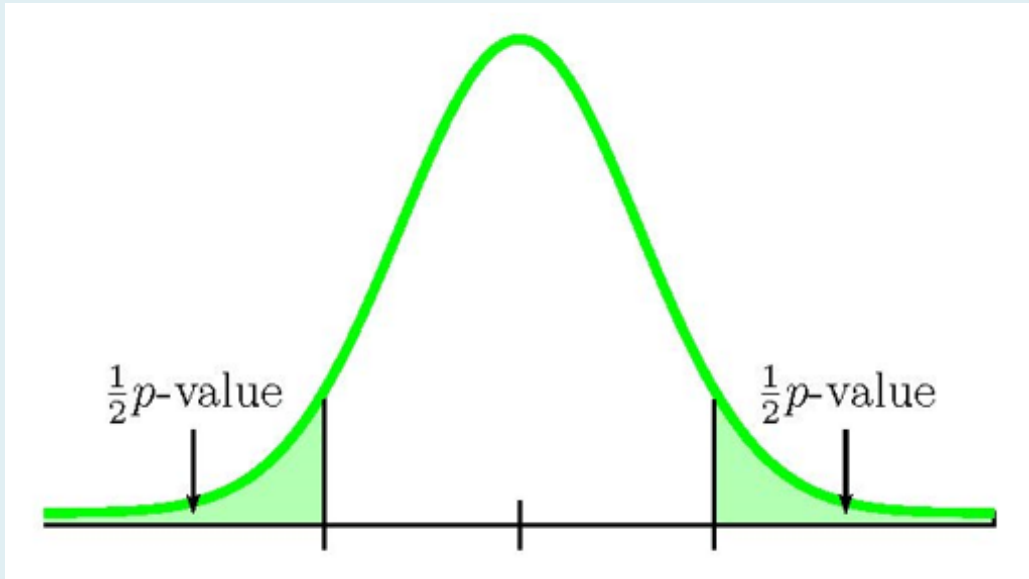
$$\begin{array}{l} H_0: p_1 - p_2 = 0 \\ H_a: p_1 - p_2 \neq 0 \end{array}$$

p-value:

Before calculating the p -value, we check that the normal distribution applies:

$$\begin{aligned} n_1 \times \hat{p}_1 &= 200 \times 0.1 = 20 \geq 5 \\ n_1 \times (1 - \hat{p}_1) &= 200 \times (1 - 0.1) = 180 \geq 5 \\ n_2 \times \hat{p}_2 &= 200 \times 0.06 = 12 \geq 5 \\ n_2 \times (1 - \hat{p}_2) &= 200 \times (1 - 0.06) = 188 \geq 5 \end{aligned}$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies and so we use a normal distribution to calculate the p -value. Because the alternative hypothesis is a \neq , the p -value is the sum of the area in the two tails of the distribution.



We need to know if the sample information relates to the left or right tail because that will determine how we calculate out the area of that tail using the normal distribution. In this case, $\hat{p}_1 > \hat{p}_2$ ($0.1 > 0.06$), so the sample information relates to the right tail of the normal distribution. This means that we will calculate out the area in the right tail using **1-norm.dist**. However, this is a two-tailed test where the p -value is the sum of the area in the two tails and the area in the right tail is only one half of the p -value. The area in the right tail equals the area in the left tail and the p -value is the sum of these two areas.

The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{20 + 12}{200 + 200} \\ &= \frac{32}{400} \\ &= 0.08\end{aligned}$$

Function	1-norm.dist	Answer
Field 1	0.1-0.06	0.0702
Field 2	0	
Field 3	sqrt(0.08*(1-0.08)*(1/200+1/200))	
Field 4	true	

So the area in the right tail is 0.0702, which means $\frac{1}{2}(p\text{-value}) = 0.0702$. This is also the area in the left tail, so

$$p\text{-value} = 0.0702 + 0.0702 = 0.1404$$

Conclusion:

Because $p\text{-value} = 0.1404 > 0.01 = \alpha$, we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that there is a difference in the proportion of adults who still have hives 30 minutes after taking medication.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that there is no difference in the proportion of adults with hives 30 minutes after taking the medications. That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 \neq 0$ is the claim that there is a difference in the proportion of adults with hives 30 minutes after taking the medications ($p_1 \neq p_2$).
3. In a two-tailed hypothesis test that uses the normal distribution, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.
 - If the sample proportion \hat{p}_1 is less than the sample proportion \hat{p}_2 ($\hat{p}_1 < \hat{p}_2$), the sample information belongs to the **left tail**.

- We use $\text{norm.dist}(\hat{p}_1 - \hat{p}_2, 0, \sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{true})$

to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.

- If the sample proportion \hat{p}_1 is greater than the sample proportion \hat{p}_2 ($\hat{p}_1 > \hat{p}_2$), the sample information belongs to the **right tail**.

- We use $\text{1-norm.dist}(\hat{p}_1 - \hat{p}_2, 0, \sqrt{\bar{p} \times (1 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{true})$

to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.

4. The p -value of 0.1404 is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, there is no difference in the proportion of adults with hives 30 minutes after taking the medications.

EXAMPLE

A valve manufacturer recently launched a new valve, Valve A, and they want to claim that the proportion of their valves that fail under 4500 psi is the smallest of all the other valves on the market. The manufacturer decides to compare Valve A with the most popular valve on the market, Valve B. In a sample of 100 Valve A's, 6 failed at 4500 psi. In a sample of 150 Valve B's, 16 failed at

4500 psi. At the 5% significance level, is the proportion of Valve As that fail under 4500 psi less than the proportion of Valve Bs that fail under 4500 psi?

Solution:

Let Valve A be population 1 and Valve B be population 2. From the question, we have the following information:

Valve A	Valve B
$n_1 = 100$	$n_2 = 150$
$x_1 = 6$	$x_2 = 16$
$\hat{p}_1 = \frac{6}{100} = 0.06$	$\hat{p}_2 = \frac{16}{150} = 0.1066\dots$

Hypotheses:

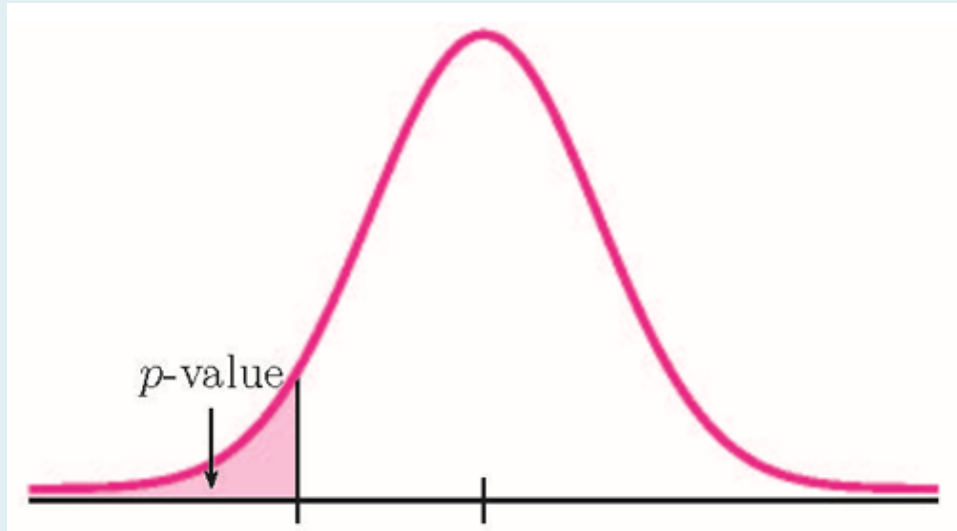
$$\begin{array}{l} H_0: p_1 - p_2 = 0 \\ H_a: p_1 - p_2 < 0 \end{array}$$

p-value:

Before calculating the p -value, we check that the normal distribution applies:

$$\begin{aligned} n_1 \times \hat{p}_1 &= 100 \times 0.06 = 6 \geq 5 \\ n_1 \times (1 - \hat{p}_1) &= 100 \times (1 - 0.06) = 94 \geq 5 \\ n_2 \times \hat{p}_2 &= 150 \times 0.1066\dots = 16 \geq 5 \\ n_2 \times (1 - \hat{p}_2) &= 150 \times (1 - 0.1066\dots) = 134 \geq 5 \end{aligned}$$

Because $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$, the normal distribution applies and so we use a normal distribution to calculate the p -value. Because the alternative hypothesis is a $<$, the p -value is the area in the left tail of the distribution.



The pooled sample proportion is:

$$\begin{aligned}\bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{6 + 16}{100 + 150} \\ &= \frac{22}{250} \\ &= 0.088\end{aligned}$$

Function	norm.dist	Answer
Field 1	0.06-0.1066...	0.1010
Field 2	0	
Field 3	$\text{sqrt}(0.088 * (1 - 0.088) * (1/100 + 1/150))$	
Field 4	true	

So the $p\text{-value} = 0.1010$.

Conclusion:

Because $p\text{-value} = 0.1010 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that the proportion of Valve As that fail under 4500 psi less than the proportion of Valve Bs that fail under 4500 psi.

NOTES

1. The null hypothesis $p_1 - p_2 = 0$ is the claim that the proportion of valves that fail under 4500 psi is the same for both valves. That is, the two populations have the same proportion.
2. The alternative hypothesis $p_1 - p_2 < 0$ is the claim that the proportion of Valve As that fail under 4500 psi is less than the proportion of Valve Bs that fail under 4500 psi ($p_1 < p_2$).
3. Make sure to keep all of the decimal places throughout the calculation to avoid any round-off error in the p -value. Perform the calculations of the sample proportions and the pooled sample proportion \bar{p} in Excel and then click on the corresponding cells when completing the fields in the norm.dist function.
4. The p -value of 0.1010 is a large probability compared to the significance level, and so is likely to happen assuming that the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the proportion of Valve As that fail under 4500 psi equals the proportion of Valve Bs that fail under 4500 psi. For the company, this means that they could not claim that the proportion of their valves that fail under 4500 psi is the smallest of all the other valves on the market.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=213#oembed-1>

Watch this video: Excel 2013 Statistical Analysis #71: Inference About Difference Between 2 Pop. Proportions Z Method by ExcelIsFun [28:03]

Concept Review

The general form of a confidence interval for the difference in two population proportions is

$$\text{Lower Limit} = \hat{p}_1 - \hat{p}_2 - z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

$$\text{Upper Limit} = \hat{p}_1 - \hat{p}_2 + z \times \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

where z is the positive z -score of the standard normal distribution so the area under the normal distribution in between $-z$ and z is C .

The hypothesis test for the difference in two population proportions with is a well established process:

1. Write down the null and alternative hypotheses in terms of the differences in the population proportions $p_1 - p_2$.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Check that $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1 - \hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$ and $n_2 \times (1 - \hat{p}_2) \geq 5$ to verify that the normal distribution applies.
5. Find the p -value (the area in the corresponding tail) for the test using the normal distribution.
6. Compare the p -value to the significance level and state the outcome of the test.
7. Write down a concluding sentence specific to the context of the question.

Attribution

“10.3 Comparing Two Independent Population Proportions“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

9.6 EXERCISES

1. The known standard deviation in salary for all mid-level professionals in the financial industry is \$11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is \$80,000. The sample mean salary for mid-level professionals in Company B is \$96,000.

- a. Construct a 99% confidence interval for the difference in the mean salary for mid-level professionals at the two companies.
- b. Interpret the confidence interval in part (a).
- c. Is it reasonable to claim that mean salary for mid-level professionals the same at the two companies? Explain.

2. It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four. At the 5% significance level test if the average grade on an English essay is higher for females than males.

3. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. At the 5% significance level, is the proportion of conifers in the United States greater than the proportion of conifers in Mexico?

4. A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. Both populations have normal distributions. At the 5% significance level, determine if the average amount of sugar in Beverage B is greater than Beverage A.

5. The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same? Use a 5% significance level.

6. A student at a four-year college claims that mean enrollment at four-year colleges is higher than at two-year colleges in the United States. Two surveys are conducted. Of the 35 two-year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191. At the 5% significance level, is the mean enrollment at four-year colleges higher than at two-year colleges?

7. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were \$46,100 and \$46,700, respectively. Their standard deviations were \$3,450 and \$4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary. Use a 5% significance level.

8. Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean. Use a 5% significance level.

9. Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The

mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

10. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds. Test at a 5% significance level.

- Construct a 94% confidence interval for the difference in the mean weight loss for the powder and liquid diets.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the mean weight loss with the powder diet is less than the liquid diet? Explain.

11. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. The table shows the result. Scouters believe that Rodriguez pitches a speedier fastball. At the 1% significance level, what is your conclusion?

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

12. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller. At the 1% significance level, what is your conclusion?

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

13. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. 15 pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point. At the 1% significance level, what is your conclusion?

	Sample Mean Melting Temperatures (°F)	Population Standard Deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

14. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. At the 5% significance level, what is your conclusion?

15. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was \$679. For 23 teenage girls, it was \$559. From past years, it is known that the population standard deviation for each group is \$180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls. Use a 5% significance level.

16. A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were \$947 and \$1,011, respectively. The population standard deviations are known to be \$254 and \$87, respectively.

- Construct a 96% confidence interval for the difference in the mean amount students spend on texts at university and community college.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the mean amount students spend on texts is the same at university and community college? Explain.

17. Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg.

- Construct a 95% confidence interval for the difference in the average miles-per-gallon in hybrid and non-hybrid cars.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the average mpg for hybrid cars is higher than non-hybrid cars? Explain.

18. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement “I’m pleased with the way we divide the responsibilities for childcare.” The ratings went from one (strongly agree) to five (strongly disagree). The table below contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband’s versus the wife’s satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife). Use a 5% significance level.

Wife’s Score	2	2	3	3	4	2	1	1	2	4
Husband’s Score	2	2	1	3	2	1	1	1	2	4

19. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS₁ had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS₂ had system failures within the first eight hours of operation. OS₂ is believed to be more stable (have fewer crashes) than OS₁. At the 5% significance level, is there a difference in the proportions of system failures?

20. A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them. At the 5% significance level, is the proportion of drug and alcohol abuse higher locally than nationally?

21. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system . It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

22. Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in table. Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

23. Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. The table details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 year olds	30 years old and older
Own a Tablet	69	231
Sample Size	628	2,309

24. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

- Construct a 93% confidence interval for the difference in the proportion of men and women who use smartphones.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the proportion of men who use smartphones is higher than the proportion of women? Explain.

25. We are interested in whether children’s educational computer software costs less, on average, than children’s entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was \$31.14 with a standard deviation of \$4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was \$33.86 with a standard deviation of \$10.87. At the 5% significance level, determine if children’s educational software costs less, on average, than children’s entertainment software.

26. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear.

- Construct a 98% confidence interval for the difference in the proportion of college-age males and females with at least one pierced ear.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the proportion of college-age males with at least one pierced ear equals the proportion of college-age females? Explain.

27. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in the table below. The “before” value is matched to an “after” value, and the differences are calculated. The differences have a normal distribution.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

- Construct a 97% confidence interval for the mean difference in the number of failures before and after the software patch was installed.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the average number of failures did not change after the software patch was installed? Explain.

28. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

- Construct a 99% confidence interval for the mean difference in the number of balls a subject can juggle after the class.
- Interpret the confidence interval in part (a).
- Is it reasonable to claim that the average number of balls a subject can juggle higher after the class? Explain.

29. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. At the 1% significance level, did the medication, on average, lower the patients blood pressure?

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

30. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in the table below. Do you think that their cholesterol levels were significantly lowered? Use a 5% significance level.

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

31. A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are in the table. At the 5% significance level, determine if the average number of breast cancer cases is higher in 2013 than in 2012.

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

32. A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in the table. Test at the 1% level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

Attribution

“Chapter 10 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART X

STATISTICAL INFERENCES USING THE CHI-SQUARE DISTRIBUTION

Chapter Outline

10.1 Introduction to Statistical Inferences Using the Chi-Square Distribution

10.2 The Chi-Square Distribution

10.3 Statistical Inference for a Single Population Variance

10.4 The Goodness-of-Fit Test

10.5 The Test of Independence

10.6 Exercises

10.1 INTRODUCTION TO STATISTICAL INFERENCES USING THE CHI-SQUARE DISTRIBUTION



The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. Photo by Pete, CC BY 4.0.

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

We will now study a new distribution called the χ^2 -distribution. Statistical inferences that use the χ^2 -distribution can help us answer the types of questions posed above. In this chapter, we will

learn the three major applications of the χ^2 -distribution: testing a single population variance, the goodness-of-fit test, and the test of independence.

Attribution

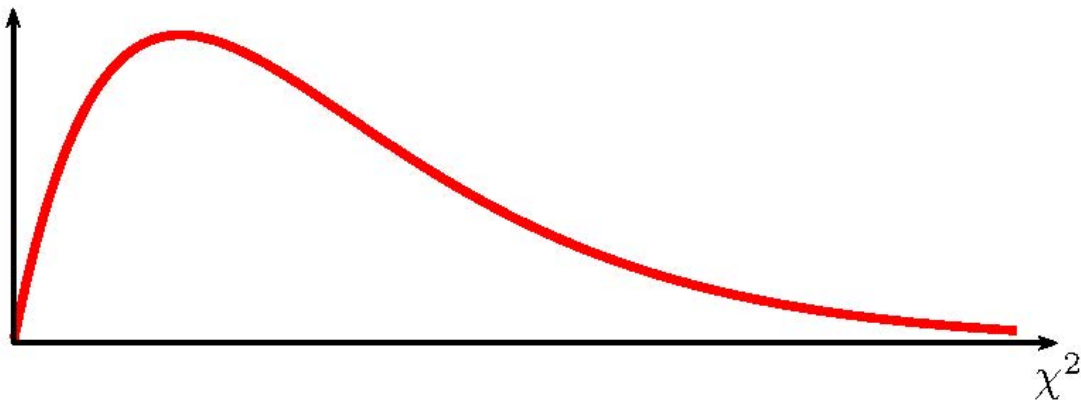
“Chapter 11 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

10.2 THE CHI SQUARE DISTRIBUTION

LEARNING OBJECTIVES

- Find the area under a χ^2 -distribution.
- Find the χ^2 -score for a given area under the curve of a χ^2 -distribution.

The χ^2 -distribution is a continuous probability distribution. The graph of a χ^2 -distribution is shown below.



Properties of the χ^2 -distribution:

- The graph of a χ^2 -distribution is positively-skewed and asymmetrical with a minimum value of 0 and no maximum value.
- A χ^2 -distribution is determined by its degrees of freedom, df . The value of the degrees of freedom depends on how the χ^2 -distribution is used. There is a different χ^2 -distribution for every value of df . As the degrees of freedom increases, the χ^2 -distribution approaches a

normal distribution.

- The total area under the graph of a χ^2 -distribution is 1.
- The mean of a χ^2 -distribution is its degrees of freedom: $\mu = df$.
- The variance of a χ^2 -distribution is twice its degrees of freedom: $\sigma^2 = 2 \times df$.
- The mode of a χ^2 -distribution is $df - 2$. The peak of the graph occurs at the mode.
- Probabilities associated with a χ^2 -distribution are given by the area under the curve of the χ^2 -distribution.

USING EXCEL TO CALCULATE THE AREA UNDER A Formula does not parse -DISTRIBUTION

To find the area in the left tail:

- To find the area under a χ^2 -distribution to the left of a given χ^2 -score, use the **chisq.dist(χ^2 , degrees of freedom, logic operator)** function.
 - For χ^2 , enter the χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
 - For **logic operator**, enter **true**.
- The output from the **chisq.dist** function is the area to the left of the entered χ^2 -score.
- Visit the Microsoft page for more information about the **chisq.dist** function.

To find the area in the right tail:

- To find the area under a χ^2 -distribution to the right of a given χ^2 -score, use the **chisq.dist.rt(χ^2 , degrees of freedom)** function.
 - For χ^2 , enter the χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
- The output from the **chisq.dist.rt** function is the area to the right of the entered χ^2

-score.

- Visit the Microsoft page for more information about the **chisq.dist.rt** function.

EXAMPLE

Consider a χ^2 -distribution with 12 degrees of freedom.

1. Find the area under the χ^2 -distribution to the left of $\chi^2 = 3.71$.
2. Find the area under the χ^2 -distribution to the right of $\chi^2 = 6.29$.

Solution:

1.

Function	chisq.dist	Answer
Field 1	3.71	0.0119
Field 2	12	
Field 3	true	

2.

Function	chisq.dist.rt	Answer
Field 1	6.72	0.8755
Field 2	12	

USING EXCEL TO CALCULATE Formula does not parse -SCORES

To find the χ^2 -score for a given left-tail area:

- To find the χ^2 -score for a given area under the χ^2 -distribution to the left of the χ^2 -score, use the **chisq.inv(area to the left, degrees of freedom)** function.
 - For **area to the left**, enter the area to the left of required χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
- The output from the **chisq.inv** function is the value of χ^2 -score so that the area to the left of the χ^2 -score is the entered area.
- Visit the Microsoft page for more information about the **chisq.inv** function.

To find the χ^2 -score for a given right-tail area:

- To find the χ^2 -score for a given area under the χ^2 -distribution to the right of the χ^2 -score, use the **chisq.inv.rt(area to the right, degrees of freedom)** function.
 - For **area to the right**, enter the area to the right of required χ^2 -score.
 - For **degrees of freedom**, enter the value of the degrees of freedom for the χ^2 -distribution.
- The output from the **chisq.inv.rt** function is the value of χ^2 -score so that the area to the right of the χ^2 -score is the entered area.
- Visit the Microsoft page for more information about the **chisq.inv.rt** function.

EXAMPLE

Consider a χ^2 -distribution with 37 degrees of freedom.

1. Find the χ^2 -score so that the area under the χ^2 -distribution to the left of χ^2 is 0.25.
2. Find the χ^2 -score so that the area under the χ^2 -distribution to the right of χ^2 is 0.148.

Solution:

1.

Function	chisq.inv	Answer
Field 1	0.25	30.89
Field 2	37	

2.

Function	chisq.dist.rt	Answer
Field 1	0.148	45.97
Field 2	37	

Concept Review

The χ^2 -distribution is a useful tool for assessment in a series of problem categories. These problem categories include: determining if a data set fits a particular distribution, determining if the distributions of two populations are the same, determining if two categorical variables are independent or dependent, and determining if there is a different variability than expected within a population.

An important parameter in a χ^2 -distribution is the degrees of freedom in a given problem. The χ^2 -distribution curve is skewed to the right, and its shape depends on the degrees of freedom. As the degrees of freedom increases, the curve of a χ^2 -distribution approaches a normal distribution.

Attribution

“11.1 Facts About the Chi-Square Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

10.3 STATISTICAL INFERENCE FOR A SINGLE POPULATION VARIANCE

LEARNING OBJECTIVES

- Calculate and interpret a confidence interval for a population variance.
- Conduct and interpret a hypothesis test on a single population variance.

The mean of a population is important, but in many cases the variance of the population is just as important. In most production processes, quality is measured by how closely the process matches the target (i.e. the mean) and by the variability (i.e. the variance) of the process. For example, if a process is to fill bags of coffee beans, we are interested in both the average weight of the bag and how much variation there is in the weight of the bags. The quality is considered poor if the average weight of the bags is accurate but the variance of the weight of the bags is too high—a variance that is too large means some bags would be too full and some bags would be almost empty.

As with other population parameters, we can construct a confidence interval to capture the population variance and conduct a hypothesis test on the population variance. In order to construct a confidence interval or conduct a hypothesis test on a population variance σ^2 , we need to use the distribution of $\frac{(n-1) \times s^2}{\sigma^2}$. Suppose we have a normal population with population variance σ^2 and a sample of size n is taken from the population. The sampling distribution of $\frac{(n-1) \times s^2}{\sigma^2}$ follows a χ^2 -distribution with $n - 1$ degrees of freedom.

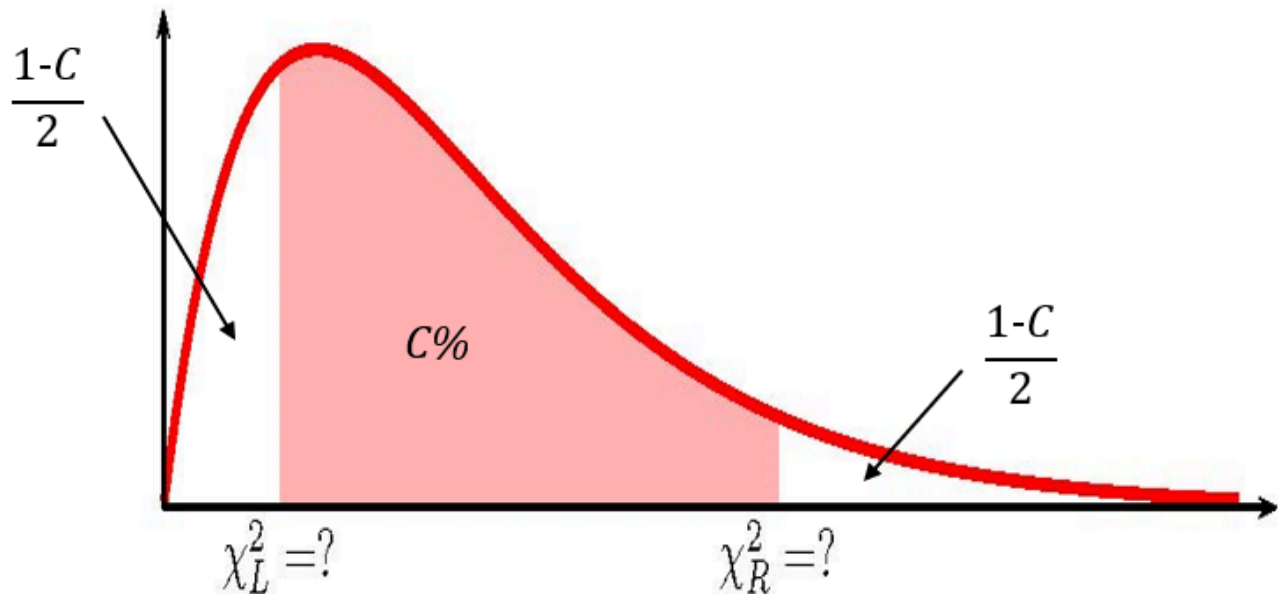
Constructing a Confidence Interval for a Population Variance

To construct the confidence interval, take a random sample of size n from a normally distributed

population. Calculate the sample variance s^2 . The limits for the confidence interval with confidence level C for an unknown population variance σ^2 are

$$\begin{array}{l} \text{Lower Limit} \quad = \quad \frac{(n-1) \times s^2}{\chi^2_R} \\ \text{Upper Limit} \quad = \quad \frac{(n-1) \times s^2}{\chi^2_L} \end{array}$$

where χ^2_L is the χ^2 -score so that the area in the left-tail of the χ^2 -distribution is $\frac{1-C}{2}$, χ^2_R is the χ^2 -score so that the area in the right-tail of the χ^2 -distribution is $\frac{1-C}{2}$ and the χ^2 -distribution has $n - 1$ degrees of freedom.



NOTES

1. Like the other confidence intervals we have seen, the χ^2 -scores are the values that trap $C\%$ of the observations in the middle of the distribution so that the area of each tail is $\frac{1-C}{2}$.
2. Because the χ^2 -distribution is not symmetrical, the confidence interval for a population variance requires that we calculate **two** different χ^2 -scores: one for the left tail and one for the right tail. In Excel, we will need to use both the **chisq.inv** function (for the left tail) and the **chisq.inv.rt** function (for the right tail) to find the two different χ^2 -scores.

- The χ^2 -score for the left tail is part of the formula for the upper limit and the χ^2 -score for the right tail is part of the formula for the lower limit. **This is not a mistake.** It follows from the formula used to determine the limits for the confidence interval.

EXAMPLE

A local telecom company conducts broadband speed tests to measure how much data per second passes between a customer's computer and the internet compared to what the customer pays for as part of their plan. The company needs to estimate the variance in the broadband speed. A sample of 15 ISPs is taken and amount of data per second is recorded. The variance in the sample is 174.

- Construct a 97% confidence interval for the variance in the amount of data per second that passes between a customer's computer and the internet.
- Interpret the confidence interval found in part 1.

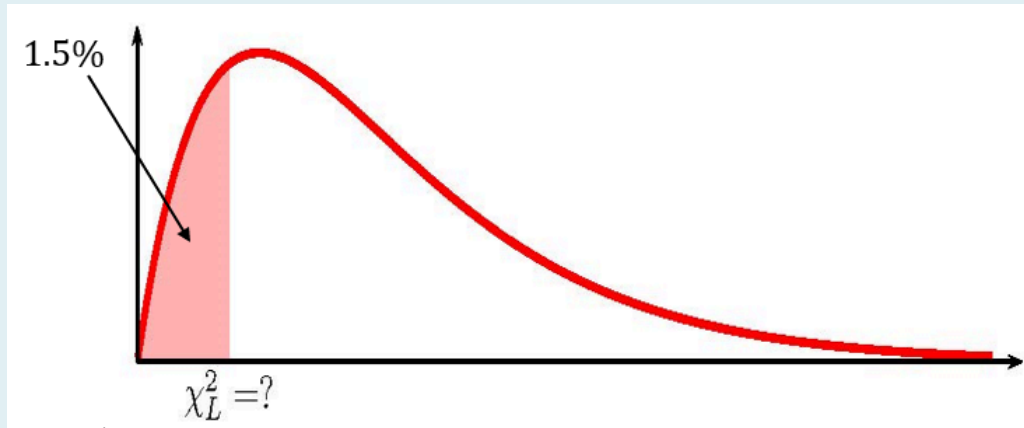
Solution:

- To find the confidence interval, we need to find the χ_L^2 -score for the 97% confidence interval.

This means that we need to find the χ_L^2 -score so that the area in the left tail is

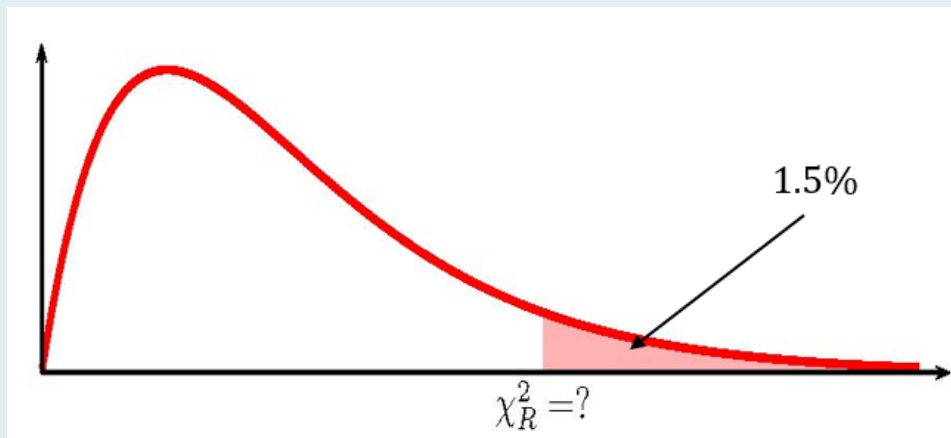
$$\frac{1 - 0.97}{2} = 0.015. \text{ The degrees of freedom for the } \chi^2\text{-distribution is}$$

$$n - 1 = 15 - 1 = 14.$$



Function	chisq.inv	Answer
Field 1	0.015	5.0572...
Field 2	14	

We also need find the χ^2_R -score for the 97% confidence interval. This means that we need to find the χ^2_R -score so that the area in the right tail is $\frac{1 - 0.97}{2} = 0.015$. The degrees of freedom for the χ^2 -distribution is $n - 1 = 15 - 1 = 14$.



Function	chisq.inv.rt	Answer
Field 1	0.015	27.826...
Field 2	14	

So $\chi_L^2 = 5.0572\dots$ and $\chi_R^2 = 27.826\dots$. From the sample data supplied in the question $s^2 = 174$ and $n = 15$. The 97% confidence interval is

$$\begin{array}{l} \boxed{\text{Lower Limit}} \ \&= \ \& \ \frac{(n-1) \times s^2}{\chi^2_R} \ \& \\ \&= \ \& \ \frac{(15-1) \times 174}{27.826\dots} \ \&= \ \& \ 87.54 \ \& \\ \boxed{\text{Upper Limit}} \ \&= \ \& \ \frac{(n-1) \times s^2}{\chi^2_L} \ \&= \ \& \ \frac{(15-1) \times 174}{5.0572\dots} \ \&= \ \& \ 481.69 \ \& \end{array}$$

- We are 97% confident that the variance in the amount of data per second that passes between a customer's computer and the internet is between 87.54 and 481.69.

NOTES

- When calculating the limits for the confidence interval keep all of the decimals in the χ^2 -scores and other values throughout the calculation. This will ensure that there is no round-off error in the answer. You can use Excel to do the calculations of the limits, clicking on the cells containing the χ^2 -scores and any other values.
- When writing down the interpretation of the confidence interval, make sure to include the confidence level and the actual population variance captured by the confidence interval (i.e. be specific to the context of the question). In this case, there are no units for the limits because variance does not have any limits.

Steps to Conduct a Hypothesis Test for a Population Variance

- Write down the null and alternative hypotheses in terms of the population variance σ^2 .
- Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
- Collect the sample information for the test and identify the significance level α .
- Use the χ^2 -distribution to find the p -value (the area in the corresponding tail) for the test. The χ^2 -score and degrees of freedom are

$$\chi^2 = \frac{(n - 1) \times s^2}{\sigma^2} \quad df = n - 1$$

5. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

EXAMPLE

A statistics instructor at a local college claims that the variance for the final exam scores was 25. After speaking with his classmates, one of the class's best students thinks that the variance for the final exam scores is higher than the instructor claims. The student challenges the instructor to prove her claim. The instructor takes a sample of 30 final exams and finds the variance of the scores is 28. At the 5% significance level, test if the variance of the final exam scores is higher than the instructor claims.

Solution:

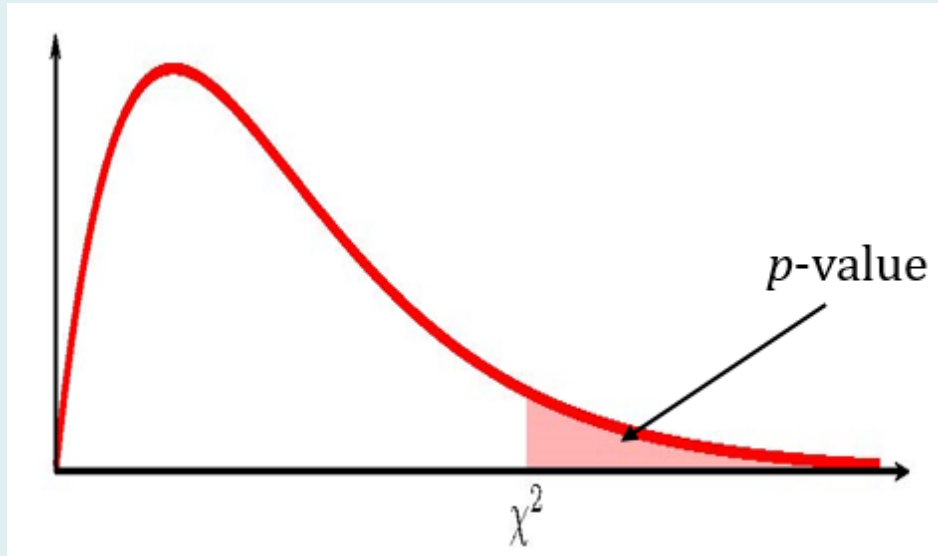
Hypotheses:

$$\begin{array}{l} H_0: \sigma^2 = 25 \\ H_a: \sigma^2 > 25 \end{array}$$

p -value:

From the question, we have $n = 30$, $s^2 = 28$, and $\alpha = 0.05$.

Because the alternative hypothesis is a $>$, the p -value is the area in the right tail of the χ^2 -distribution.



To use the **chisq.dist.rt** function, we need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned}\chi^2 &= \frac{(n-1) \times s^2}{\sigma^2} \\ &= \frac{(30-1) \times 28}{25} \\ &= 32.48\end{aligned}$$

$$\begin{aligned}df &= n - 1 \\ &= 30 - 1 \\ &= 29\end{aligned}$$

Function	chisq.dist.rt	Answer
Field 1	32.48	0.2992
Field 2	29	

So the p -value = 0.2992.

Conclusion:

Because p -value = 0.2992 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that the variance of the final exam scores is higher than 25.

NOTES

1. The null hypothesis $\sigma^2 = 25$ is the claim that the variance on the final exam is 25.
2. The alternative hypothesis $\sigma^2 > 25$ is the claim that the variance on the final exam is greater than 25.
3. There are no units included with the hypotheses because variance does not have any units.
4. The p -value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 32.84$.
In the calculation of the p -value:
 - The function is `chisq.dist.rt` because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the degrees of freedom.
5. The p -value of 0.2992 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variance of the scores on the final exam is most likely 25.

EXAMPLE

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers the waiting times for customers have a standard deviation of 4.5 minutes. At the 5% significance level, determine if the single line changed the variation among the wait times for customers.

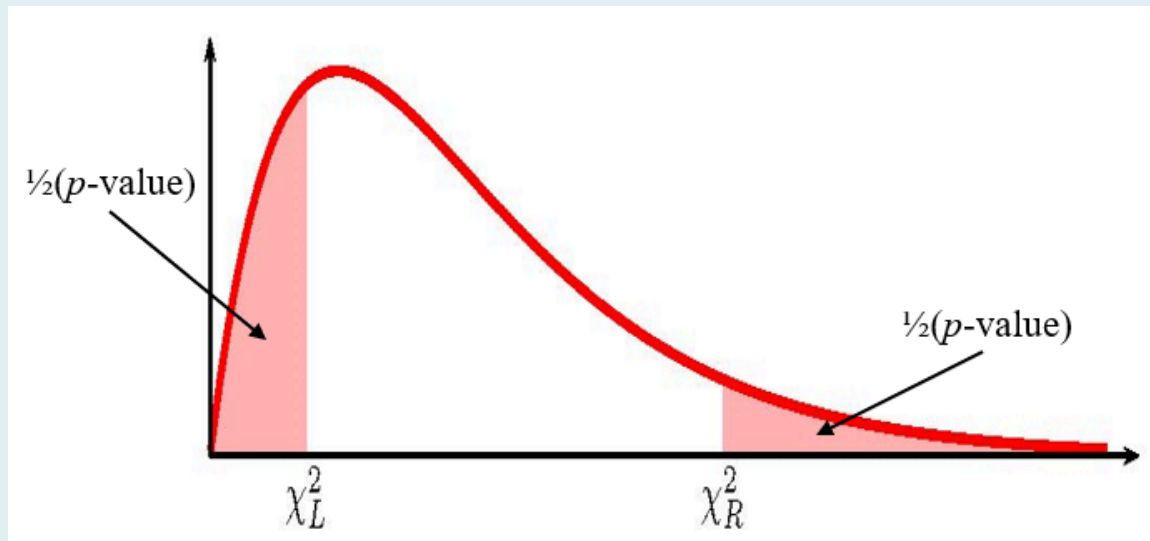
Solution:**Hypotheses:**

$$\begin{array}{l} H_0: \sigma^2 = 51.84 \\ H_a: \sigma^2 \neq 51.84 \end{array}$$

p-value:

From the question, we have $n = 25$, $s^2 = 20.25$, and $\alpha = 0.05$.

Because the alternative hypothesis is a \neq , the p-value is the sum of the areas in the tails of the χ^2 -distribution.



We need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned} \chi^2 &= \frac{(n-1) \times s^2}{\sigma^2} \\ &= \frac{(25-1) \times 20.25}{51.84} \\ &= 9.375 \end{aligned}$$

$$\begin{aligned} df &= n - 1 \\ &= 25 - 1 \\ &= 24 \end{aligned}$$

Because this is a two-tailed test, we need to know which tail (left or right) we have the χ^2 -score for so that we can use the correct Excel function. If $\chi^2 > df - 2$, the χ^2 -score corresponds to the

right tail. If the $\chi^2 < df - 2$, the χ^2 -score corresponds to the left tail. In this case, $\chi^2 = 9.375 < 22 = df - 2$, so the χ^2 -score corresponds to the left tail. We need to use **chisq.dist** to find the area in the left tail.

Function	chisq.dist	Answer
Field 1	9.375	0.0033
Field 2	24	

So the area in the left tail is 0.0033, which means that $\frac{1}{2}(p\text{-value})=0.0033$. This is also the area in the right tail, so

$$p\text{-value}=0.0033 + 0.0033 = 0.0066$$

Conclusion:

Because $p\text{-value}= 0.0066 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the variation among the wait times for customers has changed.

NOTES

- The null hypothesis $\sigma^2 = 51.84$ is the claim that the variance in the wait times is 51.84. Note that we were given the standard deviation ($\sigma = 7.2$) in the question. But this is a test on variance, so we must write the hypotheses in terms of the variance $\sigma^2 = 7.2^2 = 51.84$.
- The alternative hypothesis $\sigma^2 \neq 51.84$ is the claim that the variance in the wait times has changed from 51.84.
- There are no units included with the hypotheses because variance does not have any units.
- In a two-tailed hypothesis test for population variance, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.
 - If $\chi^2 < df - 2$, the sample information belongs to the **left tail**.

- We use **chisq.dist** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.
 - If $\chi^2 > df - 2$, the sample information belongs to the **right tail**.
 - We use **chisq.dist.rt** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.
5. The p -value of 0.0066 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the variance in the wait times has most likely changed.

TRY IT

A scuba instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation of the depths is 1.2 meters. But his assistant thinks the standard deviation is less than 1.2 meters. The instructor wants to test this claim. The scuba instructor uses his most recent class of 20 students as a sample and finds that the standard deviation of the depths is 0.85 meters. At the 1% significance level, test if the variability in the depths of the student scuba divers is less than claimed.

Click to see Solution

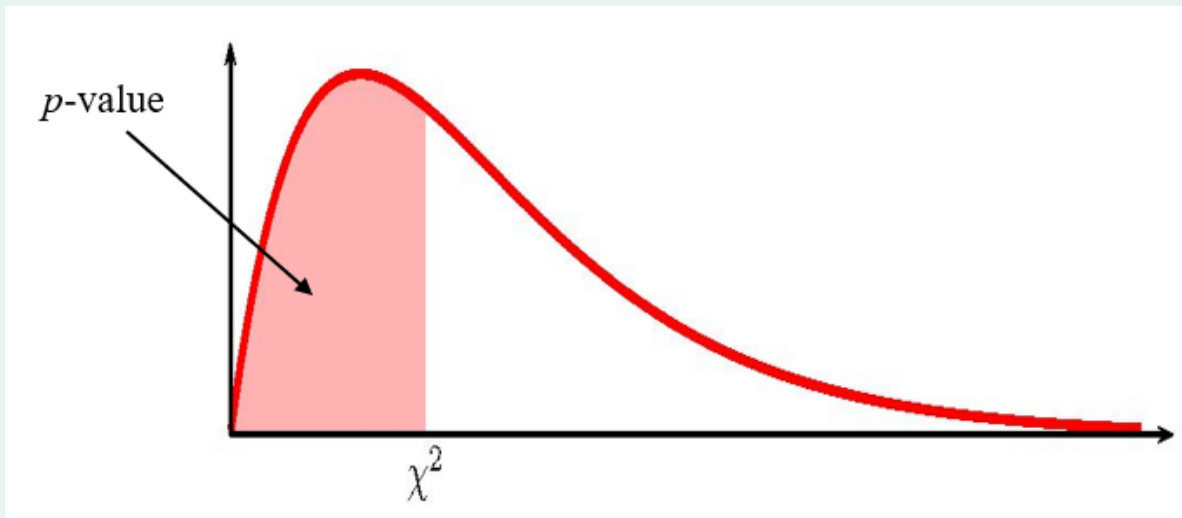
Hypotheses:

$$\begin{array}{l} H_0: \sigma^2 = 1.44 \\ H_a: \sigma^2 < 1.44 \end{array}$$

p-value:

From the question, we have $n = 20$, $s^2 = 0.7225$, and $\alpha = 0.01$.

Because the alternative hypothesis is a $<$, the p-value is the area in the left tail of the χ^2 -distribution.



To use the **chisq.dist** function, we need to calculate out the χ^2 -score and the degrees of freedom:

$$\begin{aligned} \chi^2 &= \frac{(n - 1) \times s^2}{\sigma^2} \\ &= \frac{(20 - 1) \times 0.7225}{1.44} \\ &= 9.5329... \end{aligned}$$

$$\begin{aligned} df &= n - 1 \\ &= 20 - 1 \\ &= 19 \end{aligned}$$

Function	chisq.dist	Answer
Field 1	9.5329...	0.0365
Field 2	19	
Field 3	true	

So the p -value = 0.0365.

Conclusion:

Because p -value = 0.0365 > 0.01 = α , we do not reject the null hypothesis. At the 1% significance level there is not enough evidence to suggest that the variation in the depths of the students is less than claimed.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=229#oembed-1>

Watch this video: Hypothesis Tests for One Population Variance by jbstatistics [8:51]

Concept Review

To construct a confidence interval or conduct a hypothesis test on a population variance, we use the sampling distribution of $\frac{(n-1) \times s^2}{\sigma^2}$, which follows a χ^2 -distribution with $n - 1$ degrees of freedom.

The hypothesis test for a population variance is a well established process:

1. Write down the null and alternative hypotheses in terms of the population variance σ^2 .
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or

two-tailed.

3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the χ^2 -distribution where $\chi^2 = \frac{(n-1) \times s^2}{\sigma^2}$ and $df = n - 1$.
5. Compare the p -value to the significance level and state the outcome of the test.
6. Write down a concluding sentence specific to the context of the question.

The general form of a confidence interval for an unknown population variance σ^2 is

$$\text{Lower Limit} = \frac{(n-1) \times s^2}{\chi_R^2}$$

$$\text{Upper Limit} = \frac{(n-1) \times s^2}{\chi_L^2}$$

where χ_L^2 is the χ^2 -score so that the area in the left-tail of of the χ^2 -distribution is $\frac{1-C}{2}$, χ_R^2 is the χ^2 -score so that the area in the right-tail of of the χ^2 -distribution is $\frac{1-C}{2}$ and the χ^2 -distribution has $n - 1$ degrees of freedom.

Attribution

“11.6 Test of a Single Variance“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

10.4 THE GOODNESS-OF-FIT TEST

LEARNING OBJECTIVES

- Conduct and interpret χ^2 -goodness-of-fit hypothesis tests.

Recall that a **categorical** (or qualitative) variable is a variable where the data can be grouped by specific categories. Examples of categorical variables include eye colour, blood type, or brand of car. A categorical variable is a random variable that takes on categories. Suppose we want to determine whether the data from a categorical variable “fit” a particular distribution or not. That is, for a categorical variable with a historical or assumed probability distribution, does a new sample from the population support the assumed probability distribution or does the sample indicate that there has been a change in the probability distribution?

The χ^2 -goodness-of-fit test allows us the test if the sample data from a categorical variable fits the pattern of **expected probabilities** for the variable. In a χ^2 -goodness-of-fit test, we are analyzing the distribution of the frequencies for one categorical variable. This is a hypothesis test where the hypotheses state that the categorical variable does or does not follow an assumed probability distribution and a χ^2 -distribution is used to determine the p -value for the test.

Steps to Conduct a χ^2 -Goodness-of-Fit Test

Suppose a categorical variable has k possible outcomes (categories) with probabilities p_1, p_2, \dots, p_k . Suppose n independent observations are taken from this categorical variable.

1. Write down the null and alternative hypotheses:

$$H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, \dots, p_k = p_{k_0}$$

$$H_a : \text{at least one } p_i \neq p_{i_0}$$

2. Collect the sample information for the test and identify the significance level α .
3. Use the χ^2 -distribution to find the p -value, which is the **area in the right tail** of the distribution. The χ^2 -score and degrees of freedom are

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad \text{df} = k - 1$$

χ^2 = observed frequency from the sample data / expected frequency from assumed distribution
 k = number of categories

4. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
5. Write down a concluding sentence specific to the context of the question.

NOTES

1. The null hypothesis is the claim that the categorical variable follows the assumed distribution. That is, the probability p_i of each possible outcome of the categorical variable equals a hypothesized probability p_{i_0} .
2. The alternative hypothesis is the claim that the categorical variable does not follow the assumed distribution. That is, for at least one possible outcome of the categorical variable the probability p_i does not equal the claimed probability p_{i_0} .

3. In order to use the χ^2 -goodness-of-fit test, the expected frequency for each category must be at least 5.
4. The p -value for a χ^2 -goodness-of-fit test is always the area in the right tail of the χ^2 -distribution. So, we use **chisq.dist.rt** to find the p -value for a χ^2 -goodness-of-fit test.
5. To calculate the χ^2 -score:
 1. For each of the possible outcomes of the categorical variable, calculate $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$:
 - i. Find the difference between the observed frequency (from the sample) and the expected frequency (from the null hypothesis). The expected frequency equals $n \times p_{i_0}$ where n is the sample size and p_{i_0} is the assumed probability for the i th outcome claimed in the null hypothesis.
 - ii. Square the difference in step (i).
 - iii. Divide the value found in step (iii) by the expected frequency.
 2. Add up the values of $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ for each of the outcomes.
6. We expect that there will be a discrepancy between the observed frequency and the expected frequency. If this discrepancy is very large, the value of χ^2 will be very large and result in a small p -value.

EXAMPLE

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the

actual student absenteeism rate follows faculty perception. The faculty believe that the distribution of the number of absences per term is as follows:

Number of Absences per Term	Expected Percent of Students
0–2	50%
3–5	30%
6–8	12%
9–11	6%
12+	2%

At the end of the semester, a random survey of 300 students across all mathematics courses was taken and the actual (observed) number of absences for the 300 students is recorded.

Number of Absences per Term	Observed Number of Students
0–2	120
3–5	100
6–8	55
9–11	15
12+	10

At the 5% significance level, determine if the number of absences per term follow the distribution assumed by the faculty.

Solution:

Let p_1 be the probability a student has 0-2 absences, p_2 be the probability a student has 3-5 absences, p_3 be the probability a student has 6-8 absences, p_4 be the probability a student has 9-11 absences, and p_5 be the probability a student has 12 or more absences.

Hypotheses:

$$\begin{array}{l} H_0: p_1=50\%, p_2=30\%, p_3=12\%, p_4=6\%, p_5=2\% \\ H_a: \text{at least one of the } p_i\text{'s does not equal its stated probability} \end{array}$$

p-value:

From the question, we have $n = 300$ and $k = 5$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each category is the number of observations in the sample that fall into that category. This is the information provided in the sample above.

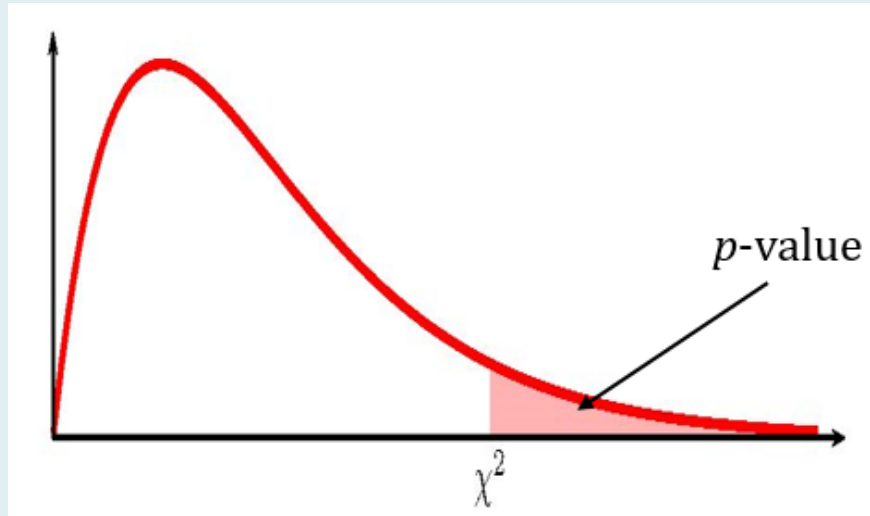
Next, we must calculate out the expected frequencies. The expected frequency is the number of observations we would expect to see in the sample, assuming the null hypothesis is true. To calculate the expected frequency for each category, we multiply the sample size n by the probability associated with that category claimed in the null hypothesis.

Number of Absences per Term	Observed Frequency	Expected Frequency
0-2	120	$0.5 \times 300 = 150$
3-5	100	$0.3 \times 300 = 90$
6-8	55	$0.12 \times 300 = 36$
9-11	15	$0.06 \times 300 = 18$
12+	10	$0.02 \times 300 = 6$

To calculate the χ^2 -score, we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ for each category and then add up these quantities.

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\ &= \frac{(120 - 150)^2}{150} + \frac{(100 - 90)^2}{90} + \frac{(55 - 36)^2}{36} + \frac{(15 - 18)^2}{18} + \frac{(10 - 6)^2}{6} \\ &= 20.305\dots\end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$. The χ^2 -goodness-of-fit test is a right tailed test, so we use the **chisq.dist.rt** function to find the p -value:



Function	chisq.dist.rt	Answer
Field 1	20.305....	0.0004
Field 2	4	

So the p -value = 0.0004.

Conclusion:

Because p -value = 0.0004 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the number of absences per term does not follow the distribution assumed by faculty.

NOTES

1. The null hypothesis is the claim that the percent of students that fall into each category is as stated. That is, 50% students miss between 0 and 2 classes, 30% of the students miss between 3 and 5 students, etc.
2. The alternative hypothesis is the claim that at least one of the percent of students that fall into each category is not as stated. The alternative hypothesis does not say that every p_i does not equal its stated probabilities, only that one of them does not equal its stated probability.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to

avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value. You can use Excel to calculate the expected frequencies and the χ^2 -score.

4. The p -value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 20.305\dots$. In the calculation of the p -value:
 - The function is `chisq.dist.rt` because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the value of the degrees of freedom df .
5. The p -value of 0.0004 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, student absenteeism does not fit faculty perception.

EXAMPLE

Employers want to know which days of the week employees have the highest number of absences in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers are asked on which day of the week they had the highest number of employee absences. The results are recorded in the table below. At the 5% significance level, test if the day of the week with the highest number of absences occur with equal frequency during a five-day work week.

Day of the Week	Observed Frequency
Monday	15
Tuesday	11
Wednesday	10
Thursday	9
Friday	15

Solution:

Let p_1 be the probability the highest number of absences occurs on Monday, p_2 be probability the highest number of absences occurs on Tuesday, p_3 be the probability the highest number of absences occurs on Wednesday, p_4 be the probability the highest number of absences occurs on Thursday, and p_5 be the probability the highest number of absences occurs on Friday.

If the day of the week with the highest number of absences occurs with equal frequency, then the probability that any day has the highest number of absences is the same as any other day. Because there are 5 days (categories), if the frequencies are equal then each day would have a probability of

$$20\% \left(\text{or } \frac{1}{5} \right).$$

Hypotheses:

$$\begin{array}{l} H_0: p_1 = p_2 = p_3 = p_4 = p_5 = 20\% \\ H_a: \text{at least one of the } p_i \neq 20\% \end{array}$$

p-value:

From the question, we have $n = 60$ and $k = 5$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each category is the number of observations in the sample that fall into that category. This is the information provided in the sample above.

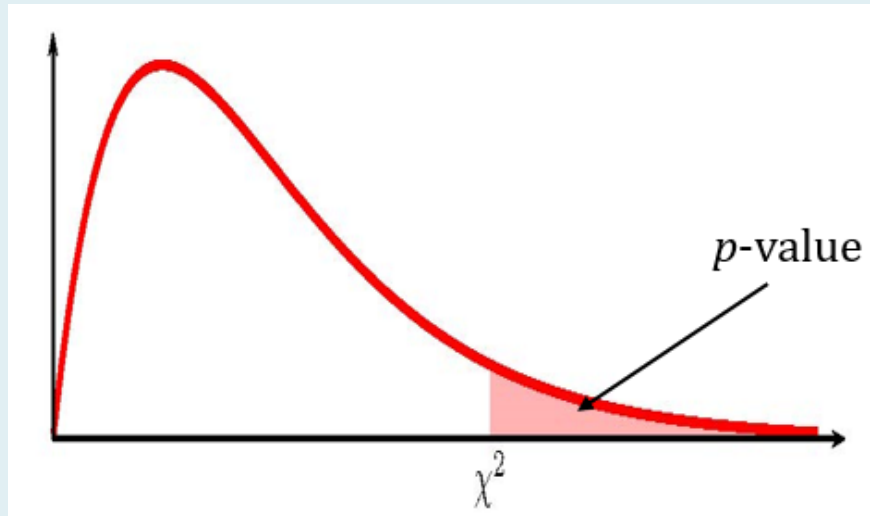
Next, we must calculate out the expected frequencies. The expected frequency is the number of observations we would expect to see in the sample, assuming the null hypothesis is true. To calculate the expected frequency for each category, we multiply the sample size n by the probability associated with that category claimed in the null hypothesis.

Day of the Week	Observed Frequency	Expected Frequency
Monday	15	$0.2 \times 60 = 12$
Tuesday	11	$0.2 \times 60 = 12$
Wednesday	10	$0.2 \times 60 = 12$
Thursday	9	$0.2 \times 60 = 12$
Friday	15	$0.2 \times 60 = 12$

To calculate the χ^2 -score, we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ for each category and then add up these quantities.

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\ &= \frac{(15-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(15-12)^2}{12} \\ &= 2.666\dots\end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$. The χ^2 -goodness-of-fit test is a right tailed test, so we use the **chisq.dist.rt** function to find the p -value:



Function	chisq.dist.rt	Answer
Field 1	2.666....	0.6151
Field 2	4	

So the p -value = 0.6151.

Conclusion:

Because p -value = 0.6151 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is enough evidence to suggest that the day of the week with the highest number of absences occur with equal frequency during a five-day work week.

NOTES

1. The null hypothesis is the claim that the probability each day of the week has the highest number of absences is 20%.
2. The alternative hypothesis is the claim that at least one of the probabilities is not 20%. The alternative hypothesis does not say that every p_i does not equal 20%, only that one of them does not equal 20%.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value.
4. The p -value of 0.6151 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis.

TRY IT

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students are asked on which night of the week they did the most homework. The results are shown in the table below. At the 5% significance level, are the nights that students do most of their homework equally distributed?

Day of Week	Number of Students
Sunday	11
Monday	8
Tuesday	10
Wednesday	7
Thursday	10
Friday	5
Saturday	5

Click to see Solution

Let p_1 be the probability students do their homework on Sunday, p_2 be the probability students do their homework on Monday, p_3 be the probability students do their homework on Tuesday, p_4 be the probability students do their homework on Wednesday, p_5 be the probability students do their homework on Thursday, p_6 be the probability students do their homework on Friday, and p_7 be the probability students do their homework on Saturday.

Hypotheses:

$$\begin{array}{l} H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7} \\ H_a: \\ \text{at least one of the } p_i \neq \frac{1}{7} \end{array}$$

p -value:

From the question, we have $n = 49$ and $k = 7$.

Day of the Week	Observed Frequency	Expected Frequency
Sunday	11	$1/7 \times 49 = 7$
Monday	8	$1/7 \times 49 = 7$
Tuesday	10	$1/7 \times 49 = 7$
Wednesday	7	$1/7 \times 49 = 7$
Thursday	10	$1/7 \times 49 = 7$
Friday	5	$1/7 \times 49 = 7$
Saturday	5	$1/7 \times 49 = 7$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(11-7)^2}{7} + \frac{(8-7)^2}{7} + \frac{(10-7)^2}{7} + \frac{(7-7)^2}{7} + \frac{(10-7)^2}{7} + \frac{(5-7)^2}{7} + \frac{(5-7)^2}{7} = 6.142\dots$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 7 - 1 = 6$.

Function	chisq.dist.rt	Answer
Field 1	6.142....	0.4074
Field 2	6	

So the p -value = 0.4074.

Conclusion:

Because p -value = 0.4074 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is enough evidence to suggest that the nights students do most of their homework are equally distributed.

TRY IT

One study indicates that the number of televisions that American families have is distributed as shown in this table:

Number of Televisions	Percent
0	10%
1	16%
2	55%
3	11%
4 or more	8%

A researcher wants to determine if the number of televisions that families in the far western part of the U.S. have the same distribution as the above study. A random sample of 600 families in the far western U.S. is taken and the results are recorded in the following table:

Number of Televisions	Observed Frequency
0	66
1	119
2	340
3	60
4 or more	15

At the 1% significance level, does it appear that the distribution of the number of televisions for families in the far western U.S is different from the distribution for the American population as a whole?

Click to see Solution

Let p_1 be the probability a family owns 0 televisions, p_2 be the probability a family owns 1

television, p_3 be the probability a family owns 2 televisions, p_4 be the probability a family owns 3 televisions, and p_5 be the probability a family owns 4 or more televisions.

Hypotheses:

$$\begin{array}{l} H_0: p_1=10\%, p_2=16\%, p_3=55\%, p_4=11\%, p_5=8\% \\ H_a: \text{at least one of the } p_i\text{'s does not equal its stated probability} \end{array}$$

***p*-value:**

From the question, we have $n = 600$ and $k = 5$.

Number of Televisions	Observed Frequency	Expected Frequency
0	66	$0.1 \times 600 = 60$
1	119	$0.16 \times 600 = 96$
2	340	$0.55 \times 600 = 330$
3	60	$0.11 \times 600 = 66$
4 or more	15	$0.08 \times 600 = 48$

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(66-60)^2}{60} + \frac{(119-96)^2}{96} + \frac{(340-330)^2}{330} + \frac{(60-66)^2}{66} \\ &\quad + \frac{(15-48)^2}{48} \\ &= 29.646... \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df = k - 1 = 5 - 1 = 4$.

Function	chisq.dist.rt	Answer
Field 1	29.646....	0.000006
Field 2	4	

So the $p\text{-value} = 0.000006$.

Conclusion:

Because $p\text{-value} = 0.000006 < 0.01 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that the distribution of the number of televisions for families in the far western U.S is different from the distribution for the American population as a whole.

TRY IT

The expected percentage of the number of pets students in the United States have in their homes is distributed as follows:

Number of Pets	Percent
0	18%
1	25%
2	30%
3	18%
4 or more	9%

A researcher wants to find out if the distribution of the number of pets students in Canada have is the same as the distribution shown in the U.S. A random sample of 1,000 students from Canada is taken and the results are shown in the table below:

Number of Pets	Observed Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, is the distribution of the number of pets students in Canada have different from the distribution for the United States?

Click to see Solution

Let p_1 be the probability a student owns 0 pets, p_2 be the probability a student owns 1 pet, p_3 be

the probability a student owns 2 pets, p_4 be the probability a student owns 3 pets, and p_5 be the probability a student owns 4 or more pets.

Hypotheses:

$$\begin{array}{l} H_0: p_1=18\%, p_2=25\%, p_3=30\%, p_4=18\%, p_5=9\% \\ H_a: \text{at least one of the } p_i\text{'s does not equal its stated probability} \end{array}$$

***p*-value:**

From the question, we have $n = 1000$ and $k = 5$.

Number of Pets	Observed Frequency	Expected Frequency
0	210	$0.18 \times 1000 = 180$
1	240	$0.25 \times 1000 = 250$
2	320	$0.30 \times 1000 = 300$
3	140	$0.18 \times 1000 = 180$
4 or more	90	$0.09 \times 1000 = 90$

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(210-180)^2}{180} + \frac{(240-250)^2}{250} + \frac{(320-300)^2}{300} + \frac{(140-180)^2}{180} \\ &\quad + \frac{(90-90)^2}{90} \\ &= 15.622... \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is $df - k - 1 = 5 - 1 = 4$.

Function	chisq.dist.rt	Answer
Field 1	15.622....	0.0036
Field 2	4	

So the $p\text{-value} = 0.0036$.

Conclusion:

Because $p\text{-value} = 0.0036 < 0.01 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that the distribution of the number of pets students in Canada have is different from the distribution for the United States.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=231#oembed-1>

Watch this video: Pearson's chi square test (goodness of fit) | Probability and Statistics | Khan Academy by Khan Academy
[11:47]

Concept Review

The χ^2 -goodness-of-fit test is used to determine if a categorical variable follows a hypothesized distribution. The goodness-of-fit test is a well established process:

1. Write down the null and alternative hypotheses. The null hypothesis is the claim that the categorical variable follows the hypothesized distribution and the alternative hypothesis is the claim that the categorical variable does not follow the hypothesized distribution.
 2. Collect the sample information for the test and identify the significance level.
 3. The p -value is the area in the right tail of the χ^2 -distribution where

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
and

$$df = k - 1.$$
 4. Compare the p -value to the significance level and state the outcome of the test.
 5. Write down a concluding sentence specific to the context of the question.
-

Attribution

“11.2 Goodness-of-Fit Test“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

10.5 THE TEST OF INDEPENDENCE

LEARNING OBJECTIVES

- Conduct and interpret the χ^2 test of independence.

Given two categorical variables, is there some relationship between the two categorical variables or are the two categorical variables independent. The χ^2 test of independence allows us to test if two categorical variables are independent (not related) or dependent (related). The test of independence can only show if a relationship exists between two variables, but the test does not show if one variable causes changes in the other variable.

The test of independence uses a contingency table to analyze the data. As we saw previously in probability, a contingency table lists the categories of one variables as the rows and the categories of the other variable as the columns. The frequency of a row-column combination is the number of items that occur in both categories.

Steps to Conduct a χ^2 Test of Independence

Suppose one categorical variable has r possible outcomes (categories) and the other categorical variable has c possible outcomes (categories).

1. Write down the null and alternative hypotheses:

$$\begin{array}{l} H_0: \text{The two categorical variables are} \\ \text{independent} \\ H_a: \text{The two categorical variables are dependent} \end{array}$$

2. Collect the sample information for the test and identify the significance level α .
3. Use the χ^2 -distribution to find the p -value, which is the **area in the right tail** of the

distribution. The χ^2 -score and degrees of freedom are

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ df &= (r-1) \times (c-1) \\ \text{observed} &= \text{observed frequency from the sample data} \\ \text{expected} &= \frac{\text{row total} \times \text{column total}}{\text{table total}} \\ r &= \text{number of rows in the contingency table} \\ c &= \text{number of columns in the contingency table} \end{aligned}$$

4. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
5. Write down a concluding sentence specific to the context of the question.

NOTES

1. The null hypothesis is the claim that the two categorical variables are independent. That is, there is no relationship between the two categorical variables.
2. The alternative hypothesis is the claim that the two categorical variables are dependent. That is, there is some relationship between the two categorical variables.
3. The test can only show if a relationship exists between the two categorical variables. The test cannot show any type of causal relationship between the two categorical variables.
4. The formula to find the expected frequencies follows from the assumption that the null hypothesis is true and how we calculate joint probabilities for independent events. Assuming the null hypothesis is true means that we assume the variables are independent. This means that we assume that the events in any row and column combination of the contingency tables are independent. As we saw in probability, when

two events A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$. Using this fact, we get the formula for the expected frequency:

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

5. In order to use the χ^2 test of independence, the expected frequency for a cell in the contingency table must be at least 5.
6. The p -value for a χ^2 test of independence is always the area in the right tail of the χ^2 -distribution. So, we use **chisq.dist.rt** to find the p -value for a χ^2 test of independence.
7. To calculate the χ^2 -score:
 1. For each of the possible outcomes of the categorical variables, calculate $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$:
 - i. Find the difference between the observed frequency (from the sample) and the expected frequency (from the null hypothesis). The expected frequency of any cell of the contingency table when the null hypothesis is true is:

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$
 - ii. Square the difference in step (i).
 - iii. Divide the value found in step (iii) by the expected frequency.
 2. Add up the values of $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ for each of the outcomes.

EXAMPLE

A researcher is studying the relationship between the drivers who commit speeding violations and

drivers who use cell phones while driving. The researcher took a sample of 755 drivers and obtained the information shown in the table below.

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

At the 5% significance level, is there a relationship between drivers who commit speeding violations and drivers who use cell phones while driving?

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \text{The two variables are independent} \\ H_a: \text{The two variables are dependent} \end{array}$$

p-value:

From the question, we have $r = 2$ and $c = 2$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each cell is the number of observations in the sample that fall into that cell. This is the information provided in the sample above.

Observed Frequencies (Sample Data)			
	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Next, we must calculate out the expected frequencies. Because we assume the null hypothesis is true (i.e. the variables are independent), the expected frequency in each cell is

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

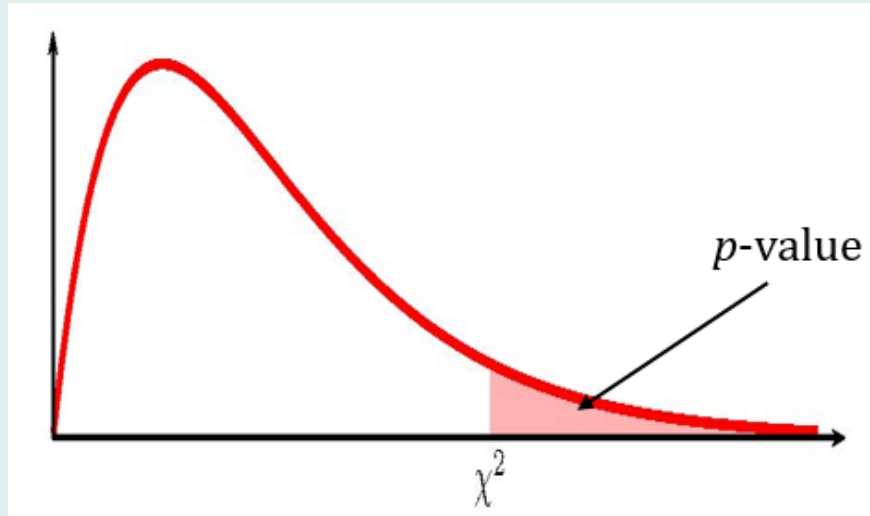
Expected Frequencies			
	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	$\frac{305 \times 70}{755} = 28.27\dots$	$\frac{305 \times 685}{755} = 276.72\dots$	305
Not a cell phone user	$\frac{450 \times 70}{755} = 41.72\dots$	$\frac{450 \times 685}{755} = 408.27\dots$	450
Total	70	685	755

To calculate the χ^2 -score, for each cell we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and then add up these quantities.

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} \\ &= \frac{(25 - 28.27\dots)^2}{28.27\dots} + \frac{(280 - 276.72\dots)^2}{276.72\dots} + \frac{(45 - 41.72\dots)^2}{41.72\dots} + \frac{(405 - 408.27\dots)^2}{408.27\dots} \\ &= 0.7027\dots \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is

$df = (r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$. The χ^2 test of independence is a right tailed test, so the we use **chisq.dist.rt** function to find the p -value:



Function	chisq.dist.rt	Answer
Field 1	0.7027....	0.4019
Field 2	1	

So the $p\text{-value} = 0.4019$.

Conclusion:

Because $p\text{-value} = 0.4019 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is enough evidence to suggest that the two variables are independent.

NOTES

1. The null hypothesis is the claim that the variables are independent. That is, there is no relationship between drivers who commit speeding violations and drivers who use cell phones while driving.
2. The alternative hypothesis is the claim that the variables are dependent. That is, there is a relationship between drivers who commit speeding violations and drivers who use cell phones while driving.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most

accurate value for the p -value.

4. The p -value is the area in the right tail of the χ^2 -distribution, to the right of $\chi^2 = 0.7027\dots$. In the calculation of the p -value:
 - The function is `chisq.dist.rt` because we are finding the area in the right tail of a χ^2 -distribution.
 - Field 1 is the value of χ^2 .
 - Field 2 is the value of the degrees of freedom df .
5. The p -value of 0.4019 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the two variables are independent.

EXAMPLE

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among college students, university students, and non students. The table below is a sample of the adult volunteers and the number of hours they volunteer per week.

	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	111	96	48	255
University Students	96	133	61	290
Non Students	91	150	53	294
Total	298	379	162	839

At the 5% significance level, is the number of hours volunteered independent of the type of volunteer?

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \text{The two variables are independent} \\ H_a: \text{The two variables are dependent} \end{array}$$

p-value:

From the question, we have $r = 3$ and $c = 3$. Now we need to calculate out the χ^2 -score for the test.

The observed frequency for each cell is the number of observations in the sample that fall into that cell. This is the information provided in the sample above.

Observed Frequencies (Sample Data)				
	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	111	96	48	255
University Students	96	133	61	290
Non Students	91	150	53	294
Total	298	379	162	839

Next, we must calculate out the expected frequencies. Because we assume the null hypothesis is true (i.e. the variables are independent), the expected frequency in each cell is

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

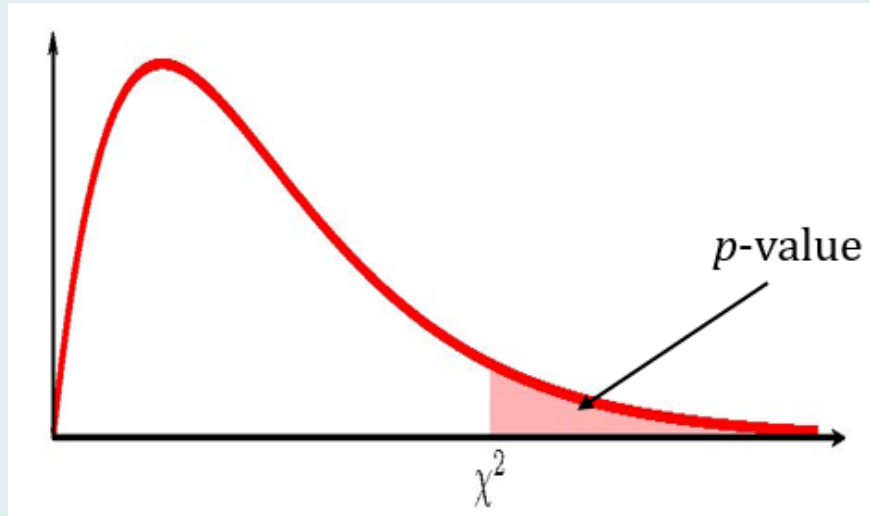
Expected Frequencies				
	1-3 Hours	4-6 Hours	7-9 Hours	Total
College Students	$\frac{255 \times 298}{839} = 90.57\dots$	$\frac{255 \times 379}{839} = 115.19\dots$	$\frac{255 \times 162}{839} = 49.23\dots$	255
University Students	$\frac{290 \times 298}{839} = 103.00\dots$	$\frac{290 \times 379}{839} = 131.00\dots$	$\frac{290 \times 162}{839} = 55.99\dots$	290
Non Students	$\frac{294 \times 298}{839} = 104.42\dots$	$\frac{294 \times 379}{839} = 132.80\dots$	$\frac{294 \times 162}{839} = 56.76\dots$	294
Total	298	379	162	839

To calculate the χ^2 -score, for each cell we work out the quantity $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and then add up these quantities.

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(111-90.57\dots)^2}{90.57\dots} + \frac{(96-115.19\dots)^2}{115.19\dots} + \frac{(48-49.23\dots)^2}{49.23\dots} \\ &\quad + \frac{(96-103.00\dots)^2}{103.00\dots} + \frac{(133-131.00\dots)^2}{131.00\dots} + \frac{(61-55.99\dots)^2}{55.99\dots} \\ &\quad + \frac{(91-104.42\dots)^2}{104.42\dots} + \frac{(150-132.80\dots)^2}{132.80\dots} + \frac{(53-56.76\dots)^2}{56.76\dots} \\ &= 12.99\dots \end{aligned}$$

The degrees of freedom for the χ^2 -distribution is

$df = (r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4$. The χ^2 test of independence is a right tailed test, so we use **chisq.dist.rt** function to find the p -value:



Function	chisq.dist.rt	Answer
Field 1	12.99....	0.0113
Field 2	4	

So the p -value = 0.0113.

Conclusion:

Because p -value = 0.0113 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the number of hours volunteered and type of volunteer are dependent.

NOTES

1. The null hypothesis is the claim that the variables are independent. That is, there is no relationship between the number of hours volunteered and type of volunteer.
2. The alternative hypothesis is the claim that the variables are dependent. That is, there is a relationship between the number of hours volunteered and type of volunteer.
3. Keep all of the decimals throughout the calculation (i.e. in the calculation of the χ^2 -score) to avoid any round-off error in the calculation of the p -value. This ensures that we get the most accurate value for the p -value.

4. The p -value of 0.0113 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the two variables are dependent.

TRY IT

In a local school district, a music teacher wants to study the relationship between students who take music and students on the honour roll. The teacher took a sample of 300 students and obtained the information shown in the table below.

	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	24	26	50
Non-Music Student	67	183	250
Total	97	203	300

At the 5% significance level, is there a relationship between music/non-music students and honour roll/non-honour roll students?

Click to see Solution

Hypotheses:

$$H_0: \text{The two variables are independent} \\ H_a: \text{The two variables are dependent}$$

p-value:

From the question, we have $r = 2$ and $c = 2$.

Observed Frequencies (Sample Data)			
	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	24	26	50
Non-Music Student	67	183	250
Total	97	203	300

Expected Frequencies			
	Honour Roll Student	Non-Honour Roll Student	Total
Music Student	16.166...	33.833...	50
Non-Music Student	80.833...	169.166...	250
Total	97	203	300

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(24 - 16.166\dots)^2}{16.166\dots} + \frac{(26 - 33.833\dots)^2}{33.833\dots} + \frac{(67 - 80.833\dots)^2}{80.833\dots} + \frac{(183 - 169.166\dots)^2}{169.166\dots} \\ &= 9.107\dots \end{aligned}$$

$$\begin{aligned} df &= (r - 1) \times (c - 1) \\ &= (2 - 1) \times (2 - 1) \\ &= 1 \end{aligned}$$

Function	chisq.dist.rt	Answer
Field 1	9.107...	0.0025
Field 2	1	

So the $p\text{-value} = 0.0025$.

Conclusion:

Because $p\text{-value} = 0.0025 < 0.05 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the two variables are dependent.

TRY IT

A local college is interested in the relationship between student anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and the need to succeed in school. The results are shown in the table below.

	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Total	57	95	127	63	58	400

At the 5% significance level, is there a relationship between student anxiety level and the need to succeed in school?

Click to see Solution

Hypotheses:

$$\begin{array}{l} H_0: \text{The two variables are independent} \\ H_a: \text{The two variables are dependent} \end{array}$$

p-value:

From the question, we have $r = 3$ and $c = 5$.

Observed Frequencies (Sample Data)						
	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Total	57	95	127	63	58	400

Expected Frequencies						
	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Total
High Need	22.08...	36.81...	49.21...	24.41...	22.47...	155
Medium Need	27.50...	45.83...	61.27...	30.39...	27.98...	193
Low Need	7.41	12.35	16.51	8.19	7.54	52
Total	57	95	127	63	58	400

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(35-22.08)^2}{22.08} + \frac{(18-27.50)^2}{27.50} + \frac{(4-7.41)^2}{7.41} \\ &\quad + \frac{(42-36.81)^2}{36.81} + \frac{(48-45.83)^2}{45.83} + \frac{(5-12.35)^2}{12.35} \\ &\quad + \frac{(53-49.21)^2}{49.21} + \frac{(63-61.27)^2}{61.27} + \frac{(11-16.51)^2}{16.51} \\ &\quad + \frac{(15-24.41)^2}{24.41} + \frac{(33-30.39)^2}{30.39} + \frac{(15-8.19)^2}{8.19} \\ &\quad + \frac{(10-22.47)^2}{22.47} + \frac{(31-27.98)^2}{27.98} + \frac{(17-7.54)^2}{7.54} \\ &= 48.419 \end{aligned}$$

df = (r-1) × (c-1) = (3-1) × (5-1) = 8

Function	chisq.dist.rt	Answer
Field 1	48.419...	0.00000008
Field 2	8	

So the p -value = 0.00000008.

Conclusion:

Because p -value = 0.00000008 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that the two variables are dependent.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=233#oembed-1>

Watch this video: Chi-Square Test of Independence by Khan Academy [10:27]

Concept Review

The χ^2 test of independence is used to determine if two categorical variables are independent or dependent. The test of independence is a well established process:

1. Write down the null and alternative hypotheses. The null hypothesis is the claim that the categorical variables are independent and the alternative hypothesis is the claim that the categorical variables are dependent.
2. Collect the sample information for the test and identify the significance level.
3. The p -value is the area in the right tail of the χ^2 -distribution where

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
 and
$$df = (r - 1) \times (c - 1).$$

4. Compare the p -value to the significance level and state the outcome of the test.
 5. Write down a concluding sentence specific to the context of the question.
-

Attribution

“11.3 Test of Independence“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

10.6 EXERCISES

1. If the number of degrees of freedom for a χ^2 -distribution is 25, what is the population mean and standard deviation?

2. Where is μ located on a chi-square curve?

3. A teacher predicts that the distribution of grades on the final exam will be and they are recorded in the table.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 20 is in the table below.

Grade	Frequency
A	7
B	7
C	5
D	1

At the 5% significance level, do the actual grades match the teacher's assumed distribution?

4. The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in the table below.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is as in the table below.

Ethnicity	Percentage of total county population
White	42.9%
Hispanic	26.7%
Black/African-American	2.6%
Asian, Pacific Islander	27.8%
	Total = 100%

At the 5% significance level, does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county?

5. A six-sided die is rolled 120 times and the results are recorded in the table below. At the 5% significance level, determine if the die is fair. (Hint: in a fair die, each of the faces is equally likely to occur.)

Face Value	Frequency
1	15
2	29
3	16
4	15
5	30
6	15

6. The marital status distribution of the U.S. male population, ages 15 and older, is as shown in the table below.

Marital Status	Percent
never married	31.3
married	56.1
widowed	2.5
divorced/separated	10.1

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. At the 5% significance level, test if this age group of males fits the distribution of the U.S. adult population.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

7. The columns in the table below contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African-American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

- a. At the 5% significance level, determine if the local results follow the distribution of the U.S. overall student population based on ethnicity.
- b. At the 5% significance level, determine if the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

8. UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in THE CHRONICLE OF HIGHER EDUCATION (2/2/2006). Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables below. The second column in each table does not add to 100% because of rounding.

- a. At the 5% significance level, determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Women – Expected Major	Women – Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

- b. At the 5% significance level determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Men – Expected Major	Men – Actual Major
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

9. The table below contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. At the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (Years)	Obese (Percentage)	Expected USA average (Percentage)
20–30	75.0	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21.0	39.7

10. Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. The table below shows the results. At the 5% significance level determine if a passenger's choice in ticket class is independent of the distance they must travel.

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

11. A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. At the 5% significance level test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

12. Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in the table. Conduct a test of independence. Use a 5% significance level.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

13. College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. The table below shows the data. Conduct a test of independence. Use a 5% significance level.

Major	< \$50,000	\$50,000 – \$68,999	\$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

14. Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is recorded in the table below. Conduct a test of independence. Use a 5% significance level.

Location	20–29	30–39	40–49	50 and over
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

15. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence. Use a 5% significance level

Sport	18 – 25	26 – 30	31 – 40	41 and over
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

16. A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in the table. Conduct a test of independence. Use a 5% significance level.

Type of Fries	Northeast	South	Central	West
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

17. According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence. Use a 5% significance level.

Age of Males	None	< \$200,000	\$200,000–\$400,000	\$401,001–\$1,000,000	\$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

18. Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence. Use a 5% significance level.

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< \$30,000	15	25	10	5
\$30,000–\$40,000	20	40	70	30
\$40,000–\$50,000	10	20	40	55
\$50,000–\$60,000	5	10	20	60
\$60,000+	0	5	10	150

19. An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on the table, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

U.S. region/ Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Row total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Column Total	45	112	101	46	60	27	391

20. The table provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Age Group \ Net Worth Value (in millions of US dollars)	1–5	6–24	≥25	Row Total
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

21. A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in the table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/ Ethnicity	Asian-American	White/ Non-Hispanic	African-American	Latino	Row Total
Against tax	48	433	41	160	628
In Favor of tax	54	234	24	147	459
No opinion	16	43	16	19	84
Column Total	118	710	71	272	1171

22. An archer's standard deviation for his hits is six (data is measured in distance from the center of

the target). An observer claims the standard deviation is less. At the 5% significance level, test the observer's claim.

23. The variance of heights for students in a school is 0.66. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the variation of heights for the school is greater than 0.66. At the 5% significance level, determine if the variance in the heights for students in the school is greater than 0.66.

24. The average waiting time in a doctor's office varies. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

- a. Construct a 96% confidence interval for the variation in the wait times at the doctor's office.
- b. Interpret the confidence interval found in part (a).
- c. One of the doctors believes that the variance in the wait times is greater than 12. Is the doctor's claim reasonable? Explain.

25. Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes. At the 5% significance level, determine if variance in the delay times is greater than 150.

26. A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54.

- a. Construct a 99% confidence interval for the variance in the weight of the cereal boxes.
- b. Interpret the confidence interval found in part (a).
- c. If the variance in the weight of the cereal boxes is supposed to be at most 25, does the machine need to be recalibrated?

27. Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of \$84 and a sample standard

deviation of \$12, test the claim that the standard deviation is greater than \$15. Use a 5% significance level.

28. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief. Use a 5% significance level.

29. According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a variance of four. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the variance is different from four? Use a 5% significance level. Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

30. The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the variation for a ten-ounce order of fries is 2.25, but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of two oz. At the 5% significance level, determine if the variation in the amount of fries per order is higher than claimed.

31. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of \$1,249 with a variance of 625. You find a website that has a price comparison for the same computer at a series of stores as follows: \$1,299; \$1,229.99; \$1,193.08; \$1,279; \$1,224.95; \$1,229.99; \$1,269.95; \$1,249. Can you argue that pricing has a larger variation than claimed by the manufacturer? Use the 5% significance level. As a potential buyer, what would be the practical conclusion from your analysis?

32. A company packages apples by weight. One of the weight grades is Class A apples. A batch of apples is selected to be included in a Class A apple package.

- a. Construct a 95% confidence interval for the variation in the weight of apples in the package.
- b. Interpret the confidence interval found in part (a).

Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; 172;

Attribution

“Chapter 11 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART XI

STATISTICAL INFERENCE USING THE F-DISTRIBUTION

Chapter Outline

11.1 Introduction to Statistical Inferences Using the F-Distribution

11.2 The F-Distribution

11.3 Statistical Inference for Two Population Variances

11.4 One-Way ANOVA and Hypothesis Tests for Three or More Population Means

11.5 Exercises

11.1 INTRODUCTION TO STATISTICAL INFERENCES USING THE F-DISTRIBUTION



One-way ANOVA is used to measure information from several groups. Photo by OpenStax, CC BY 4.0.

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests that compare averages between more than two groups, statisticians have developed a method called “Analysis of Variance” (abbreviated ANOVA). In this chapter, we will study the simplest form of ANOVA called single factor or one-way ANOVA. We will also study the

F-distribution, used in a one-way ANOVA and the test of two population variances. This is just a very brief overview of one-way ANOVA.

Attribution

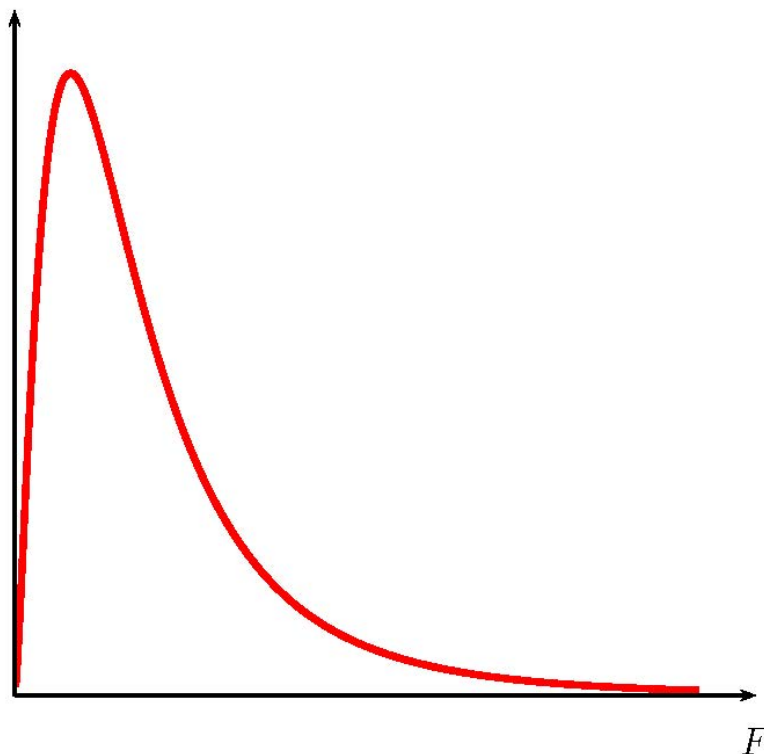
“Chapter 13 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

11.2 THE F-DISTRIBUTION

LEARNING OBJECTIVES

- Find the area under an F -distribution.
- Find the F -score for a given area under the curve of an F -distribution.

The F -distribution is a continuous probability distribution. The graph of an F -distribution is shown below. The F -distribution is used in statistical inference to test the equality of population variances, test the difference in three or more population means, or to test the overall multiple regression model.



Properties of the F -distribution:

- The graph of an F -distribution is positively-skewed and asymmetrical with a minimum value of 0 and no maximum value.
- An F -distribution is determined by two different degrees of freedom, df_1 and df_2 . df_1 is the degrees of freedom for the numerator of the F -score and df_2 is the degrees of freedom for the denominator of the F -score. The values of the degrees of freedom depends on how the F -distribution is used. There is a different F -distribution for every set of degrees of freedom. As the values of df_1 and df_2 get larger, the F -distribution approaches a normal distribution.
- The total area under the graph of an F -distribution is 1.
- Probabilities associated with an F -distribution are given by the area under the curve of the F -distribution.

USING EXCEL TO CALCULATE THE AREA UNDER AN Formula does not parse -DISTRIBUTION

To find the area in the left tail:

- To find the area under an F -distribution to the left of a given F -score, use the **f.dist(F , degrees of freedom 1, degrees of freedom 2, logic operator)** function.
 - For F , enter the F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution.
 - For **logic operator**, enter **true**.
- The output from the **f.dist** function is the area to the left of the entered F -score.
- Visit the Microsoft page for more information about the **f.dist** function.

To find the area in the right tail:

- To find the area under an F -distribution to the right of a given F -score, use the **f.dist.rt(F ,**

degrees of freedom 1, degrees of freedom 2) function.

- For F , enter the F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution.
- The output from the **f.dist.rt** function is the area to the right of the entered F -score.
 - Visit the Microsoft page for more information about the **f.dist.rt** function.

EXAMPLE

Consider an F -distribution with $df_1 = 12$ and $df_2 = 27$.

1. Find the area under the F -distribution to the left of $F = 0.69$.
2. Find the area under the F -distribution to the right of $F = 1.53$.

Solution:

1.	Function	f.dist	Answer
	Field 1	0.69	0.2535
	Field 2	12	
	Field 3	27	
	Field 4	true	

2.	Function	f.dist.rt	Answer
	Field 1	1.53	0.1738
	Field 2	12	
	Field 3	27	

USING EXCEL TO CALCULATE Formula does not parse -SCORES

To find the F -score for the a given left-tail area:

- To find the F -score for a given area under an F -distribution to the left of the F -score, use the **f.inv(area to the left, degrees of freedom 1, degrees freedom 2)** function.
 - For **area to the left**, enter the area to the left of required F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution.
- The output from the **f.inv** function is the value of F -score so that the area to the left of the F -score is the entered area.
- Visit the Microsoft page for more information about the **f.inv** function.

To find the F -score for the a given right-tail area:

- To find the F -score for a given area under an F -distribution to the right of the F -score, use the **f.inv.rt(area to the right, degrees of freedom 1, degrees of freedom 2)** function.
 - For **area to the right**, enter the area to the right of required F -score.
 - For **degrees of freedom 1**, enter the value of df_1 for the F -distribution.
 - For **degrees of freedom 2**, enter the value of df_2 for the F -distribution.
- The output from the **f.inv.rt** function is the value of F -score so that the area to the right of the F -score is the entered area.
- Visit the Microsoft page for more information about the **f.inv.rt** function.

EXAMPLE

Consider an F -distribution with $df_1 = 37$ and $df_2 = 15$.

1. Find the F -score so that the area under the F -distribution to the left of F is 0.413.
2. Find the F -score so that the area under the F -distribution to the right of F is 0.148.

Solution:

1.	Function	f.inv	Answer
	Field 1	0.413	0.934
	Field 2	37	
	Field 3	15	

2.	Function	f.dist.rt	Answer
	Field 1	0.269	1.354
	Field 2	37	
	Field 3	15	

Concept Review

The F -distribution is a useful tool for assessment in a series of problem categories. These problem categories include: statistical inference for two population variances, testing the equality of three or more population means (one-way ANOVA), and testing the overall significance of the multiple regression model.

Important parameters in an F -distribution are the degrees of freedom in a given problem. The F -distribution curve is skewed to the right, and its shape depends on the degrees of freedom. As the degrees of freedom increase, the curve of an F -distribution approaches a normal distribution.

Attribution

“13.3 Facts About the F Distribution“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

11.3 STATISTICAL INFERENCE FOR TWO POPULATION VARIANCES

LEARNING OBJECTIVES

- Construct and interpret a confidence interval for two population variances.
- Conduct and interpret a hypothesis test for two population variances.

Sometimes we want to compare the variability between two populations instead of comparing the means of the populations. For example, college administrators would like two college professors grading exams to have the same variation in their grading or a supermarket might be interested in the variability of the check-out times for two checkers.

As with comparing other population parameters, we can construct confidence intervals and conduct hypothesis tests to study the relationship between two population variances. However, because of the distribution we need to use, we study the **ratio of two population variances**, not the difference in the variances.

Throughout this section, we will use subscripts to identify the values for the sample sizes, variances, and standard deviations for the two populations:

Symbol for:	Population 1	Population 2
Population Variance	σ_1^2	σ_2^2
Population Standard Deviation	σ_1	σ_2
Sample Size	n_1	n_2
Sample Variance	s_1^2	s_2^2
Sample Standard Deviation	s_1	s_2

In order to construct a confidence interval or conduct a hypothesis test on the ratio of two population variances, $\frac{\sigma_1^2}{\sigma_2^2}$, we need to use the distribution of $\frac{s_1^2}{s_2^2}$ when the population variances are equal ($\sigma_1^2 = \sigma_2^2$). Suppose we have two normal populations with equal variances $\sigma_1^2 = \sigma_2^2$. A sample of size n_1 with sample variance s_1^2 is taken from population 1 and a sample of size n_2 with sample variance s_2^2 is taken from population 2. The sampling distribution of the ratio of the sample variances $\frac{s_1^2}{s_2^2}$ follows an F -distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

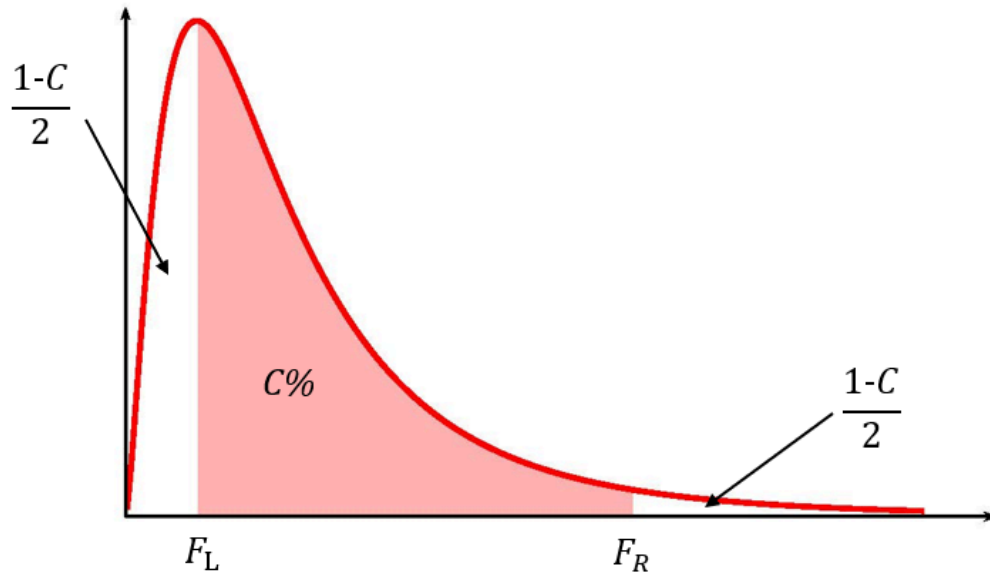
Constructing a Confidence Interval for the Ratio of Two Population Variances

Suppose a sample of size n_1 with sample variance s_1^2 is taken from population 1 and a sample of size n_2 with sample variance s_2^2 is taken from population 2, where the populations are independent and normally distributed. The limits for the confidence interval with confidence level C for the ratio of the population variances $\frac{\sigma_1^2}{\sigma_2^2}$ are

$$\text{Lower Limit} = \frac{1}{F_R} \times \frac{s_1^2}{s_2^2}$$

$$\text{Upper Limit} = \frac{1}{F_L} \times \frac{s_1^2}{s_2^2}$$

where F_L is the F -score so that the area in the left-tail of the F -distribution is $\frac{1-C}{2}$, F_R is the F -score so that the area in the right tail of the F -distribution is $\frac{1-C}{2}$ and the F -distribution has degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.



NOTES

1. Like the other confidence intervals we have seen, the F -scores are the values that trap $C\%$ of the observations in the middle of the distribution so that the area of each tail is $\frac{1-C}{2}$.
2. Because the F -distribution is not symmetrical, the confidence interval for the ratio of the population variances requires that we calculate two different F -scores: one for the left tail and one for the right tail. In Excel, we will need to use both the **f.inv** function (for the left tail) and the **f.inv.rt** function (for the right tail) to find the two different F -scores.
3. The F -score for the left tail is part of the formula for the upper limit and the F -score for the right tail is part of the formula for the lower limit. **This is not a mistake.** It follows from the formula used to determine the limits for the confidence interval.
4. It is important that the populations are independent and normally distributed. If the populations are not normal, the confidence interval will not give an accurate result.

EXAMPLE

Two local walk-in medical clinics want to determine if there is any variability in the time patients wait to see a doctor at each clinic. In a sample of 30 patients at Clinic 1, the standard deviation for the wait time to see a doctor was 45 minutes. In a sample of 40 patients at Clinic 2, the standard deviation for the wait time to see a doctor was 27 minutes. Assume the population of wait times at the two clinics are independent and normally distributed.

1. Construct a 95% confidence interval for the ratio of the variances for the wait times at the two clinics.
2. Interpret the confidence interval found in part 1.
3. Is there evidence to suggest that there is a difference in the variances of the wait times at the two clinics? Explain.

Solution:

1. Let Clinic 1 be population 1 and Clinic 2 be population 2. From the question we have the following information:

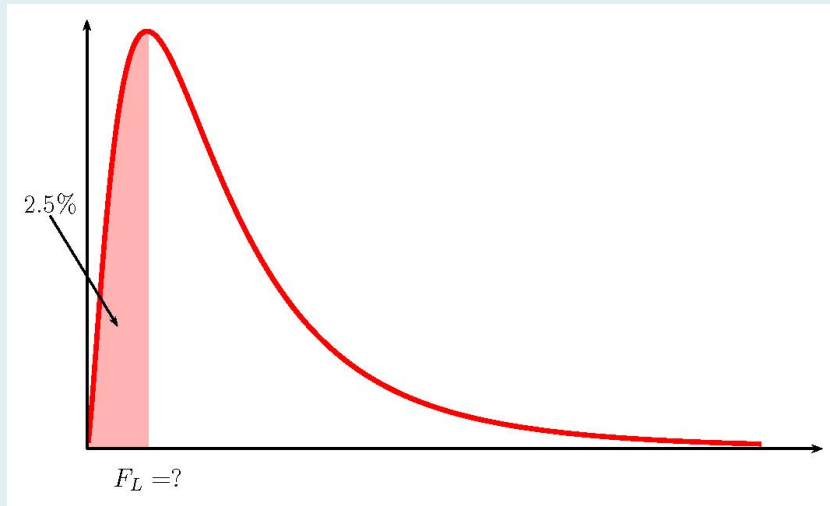
Clinic 1	Clinic 2
$n_1 = 30$	$n_2 = 40$
$s_1^2 = 45^2 = 2025$	$s_2^2 = 27^2 = 729$

To find the confidence interval, we need to find the F_L -score for the 95% confidence interval.

This means that we need to find the F_L -score so that the area in the left tail is

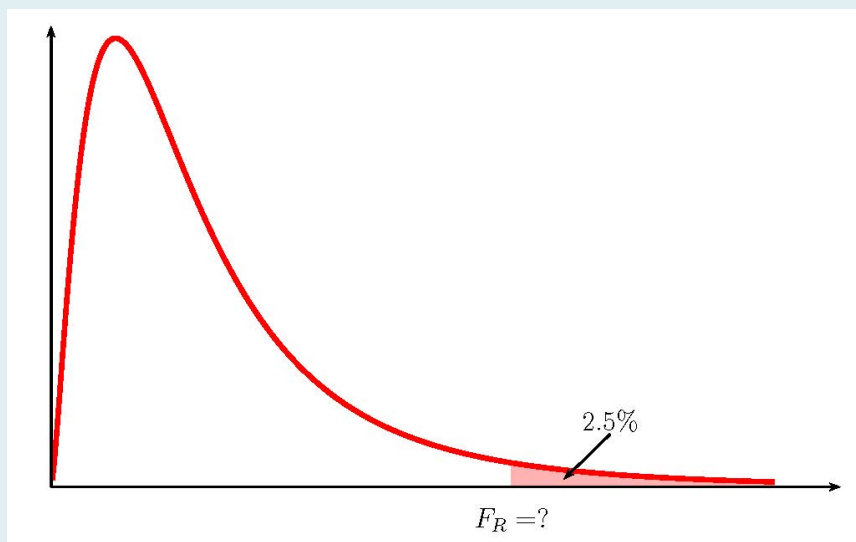
$$\frac{1 - 0.95}{2} = 0.025. \text{ The degrees of freedom for the } F\text{-distribution are}$$

$$df_1 = n_1 - 1 = 30 - 1 = 29 \text{ and } df_2 = n_2 - 1 = 40 - 1 = 39.$$



Function	f.inv	Answer
Field 1	0.025	0.4919...
Field 2	29	
Field 3	39	

We also need find the F_R -score for the 95% confidence interval. This means that we need to find the F_R -score so that the area in the right tail is $\frac{1 - 0.95}{2} = 0.025$. The degrees of freedom for the F -distribution are $df_1 = n_1 - 1 = 30 - 1 = 29$ and $df_2 = n_2 - 1 = 40 - 1 = 39$.



Function	f.inv.rt	Answer
Field 1	0.025	1.9618...
Field 2	29	
Field 3	39	

So $F_L = 0.4919\dots$ and $F_R = 1.9618\dots$. The 95% confidence interval is

$$\begin{aligned} \text{Lower Limit} &= \frac{1}{F_R} \times \frac{s_1^2}{s_2^2} = \frac{1}{1.9618\dots} \times \frac{2025}{729} = 1.416 \\ \text{Upper Limit} &= \frac{1}{F_L} \times \frac{s_1^2}{s_2^2} = \frac{1}{0.4919\dots} \times \frac{2025}{729} = 5.646 \end{aligned}$$

- We are 95% confident that the ratio of the variances in the wait times at the two clinics is between 1.416 and 5.646.
- Because 1 is outside the confidence interval, it suggests that the ratio of the variances $\frac{\sigma_1^2}{\sigma_2^2}$ is not 1. If the ratio of the variances cannot equal 1, then the variances cannot be equal. So there is a difference in the variances of the wait times at the two clinics.

NOTES

- When calculating the limits for the confidence interval keep all of the decimals in the F -scores and other values throughout the calculation. This will ensure that there is no round-off error in the answer. You can use Excel to do the calculations of the limits, clicking on the cells containing the F -scores and any other values.
- When writing down the interpretation of the confidence interval, make sure to include the confidence level and the actual ratio of population variances captured by the confidence interval (i.e. be specific to the context of the question). In this case, there are no units for the limits because variance does not have any limits.

Steps to Conduct a Hypothesis Test for Two Population

Variations

1. Write down the null hypothesis that there is no difference in the population variances:

$$\begin{array}{l} H_0: \sigma^2_1 = \sigma^2_2 \end{array}$$

The null hypothesis is always the claim that the two population variances are equal.

2. Write down the alternative hypotheses in terms of the difference in the population variances. The alternative hypothesis will be one of the following:

$$\begin{array}{l} H_a: \sigma^2_1 < \sigma^2_2 \\ H_a: \sigma^2_1 > \sigma^2_2 \\ H_a: \sigma^2_1 \neq \sigma^2_2 \end{array}$$

3. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
4. Collect the sample information for the test and identify the significance level α .
5. Use the F -distribution to find the p -value (the area in the corresponding tail) for the test. The F -score and degrees of freedom are

$$F = \frac{s_1^2}{s_2^2} \quad df_1 = n_1 - 1 \quad df_2 = n_2 - 1$$

6. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

7. Write down a concluding sentence specific to the context of the question.

EXAMPLE

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. At the 5% significance level, test the claim that the first instructor's variance is smaller.

Solution:

Let the first instructor's grades be population 1 and the second instructor's grades be population 2. From the question we have the following information:

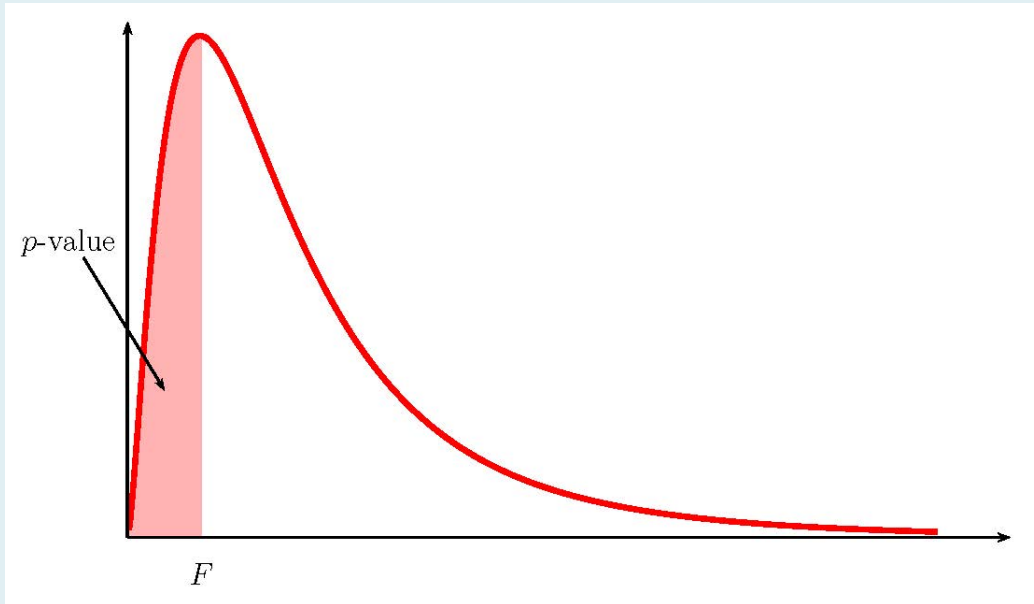
Instructor 1	Instructor 2
$n_1 = 30$	$n_2 = 30$
$s_1^2 = 52.3$	$s_2^2 = 89.9$

Hypotheses:

$$\begin{array}{l} H_0: \sigma_1^2 = \sigma_2^2 \\ H_a: \sigma_1^2 < \sigma_2^2 \end{array}$$

p-value:

Because the alternative hypothesis is a $<$, the p -value is the area in the left tail of the F -distribution.



To use the **f.dist** function, we need to calculate out the F -score and the degrees of freedom:

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ &= \frac{52.3}{89.9} \\ &= 0.58175\dots \end{aligned}$$

$$\begin{aligned} df_1 &= n_1 - 1 \\ &= 30 - 1 \\ &= 29 \end{aligned}$$

$$\begin{aligned} df_2 &= n_2 - 1 \\ &= 30 - 1 \\ &= 29 \end{aligned}$$

Function	f.dist	Answer
Field 1	0.58175...	0.0753
Field 2	29	
Field 3	29	
Field 4	true	

So the p -value = 0.0753.

Conclusion:

Because p -value = 0.0753 > 0.05 = α , we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that the first instructor's variance is smaller.

NOTES

- The null hypothesis $\sigma_1^2 = \sigma_2^2$ is the claim that the variances for the two instructors are equal.
- The alternative hypothesis $\sigma_1^2 < \sigma_2^2$ is the claim that the variance for the first instructor's grades is less than the variance for the second instructor's grades.
- The p -value is the area in the left tail of the F -distribution, to the left of $F = 0.5817\dots$.
In the calculation of the p -value:
 - The function is f.dist because we are finding the area in the left tail of an F -distribution.
 - Field 1 is the value of F .
 - Field 2 is the value of df_1 .
 - Field 3 is the value of df_2 .
 - Field 4 is true.
- The p -value of 0.0753 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variances for the two instructors are most likely equal.

EXAMPLE

A local choral society divides the male singers into tenors and basses. The choral society director wants to know if the variance in the heights of the two groups of singers is the same or different. The director takes a sample from each group and records their height in inches. In a sample of 22 tenors, the sample variance is 3.89. In a sample of 27 basses, the sample variance is 2.72. At the 5% significance level, is there a difference in the heights of the two groups of singers?

Solution:

Let the tenors be population 1 and the basses be population 2. From the question we have the following information:

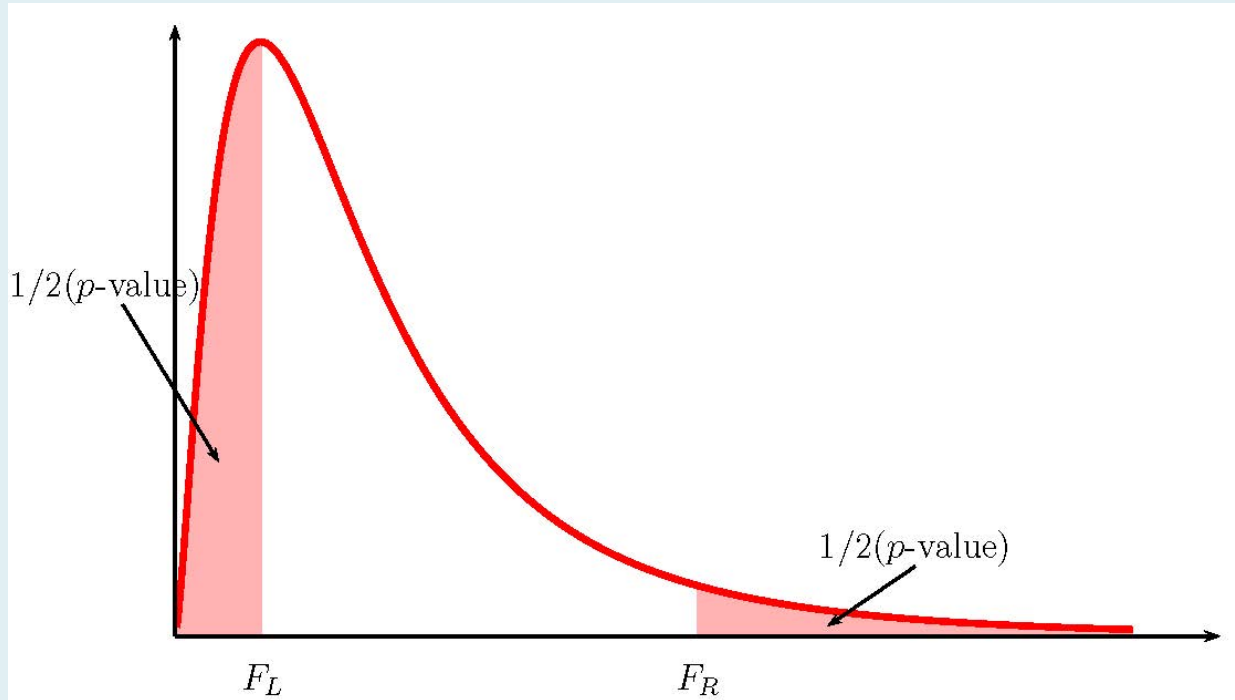
Tenors	Basses
$n_1 = 22$	$n_2 = 27$
$s_1^2 = 3.89$	$s_2^2 = 2.72$

Hypotheses:

$$\begin{array}{l} H_0: \sigma^2_1 = \sigma^2_2 \\ H_a: \sigma^2_1 \neq \sigma^2_2 \end{array}$$

p-value:

Because the alternative hypothesis is \neq , the p -value is the sum of the areas in the tails of the F -distribution.



We need to calculate out the F -score and the degrees of freedom:

$$F = \frac{s_1^2}{s_2^2} = \frac{3.89}{2.72} = 1.430... \quad \text{df}_1 = n_1 - 1 = 22 - 1 = 21 \quad \text{df}_2 = n_2 - 1 = 27 - 1 = 26$$

Because this is a two-tailed test, we need to know which tail (left or right) we have the F -score for so that we can use the correct Excel function. If $F > 1$, the F -score corresponds to the right tail. If the $F < 1$, the F -score corresponds to the left tail. In this case $F = 1.430... > 1$, so the F -score corresponds to the right tail. We need to use **f.dist.rt** to find the area in the right tail.

Function	f.dist.rt	Answer
Field 1	1.430....	0.1919
Field 2	21	
Field 3	26	

So the area in the right tail is 0.1919, which means that $\frac{1}{2}(p\text{-value})=0.1919$. This is also the area in the left tail, so

$$p\text{-value}=0.1919 + 0.1919 = 0.3838$$

Conclusion:

Because $p\text{-value} = 0.3838 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is not enough evidence to suggest that there is a difference in the variation in the heights of the two groups of singers.

NOTES

1. The null hypothesis $\sigma_1^2 = \sigma_2^2$ is the claim that the variances of the heights for the two groups of singers are equal.
2. The alternative hypothesis $\sigma_1^2 \neq \sigma_2^2$ is the claim that the variances of the heights for the two groups of singers are not equal.
3. In a two-tailed hypothesis test for two population variance, we will only have sample information relating to **one** of the two tails. We must determine which of the tails the sample information belongs to, and then calculate out the area in that tail. The area in each tail represents exactly half of the p -value, so the p -value is the sum of the areas in the two tails.
 - If $F < 1$, the sample information belongs to the **left tail**.
 - We use **f.dist** to find the area in the left tail. The area in the right tail equals the area in the left tail, so we can find the p -value by adding the output from this function to itself.
 - If $F > 1$, the sample information belongs to the **right tail**.
 - We use **f.dist.rt** to find the area in the right tail. The area in the left tail equals the area in the right tail, so we can find the p -value by adding the output from this function to itself.
4. The p -value of 0.3838 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the variances in the heights of the two groups of singers are the same.

NOTES

- When two populations have equal variances, the values of s_1^2 and s_2^2 are close in value. So, the value of $\frac{s_1^2}{s_2^2}$ is close to 1. This will result in a large p -value in the hypothesis test and the evidence favours the null hypothesis.
- When two populations have unequal variances, then the values of s_1^2 and s_2^2 are not close in value. So, the value of $\frac{s_1^2}{s_2^2}$ will either be larger than 1 or smaller than 1 (depending on which sample variance is smaller and which is larger). This will result in a small p -value in the hypothesis test and the evidence favours the alternative hypothesis.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=248#oembed-1>

Watch this video: Hypothesis Tests for Equality of Two Variances by jbstatistics [11:39]

Concept Review

To construct a confidence interval or conduct a hypothesis test on two population variances, we use the sampling distribution of the ratio of the sample variances $\frac{s_1^2}{s_2^2}$, which follows an F -distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

The hypothesis test for two population variances is a well established process:

1. Write down the null and alternative hypotheses in terms of the population variances.
2. Use the form of the alternative hypothesis to determine if the test is left-tailed, right-tailed, or two-tailed.
3. Collect the sample information for the test and identify the significance level.
4. Find the p -value (the area in the corresponding tail) for the test using the F -distribution where $F = \frac{s_1^2}{s_2^2}$, $df_1 = n_1 - 1$, and $df_2 = n_2 - 1$.
5. Compare the p -value to the significance level and state the outcome of the test.
6. Write down a concluding sentence specific to the context of the question.

The limits for the confidence interval for the ratio of the population variances $\frac{\sigma_1^2}{\sigma_2^2}$ are

$$\text{Lower Limit} = \frac{1}{F_R} \times \frac{s_1^2}{s_2^2}$$

$$\text{Upper Limit} = \frac{1}{F_L} \times \frac{s_1^2}{s_2^2}$$

where F_L is the F -score so that the area in the left-tail of the F -distribution is $\frac{1 - C}{2}$, F_R is the F -score so that the area in the right tail of the F -distribution is $\frac{1 - C}{2}$, and the F -distribution has degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

Attribution

“13.4 Test of Two Variances“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

11.4 ONE-WAY ANOVA AND HYPOTHESIS TESTS FOR THREE OR MORE POPULATION MEANS

LEARNING OBJECTIVES

- Conduct and interpret hypothesis tests for three or more population means using one-way ANOVA.

The purpose of a one-way ANOVA (analysis of variance) test is to determine the existence of a statistically significant difference among the means of three or more populations. The test actually uses variances to help determine if the population means are equal or not.

Throughout this section, we will use subscripts to identify the values for the means, sample sizes, and standard deviations for the populations:

Symbol for:	Population k
Population Mean	μ_k
Population Standard Deviation	σ_k
Sample Size	n_k
Sample Mean	\bar{x}_k
Sample Standard Deviation	s_k

k is the number of populations under study, n is the total number of observations in all of the samples combined, and $\bar{\bar{x}}$ is the mean of the sample means.

$$n = n_1 + n_2 + \cdots + n_k$$

$$\bar{x} = \frac{n_1 \times \bar{x}_1 + n_2 \times \bar{x}_2 + \cdots + n_k \times \bar{x}_k}{n}$$

One-Way ANOVA

A predictor variable is called a **factor** or **independent variable**. For example age, temperature, and gender are factors. The groups or samples are often referred to as **treatments**. This terminology comes from the use of ANOVA procedures in medical and psychological research to determine if there is a difference in the effects of different treatments.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball, baseball, hockey, and lacrosse. A random sample of players was taken from each team and their GPA recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

In this example, the factor is the sports team.

	Basketball	Baseball	Hockey	Lacrosse
	Population 1	Population 2	Population 3	Population 4
Sample Size (n_i)	5	5	5	5
Sample Mean (\bar{x}_i)	3.22	3.02	3	2.94

$$\begin{aligned} k &= 4 \\ n &= n_1 + n_2 + n_3 + n_4 \\ &= 5 + 5 + 5 + 5 \\ &= 20 \\ \overline{\overline{x}} &= \frac{n_1 \overline{x}_1 + n_2 \overline{x}_2 + n_3 \overline{x}_3 + n_4 \overline{x}_4}{n} \\ &= \frac{5 \times 3.22 + 5 \times 3.02 + 5 \times 3 + 5 \times 2.94}{20} \\ &= 3.045 \end{aligned}$$

The following assumptions are required to use a one-way ANOVA test:

1. Each population from which a sample is taken is normally distributed.
2. All samples are randomly selected and independently taken from the populations.
3. The populations are assumed to have **equal variances**.
4. The population data is numerical (interval or ratio level).

The logic behind one-way ANOVA is to compare population means based on two independent estimates of the (assumed) equal variance σ^2 between the populations:

- One estimate of the equal variance σ^2 is based on the variability among the sample means themselves (called the between-groups estimate of population variance).
- One estimate of the equal variance σ^2 is based on the variability of the data within each sample (called the within-groups estimate of population variance).

The one-way ANOVA procedure compares these two estimates of the population variance σ^2 to determine if the population means are equal or if there is a difference in the population means. Because ANOVA involves the comparison of two estimates of variance, an F -distribution is used to conduct the ANOVA test. The test statistic is an F -score that is the ratio of the two estimates of population variance:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

The degrees of freedom for the F -distribution are $df_1 = k - 1$ and $df_2 = n - k$ where k is the number of populations and n is the total number of observations in all of the samples combined.

The **variance between groups** estimate of the population variance is called the **mean square due to treatment**, MST . The MST is the estimate of the population variance determined by the variance of the sample means from the overall sample mean $\bar{\bar{x}}$. When the population means are equal, MST provides an unbiased estimate of the population variance. When the population means are not equal, MST provides an overestimate of the population variance.

$$\begin{aligned} SST &= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2 \\ MST &= \frac{SST}{k-1} \end{aligned}$$

The **variance within groups** estimate of the population variance is called the **mean square due to error**, MSE . The MSE is the pooled estimate of the population variance using the sample variances as estimates for the population variance. The MSE always provides an unbiased estimate of the population variance because it is not affected by whether or not the population means are equal.

$$\begin{aligned} SSE &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2 \\ MSE &= \frac{SSE}{n - k} \end{aligned}$$

The one-way ANOVA test depends on the fact that the variance between groups MST is influenced by differences between the population means, which results in MST being either an unbiased or overestimate of the population variance. Because the variance within groups MSE compares values of each group to its own group mean, MSE is not affected by differences between the population means and is always an unbiased estimate of the population variance.

The null hypothesis in a one-way ANOVA test is that the population means are all equal and the alternative hypothesis is that there is a difference in the population means. The F -score for the one-way ANOVA test is $F = \frac{MST}{MSE}$ with $df_1 = k - 1$ and $df_2 = n - k$. The p -value for the test is the area in the right tail of the F -distribution, to the right of the F -score.

- When the variance between groups MST and variance within groups MSE are close in value, the F -score is close to 1 and results in a large p -value. In this case, the conclusion is that the population means are equal.
- When the variance between groups MST is significantly larger than the variability within groups MSE , the F -score is large and results in a small p -value. In this case, the conclusion is that there is a difference in the population means.

Steps to Conduct a Hypothesis Test for Three or More

Population Means

1. Verify that the one-way ANOVA assumptions are met.
2. Write down the null hypothesis that there is no difference in the population means:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \dots = \mu_k \end{array}$$

The null hypothesis is always the claim that the population means are equal.

3. Write down the alternative hypotheses that there is some difference in the population means:

$$H_a: \text{at least one population mean is different from the others}$$

4. Collect the sample information for the test and identify the significance level α .
5. The p -value is the area in the right tail of the F -distribution. The F -score and degrees of freedom are

$$F = \frac{MST}{MSE} \quad df_1 = k - 1 \quad df_2 = n - k$$

6. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.
7. Write down a concluding sentence specific to the context of the question.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball,

baseball, hockey, and lacrosse. A random sample of players was taken from each team and their GPA recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Assume the populations are normally distributed and have equal variances. At the 5% significance level, is there a difference in the average GPA between the sports team.

Solution:

Let basketball be population 1, let baseball be population 2, let hockey be population 3, and let lacrosse be population 4. From the question we have the following information:

Basketball	Baseball	Hockey	Lacrosse
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$n_4 = 5$
$\bar{x}_1 = 3.22$	$\bar{x}_2 = 3.02$	$\bar{x}_3 = 3$	$\bar{x}_4 = 2.94$
$s_1^2 = 0.277$	$s_2^2 = 0.487$	$s_3^2 = 0.56$	$s_4^2 = 0.613$

Previously, we found $k = 4$, $n = 20$, and $\bar{\bar{x}} = 3.045$.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one population mean is different from the others} \end{array}$$

p-value:

To calculate out the F -score, we need to find MST and MSE .

$$\begin{aligned}
 SST &= n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 + n_3 (\bar{x}_3 - \bar{x})^2 + n_4 (\bar{x}_4 - \bar{x})^2 \\
 &= 5(3.22 - 3.045)^2 + 5(3.02 - 3.045)^2 + 5(3 - 3.045)^2 + 5(2.94 - 3.045)^2 \\
 &= 0.2215 \\
 MST &= \frac{SST}{k-1} \\
 &= \frac{0.2215}{4-1} \\
 &= 0.0738... \\
 SSE &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2 \\
 &= (5-1)0.277 + (5-1)0.487 + (5-1)0.56 + (5-1)0.623 \\
 &= 7.788 \\
 MSE &= \frac{SSE}{n-k} \\
 &= \frac{7.788}{20-4} \\
 &= 0.48675
 \end{aligned}$$

The p -value is the area in the right tail of the F -distribution. To use the **f.dist.rt** function, we need to calculate out the F -score and the degrees of freedom:

$$\begin{aligned}
 F &= \frac{MST}{MSE} \\
 &= \frac{0.0738...}{0.48675} \\
 &= 0.15168...
 \end{aligned}$$

$$\begin{aligned}
 df_1 &= k - 1 \\
 &= 4 - 1 \\
 &= 3
 \end{aligned}$$

$$\begin{aligned}
 df_2 &= n - k \\
 &= 20 - 4 \\
 &= 16
 \end{aligned}$$

Function	f.dist.rt	Answer
Field 1	0.15168...	0.9271
Field 2	3	
Field 3	16	

So the p -value = 0.9271.

Conclusion:

Because p -value = 0.9271 > 0.05 = α , we do not reject the null hypothesis. At the 5%

significance level there is enough evidence to suggest that the mean GPA for the sports teams are the same.

NOTES

1. The null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ is the claim that the mean GPA for the sports teams are all equal.
2. The alternative hypothesis is the claim that at least one of the population means is not equal to the others. The alternative hypothesis does not say that all of the population means are not equal, only that at least one of them is not equal to the others.
3. The p -value is the area in the right tail of the F -distribution, to the right of $F = 0.15168\dots$. In the calculation of the p -value:
 - The function is `f.dist.rt` because we are finding the area in the right tail of an F -distribution.
 - Field 1 is the value of F .
 - Field 2 is the value of df_1 .
 - Field 3 is the value of df_2 .
4. The p -value of 0.9271 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the population means are all equal.

ANOVA Summary Tables

The calculation of the MST , MSE , and the F -score for a one-way ANOVA test can be time consuming, even with the help of software like Excel. However, Excel has a built-in one-way ANOVA summary table that not only generates the averages, variances, MST and MSE , but also calculates the required F -score and p -value for the test.

USING EXCEL TO CREATE A ONE-WAY ANOVA SUMMARY TABLE

In order to create a one-way ANOVA summary table, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter the data into an Excel worksheet.
2. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
3. In the **Data Analysis** window, select **Anova: Single Factor**. Click **OK**.
4. In the **Input** range, enter the cell range for the data.
5. In the **Grouped By** box, select rows if your data is entered as rows (the default is columns).
6. Click on **Labels in first row** if the you included the column headings in the input range.
7. In the **Alpha** box, enter the significance level for the test.
8. From the **Output Options**, select the location where you want the output to appear.
9. Click **OK**.

This website provides additional information on using Excel to create a one-way ANOVA summary table.

NOTE

Because we are using the p -value approach to hypothesis testing, it is not crucial that we enter the actual significance level we are using for the test. The p -value (the area in the right tail of the F -distribution) is not affected by significance level. For the critical-value approach to hypothesis testing, we must enter the correct significance level for the test because the critical value does depend on the significance level.

EXAMPLE

A local college wants to compare the mean GPA for players on four of its sports teams: basketball, baseball, hockey, and lacrosse. A random sample of players was taken from each team and their GPA recorded in the table below.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Assume the populations are normally distributed and have equal variances. At the 5% significance level, is there a difference in the average GPA between the sports team.

Solution:

Let basketball be population 1, let baseball be population 2, let hockey be population 3, and let lacrosse be population 4.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one population mean is different from the others} \end{array}$$

p-value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Basketball	5	16.1	3.22	0.277		
Baseball	5	15.1	3.02	0.487		
Hockey	5	15	3	0.56		
Lacrosse	5	14.7	2.94	0.623		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.2215	3	0.073833	0.151686	0.927083	3.238872
Within Groups	7.788	16	0.48675			
Total	8.0095	19				

The p -value for the test is in the **P -value column** of the **between groups row**. So the p -value = 0.9271.

Conclusion:

Because $p\text{-value} = 0.9271 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5% significance level there is enough evidence to suggest that the mean GPA for the sports teams are the same.

NOTES

1. In the top part of the ANOVA summary table (under the Summary heading), we have the averages and variances for each of the groups (basketball, baseball, hockey, and lacrosse).
2. In the bottom part of the ANOVA summary table (under the ANOVA heading), we have

- The value of SST (in the SS column of the **between groups** row).
- The value of MST (in the MS column of the **between groups** row).
- The value of SSE (in the SS column of the **within groups** row).
- The value of MSE (in the MS column of the **within groups** row).
- The value of the F -score (in the F column of the **between groups** row).
- The p -value (in the p -value column of the **between groups** row).

EXAMPLE

A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in the table below.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Assume the heights of the plants are normally distribution and have equal variance. At the 5%

significance level, does it appear that the three media in which the bean plants were grown produced the same mean height?

Solution:

Let Tommy's plants be population 1, let Tara's plants be population 2, and let Nick's plants be population 3.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_a: \text{at least one population mean is different from the others} \end{array}$$

p-value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Tommy's Plants	5	121	24.2	11.7		
Tara's Plants	5	127	25.4	18.3		
Nick's Plants	5	122	24.4	16.3		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4.133333	2	2.066667	0.133909	0.875958	3.885294
Within Groups	185.2	12	15.43333			
Total	189.3333	14				

So the $p\text{-value} = 0.8760$.

Conclusion:

Because $p\text{-value} = 0.8760 > 0.05 = \alpha$, we do not reject the null hypothesis. At the 5%

significance level there is enough evidence to suggest that the mean heights of the plants grown in three media are the same.

NOTES

1. The null hypothesis $\mu_1 = \mu_2 = \mu_3$ is the claim that the mean heights of the plants grown in the three different media are all equal.
2. The alternative hypothesis is the claim that at least one of the population means is not equal to the others. The alternative hypothesis does not say that all of the population means are not equal, only that at least one of them is not equal to the others.
3. The p -value of 0.8760 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the population means are all equal.

TRY IT

A statistics professor wants to study the average GPA of students in four different programs: marketing, management, accounting, and human resources. The professor took a random sample of GPAs of students in those programs at the end of the past semester. The data is recorded in the table below.

Marketing	Management	Accounting	Human Resources
2.17	2.63	3.21	3.27
1.85	1.77	3.78	3.45
2.83	3.25	4.00	2.85
1.69	1.86	2.95	2.26
3.33	2.21	2.65	3.18

Assume the GPAs of the students are normally distributed and have equal variance. At the 5% significance level, is there a difference in the average GPA of the students in the different programs?

Click to see Solution

Let marketing be population 1, let management be population 2, let accounting be population 3, and let human resources be population 4.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one population mean is different from the others} \end{array}$$

p-value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Marketing	5	11.87	2.374	0.47648		
Management	5	11.72	2.344	0.37108		
Accounting	5	16.59	3.318	0.31797		
Human Resources	5	15.01	3.002	0.21947		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3.459895	3	1.153298	3.330826	0.046214	3.238872
Within Groups	5.54	16	0.34625			
Total	8.999895	19				

So the p -value = 0.0462.

Conclusion:

Because p -value = 0.0462 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that there is a difference in the average GPA of the students in the different programs.

TRY IT

A manufacturing company runs three different production lines to produce one of its products. The company wants to know if the average production rate is the same for the three lines. For each

production line, a sample of eight hour shifts was taken and the number of items produced during each shift was recorded in the table below.

Line 1	Line 2	Line 3
35	21	31
35	36	34
36	22	24
39	38	21
37	28	27
36	34	29
31	35	33
38	39	20
33	40	24

Assume the numbers of items produced on each line during an eight hour shift are normally distributed and have equal variance. At the 1% significance level, is there a difference in the average production rate for the three lines?

Click to see Solution

Let Line 1 be population 1, let Line 2 be population 2, and let Line 3 be population 3.

Hypotheses:

$$\begin{array}{l} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_a: \text{at least one population mean is different from the others} \end{array}$$

p-value:

The ANOVA summary table generated by Excel is shown below:

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Line 1	9	320	35.55556	6.027778		
Line 2	9	293	32.55556	51.52778		
Line 3	9	243	27	26		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	339.1852	2	169.5926	6.089096	0.007264	5.613591
Within Groups	668.4444	24	27.85185			
Total	1007.63	26				

So the $p\text{-value} = 0.0073$.

Conclusion:

Because $p\text{-value} = 0.0073 < 0.01 = \alpha$, we reject the null hypothesis in favour of the alternative hypothesis. At the 1% significance level there is enough evidence to suggest that there is a difference in the average production rate of the three lines.

Concept Review

A one-way ANOVA hypothesis test determines if several population means are equal. In order to conduct a one-way ANOVA test, the following assumptions must be met:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal variances.

The analysis of variance procedure compares the variation between groups MST to the variation within groups MSE . The ratio of these two estimates of variance is the F -score from an F

-distribution with $df_1 = k - 1$ and $df_2 = n - k$. The p -value for the test is the area in the right tail of the F -distribution. The statistics used in an ANOVA test are summarized in the ANOVA summary table generated by Excel.

The one-way ANOVA hypothesis test for three or more population means is a well established process:

1. Write down the null and alternative hypotheses in terms of the population means. The null hypothesis is the claim that the population means are all equal and the alternative hypothesis is the claim that at least one of the population means is different from the others.
 2. Collect the sample information for the test and identify the significance level.
 3. The p -value is the area in the right tail of the F -distribution. Use the ANOVA summary table generated by Excel to find the p -value.
 4. Compare the p -value to the significance level and state the outcome of the test.
 5. Write down a concluding sentence specific to the context of the question.
-

Attribution

“13.1 One-Way ANOVA“ and “13.2 The F Distribution and the F-Ratio“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

11.5 EXERCISES

1. Three different traffic routes are tested for mean driving time. The entries in the table are the driving times in minutes on the three different routes. At the 5% significance level, test if the mean driving time for the three routes are the same.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

2. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses. At the 5% significance level, determine if the mean age is the same in the different regions of the country.

Northeast	South	West	Central	East
16.3	16.9	16.4	16.2	17.1
16.1	16.5	16.5	16.6	17.2
16.4	16.4	16.6	16.5	16.6
16.5	16.2	16.1	16.4	16.8

3. Groups of men from three different areas of the country are to be tested for mean weight. The entries in the table are the weights for the different groups. At the 5% significance level, test if the average weight for men is the same for the three groups.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

4. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in the table are the goals per game for the different teams. At the 5% significance level, test if the mean goal scored per game is the same for the four teams.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

5. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). At the 5% significance level, is there a difference in the mean jump heights among the teams?

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

6. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. At the 5% significance level, are the scores among the different groups different?

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

7. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 5%, determine if the three formulas produce the same mean weight gain.

Weights of Student Lab Rats

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

8. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. Using a 5% significance level, test if the three mean commuting mileages are the same.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

9. The following table lists the number of pages in four different types of magazines. Using a significance level of 5%, test if the four magazine types have the same mean length. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

10. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. At the 5% significance level, test if the mean times that people watch their favorite news station are the same. Assume that all distributions are normal, the three population standard deviations are approximately the same, and the data were collected independently and randomly

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

11. Are the means for the final exams the same for all statistics class delivery types? The table shows the scores on final exams from several randomly selected classes that used the different delivery types. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a 5% significance level.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

12. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? The table shows the results of a study. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a 5% significance level.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

13. Two coworkers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. At the 5% significance level, test if the variation in the first worker's commute time is smaller than the second worker's.
14. Two students are interested in whether or not there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. At the 5% significance level, determine if the variation in the second student's scores are lower than the first student's.
15. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8 and the second cyclist has a variance of 32.1. At the 5% significance level, is there a difference in the variance in the cyclists' speeds?
16. Students Linda and Tuan are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A and Tuan feeds his rats Formula B. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's rats	Tuan's rats
43.5	47.0
39.4	40.5
41.3	38.9
46.0	46.3
38.2	44.2

- Construct a 98% confidence interval for the ratio of the variance in the net weight gain for Linda's and Tuan's rats.
- Interpret the confidence interval found in part (a).
- Is there evidence to suggest that the variance in the net weight gain for Linda and Tuan's rats is the same? Explain.

17. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 16 individuals and asked them their daily one-way commuting mileage. The results are as follows. Determine whether or not the variance in mileage driven is statistically the same among the working class and professional (middle income) groups. Use a 5% significance level.

working-class	professional (middle incomes)
17.8	16.5
26.7	17.4
49.4	22.0
9.4	7.4
65.4	9.4
47.1	2.1
19.5	6.4
51.2	13.9

18. A researcher wants to study the amount of money, in dollars, that shoppers spend on Saturdays and Sundays at the mall. A sample of shoppers is taken, and the amount of money they spent at the mall on Saturday or Sunday is recorded in the table below.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

- Construct a 93% confidence interval for the ratio of the variances for the amount of money spent on Saturdays and Sundays at the mall.
- Interpret the confidence interval found in part (a).
- Is there evidence to suggest that variance in the amount of money spent on Saturdays and Sundays at the mall is different? Explain.

19. Are the variances for incomes on the East Coast and the West Coast the same? The table shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a 5% level of significance.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

Attribution

“Chapter 13 Homework” and “Chapter 13 Practice” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

PART XII

SIMPLE LINEAR REGRESSION AND CORRELATION

Chapter Outline

12.1 Introduction to Linear Regression and Correlation

12.2 Linear Equations

12.3 Scatter Diagrams

12.4 Correlation

12.5 The Regression Equation

12.6 Coefficient of Determination

12.7 Standard Error of the Estimate

12.8 Exercises

12.1 INTRODUCTION TO LINEAR REGRESSION AND CORRELATION



Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. Photo by Joshua Rothhaas, CC BY 4.0.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it? In another example, the amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

In this chapter, we will be studying the simplest form of regression, **simple linear regression**, with one independent variable x . This involves data that fits a line in two dimensions. We will also study correlation which measures how strong the relationship is.

Attribution

“Chapter 12 Introduction” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.2 LINEAR EQUATIONS

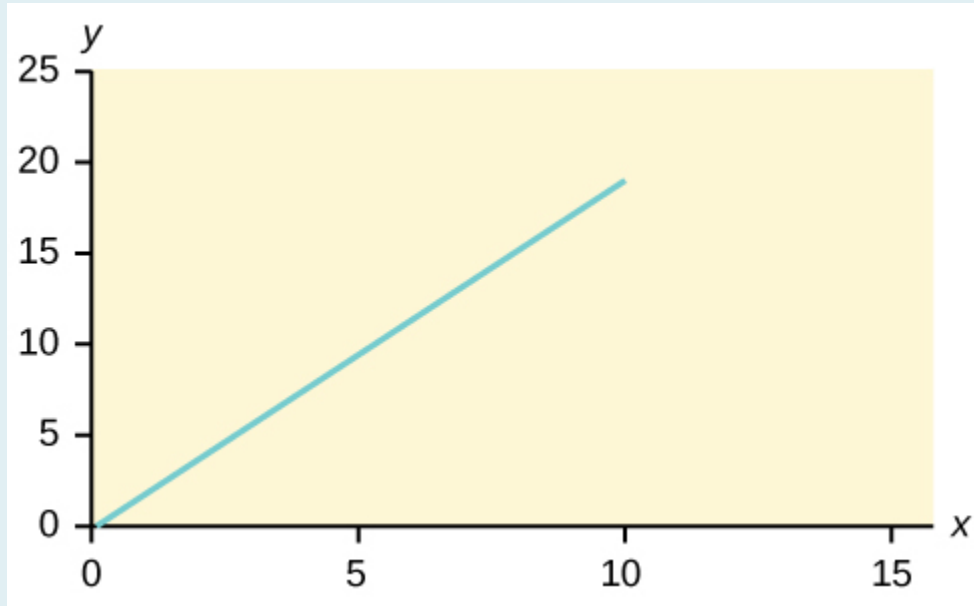
LEARNING OBJECTIVES

- Identify a linear equation, graphically or algebraically.

In this chapter we will be studying simple linear regression, which models the linear relationship between two variables x and y . A linear equation has the form $y = b_0 + b_1x$ where b_0 is the y -intercept of the line and b_1 is the slope of the line. For example, $y = 3 + 2x$ and $y = 1 - 4x$ are examples of linear equations. The graph of linear equation is a straight line.

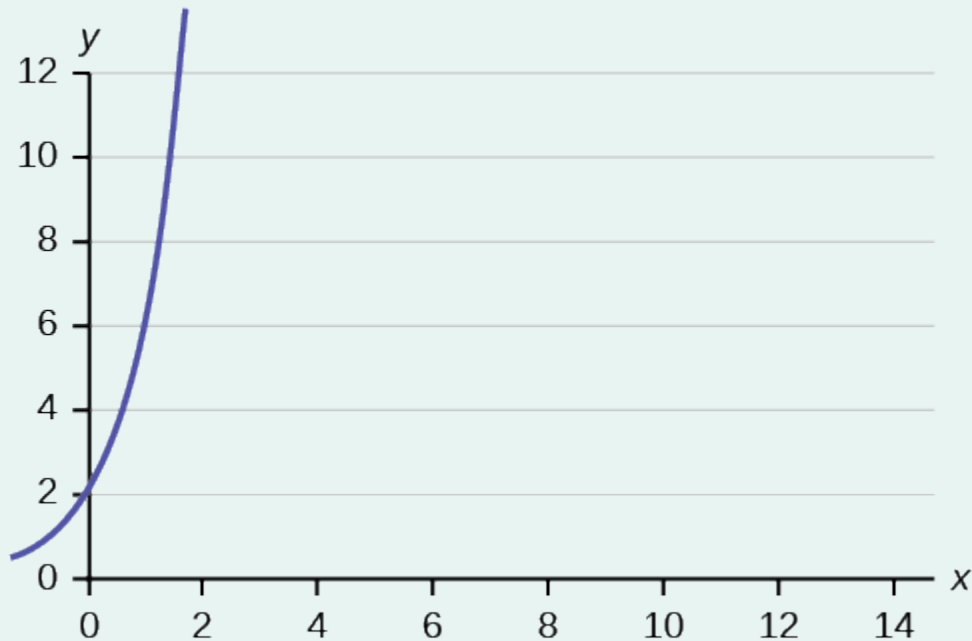
EXAMPLE

The equation $y = -1 + 2x$ is a linear equation. The slope is 2 and the y -intercept is -1 . The graph of $y = -1 + 2x$ is shown below.



TRY IT

Is the graph shown below the graph of a linear equation? Why or why not?



Click to see Solution

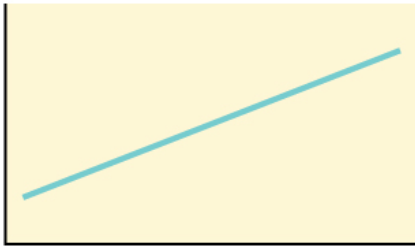
This is not a linear equation because the graph is not a straight line.

The **slope** b_1 is a number that describes the steepness of a line. The slope tells us how the value of the y variable will change for every one unit increase in the value of the x variable.

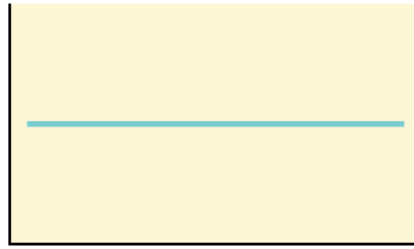
The **y -intercept** is the value of the y -coordinate where the line crosses the y -axis. Algebraically, the y -intercept is the value of y when $x = 0$.

Consider the figure below, which illustrates three different linear equations:

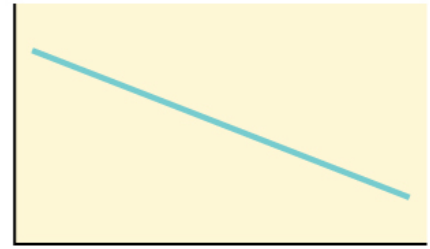
- In (a), the line rises from left to right across the graph. This means that the slope b_1 is a positive number ($b_1 > 0$).
- In (b), the line is horizontal (parallel to the x -axis). This means that the slope b_1 is zero ($b_1 = 0$).
- In (c), the line falls from left to right across the graph. This means that the slope b_1 is a negative number ($b_1 < 0$).



(a)



(b)



(c)

0 and so the line slopes upward to the right. For the second, $b = 0$ and the graph of the equation is a horizontal line. In the third graph, (c), b

EXAMPLE

Consider the linear equation $y = -25 + 15x$.

- The slope is 15. This tells us that when the value of x increases by 1, the value of y increases by 15. Because the slope is positive, the graph of $y = -25 + 15x$ rises from left to right.
- The y -intercept is -25 . This tells us that when $x = 0$, $y = -25$. On the graph of $y = -25 + 15x$, the line crosses the y -axis at -25 .

TRY IT

Consider the linear equation $y = 17 - 10x$. Identify the slope and y -intercept. Describe the slope and y -intercept in sentences.

Click to see Solution

- The slope is -10 . This tells us that when the value of x increases by 1 , the value of y decreases by 10 . Because the slope is negative, the graph of $y = 17 - 10x$ falls from left to right.
- The y -intercept is 17 . This tells us that when $x = 0$, $y = 17$. On the graph of $y = 17 - 10x$, the line crosses the y -axis at 17 .

Concept Review

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equation used, numerically with actual or predicted data values, or graphically from a plotted curve (lines are classified as straight curves). Algebraically, a linear equation typically takes the form $y = b_0 + b_1 x$, where b_0 is the y -intercept and b_1 is the slope.

The **slope** is a value that describes the rate of change of the y variable for a single unit increase in the x variable. The **y -intercept** is the value of y when $x = 0$.

Attribution

“12.1 Linear Equations” in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.3 SCATTER DIAGRAMS

LEARNING OBJECTIVES

- Define independent and dependent variables.
- Create and analyze scatter diagrams.

Independent and Dependent Variables

An **independent variable** (or the x -variable) is called the **explanatory** or **predictor** variable. The independent variable is used for prediction and provides the basis for estimation. The independent variable may be thought of as the input value and is used to determine the output value (the value of the dependent variable).

A **dependent variable** (or the y -variable) is called the **response** or **outcome** variable. The dependent variable is the variable being predicted or estimated based on the value of the independent variable. The dependent variable may be thought of as the output value and is determined by the input value (the value of the independent variable).

EXAMPLE

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time

fee of \$25 plus \$25 per hour of tutoring. Here, there are two variables: the number of hours per session and the amount of money earned per session.

- The number of hours per session is the independent variable because it can be used to predict the value of the other variable (the amount of money earned per session).
- The amount of money earned per session is the dependent variable because its value can be determined from the value of the other variable (the number of hours per session).

Scatter Diagrams

Before we begin the discussion about correlation and linear regression, we need to consider ways to display the relationship between the independent variable x and the dependent variable y . The most common and easiest way to illustrate the relationship between the two variables is with a scatter diagram.

A **scatter diagram** (or scatter plot) is a graphical presentation of the relationship between two numerical variables. Each point on the scatter diagram represents the values of two variables. The x -coordinate is the value of the independent variable and the y -coordinate is the value of the corresponding dependent variable.

To construct a scatter diagram:

1. Identify the independent and dependent variables.
2. Assign the independent variable to the horizontal or x -axis. Assign the dependent variable to the vertical or y -axis.
3. Plot the points on an (x, y) -grid.
4. Label the axes, including both the variable names and units.
5. Include a chart title. A common chart title is ***independent variable vs dependent variable***, using the actual names of the variables.

EXAMPLE

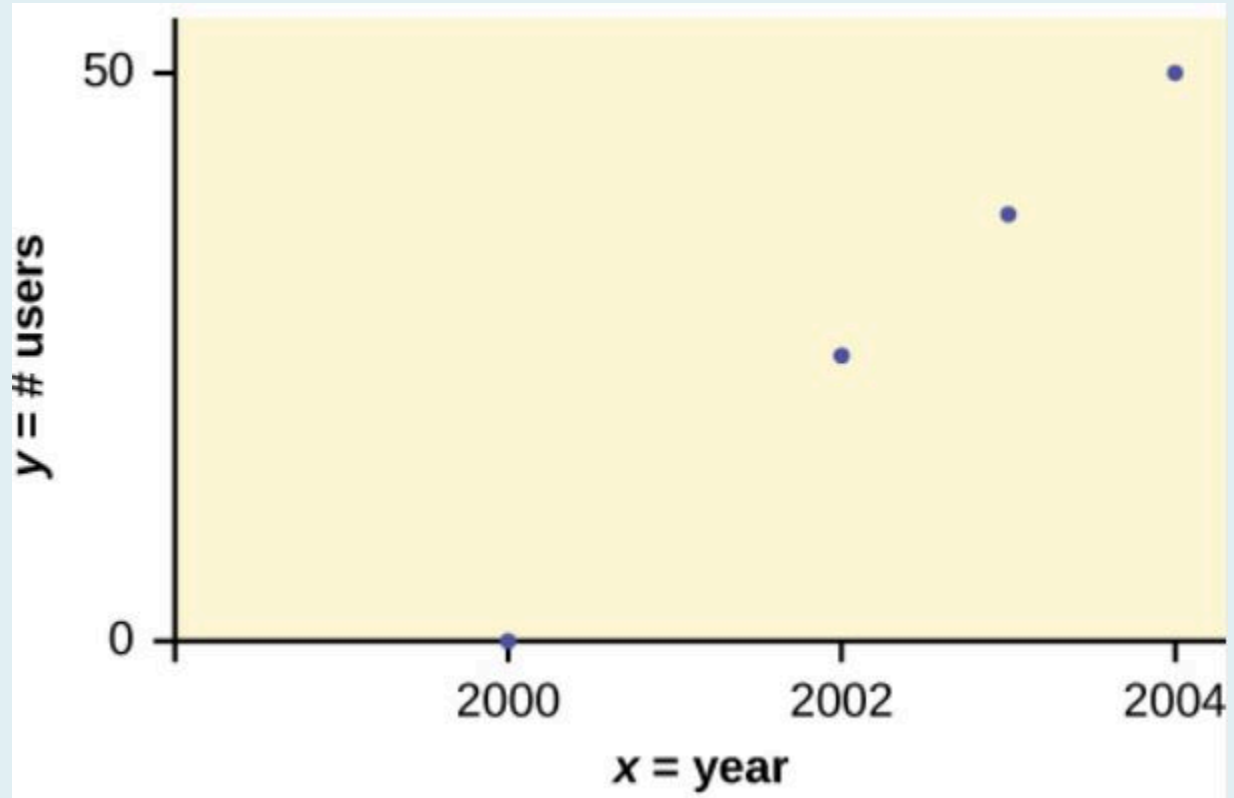
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the internet. Data for the number of user from years 2000 through 2004 is given in the table below.

Year	Number of Users (in millions)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Which variable is the independent variable? Which variable is the dependent variable? Construct a scatter diagram for this data.

Solution:

- The year is the independent variable because it can be used to predict the value of the other variable (the number of users).
- The number of users is the dependent variable because its value can be determined from the value of the other variable (year).



TRY IT

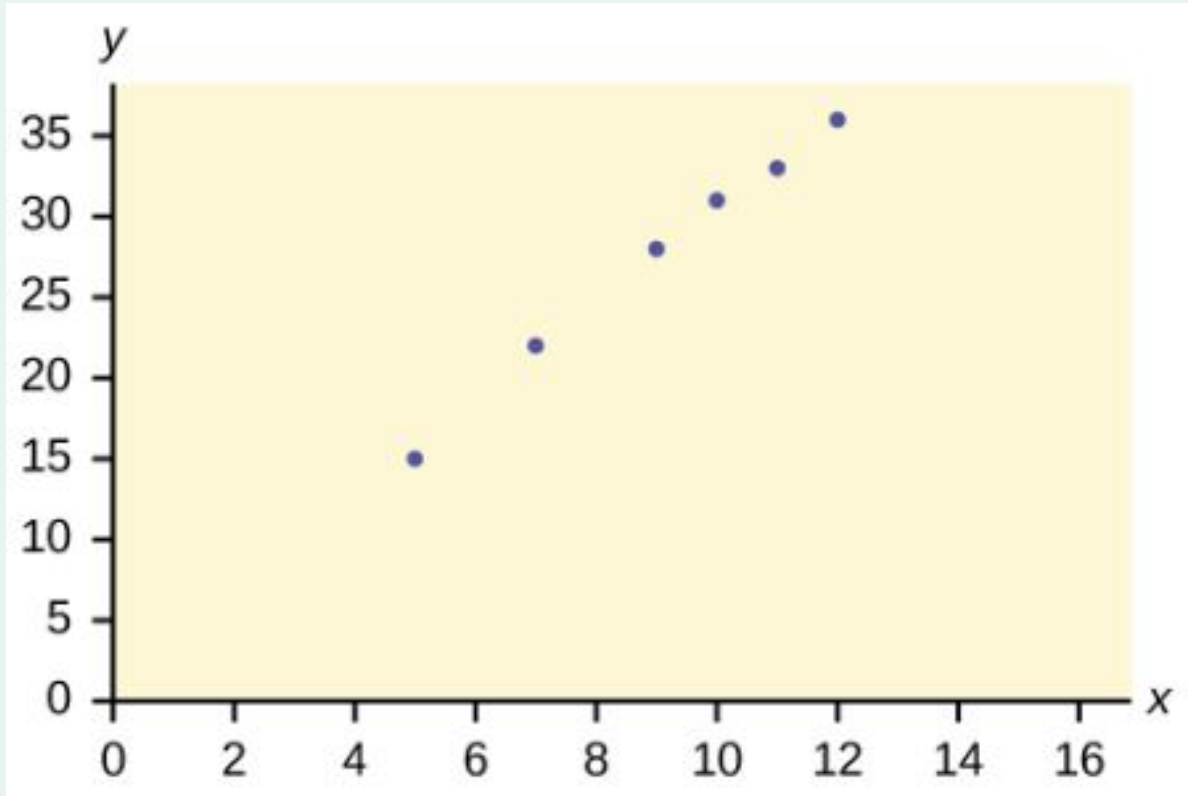
Amelia plays basketball for her high school. She wants to improve her play so she can compete at the college level. The table below records the number of hours she spends practicing her jump shot before a game and the number of points she scored in the following game.

Hours Spent Practicing Jump Shot	Points Scored in Game
5	15
7	22
9	28
10	31
11	33
12	36

Which variable is the independent variable? Which variable is the dependent variable? Construct a scatter diagram for this data.

Click to see Solution

- The hours spent practicing jump shot is the independent variable because it can be used to predict the value of the other variable (points scored in game).
- The points scored in game is the dependent variable because its value can be determined from the value of the other variable (hours spent practicing jump shot).



CONSTRUCTING A SCATTER DIAGRAM IN EXCEL

To create a scatter diagram in Excel:

1. Identify the independent and dependent variables.
2. If necessary, rearrange the columns so that the column containing the independent variable data is on the left and the dependent variable is on the right. (Excel always places the variable on the left on the horizontal axis.)
3. Go to the **Insert** tab. In the **Charts** group, click on **Scatter**. Select the scatter diagram with only markers (points).

4. Using the chart tools, add axis titles, including both the variable names and units on the axes.
5. Using the chart tools, add a chart title. A common chart title is ***independent variable vs dependent variable***, using the actual names of the variables.

Visit the Microsoft page for more information about creating a scatter diagram in Excel.

A scatter diagram shows the **direction** of the relationship between the independent and dependent variables. That is, a scatter diagram shows if the points are, in general, rising or falling as we read from left to right across the graph.

We can determine the **strength** of the relationship by looking at the scatter diagram to see how close the points are to a line, a power function, an exponential function, or to some other type of function. The stronger the relationship, the better the corresponding regression model (linear, exponential, etc.) will be at predicting values of the dependent variable.

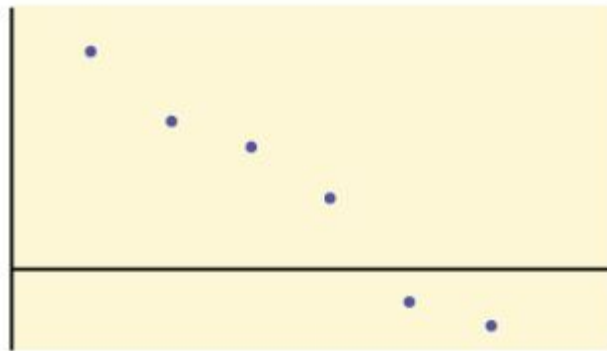
When we look at a scatter diagram, we want to notice the **overall pattern** and any **deviations** from the pattern. The scatter diagrams shown below illustrate these concepts.



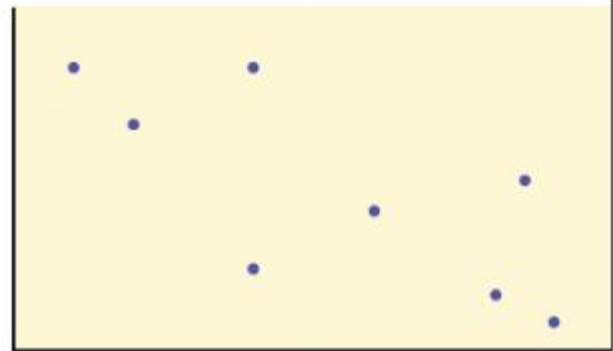
(a) Positive linear pattern (strong)



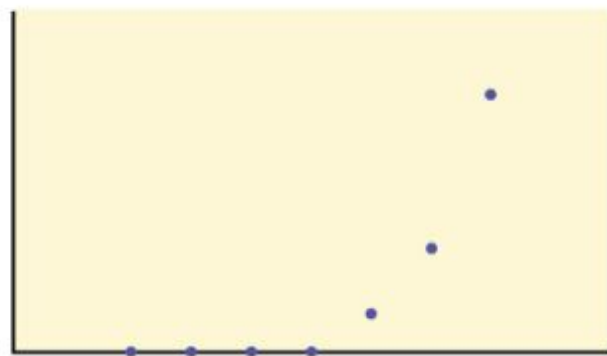
(b) Linear pattern w/ one deviation



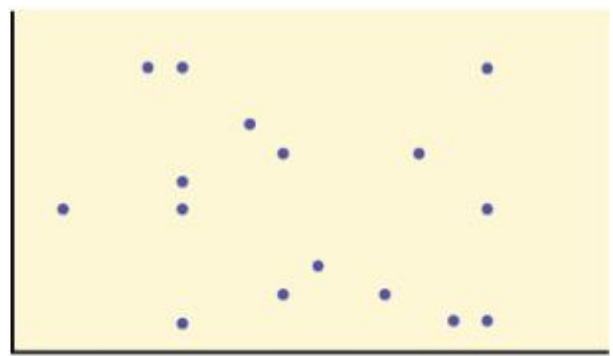
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

In this chapter, we are only concerned with the strength and direction of the **linear** relationship between the independent and dependent variables. In the next section, we will learn about a numerical measure, the correlation coefficient, that measures the strength and direction of the linear relationship.

Because linear patterns are quite common, we are interested in scatter diagrams that show a linear pattern. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If a scatter diagram shows a linear relationship, we would like to create a model based on this apparent linear relationship. This model is constructed through a process called **simple linear regression**. However, we only calculate a regression line if one of the variables, x , helps to explain or predict the other variable, y . If x is the independent variable and y is the dependent variable, then we can use a regression line to predict a value for y for a given value of x .



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=267#oembed-1>

Watch this video: Introduction to Linear Regression and Scatter Diagrams by ExcelIsFun [15:45]

Concept Review

Scatter diagrams are particularly helpful graphs when we want to see if there is a linear relationship between two variables. They indicate both the direction of the relationship between the independent variable x and the dependent variable y , and the strength of the relationship.

Attribution

“12.2 Scatter Plots“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.4 CORRELATION

LEARNING OBJECTIVES

- Calculate and interpret the correlation coefficient.

The purpose of simple linear regression is to build a linear model that can be used to make predictions for the y variable for given value of the x variable. Of course, we want the model to give us good predictions—there is no point in using a model that gives bad or inaccurate predictions. But how can we tell if the linear model will provide accurate predictions? As we have seen, we can examine the scatter diagram for a set of data to get a sense of the strength and direction of the linear relationship between the independent variable x and the dependent variable y . But we would like a **numerical measure** of the strength and direction of the linear relationship we observe on the scatter diagram. This numerical measure is called the **correlation coefficient**.

The correlation coefficient was developed by Karl Pearson in the early 1900s, and is sometimes referred to as Pearson's correlation coefficient. Denoted by r , the **correlation coefficient** is a numerical measure of the strength and direction of the linear relationship between the independent variable x and the dependent variable y . Although there is an algebraic formula to find the value of r , we will perform the calculation using the built-in function in Excel.

Interpreting the Correlation Coefficient

The value of r :

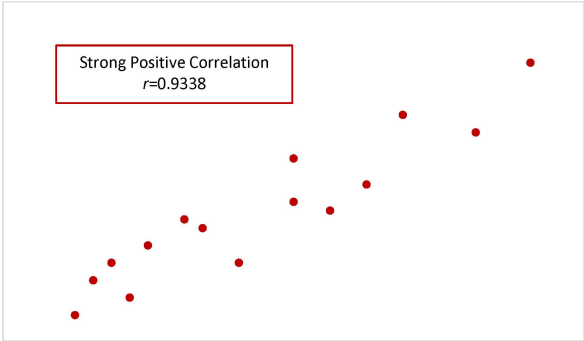
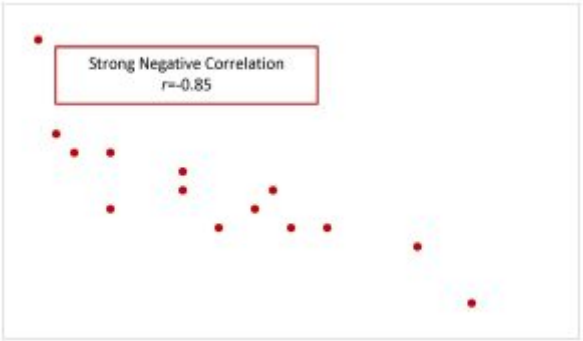
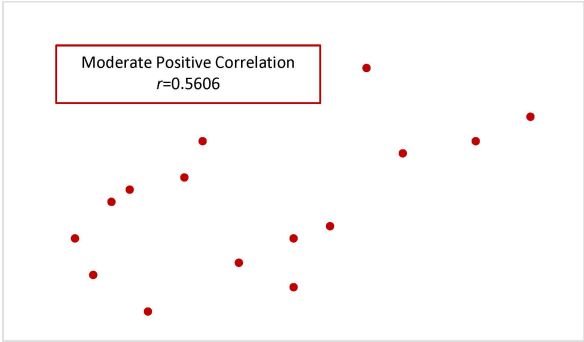
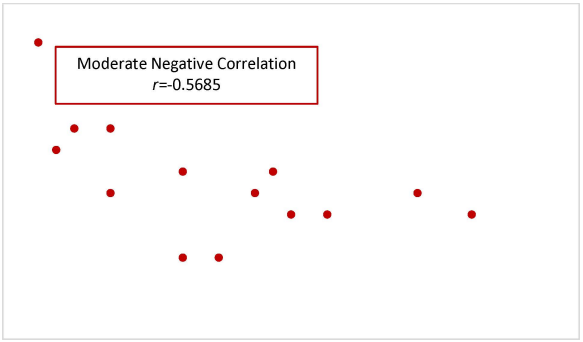
- The value of the correlation coefficient r is always a number between -1 and 1 .
- Values of r close to 1 or -1 indicate a strong linear relationship between x and y . If $r = 1$, then there is a perfect, positive correlation between x and y , in which case the points on the scatter diagram would all lie on a straight line that rises from left to right. If $r = -1$, then

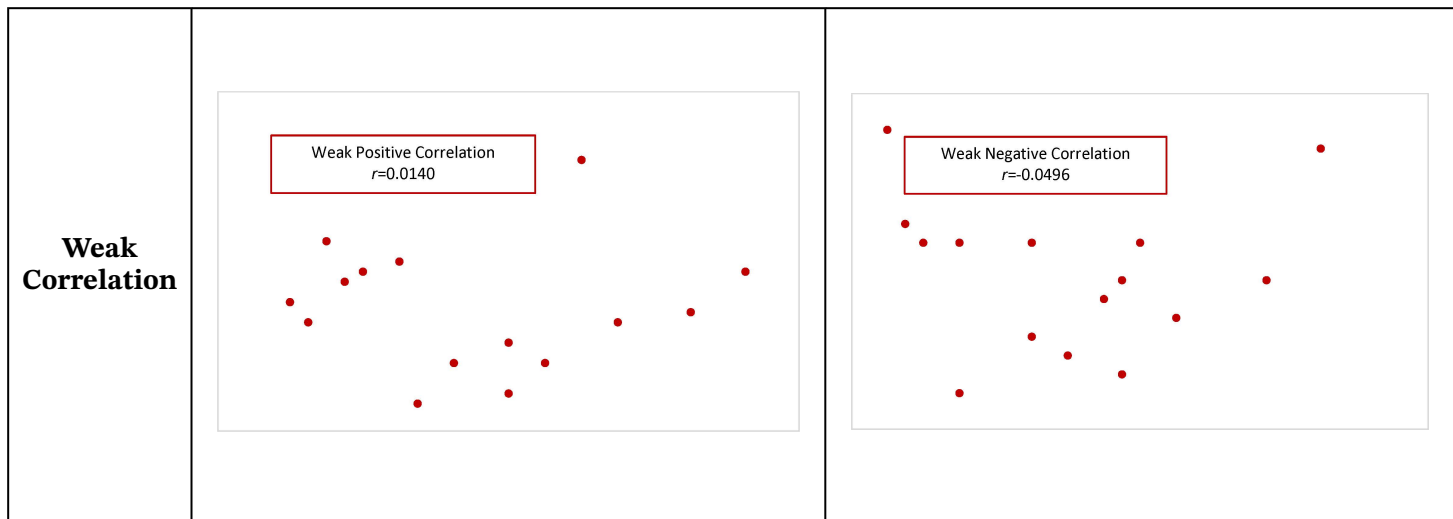
there is a perfect, negative correlation between x and y , in which case the points on the scatter diagram would all line on a straight line that falls from left to right.

- Values of r close to 0.5 or -0.5 indicate a moderate linear relationship between x and y .
- Values of r close to 0 indicate a negative linear relationship between x and y . If $r = 0$, then there is no correlation between x and y .

The sign of r :

- A positive value of r means that the points on the scatter diagram have the general tendency to rise from left to right. In other words, when x increases, y tends to increase and when x decreases, y tends to decrease.
- A negative value of r means that the points on the scatter diagram have the general tendency to fall from left to right. In other words, when x increases, y tends to decrease and when x decreases, y tends to increase.

	Positive Correlation	Negative Correlation
Strong Correlation	 <p>Strong Positive Correlation $r=0.9338$</p> <p>A scatter plot with 15 red data points showing a very tight, upward-sloping linear relationship. The points are clustered closely around a diagonal line from the bottom-left to the top-right.</p>	 <p>Strong Negative Correlation $r=-0.85$</p> <p>A scatter plot with 15 red data points showing a very tight, downward-sloping linear relationship. The points are clustered closely around a diagonal line from the top-left to the bottom-right.</p>
Moderate Correlation	 <p>Moderate Positive Correlation $r=0.5606$</p> <p>A scatter plot with 15 red data points showing a clear upward-sloping linear relationship, but with more spread around the trend line compared to the strong correlation plot.</p>	 <p>Moderate Negative Correlation $r=-0.5685$</p> <p>A scatter plot with 15 red data points showing a clear downward-sloping linear relationship, but with more spread around the trend line compared to the strong correlation plot.</p>



CALCULATING THE CORRELATION COEFFICIENT IN EXCEL

To calculate the correlation coefficient, use the **correl(array,array)** function. Enter the cell array containing the independent variable data into one of the arrays and enter the cell array containing the dependent variable data into the other array.

The output from the **correl** function is the value of the correlation coefficient.

Visit the Microsoft page for more information about the **correl** function.

NOTE

The arrays containing the independent and dependent variable data may be entered into the **correl** function in either order. The output from the **correl** function does not depend on the order in which the arrays are entered.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. Find the correlation coefficient for this data.
2. Interpret the correlation coefficient found in part 1.

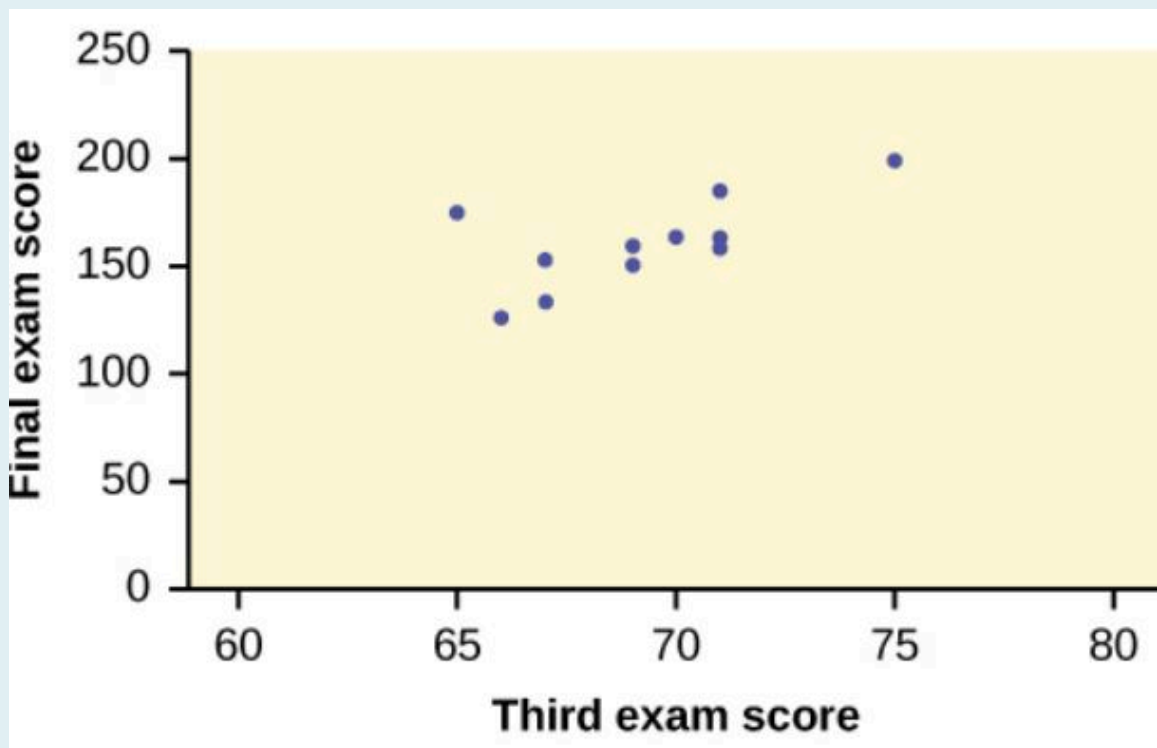
Solution:

1. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the third exam score is in column B from B1 to B11, and the final exam score is in column C from C1 to C11.

Function	correl	Answer
Field 1	B1:B11	0.6631
Field 2	C1:C11	

The value of the correlation coefficient is $r = 0.6631$.

By examining the scatter diagram for this data, shown below, we can see that the points are rising from left to right (corresponding to the fact that r is positive) and the general pattern of points corresponds to a moderate linear relationship (corresponding to the fact that r is close to 0.5).



- There is a moderate, positive linear relationship between the third test score and the final exam score.

NOTES

1. In this case the value of r is close to 0.5 , so we would consider this a moderate linear relationship.
2. When writing down the interpretation of the correlation coefficient, remember to be specific to the question using the actual names of the independent and dependent variables. Also make sure to include in the sentence the strength of the linear relationship (strong, moderate, or weak) and the direction of the linear relationship (positive or negative).

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the correlation coefficient for this data.
2. Interpret the correlation coefficient found in part 1.

Click to see Solution

1. $r = -0.9629$
2. There is a strong, negative linear relationship between depth and maximum dive time.

Correlation versus Causation

The correlation coefficient only measures the **correlation** between two variables, not **causation**. A strong correlation between two variables does not mean that changes in one variable actually cause changes in the other variable. The correlation coefficient can only tell us that changes in the independent variable and dependent variable are related. In general remember “correlation does not equal causation.”



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=276#oembed-1>

Watch this video: Using Excel to Calculate a Correlation Coefficient by Matt Macarty [5:21]

Concept Review

The correlation coefficient r measures the strength and direction of the linear relationship between x and y . The value of r is between -1 and 1 . When r is positive, the values of x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase.

Attribution

“12.3 The Regression Equation“ and “12.4 Testing the Significance of the Correlation Coefficient“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.5 THE REGRESSION EQUATION

LEARNING OBJECTIVES

- Find the equation of the line-of-best fit.
- Use the line-of-best-fit to make predictions.

We often want to use values of the independent variable to make predictions about the value of the dependent variable. For example, we might want to use the amount a business spends on advertising each quarter to make a prediction about the revenue the business will generate that quarter. When a linear relationship exists between an independent and dependent variable, we can build a linear model of that relationship, and then we can use that model to make predictions about the dependent variable.

Simple linear regression is a modeling technique in which the linear relationship between one independent variable x and one dependent variable y is approximated by a straight line, called the **line-of-best-fit** or **least squares line**. It is important to note that the line-of-best-fit only models the linear relationship between the independent and dependent variables.

The equation for the regression line is:

$$\hat{y} = b_0 + b_1 x$$

\hat{y} = predicted value of y

x = value of the independent variable

b_0 = y -intercept of the line

b_1 = slope of the line

The value of \hat{y} is the **estimated value of y** . It is the value of y obtained using the regression line. It is not generally equal to the value of y from the sample data. The values for the slope b_1 and the y -intercept b_0 in the line-of-best-fit are calculated using the sample data and the **least squares**

method. Although there are formulas to calculate the values of the slope and y -intercept in the regression line, we will calculate the slope and y -intercept using the built-in functions in Excel.

The slope of the linear regression equation:

- The slope of the line-of-best-fit b_1 and the correlation coefficient r have the same sign. That is, b_1 and r are either both positive or both negative.
- The slope b_1 of the regression equation tells us how the dependent variable y changes for a one unit increase in the independent variable x .
- When interpreting the slope, be specific to the context of the question, using the actual names of the variable and correct units.

The y -intercept of the linear regression equation:

- The y -intercept b_0 of the line-of-best-fit is the predicted value of the dependent variable y when $x = 0$.
- When interpreting the y -intercept, be specific to the context of the question, using the actual names of the variable and correct units.

CALCULATING THE SLOPE AND Formula does not parse -INTERCEPT OF THE LINEAR REGRESSION EQUATION IN EXCEL

To calculate the slope of the linear regression equation, use the **slope(array for y's,array for x's)** function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **slope** function.

To calculate the y -intercept of the linear regression equation, use the **intercept(array for y's,array for x's)** function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **intercept** function.

NOTE

The order in which the data is entered into these functions is important. In both the slope and intercept functions, the data for the **dependent** variable is entered in the **first** array and the data for the **independent** variable is entered in the **second** array. The output from the **slope** and **intercept** function will be different when the order of the inputs are switched.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. Find the equation for the line-of-best-fit.
2. Interpret the slope of the line-of-best fit.

Solution:

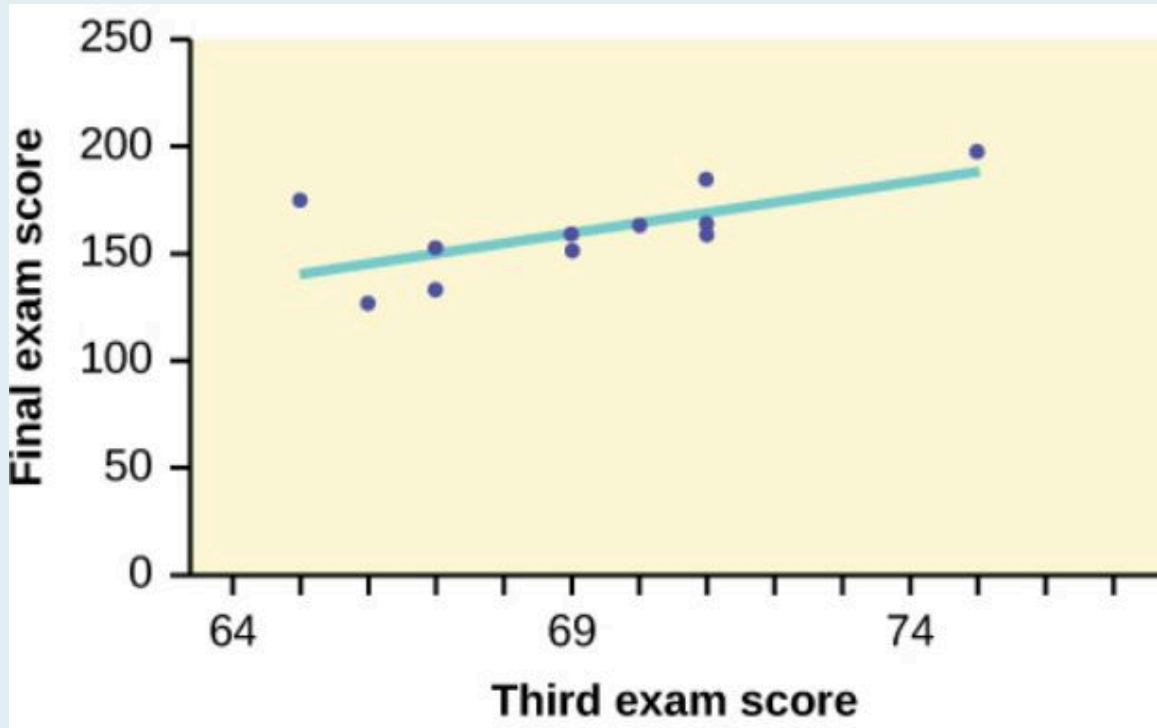
1. Because we want to predict the final exam score from the third exam score, the independent variable x is the third exam score and the dependent variable y is the final exam score. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the third exam score is in column B from B1 to B11, and the final exam score is in column C from C1 to C11.

Function	slope	Answer
Field 1	C1:C11	4.83
Field 2	B1:B11	

Function	intercept	Answer
Field 1	C1:C11	-173.51
Field 2	B1:B11	

The equation for the line-of-best-fit is $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

The graph below shows the scatter diagram with the line-of-best-fit.



2. The slope is $b_1 = 4.83$. Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points.

NOTE

1. When writing down the linear regression equation, remember to define what the variables represent in the context of the question. That is, state what x and \hat{y} represent in relation to the question.
2. When writing down the interpretation of the slope, remember to be specific to the question using the actual names of the independent and dependent variables and appropriate units.

Making Predictions with the Linear Regression Equation

Given a specific value of the independent variable x , the linear regression equation may be used to predict/estimate the value of the dependent variable y . To make predictions, the following condition must be met:

- There must be a linear relationship between the variables. The stronger the linear relationship, the better the prediction will be.
- The linear regression equation is only valid to predict values of the dependent variable. That is, we may only use the equation to solve for \hat{y} for a given value of x , and not the other way around.
- The linear regression equation should only be used to make predictions for y for values of x within the domain of the x values in the sample data used to construct the regression equation. The regression equation does not provide reliable predictions for values of x that fall outside the domain of the x values in the sample data.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor developed the linear regression model $\hat{y} = -173.51 + 4.83x$ to predict a student's final exam score (\hat{y}) from a student's third exam score (x).

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

1. What is the professor's final exam prediction for a student that scored 66 on the third exam?
2. What is the professor's final exam prediction for a student that scored 73 on the third exam?
3. Should the professor use the linear regression model to predict the final exam score for a student that scored 90 on the third exam? Why?

Solution:

1. Substitute $x = 66$ into the linear regression equation:

$$\begin{aligned}\hat{y} &= -173.51 + 4.83 * 66 \\ &= 145.27\end{aligned}$$

A student that scored 66 on the third exam has a predicted score of 145.27 on the final exam.

2. Substitute $x = 73$ into the linear regression equation:

$$\begin{aligned}\hat{y} &= -173.51 + 4.83 * 73 \\ &= 179.08\end{aligned}$$

A student that scored 73 on the third exam has a predicted score of 179.08 on the final exam.

3. The x values (third exam score) in the sample data are between 65 and 75. An x value of 90 is outside the domain of the observed x values in the data. So, we cannot **reliably** predict the final exam score for a student that scored 90 on the third exam. Of course, it is possible to

enter $x = 90$ into the linear regression equation and calculate the corresponding value of \hat{y} , but this value is not a reliable prediction. If we calculate out the value of \hat{y} in the regression equation for $x = 90$, we get $\hat{y} = 261.19$, a value that makes no sense in the context of the question because the maximum score on the final exam is 200.

NOTES

1. The values obtained for the linear regression equation are predictions only. Here, 145.27 is the **predicted** final exam score for a student that scored 66 on the third exam. This does not mean that a student that actually scored 66 on the third exam will score 145.27 on the final exam.
2. Remember that the linear regression only gives reliable predictions for values of x that fall within the domain of x values in the sample data.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the linear regression equation to predict the maximum dive time from the depth.
2. Interpret the slope of the regression equation found in part 1.
3. Predict the maximum dive time for a depth of 75 feet.

Click to see Solution

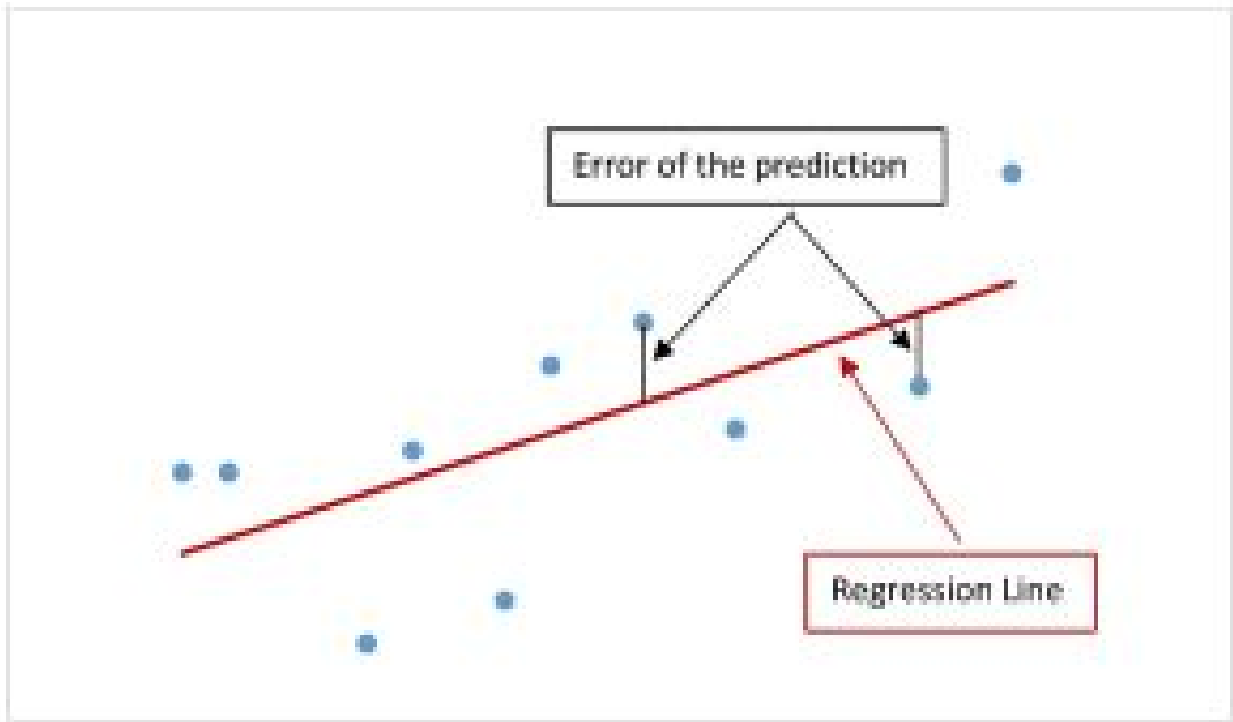
1. $\hat{y} = 127.24 - 1.11x$ where x is the depth in feet and \hat{y} is the (predicted) maximum dive time in minutes.
2. For each one foot increase in depth, the maximum dive time decreases by 1.11 minutes.
3. $\hat{y} = 127.24 - 1.11 * 75 = 43.99$ minutes

Errors and The Least Squares Method

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the linear regression equation is called the **error** or **residual**.

$$\begin{aligned} \text{Error} &= \text{Actual Value} - \text{Predicted Value} \\ &= y - \hat{y} \end{aligned}$$

Graphically, the absolute value of the error is the vertical distance between the actual value of y (the point on the scatter diagram) and the predicted value of \hat{y} (the point on the linear regression line). In other words, the absolute value of the error measures the vertical distance between the actual data point and the line.



The slope and y -intercept for the linear regression equation are generated using the errors and the **least squares method**. The idea behind finding the line-of-best-fit is based on the assumption that the data are scattered about a straight line. For any line, the errors can be calculated, squared, and then these squared errors can be added up. Of all of the possible lines, the line-of-best-fit is the **one** line that **minimizes** this sum of the squared errors. Any other line will have a higher sum of the squared errors compared to the sum of the squared errors for the line-of-best-fit.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=280#oembed-1>

Watch this video: Slope and Intercept for Linear Regression in Excel by ExcelIsFun [18:29]

Concept Review

A regression line, or a line-of best-fit, can be drawn on a scatter diagram and used to predict outcomes for the y variable in a given data set or sample data. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

Attribution

“12.3 The Regression Equation“ and “12.5 Prediction“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.6 COEFFICIENT OF DETERMINATION

LEARNING OBJECTIVES

- Calculate and interpret the coefficient of determination.

Previously, we saw how to use the correlation coefficient to measure the strength and direction of the linear relationship between the independent and dependent variables. The correlation coefficient gives us a way to measure how good a linear regression model fits the data. The coefficient of determination is another way to evaluate how well a linear regression model fits the data. Denoted r^2 , the **coefficient of determination** is the proportion of variation in the dependent variable that can be explained by the regression equation based on the independent variable. The coefficient of determination is the square of the correlation coefficient.

The coefficient of determination is a number between 0 and 1, and is the decimal form of a percent. The closer the coefficient of determination is to 1, the better the independent variable is at predicting the dependent variable. When we interpret the coefficient of determination, we use the percent form. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by the variation in the independent variable x using the regression line. When interpreting the coefficient of determination, remember to be specific to the context of the question.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

Previously we found the correlation coefficient $r = 0.6631$ and the line-of-best-fit $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

1. Find the coefficient of determination.
2. Interpret the coefficient of determination found in part 1.

Solution:

1. $r^2 = (0.6631)^2 = 0.4397$.
2. 43.97% of the variation in the final exam score can be explained by the regression line based

on the third exam score.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes. Previously, we found the correlation coefficient and the regression line to predict the maximum dive time from depth.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Find the coefficient of determination.
2. Interpret the coefficient of determination found in part 1.

Click to see Solution

1. $r^2 = (-0.9629)^2 = 0.9272$.
2. 92.72% of the variation in the maximum dive time can be explained by the regression line based on depth.

Concept Review

The coefficient of determination, r^2 , is equal to the square of the correlation coefficient. When expressed as a percent, the coefficient of determination represents the percent of variation in the dependent variable y that can be explained by the variation in the independent variable x using the regression line.

Attribution

“12.3 The Regression Equation“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

12.7 STANDARD ERROR OF THE ESTIMATE

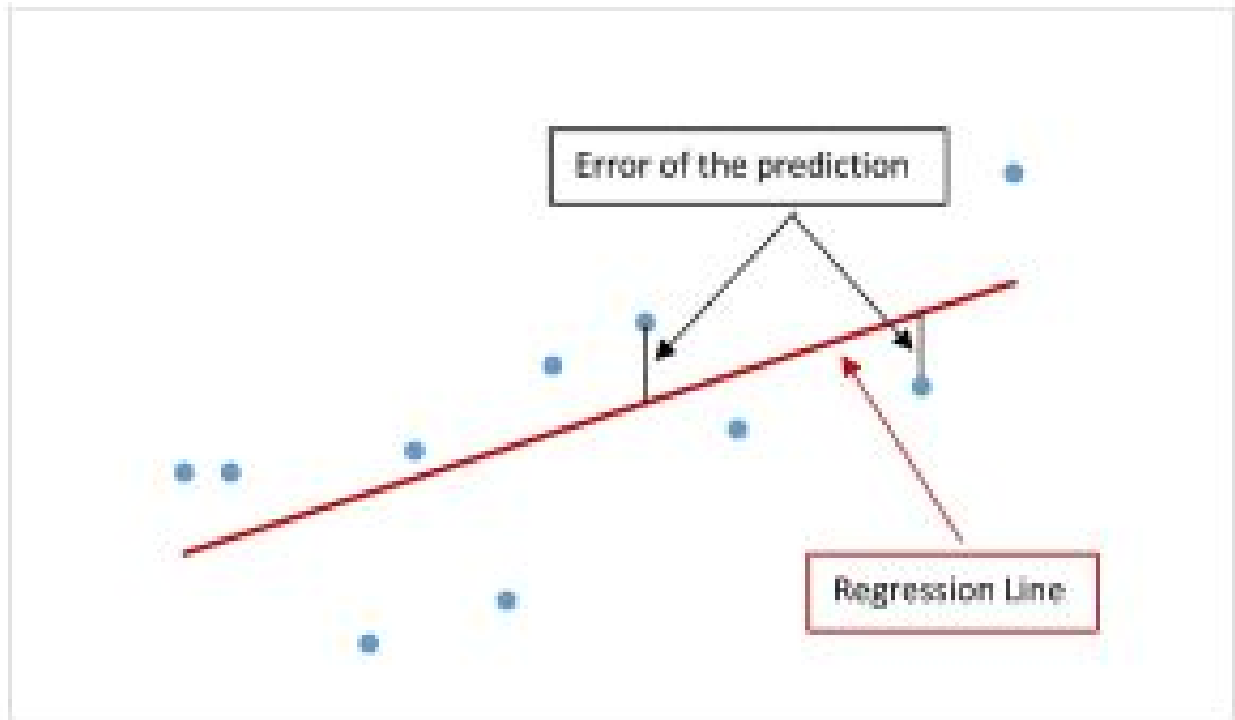
LEARNING OBJECTIVES

- Calculate and interpret the standard error of the estimate.

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the linear regression equation is called the **error** or **residual**.

$$\text{Error} = \text{Actual Value} - \text{Predicted Value}$$

Graphically, the absolute value of the error is the vertical distance between the actual value of y (the point on the scatter diagram) and the predicted value of \hat{y} (the point on the linear regression line). In other words, the absolute value of the error measures the vertical distance between the actual data point and the line.



The **standard error of the estimate**, denoted s_e , is a measure of the standard deviation of the errors in a regression model. The standard error of the estimate is a measure of the average deviation or dispersion of the points on the scatter diagram around the line-of-best-fit. The standard error of the estimate for the linear regression model is analogous to the standard deviation for a set of points, but instead of measuring the average distance from the mean we are measuring the average distance from the regression line. Graphically, the standard error of the estimate measures the average vertical distance (the absolute value of the errors) between the points on the scatter diagram and the regression line.

When the points on the scatter diagram are close to the regression line, the errors are small, and so the average of the dispersion of the points around the line will be small. In this case, the value of the standard error of the estimate will be relatively small, which reflects the fact that there is little variation between the actual data points (the points on the scatter diagram) and the linear regression model. This implies that the linear regression model is a good fit for the data and predictions made with the linear regression model will be fairly accurate.

Conversely, when the points on the scatter diagram are widely dispersed around the regression line, there errors are large, and so the average dispersion of the points around the line will be large. In this case, the value of the standard error of the estimate will be large, which reflects the greater dispersion between the actual data points and the linear regression model. This implies that the linear regression model is not a good fit for the data and predictions made with the linear regression model will be inaccurate.

The value of s_e tells us, on average, how much the dependent variable differs from the regression line based on the independent variable. When interpreting the standard error of the estimate, remember to be specific to the question, using the actual names of the dependent and independent variables, and include appropriate units. The units of the standard error of the estimate are the same as the units of the dependent variable.

Although there is a formula to calculate out the value of the standard error of the estimate, we will calculate the standard error of the estimate using the built-in function in Excel.

CALCULATING THE STANDARD ERROR OF THE ESTIMATE IN EXCEL

To calculate the standard error of the estimate, use the **ste_{yx}**(array for y's,array for x's) function.

- For **array for y's**, enter the cell array containing the **dependent** variable y data.
- For **array for x's**, enter the cell array containing the **independent** variable x data.

Visit the Microsoft page for more information about the **ste_{yx}** function.

NOTE

The order in which the data is entered into the **ste_{yx}** function is important. The data for the **dependent** variable is entered in the **first** array and the data for the **independent** variable is entered in the **second** array. The output from the **ste_{yx}** function will be different when the order of the inputs is switched.

EXAMPLE

A statistics professor wants to study the relationship between a student's score on the third exam in the course and their final exam score. The professor took a random sample of 11 students and recorded their third exam score (out of 80) and their final exam score (out of 200). The results are recorded in the table below. The professor wants to develop a linear regression model to predict a student's final exam score from the third exam score.

Student	Third Exam Score	Final Exam Score
1	65	175
2	67	133
3	71	185
4	71	163
5	66	126
6	75	198
7	67	153
8	70	163
9	71	159
10	69	151
11	69	159

Previously we found the line-of-best-fit $\hat{y} = -173.51 + 4.83x$ where x is the third exam score and \hat{y} is the (predicted) final exam score.

1. Find the standard error of the estimate.
2. Interpret the standard error of the estimate found in part 1.

Solution:

1. Enter the data into an Excel spreadsheet. For this example, suppose we entered the data (without the column headings) so that the student column is in column A from A1 to A11, the

third exam score is in column B from B1 to B11, and the final exam score is in column C from C1 to C11.

Function	steyx	Answer
Field 1	C1:C11	16.41
Field 2	B1:B11	

The value of the standard error of the estimate is $s_e = 16.41$.

- On average, the final exam score differs by 16.41 points from the regression line based on the third exam score.

TRY IT

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below shows different depths with the maximum dive times in minutes. Previously we found the regression line to predict the maximum dive time from depth.

Depth (in feet)	Maximum Dive Time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

- Find the standard error of the estimate.
- Interpret the standard error of the estimate found in part 1.

Click to see Solution

1. $s_e = 6.53$.
2. On average, the maximum dive time differs by 6.53 minutes from the regression line based on depth.

Concept Review

The standard error of the estimate, s_e , measures the average deviation or dispersion of the points on the scatter diagram around the line-of-best-fit. The smaller the value of the standard error of the estimate, the better the fit of the regression line to the data.

12.8 EXERCISES

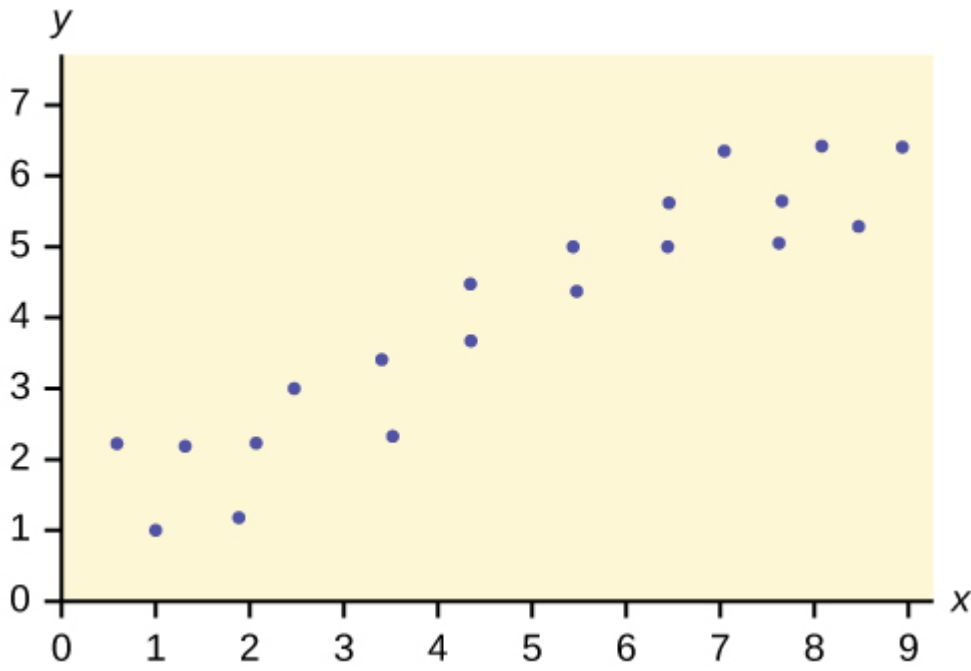
1. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of \$25 and another fee of \$12.50 an hour.
 - a. What are the dependent and independent variables?
 - b. Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.
 - c. Graph the equation from 2.
2. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?
3. Which of the following equations are linear?
 - a. $y = 6x + 8$
 - b. $y + 7 = 3x$
 - c. $y - x = 8x^2$
 - d. $4y = 8$
4. The table below contains real data for the first two decades of AIDS reporting. Use the columns “year” and “# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Adults and Adolescents only, United States

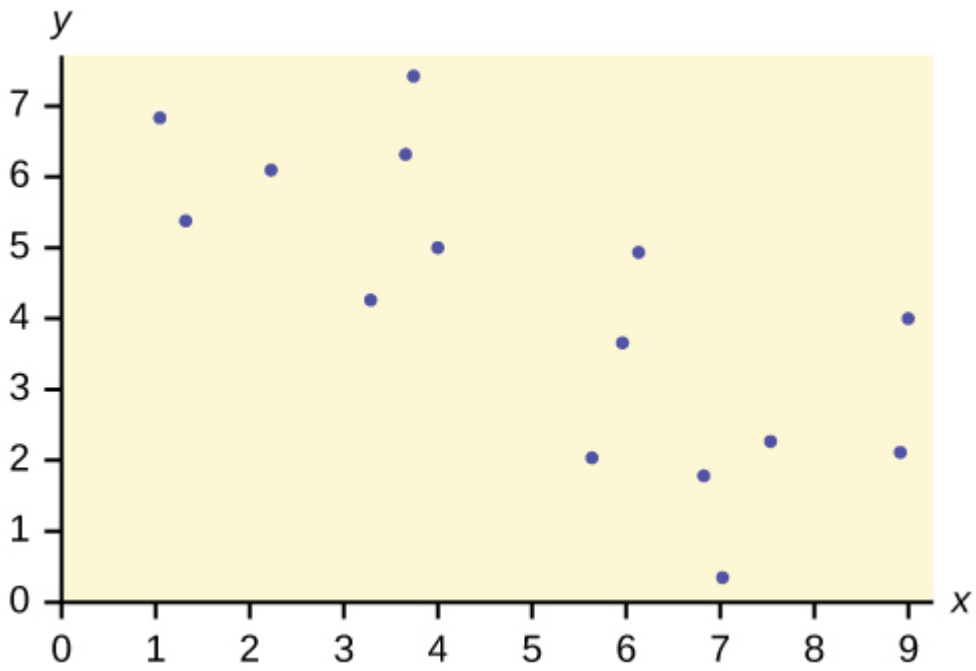
Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	802,118	489,093

5. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

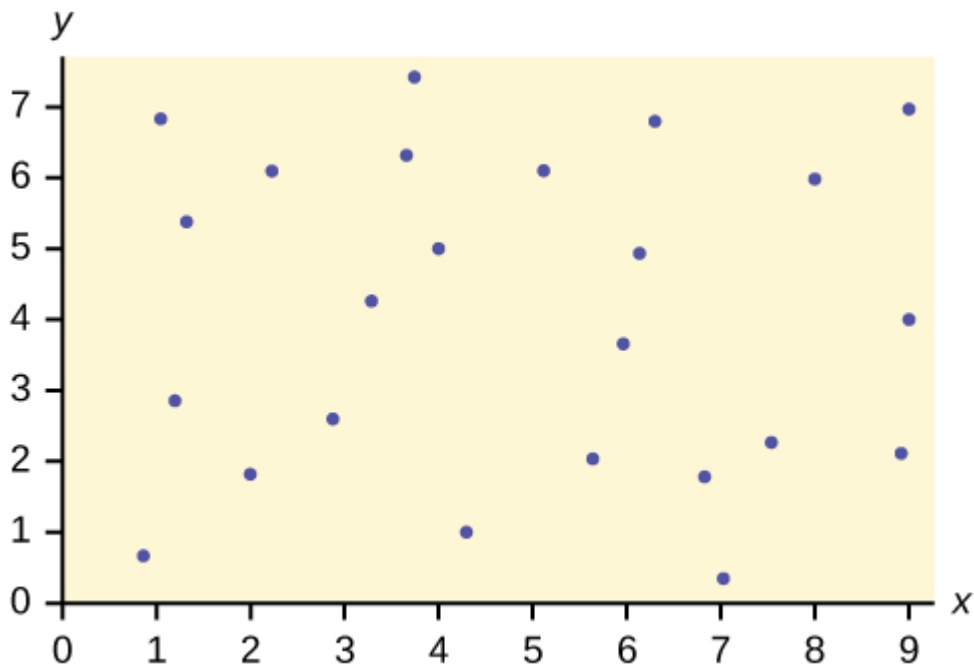
- a. What are the independent and dependent variables?
 - b. What is the y -intercept and what is the slope? Interpret them using complete sentences.
6. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.
- a. What are the independent and dependent variables?
 - b. How many pounds of soil does the shoreline lose in a year?
 - c. What is the y -intercept? Interpret its meaning.
7. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 - 1.5x$ where x is the number of hours passed in an eight-hour day of trading.
- a. What are the slope and y -intercept? Interpret their meaning.
 - b. If you owned this stock, would you want a positive or negative slope? Why?
8. For each of the following situations, state the independent variable and the dependent variable.
- a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
 - b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
 - c. Insurance companies base life insurance premiums partially on the age of the applicant.
 - d. Utility bills vary according to power consumption.
 - e. A study is done to determine if a higher education reduces the crime rate in a population.
9. Does the scatter plot appear linear? Strong, moderate, or weak? Positive or negative?



10. Does the scatter plot appear linear? Strong, moderate, or weak? Positive or negative?



11. Does the scatter plot appear linear? Strong, moderate, or weak? Positive or negative?



12. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The table below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

13. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

14. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

15. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

- a. Draw a scatter plot of the data.
 - b. Use regression to find the equation for the line of best fit.
 - c. Draw the line of best fit on the scatter plot.
 - d. What is the slope of the line of best fit? What does it represent?
 - e. What is the y -intercept of the line of best fit? What does it represent?
16. What does an r value of zero mean?
17. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?
18. Explain what it means when a correlation has an r^2 of 0.72.
19. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as $\hat{y} = 101.32 + 2.48x$ where \hat{y} is in thousands of dollars.
- a. What would you predict the sales to be on day 60?
 - b. What would you predict the sales to be on day 90?
20. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is $\hat{y} = 1350 - 1.2x$ where x is the number of hours and \hat{y} represents the number of acres left to mow.
- a. How many acres will be left to mow after 20 hours of work?
 - b. How many acres will be left to mow after 100 hours of work?
 - c. How many hours will it take to mow all of the lawns?
21. The table below contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	802,118	489,093

- a. Graph “year” versus “# AIDS cases diagnosed” (plot the scatter plot). Do not include pre-1981 data.
- b. Calculate the correlation coefficient.
- c. Interpret the correlation coefficient.

- d. Find the linear regression equation.
- e. Interpret the slope of the linear regression equation.
- f. What is the predicted number of diagnosed cases for the year 1985?
- g. What is the predicted number of diagnosed cases for the year 1970? Why doesn't this answer make sense?
- h. Calculate the coefficient of determination.
- i. Interpret the coefficient of determination.
- j. Calculate the standard error of the estimate.
- k. Interpret the standard error of the estimate.

22. Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

- a. Using “ages” as the independent variable and “Number of driver deaths per 100,000” as the dependent variable, make a scatter plot of the data.
- b. Calculate the least squares (best-fit) line.
- c. Interpret the slope of the least squares line.
- d. Predict the number of deaths people aged 40.
- e. Find the correlation coefficient.
- f. Interpret the correlation coefficient.
- g. Find the coefficient of determination.
- h. Interpret the coefficient of determination.
- i. Find the standard error of the estimate.
- j. Interpret the standard error of the estimate.

23. The table below shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- a. Decide which variable should be the independent variable and which should be the dependent variable.
- b. Draw a scatter plot of the ordered pairs.
- c. Find the correlation coefficient
- d. Interpret the correlation coefficient.
- e. Find the linear regression equation.
- f. Interpret the slope of the linear regression equation.
- g. What is the estimated life expectancy for someone born in 1950? Why doesn't this value match the life expectancy given in the table for 1950?
- h. What is the estimated life expectancy for someone born in 1982?
 - i. Using the regression equation, find the estimated life expectancy for someone born in 1850. Is this an accurate estimate for that year? Explain why or why not.
 - j. Calculate the coefficient of determination.
 - k. Interpret the coefficient of determination.
 - l. Calculate the standard error of the estimate.
 - m. Interpret the standard error of the estimate.

24. The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition ten, for various pages is given in the table below.

Page number	Maximum value (\$)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.
- Find the correlation coefficient
- Interpret the correlation coefficient.
- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated maximum value for restaurants on page 10?
- What is the estimated maximum value for restaurants on page 70?
- Using the regression equation, find the estimated maximum value for restaurants on page 200. Is this an accurate estimate for that page? Explain why or why not.
- Calculate the coefficient of determination.
- Interpret the coefficient of determination.
- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

25. The table below gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

- a. Decide which variable should be the independent variable and which should be the dependent variable.
- b. Draw a scatter plot of the ordered pairs.
- c. Find the correlation coefficient
- d. Interpret the correlation coefficient.
- e. Find the linear regression equation.
- f. Interpret the slope of the linear regression equation.
- g. What is the estimated gold medal time for 1932?
- h. What is the estimated gold medal time for 1984?
- i. Calculate the coefficient of determination.
- j. Interpret the coefficient of determination.
- k. Calculate the standard error of the estimate.
- l. Interpret the standard error of the estimate.

26. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- Using “stories” as the independent variable and “height” as the dependent variable, draw a scatter plot of the ordered pairs.
- Find the correlation coefficient
- Interpret the correlation coefficient.
- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated height for a 32 story building?
- What is the estimated height for a 94 story building?
- Using the regression equation, find the estimated height for a 6 story building. Is this an accurate estimate for the height of a 6 story building? Explain why or why not.
- Calculate the coefficient of determination.
- Interpret the coefficient of determination.
- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate

27. The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Yearly wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

- Decide which variable should be the independent variable and which should be the

dependent variable.

- Draw a scatter plot of the ordered pairs.
- Find the correlation coefficient
- Interpret the correlation coefficient.
- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- Calculate the coefficient of determination.
- Interpret the coefficient of determination.
- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

28. The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.
- Find the correlation coefficient
- Interpret the correlation coefficient.
- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated heart rate for a swim time of 34.75 minutes?
- Calculate the coefficient of determination.

- i. Interpret the coefficient of determination.
- j. Calculate the standard error of the estimate.
- k. Interpret the standard error of the estimate.

29. The table below gives percent of workers who are paid hourly rates for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- a. Decide which variable should be the independent variable and which should be the dependent variable.
- b. Draw a scatter plot of the ordered pairs.
- c. Find the correlation coefficient
- d. Interpret the correlation coefficient.
- e. Find the linear regression equation.
- f. Interpret the slope of the linear regression equation.
- g. What is the estimated percent of workers paid hourly rates in 1988?
- h. Calculate the coefficient of determination.
- i. Interpret the coefficient of determination.
- j. Calculate the standard error of the estimate.
- k. Interpret the standard error of the estimate.

30. The table below shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the ordered pairs.
- Find the correlation coefficient
- Interpret the correlation coefficient.
- Find the linear regression equation.
- Interpret the slope of the linear regression equation.
- What is the estimated average height for a one-year old?
- Using the regression equation, find the estimated average height for a 62 year old man. Do you think that your answer is reasonable? Explain why or why not.
- Calculate the coefficient of determination.
- Interpret the coefficient of determination.
- Calculate the standard error of the estimate.
- Interpret the standard error of the estimate.

Attribution

“Chapter 12 Homework” and “Chapter 12 Practice” in Introductory Statistics by OpenStax Rice University is licensed under a Creative Commons Attribution 4.0 International License.

PART XIII

MULTIPLE REGRESSION

Chapter Outline

- 13.1 Introduction to Multiple Regression
- 13.2 Multiple Regression
- 13.3 Standard Error of the Estimate
- 13.4 Coefficient of Multiple Determination
- 13.5 Testing the Significance of the Overall Model
- 13.6 Testing the Regression Coefficients
- 13.7 Multicollinearity
- 13.8 Exercises
- 13.9 Answers to Select Exercises

13.1 INTRODUCTION TO MULTIPLE REGRESSION

Previously, we studied simple linear regression, which allowed us to build a model of the linear relationship between one independent variable and one dependent variable. Then we could use the model to make predictions about the value of the dependent variable. For example, a simple linear regression model can be used to predict a person's salary (the dependent variable) from the person's age (the independent variable).

But, what if more than one independent variable impacts the value of the dependent variable? For example, a person's salary depends on more factors than just the person's age. A person's salary can also be related to their experience, their education, and their profession. We want to build a model that allows us to incorporate more than one independent variable. Because more information can be used in the model, additional independent variables can make regression models more accurate in predicting the dependent variable. A multiple regression model allows us to use two or more independent variables to predict one dependent variable.

As we saw with simple linear regression, in addition to building the model, we need ways to assess how good the multiple regression model fits the data and how good the model is at predicting values of the dependent variable.

13.2 MULTIPLE REGRESSION

LEARNING OBJECTIVES

- Develop a multiple regression model.
- Use a multiple regression model to predict values of the dependent variable.
- Interpret the partial regression coefficients.

Previously, we learned about simple linear regression, which models the linear relationship between one independent variable x and one dependent variable y . The equation for the regression line is:

$$\hat{y} = b_0 + b_1 x$$

\hat{y} = predicted value of y

x = value of the independent variable

b_0 = y -intercept of the line

b_1 = slope of the line

Multiple regression is an extension of simple linear regression where there is still only one dependent variable y but two or more independent variables x_1, x_2, \dots, x_k . Multiple regression is motivated by scenarios where many independent variables may be simultaneously connected to a dependent variable. For example, the price of product is related to demand for the product, the time of year, and the competition's price.

The equation for the multiple regression model is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

where \hat{y} is the predicted value of y , x_1, x_2, \dots, x_k are the independent variables, b_0, b_1, \dots, b_k are the regression coefficients, and k is the number of independent variables.

We will use Excel to generate the values of the regression coefficients. However, unlike simple linear regression where we used individual, built-in functions to find the slope and y -intercept, we

use a regression summary table to generate the values of the regression coefficients. As we will see, the regression summary table contains lots of information relating to the multiple regression model. For now, we will use the regression summary table to find the regression coefficients to create the multiple regression model. In later sections, we will learn about some of the other information contained on the regression summary table.

USING EXCEL TO CREATE A REGRESSION SUMMARY TABLE

In order to create a regression summary table, we need to use the Analysis ToolPak. Follow these instructions to add the Analysis ToolPak.

1. Enter the data into an Excel worksheet.
2. Go to the **Data** tab and click on **Data Analysis**. If you do not see **Data Analysis** in the **Data** tab, you will need to install the Analysis ToolPak.
3. In the **Data Analysis** window, select **Regression** and then click **OK**.
4. In the **Input Y Range**, enter the cell range for the y (dependent variable) data.
5. In the **Input X Range**, enter the cell range for the x (independent variables) data.
6. Click on **Labels in first row** if the you included the column headings in the input range.
7. From the **Output Options**, select the location where you want the output to appear. The default is a new worksheet.
8. Click **OK**. Excel will then generate a regression summary table.

NOTES

1. For the **Input X Range**, the data for the independent variables must all be together. That is, the columns (or rows) containing the data for the independent variables must all be consecutive. If the column (or row) containing data for the dependent variable is in between two columns (or rows) containing independent variables, copy the dependent variable

column and paste the dependent variable column at the beginning or end of the columns (or rows) of data. Make sure to delete the original dependent variable column after placing a copy at the beginning or end of the data.

2. There are several other options available in Regression input window, such as for confidence intervals or information about residuals. We will not need any of this other information, so leave everything else unchecked.
3. This website provides a detailed explanation of the information contained on the regression summary table.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction. Develop a multiple regression model to predict the job satisfaction score from the other variables.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Solution:

There are three independent variables: hours of unpaid work per week, age, and income (\$1000s).

Let x_1 be the hours of unpaid work per week, let x_2 be age, and let x_3 be income (\$1000s). The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The coefficients for the multiple regression model are in the **Coefficients** column in the bottom part of the table. The value of b_0 is in the Intercept row, so $b_0 = 4.7993$. The value of b_1 , the coefficient for x_1 , is in the Hours of Unpaid Work per Week row, so $b_1 = -0.3818$. The value of

b_2 , the coefficient of x_2 , is in the Age row, so $b_2 = 0.0046$. The value of b_3 , the coefficient of x_3 , is in the Income row, so $b_3 = 0.0233$.

The multiple regression equation is

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

NOTES

1. When writing down the multiple regression equation, remember to define what the variables represent in the context of the question. That is, state what \hat{y} and the independent variables represent in relation to the question.
2. A couple of the columns on the right side of the regression summary table generated by Excel were deleted in order to fit the table onto the page. These columns are not necessary for the work we will be doing.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=294#oembed-1>

Watch this video: Business Excel Business Analytics #50: Introduction to Multiple Regression, Data Analysis Regression by ExcelIsFun [13:33]

Regression Coefficients

Recall that the slope b_1 in the simple linear regression model $\hat{y} = b_0 + b_1x$ tells us how the dependent variable y changes for a single unit increase in the independent variable x . In a similar way, each regression coefficient b_i represents the change (increase or decrease) in the dependent variable for a one unit increase in the corresponding independent variable x_i , while all the other variables are held constant. When interpreting a regression coefficient, it is important to be specific to the question, using the actual names of the variables and correct units.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score
 x_1 = hours of unpaid work per week
 x_2 = age
 x_3 = income (\$1000s)

1. Interpret the regression coefficient for hours of unpaid work per week.
2. Interpret the regression coefficient for age.
3. Interpret the regression coefficient for income.

Solution:

1. $b_1 = -0.3818$. Interpretation: For a one hour increase in the hours of unpaid work per week, the job satisfaction score decreases by 0.3818, while the other variables are held constant.
2. $b_2 = 0.0046$. Interpretation: For a one year increase in the age of the employee, the job satisfaction score increases by 0.0046, while the other variables are held constant.
3. $b_3 = 0.0233$. Interpretation: For a \$1000 increase in income, the job satisfaction score increases by 0.0233, while the other variables are held constant.

NOTES

1. Remember to include “while the other variables are held constant” with the interpretation of each regression coefficient. We can only talk about how the change in one independent variable affects the dependent variable, so the values of the other variables must be kept fixed.
2. When writing down the interpretation of each regression coefficient, remember to be specific to the question using the actual names of the independent and dependent variables and appropriate units.
3. Each regression coefficient has the same units as the dependent variable.
4. Income is measured in \$1000s, so a one unit increase in income actually corresponds to a \$1000 increase in income.

Making Predictions with a Multiple Regression Model

As with simple linear regression, a multiple regression model can be used to make predictions about the dependent variable from specific values of the independent variables. To make a prediction, substitute the corresponding values of the independent variables into the multiple regression equation and calculate out the value of \hat{y} . Watch out for the units of measurement for each variable when using the multiple regression equation—the units of the values entered into the independent variable x_i in the multiple regression equation must match the units of the independent variable in the sample data.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

Predict the job satisfaction score for a 40-year old employee who works two hours of unpaid work per week and has an income of \$75,000.

Solution:

The values of the independent variables we need to enter into the multiple regression model are

$x_1 = 2$, $x_2 = 40$, and $x_3 = 75$:

$$\begin{aligned}\hat{y} &= 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3 \\ &= 4.7993 - 0.3818 \times 2 + 0.0046 \times 40 + 0.0233 \times 75 \\ &= 5.96\end{aligned}$$

The predicted job satisfaction score for a 40-year old employee who works two hours of unpaid work per week and has an income of \$75,000 is 5.96.

NOTES

1. In the sample data, income is measured in \$1000s. So an income of \$75,000 would be recorded as 75 in the sample data. So, we enter 75 for the value of x_3
2. To get the most accurate answer, use Excel to calculate out the value of \hat{y} , clicking on the corresponding cells containing the values of the coefficients in the regression summary sheet.

Assumptions about the Multiple Regression Model

The multiple regression model given above is the model we create from **sample data**—a sample is taken from the population and the sample data is used to find the regression coefficients in the model. So the regression coefficients, b_0, b_1, \dots, b_k , are estimates of the corresponding population parameters for the regression coefficients, $\beta_0, \beta_1, \dots, \beta_k$.

The population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. Because the independent variables do not account for all of the variability in the dependent variable y , the error variable ϵ captures the effects of variables other than the independent variables.

We must make certain assumptions about the regression model, in particular about the errors/residuals for the population data, in order to obtain valid conclusions about the multiple regression model. Because we do not have the population data to work with, we cannot verify if these conditions are met. However, much of regression analysis, including testing how well the data fit the model, depends on these assumptions being true.

Assumptions about the multiple regression model include:

- The model is linear.
- The errors/residuals have a normal distribution.
- The mean of the errors/residuals is 0.
- The variance of the errors/residuals is constant.
- The errors/residuals are independent.

Concept Review

In multiple regression, two or more independent variables are used to predict one dependent variable. We can find the values of the regression coefficients for the multiple regression model by generating a regression summary table in Excel. Each regression coefficient represents the change in the dependent variable y for a single unit increase in the corresponding independent variable, while the other variables are held fixed. Certain assumptions about the errors in a multiple regression model are necessary in order to test the validity of the model.

Attribution

“13.1 One-Way ANOVA“ in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License.

13.3 STANDARD ERROR OF THE ESTIMATE

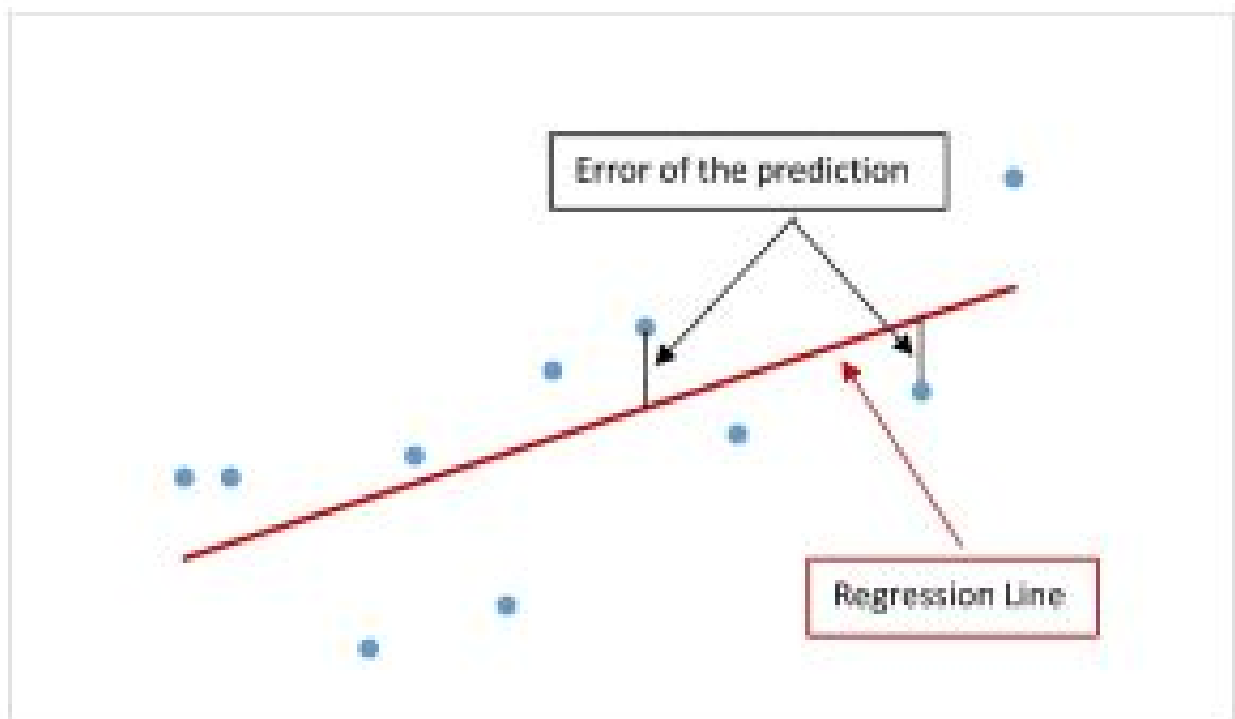
LEARNING OBJECTIVES

- Calculate and interpret the standard error of the estimate for multiple regression.

The difference between the actual value of the dependent variable y (in the sample data) and the predicted value of the dependent variable \hat{y} obtained from the multiple regression model is called the **error** or **residual**.

$$\text{Error} = \text{Actual Value} - \text{Predicted Value}$$

For the simple linear regression model, the standard error of the estimate measures the average vertical distance (the error) between the points on the scatter diagram and the regression line.



The **standard error of the estimate**, denoted s_e , is a measure of the standard deviation of the errors in a regression model. The standard error of the estimate is a measure of the average deviation of the errors, the difference between the \hat{y} -values predicted by the multiple regression model and the y -values in the sample. The standard error of the estimate for the regression model is the standard deviation of the errors/residuals.

The value of s_e tells us, on average, how much the dependent variable differs from the regression model based on the independent variables. When interpreting the standard error of the estimate, remember to be specific to the question, using the actual names of the dependent and independent variables, and include appropriate units. The units of the standard error of the estimate are the same as the units of the dependent variable.

The value of the standard error of the estimate for the regression model can be found on the regression summary table, which we learned how to generate in Excel in the previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the standard error of the estimate.
2. Interpret the standard error of the estimate.

Solution:

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The standard error of the estimate for the regression models is in the top part of the table, under the **Regression Statistics** heading in the **Standard Error** row. The value of the standard error of the estimate is $s_e = 1.5852$.

2. On average, the job satisfaction score is 1.5852 points away from the regression model based

on the independent variables “hours of unpaid work per week,” “age,” and “income.”

NOTE

The standard error of the estimate for the regression model is located in the **top** part of the table under the **Regression Statistics** heading. You will notice another standard error column at the bottom in the rows corresponding to the independent variables. These standard errors in the bottom part of the table are not related to the standard error of the estimate. In fact, the standard errors in the independent variable rows are measures of the uncertainty around the estimate of the regression coefficient for each independent variable.

Concept Review

The standard error of the estimate, s_e , measures the average deviation of the errors of the regression model. The smaller the value of the standard error of the estimate, the better the fit of the regression model to the data.

13.4 COEFFICIENT OF MULTIPLE DETERMINATION

LEARNING OBJECTIVES

- Calculate and interpret the coefficient of multiple determination.

Previously, we learned about the coefficient of determination, r^2 , for simple linear regression, which is the proportion of variation in the dependent variable that can be explained by the simple linear regression model based on the independent variable. The coefficient of determination is a good way to measure how well the simple linear regression model fits the data.

Coefficient of Multiple Determination

The **coefficient of multiple determination**, denoted R^2 , in multiple regression is similar to the coefficient of determination in simple linear regression, except in multiple regression there is more than one independent variable. The coefficient of multiple determination is the proportion of variation in the dependent variable that can be explained by the multiple regression model based on the independent variables.

The value of the coefficient of multiple determination is found on the regression summary table, which we learned how to generate in Excel in a previous section. We interpret the coefficient of multiple determination in the same way that we interpret the coefficient of determination for simple linear regression.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the coefficient of multiple determination.
2. Interpret the coefficient of multiple determination.

Solution:

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The coefficient of multiple determination for the regression model is in the top part of the table, under the **Regression Statistics** heading in the **R Square** row. The value of the coefficient of multiple determination is $R^2 = 0.5066$.

- 50.66% of the variation in the job satisfaction score can be explained by the regression model

based on the independent variables “hours of unpaid work per week,” “age,” and “income.”

Adjusted Coefficient of Multiple Determination

The value of the coefficient of multiple determination always increases as more independent variables are added to the model, even if the new independent variable has no relationship with the dependent variable. The coefficient of multiple determination is an inflated value when additional independent variables do not add any significant information to the dependent variable. Consequently, the coefficient of multiple determination is an **overestimate** of the contribution of the independent variables when new independent variables are added to the model.

Instead, we use the **adjusted coefficient of multiple determination**, denoted *adjusted R^2* , which corrects the overestimation of the coefficient of multiple determination when new independent variables are added to the model. The adjusted coefficient of multiple determination is interpreted in the same way as the coefficient of multiple determination. The adjusted coefficient of multiple determination adjusts the value of R^2 to account for the number of independent variables in the model in order to avoid overestimating the impact of adding independent variables to the model.

The adjusted coefficient of multiple determination is calculated from the value of R^2 :

$$\text{adjusted } R^2 = 1 - \left(\frac{(n-1) \times (1-R^2)}{n-k-1} \right)$$

where n is the number of observations and k is the number of independent variables. Although we can find the value of the adjusted coefficient of multiple determination using the above formula, the value of the coefficient of multiple determination is found on the regression summary table.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee’s job satisfaction from the number of hours of unpaid work per week the employee does, the employee’s age, and the employee’s income. A sample of 25 employees at the company is taken

and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

1. Find the adjusted coefficient of multiple determination.
2. Interpret the adjusted coefficient of multiple determination.

Solution:

1. The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The adjusted coefficient of multiple determination for the regression model is in the top part of the table, under the **Regression Statistics** heading in the **Adjusted R Square** row. The value of the adjusted coefficient of multiple determination is $adjusted R^2 = 0.4361$.

- 43.61% of the variation in the job satisfaction score can be explained by the regression model

based on the independent variables “hours of unpaid work per week,” “age,” and “income.”

If the addition of a new independent variable increases the value of the adjusted coefficient of multiple determination, then it is an indication that the regression model has improved as a result of adding the new independent variable. But, if the addition of a new independent variable decreases the value of the adjusted coefficient of multiple determination, then the added independent variable has not improved the overall regression model. In such cases, the new independent variable should not be added to the model.

Concept Review

The coefficient of multiple determination, R^2 , is the proportion of variation in the dependent variable that can be explained by the multiple regression model based on the independent variables. However, the addition of more independent variables into the model always causes the value of R^2 to increase, whether or not the added independent variables are actually related to the dependent variable. Instead, the adjusted coefficient of multiple determination, *adjusted R^2* , corrects for the overestimation of R^2 when new independent variables are added to the model.

13.5 TESTING THE SIGNIFICANCE OF THE OVERALL MODEL

LEARNING OBJECTIVES

- Conduct and interpret an overall model test on a multiple regression model.

Previously, we learned that the population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. The error variable ϵ accounts for the variability in the dependent variable that is not captured by the linear relationship between the dependent and independent variables. The value of ϵ cannot be determined, but we must make certain assumptions about ϵ and the errors/residuals in the model in order to conduct a hypothesis test on how well the model fits the data. These assumptions include:

- The model is linear.
- The errors/residuals have a normal distribution.
- The mean of the errors/residuals is 0.
- The variance of the errors/residuals is constant.
- The errors/residuals are independent.

Because we do not have the population data, we cannot verify that these conditions are met. We need to assume that the regression model has these properties in order to conduct hypothesis tests on the model.

Testing the Overall Model

We want to test if there is a relationship between the dependent variable and the **set** of independent variables. In other words, we want to determine if the regression model is valid or invalid.

- **Invalid Model.** There is no relationship between the dependent variable and the set of independent variables. In this case, all of the regression coefficients β_i in the population model are zero. This is the claim for the null hypothesis in the overall model test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$
- **Valid Model.** There is a relationship between the dependent variable and the set of independent variables. In this case, at least one of the regression coefficients β_i in the population model is not zero. This is the claim for the alternative hypothesis in the overall model test: $H_a : \text{at least one } \beta_i \neq 0.$

The overall model test procedure compares the means of explained and unexplained variation in the model in order to determine if the explained variation (caused by the relationship between the dependent variable and the set of independent variables) in the model is larger than the unexplained variation (represented by the error variable ϵ). If the explained variation is larger than the unexplained variation, then there is a relationship between the dependent variable and the set of independent variables, and the model is valid. Otherwise, there is no relationship between the dependent variable and the set of independent variables, and the model is invalid.

The logic behind the overall model test is based on two independent estimates of the variance of the errors:

- One estimate of the variance of the errors, MSR , is based on the mean amount of explained variation in the dependent variable y .
- One estimate of the variance of the errors, MSE , is based on the mean amount of unexplained variation in the dependent variable y .

The overall model test compares these two estimates of the variance of the errors to determine if there is a relationship between the dependent variable and the set of independent variables. Because the overall model test involves the comparison of two estimates of variance, an F -distribution is used to conduct the overall model test, where the test statistic is the ratio of the two estimates of the variance of the errors.

The **mean square due to regression**, MSR , is one of the estimates of the variance of the errors. The MSR is the estimate of the variance of the errors determined by the variance of the predicted \hat{y} -values from the regression model and the mean of the y -values in the sample, \bar{y} . If

there is no relationship between the dependent variable and the set of independent variables, then the MSR provides an unbiased estimate of the variance of the errors. If there is a relationship between the dependent variable and the set of independent variables, then the MSR provides an overestimate of the variance of the errors.

$$\begin{aligned} SSR &= \sum \left(\hat{y} - \overline{y} \right)^2 \\ MSR &= \frac{SSR}{k} \end{aligned}$$

The **mean square due to error**, MSE , is the other estimate of the variance of the errors. The MSE is the estimate of the variance of the errors determined by the error $(y - \hat{y})$ in using the regression model to predict the values of the dependent variable in the sample. The MSE always provides an unbiased estimate of the variance of errors, regardless of whether or not there is a relationship between the dependent variable and the set of independent variables.

$$\begin{aligned} SSE &= \sum (y - \hat{y})^2 \\ MSE &= \frac{SSE}{n - k - 1} \end{aligned}$$

The overall model test depends on the fact that the MSR is influenced by the explained variation in the dependent variable, which results in the MSR being either an unbiased or overestimate of the variance of the errors. Because the MSE is based on the unexplained variation in the dependent variable, the MSE is not affected by the relationship between the dependent variable and the set of independent variables, and is always an unbiased estimate of the variance of the errors.

The null hypothesis in the overall model test is that there is no relationship between the dependent variable and the set of independent variables. The alternative hypothesis is that there is a relationship between the dependent variable and the set of independent variables. The F -score for the overall model test is the ratio of the two estimates of the variance of the errors, $F = \frac{MSR}{MSE}$ with $df_1 = k$ and $df_2 = n - k - 1$. The p -value for the test is the area in the right tail of the F -distribution to the right of the F -score.

NOTES

1. If there is no relationship between the dependent variable and the set of independent variables, both the MSR and the MSE are unbiased estimates of the variance of the errors. In this case, the MSR and the MSE are close in value, which results in an F -score close to 1 and a large p -value. The conclusion of the test would be that the null hypothesis is true.

2. If there is a relationship between the dependent variable and the set of independent variables, the MSR is an overestimate of the variance of the errors. In this case, the MSR is significantly larger than the MSE , which results in a large F -score and a small p -value. The conclusion of the test would be that the alternative hypothesis is true.

Steps to Conduct a Hypothesis Test on the Overall Regression Model

1. Write down the null hypothesis that there is no relationship between the dependent variable and the set of independent variables:

$$\begin{array}{l} H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ \end{array}$$

2. Write down the alternative hypotheses that there is a relationship between the dependent variable and the set of independent variables:

$$\begin{array}{l} H_a: \text{at least one } \beta_i \text{ is not } 0 \\ \end{array}$$

3. Collect the sample information for the test and identify the significance level α .
4. The p -value is the area in the right tail of the F -distribution. The F -score and degrees of freedom are

$$F = \frac{MSR}{MSE} \quad df_1 = k \quad df_2 = n - k - 1$$

5. Compare the p -value to the significance level and state the outcome of the test:
 - If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
 - If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

The calculation of the MSR , the MSE , and the F -score for the overall model test can be time consuming, even with the help of software like Excel. However, the required F -score and p -value for the test can be found on the regression summary table, which we learned how to generate in Excel in a previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the validity of the overall model to predict the job satisfaction score.

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_a: \text{at least one } \beta_i \text{ is not } 0 \end{array}$$

p-value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p -value for the overall model test is in the middle part of the table under the **ANOVA** heading in the **Significance F column** of the **Regression row**. So the p -value=0.0017.

Conclusion:

Because p -value= 0.0017 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that there is a

relationship between the dependent variable “job satisfaction” and the set of independent variables “hours of unpaid work per week,” “age”, and “income.”

NOTES

1. The null hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ is the claim that all of the regression coefficients are zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the set of independent variables, which means that the model is not valid.
2. The alternative hypothesis is the claim that at least one of the regression coefficients is not zero. The alternative hypothesis is the claim that at least one of the independent variables is linearly related to the dependent variable, which means that the model is valid. The alternative hypothesis does not say that all of the regression coefficients are not zero, only that at least one of them is not zero. The alternative hypothesis does not tell us which independent variables are related to the dependent variable.
3. The p -value for the **overall model test** is located in the middle part of the table under the **Significance F column** heading in the **Regression row** (right underneath the **ANOVA heading**). You will notice a p -value column heading at the bottom of the table in the rows corresponding to the independent variables. These p -values in the bottom part of the table are not related to the overall model test we are conducting here. These p -values in the independent variable rows are the p -values we will need when we conduct tests on the individual regression coefficients in the next section.
4. The p -value of 0.0017 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, at least one of the regression coefficients is not zero and at least one independent variable is linearly related to the dependent variable.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://ecampusontario.pressbooks.pub/introstats/?p=300#oembed-1>

Watch this video: Basic Excel Business Analytics #51: Testing Significance of Regression Relationship with p-value by ExcelIsFun [20:44]

Concept Review

The overall model test determines if there is a relationship between the dependent variable and the **set** of independent variable. The test compares two estimates of the variance of the errors (MSR and MSE). The ratio of these two estimates of the variance of the errors is the F -score from an F -distribution with $df_1 = k$ and $df_2 = n - k - 1$. The p -value for the test is the area in the right tail of the F -distribution. The p -value can be found on the regression summary table generated by Excel.

The overall model hypothesis test is a well established process:

1. Write down the null and alternative hypotheses in terms of the regression coefficients. The null hypothesis is the claim that there is no relationship between the dependent variable and the set of independent variables. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the set of independent variables.
2. Collect the sample information for the test and identify the significance level.
3. The p -value is the area in the right tail of the F -distribution. Use the regression summary table generated by Excel to find the p -value.
4. Compare the p -value to the significance level and state the outcome of the test.
5. Write down a concluding sentence specific to the context of the question.

13.6 TESTING THE REGRESSION COEFFICIENTS

LEARNING OBJECTIVES

- Conduct and interpret a hypothesis test on individual regression coefficients.

Previously, we learned that the population model for the multiple regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where x_1, x_2, \dots, x_k are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the population parameters of the regression coefficients, and ϵ is the error variable. In multiple regression, we estimate each population regression coefficient β_i with the sample regression coefficient b_i .

In the previous section, we learned how to conduct an overall model test to determine if the regression model is valid. If the outcome of the overall model test is that the model is valid, then at least one of the independent variables is related to the dependent variable—in other words, at least one of the regression coefficients β_i is not zero. However, the overall model test does not tell us which independent variables are related to the dependent variable. To determine which independent variables are related to the dependent variable, we must test each of the regression coefficients.

Testing the Regression Coefficients

For an individual regression coefficient, we want to test if there is a relationship between the dependent variable y and the independent variable x_i .

- **No Relationship.** There is no relationship between the dependent variable y and the independent variable x_i . In this case, the regression coefficient β_i is zero. This is the claim

for the null hypothesis in an individual regression coefficient test: $H_0 : \beta_i = 0$.

- **Relationship.** There is a relationship between the dependent variable y and the independent variable x_i . In this case, the regression coefficients β_i is not zero. This is the claim for the alternative hypothesis in an individual regression coefficient test: $H_a : \beta_i \neq 0$. We are not interested if the regression coefficient β_i is positive or negative, only that it is not zero. We only need to find out if the regression coefficient is not zero to demonstrate that there is a relationship between the dependent variable and the independent variable. This makes the test on a regression coefficient a two-tailed test.

In order to conduct a hypothesis test on an individual regression coefficient β_i , we need to use the distribution of the sample regression coefficient b_i :

- The mean of the distribution of the sample regression coefficient is the population regression coefficient β_i .
- The standard deviation of the distribution of the sample regression coefficient is σ_{b_i} . Because we do not know the population standard deviation we must estimate σ_{b_i} with the sample standard deviation s_{b_i} .
- The distribution of the sample regression coefficient follows a normal distribution.

Because we are using a sample standard deviation to estimate a population standard deviation in a normal distribution, we need to use a t -distribution with $n - k - 1$ degrees of freedom to find the p -value for the test on an individual regression coefficient. The t -score for the test is $t = \frac{b_i - \beta_i}{s_{b_i}}$.

Steps to Conduct a Hypothesis Test on a Regression Coefficient

1. Write down the null hypothesis that there is no relationship between the dependent variable y and the independent variable x_i :

$$\begin{array}{l} H_0: \beta_i = 0 \end{array}$$

2. Write down the alternative hypotheses that is a relationship between the dependent variable y and the independent variable x_i :

$$\begin{array}{l} H_a: \beta_i \neq 0 \end{array}$$

3. Collect the sample information for the test and identify the significance level α .
4. The p -value is the sum of the area in the tails of the t -distribution. The t -score and degrees of freedom are

$$t = \frac{b_i - \beta_i}{s_{b_i}} \quad df = n - k - 1$$

5. Compare the p -value to the significance level and state the outcome of the test:

- If $p\text{-value} \leq \alpha$, reject H_0 in favour of H_a .
 - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H_0 is an incorrect belief and that the alternative hypothesis H_a is most likely correct.
- If $p\text{-value} > \alpha$, do not reject H_0 .
 - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis H_a may be correct.

6. Write down a concluding sentence specific to the context of the question.

The required t -score and p -value for the test can be found on the regression summary table, which we learned how to generate in Excel in a previous section.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week”.

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{array}$$

p-value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p -value for the test on the hours of unpaid work per week regression coefficient is in the bottom part of the table under the **P-value column** of the **Hours of Unpaid Work per Week row**. So the p -value=0.0082.

Conclusion:

Because p -value= 0.0082 < 0.05 = α , we reject the null hypothesis in favour of the alternative

hypothesis. At the 5% significance level there is enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week.”

NOTES

1. The null hypothesis $\beta_1 = 0$ is the claim that the regression coefficient for the independent variable X_1 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable “hours of unpaid work per week.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable X_1 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable “hours of unpaid work per week.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here the subscript on β is 1 because the “hours of unpaid work per week” is defined as X_1 in the regression model.
4. The p -value for the tests on the regression coefficients are located in the bottom part of the table under the **P-value column** heading in the corresponding independent variable row.
5. Because the alternative hypothesis is a \neq , the p -value is the sum of the area in the tails of the t -distribution. This is the value calculated out by Excel in the regression summary table.
6. The p -value of 0.0082 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the regression coefficient β_1 is not zero, and so there is a relationship between the dependent variable “job satisfaction” and the independent variable “hours of unpaid work per week.” This means that the independent variable “hours of unpaid work per week” is useful in predicting the dependent variable.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 25 employees at the company is taken and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “age”.

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \beta_2 = 0 \\ H_a: \beta_2 \neq 0 \end{array}$$

p-value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p -value for the test on the age regression coefficient is in the bottom part of the table under the **P-value column** of the **Age row**. So the p -value=0.8439.

Conclusion:

Because p -value= 0.8439 > 0.05 = α , we do not reject the null hypothesis. At the 5%

significance level there is not enough evidence to suggest that there is a relationship between the dependent variable “job satisfaction” and the independent variable “age.”

NOTES

1. The null hypothesis $\beta_2 = 0$ is the claim that the regression coefficient for the independent variable X_2 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable “age.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable X_2 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable “age.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here the subscript on β is 2 because “age” is defined as X_2 in the regression model.
4. The p -value of 0.8439 is a large probability compared to the significance level, and so is likely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely correct, and so the conclusion of the test is to not reject the null hypothesis. In other words, the regression coefficient β_2 is zero, and so there is no relationship between the dependent variable “job satisfaction” and the independent variable “age.” This means that the independent variable “age” is not particularly useful in predicting the dependent variable.

EXAMPLE

The human resources department at a large company wants to develop a model to predict an employee’s job satisfaction from the number of hours of unpaid work per week the employee does, the employee’s age, and the employee’s income. A sample of 25 employees at the company is taken

and the data is recorded in the table below. The employee's income is recorded in \$1000s and the job satisfaction score is out of 10, with higher values indicating greater job satisfaction.

Job Satisfaction	Hours of Unpaid Work per Week	Age	Income (\$1000s)
4	3	23	60
5	8	32	114
2	9	28	45
6	4	60	187
7	3	62	175
8	1	43	125
7	6	60	93
3	3	37	57
5	2	24	47
5	5	64	128
7	2	28	66
8	1	66	146
5	7	35	89
2	5	37	56
4	0	59	65
6	2	32	95
5	6	76	82
7	5	25	90
9	0	55	137
8	3	34	91
7	5	54	184
9	1	57	60
7	0	68	39
10	2	66	187
5	0	50	49

Previously, we found the multiple regression equation to predict the job satisfaction score from the other variables:

$$\hat{y} = 4.7993 - 0.3818x_1 + 0.0046x_2 + 0.0233x_3$$

\hat{y} = predicted job satisfaction score

x_1 = hours of unpaid work per week

x_2 = age

x_3 = income (\$1000s)

At the 5% significance level, test the relationship between the dependent variable “job satisfaction” and the independent variable “income”.

Solution:

Hypotheses:

$$\begin{array}{l} H_0: \beta_3 = 0 \\ H_a: \beta_3 \neq 0 \end{array}$$

p-value:

The regression summary table generated by Excel is shown below:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.711779225					
R Square	0.506629665					
Adjusted R Square	0.436148189					
Standard Error	1.585212784					
Observations	25					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	54.189109	18.06303633	7.18812504	0.001683189	
Residual	21	52.770891	2.512899571			
Total	24	106.96				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.799258185	1.197185164	4.008785216	0.00063622	2.309575344	7.288941027
Hours of Unpaid Work per Week	-0.38184722	0.130750479	-2.9204269	0.008177146	-0.65375772	-0.10993671
Age	0.004555815	0.022855709	0.199329423	0.843922453	-0.04297523	0.052086864
Income (\$1000s)	0.023250418	0.007610353	3.055103771	0.006012895	0.007423823	0.039077013

The p -value for the test on the income regression coefficient is in the bottom part of the table under the **P-value column** of the **Income row**. So the p -value=0.0060.

Conclusion:

Because p -value= 0.0060 < 0.05 = α , we reject the null hypothesis in favour of the alternative hypothesis. At the 5% significance level there is enough evidence to suggest that there is a

relationship between the dependent variable “job satisfaction” and the independent variable “income.”

NOTES

1. The null hypothesis $\beta_3 = 0$ is the claim that the regression coefficient for the independent variable X_3 is zero. That is, the null hypothesis is the claim that there is no relationship between the dependent variable and the independent variable “income.”
2. The alternative hypothesis is the claim that the regression coefficient for the independent variable X_3 is not zero. The alternative hypothesis is the claim that there is a relationship between the dependent variable and the independent variable “income.”
3. When conducting a test on a regression coefficient, make sure to use the correct subscript on β to correspond to how the independent variables were defined in the regression model and which independent variable is being tested. Here the subscript on β is 3 because “income” is defined as X_3 in the regression model.
4. The p -value of 0.0060 is a small probability compared to the significance level, and so is unlikely to happen assuming the null hypothesis is true. This suggests that the assumption that the null hypothesis is true is most likely incorrect, and so the conclusion of the test is to reject the null hypothesis in favour of the alternative hypothesis. In other words, the regression coefficient β_3 is not zero, and so there is a relationship between the dependent variable “job satisfaction” and the independent variable “income.” This means that the independent variable “income” is useful in predicting the dependent variable.

Concept Review

The test on a regression coefficient determines if there is a relationship between the dependent variable and the corresponding independent variable. The p -value for the test is the sum of the area in tails of the t -distribution. The p -value can be found on the regression summary table generated by Excel.

The hypothesis test for a regression coefficient is a well established process:

1. Write down the null and alternative hypotheses in terms of the regression coefficient being

tested. The null hypothesis is the claim that there is no relationship between the dependent variable and independent variable. The alternative hypothesis is the claim that there is a relationship between the dependent variable and independent variable.

2. Collect the sample information for the test and identify the significance level.
3. The p -value is the sum of the area in the tails of the t -distribution. Use the regression summary table generated by Excel to find the p -value.
4. Compare the p -value to the significance level and state the outcome of the test.
5. Write down a concluding sentence specific to the context of the question.

13.7 MULTICOLLINEARITY

LEARNING OBJECTIVES

- Define multicollinearity and understand its impact on multiple regression.

The term independent variable applies to any variable that is used to predict or explain the value of the dependent variable. But this does not mean that the independent variables themselves are unrelated to each other. In fact, most independent variables in multiple regression models share some degree of relatedness. For example, if “distance travelled” and “litres of gas consumed” are the independent variables in a regression model to predict the dependent variable “travel time,” the variables “distance travelled” and “litres of gas consumed” are highly correlated.

When two or more independent variables in a regression model are highly correlated to each other, **multicollinearity** exists between the independent variables. Consequently, the conclusions about the relationship between the dependent variable and the individual independent variables may be affected when the independent variables are related to each other. In addition, multicollinearity may affect the outcome of the tests on the individual regression coefficients. But multicollinearity does not affect the outcome of the overall test on the regression model.

Even though the overall model test may conclude that there is a relationship between the dependent variable and the set of independent variables, multicollinearity amongst the independent variables may cause all of the tests on the individual regression coefficients to conclude that none of the individual independent variables are related to the dependent variable. One way to address the problem of multicollinearity is to avoid including independent variables that are highly correlated or remove one of two highly correlated independent variables from the model.

Concept Review

Multicollinearity refers to the correlation that may exist between two or more independent variables in a regression model. Although multicollinearity may affect conclusions drawn about the individual regression coefficients, multicollinearity does not affect conclusions about the overall model.

13.8 EXERCISES

1. A local restaurant advocacy group wants to study the relationship between a restaurant's average weekly profit, the restaurant's seating capacity and average daily traffic that passes the restaurant's location. The group took a sample of restaurants and recording their average weekly profit (in \$1000s), the seating restaurant's seating capacity, and the average number of cars (in 1000s) that passes the restaurant's location. The data is recorded in the following table:

Seating Capacity	Traffic Count (1000s)	Weekly Net Profit (\$1000s)
120	19	23.8
180	8	29.2
150	12	22
180	15	26.2
220	16	33.5
235	10	32
115	18	22.4
110	12	20.4
165	21	23.7
220	20	34.7
140	24	27.1
145	24	23.3
140	13	20.9
200	14	29.6
210	14	31.4
175	12	23.2
175	15	31.1
190	17	28.2
100	23	25.2
145	20	20.7
135	13	37.2
25	13	26.3
140	25	20
130	14	28.2
135	10	24.6
160	23	23.7

- Find the regression model to predict the average weekly profit from the other variables.
- Interpret the coefficient for seating capacity.
- Interpret the coefficient for traffic count.

- d. Predict the average weekly profit for a restaurant with a seating capacity of 150 and a traffic count of 25,000 cars.
 - e. Find the adjusted coefficient of determination.
 - f. Interpret the adjusted coefficient of determination.
 - g. Find the standard error of the estimate.
 - h. Interpret the standard error of the estimate.
 - i. At the 5% significance level, test the validity of the model.
 - j. At the 5% significance level, test the coefficient of seating capacity.
 - k. At the 5% significance level, test the coefficient of traffic count.
2. A local university wants to study the relationship between a student's GPA, the average number of hours they spend studying each night and the average number of nights they go out each week. The university took a sample of students and recorded the following data:

GPA	Average Number of Hours Spent Studying Each Night	Average Number of Nights Go Out Each Week
3.72	5	1
3.88	3	1
3.67	2	1
3.87	3	4
2.49	1	4
1.29	1	2
1.01	2	4
2.12	1	1
1.9	1	5
3.42	3	2
1.33	1	4
1.07	0	2
2.75	3	1
3.82	4	1
3.91	5	0
2.25	2	3
2.06	1	5
2.92	3	2
3.06	3	1
3.65	2	2
3.69	4	1

- Find the regression model to predict GPA from the other variables.
- Interpret the coefficient for the average number of hours spent studying each night.
- Interpret the coefficient for the average number of nights a student goes out each week.
- Predict the GPA for a student who spends an average of 4 hours a night studying and goes out an average of 3 nights a week.
- Find the adjusted coefficient of determination.
- Interpret the adjusted coefficient of determination.
- Find the standard error of the estimate.

- h. Interpret the standard error of the estimate.
 - i. At the 1% significance level, test the validity of the model.
 - j. At the 1% significance level, test the coefficient of the average number of hours spent studying each night.
 - k. At the 1% significance level, test the coefficient of the average number of nights a student goes out each week.
3. A very large company wants to study the relationship between the salaries of employees in management positions, their age, the number of years the employee spent in college, and the number of years the employee has been with the company. A sample management employees is taken and the data recorded below:

Age	Years of College	Years with Company	Salary (\$1000s)
60	8	29	317.3
33	3	5	97.3
57	6	27	263.1
32	4	5	101.3
31	6	3	114.2
61	8	19	350.4
41	7	8	146.9
35	4	2	91.7
51	6	21	198.2
50	8	10	196.5
57	5	15	105.7
49	6	18	118.3
62	7	27	305.2
52	8	26	239.9
39	4	8	145.9
42	7	5	175.4
62	4	24	219.4
60	4	22	202.1
65	3	21	196.3
40	4	10	143.9
62	6	29	408.7
53	7	5	145.2
48	8	5	175.1
61	5	6	152.7
38	7	3	99.7
40	7	12	174.9

45	7	7	149.2
58	7	14	282.8
38	4	3	95.7
41	5	18	232.8

- a. Find the regression model to predict salary from the other variables.
- b. Interpret the coefficient for age.
- c. Interpret the coefficient for years of college
- d. Interpret the coefficient for years with the company.
- e. Predict the salary for a 47 year old management employee who spent 5 years in college and has been with the company for 15 years.
- f. Find the adjusted coefficient of determination.
- g. Interpret the adjusted coefficient of determination.
- h. Find the standard error of the estimate.
- i. Interpret the standard error of the estimate.
- j. At the 1% significance level, test the validity of the model.
- k. At the 1% significance level, test the coefficient of age.
- l. At the 1% significance level, test the coefficient of the years of college.
- m. At the 1% significance level, test the coefficient for the years with the company.

13.9 ANSWERS TO SELECT EXERCISES

1.

$$\hat{y} = 21.989 + 0.046x_1 - 0.196x_2$$

a. x_1 = seating capacity

x_2 = traffic count (1000s)

\hat{y} = average weekly profit (\$1000s)

b. For each additional seat in the restaurant, the average weekly profit increases by \$46.

c. For each additional 1000 cars that pass the restaurant, the average weekly profit decreases by \$196.

d. \$24,519.20

e. 0.2250

f. 22.50% of the variation in the average weekly profit can be explained by the regression model based on seating capacity and traffic count.

g. 4.1675.

h. On average, the average weekly profit differs by \$4167.50 from the regression model based on seating capacity and traffic count.

i. p -value=0.0205; reject the null hypothesis.

j. p -value=0.0144; reject the null hypothesis.

k. p -value=0.2645; do not reject the null hypothesis.

2.

$$\hat{y} = 1.692 + 0.524x_1 - 0.082x_2$$

a. x_1 = average number of hours spent studying a night

x_2 = average number of nights go out each week

\hat{y} = GPA

b. For each additional hour spent studying each night, the student's GPA increases by 0.524.

c. For each additional hour a student goes out each week, the student's GPA decreases by 0.082.

d. 3.54

e. 0.5833

f. 58.33% of the variation in GPA can be explained by the regression model based on the average number of hours spent studying a night and the average number of nights a student goes out

each week.

- g. 0.6613.
- h. On average, GPA differs by 0.6613 from the regression model based on the average number of hours spent studying a night and the average number of nights a student goes out each week.
- i. p -value=0.0002; reject the null hypothesis.
- j. p -value=0.0009; reject the null hypothesis.
- k. p -value=0.5083; do not reject the null hypothesis.

3.

$$\hat{y} = -42.359 + 1.436x_1 + 14.758x_2 + 5.486x_3$$

$x_1 =$ age

- a. $x_2 =$ years of college
 $x_3 =$ years with the company
 $\hat{y} =$ salary (\$1000s)
- b. For each additional year of age, the salary increases by \$1436.14.
- c. For each additional year of college, the salary increases by \$14,758.04.
- d. For each additional year with the company, the salary increases by \$5486.07.
- e. \$181,221.15
- f. 0.6959
- g. 69.59% of the variation in salary can be explained by the regression model based on age, years of college, and years with the company.
- h. 45.24522.
- i. On average, salary differs by \$45,255.22 from the regression model based on age, years of college, and years with the company.
- j. p -value=0.0000002; reject the null hypothesis.
- k. p -value=0.2373; do not reject the null hypothesis.
- l. p -value=0.0097; reject the null hypothesis.
- m. p -value=0.0005; reject the null hypothesis.

REFERENCES

1.2 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library. (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html>

1.3 Sampling and Data

Book of Odds. (n.d.). *How George Gallup Picked the President*. <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

Gallup. (n.d.). *Gallup Presidential Election Trial-Heat Trends, 1936–2008*. Retrieved May 1, 2013, from <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4>

Gallup-Healthways Well-Being Index. (n.d.). Retrieved May 1, 2013, from <http://www.well-beingindex.com/default.asp>

Gallup-Healthways Well-Being Index. (n.d.). Retrieved May 1, 2013, from <http://www.well-beingindex.com/methodology.asp>

Gallup-Healthways Well-Being Index. (n.d.) Retrieved May 1, 2013, from <http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx>

LBCC Distance Learning (DL) program data in 2010-2011. (n.d.). Retrieved May 1, 2013, from <http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus>

Lusinci, D. (2012) “President’ Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?” *Social Science History* ,36(1)1, 23-54. <http://ssh.dukejournals.org/content/36/1/23.abstract>

San Jose Mercury News. (n.d.).

The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics. (n.d.). Retrieved May 1, 2013, from <http://www.math.uah.edu/stat/data/LiteraryDigest.html>

The Data and Story Library. (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html>

1.4 Frequency, Frequency Tables, and Levels of Measurement

Lane, D. 2003, June 20. *Levels of Measurement*. OpenStax CNX. Retrieved May 1, 2013, from <http://cnx.org/content/m10809/latest/>

Levels of Measurement. (n.d.). Retrieved May 1, 2013, from http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm

State & County QuickFacts. (n.d.). Retrieved May 1, 2013, from http://quickfacts.census.gov/qfd/download_data.html

State & County QuickFacts: Quick, easy access to facts about people, business, and geography. (n.d.). U.S. Census Bureau. Retrieved from May 1, 2013, <http://quickfacts.census.gov/qfd/index.html>

Table 5: Direct hits by mainland United States Hurricanes (1851-2004). (n.d.). National Hurricane Center. Retrieved May 1, 2013, from <http://www.nhc.noaa.gov/gifs/table5.gif>

Taylor, C. 2018, February 2. *The Levels of Measurement in Statistics*. Thoughtco. <http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm>

1.5 Experimental Design and Ethics

Alden, L. (2013, May 1). *Statistics can be Misleading*. econoclass.com. Retrieved May 1, 2013, from <http://www.econoclass.com/misleadingstats.html>

America's Best Small Companies. (n.d.). Forbes. Retrieved May 1, 2013, from, <http://www.forbes.com/best-small-companies/list/>

April 2013 Air Travel Consumer Report. (2013, April 11). U.S. Department of Transportation. Retrieved, May 1, 2013, from, <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Bhattacharjee, Y. (2013, April 26). The Mind of a Con Man. *The New York Times Magazine*. http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2&

Data. (n.d.). BusinessWeek. Retrieved May 1, 2013, from <https://www.bloomberg.com/businessweek>

Data. (n.d.). Forbes. Retrieved May 1, 2013, from, <https://www.forbes.com/>

Earthquake Information by Year. (n.d.). U.S. Geological Survey. Retrieved, May 1, 2013, from <http://earthquake.usgs.gov/earthquakes/eqarchives/year/>

Jacson, M. L., Croft, R. J., Kennedy, G. A., Owens, K., & Howard, M. E. (2013). *Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors*. Accident Analysis and Prevention, *Jan(50)*, 438-44. <http://www.ncbi.nlm.nih.gov/pubmed/22721550>

Levelt, W. J. M., Drenth, P., & Noort, E. (Eds.). (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Tilburg: Commissioned by the Tilburg University,

University of Amsterdam and the University of Groningen. <http://hdl.handle.net/11858/00-001M-0000-0010-258A-9>

McClung, M., & Collins, D. (2007). "Because I know it will!": Placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*, 29(3), 382-94. <https://doi.org/10.1123/jsep.29.3.382>

Medina, de los A. (2007, November 19). *Ethics in Statistics*. OpenStax CNX. Retrieved May 1, 2013, from <http://cnx.org/contents/12a5d87b-3fe3-4606-b368-cef865e40cde@1>

Mehta, A. (2011, July 21). *Daily Dose of Aspiring Helps Reduce Heart Attacks: Study*. International Business Times. Retrieved May 1, 2013, from, <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443>

National Highway Traffic Safety Administration. (n.d.). *Fatality Analysis Report Systems (FARS) Encyclopedia*. Retrieved May 1, 2013, from <http://www-fars.nhtsa.dot.gov/Main/index.aspx>

Nutrition Source: Vitamin E. (n.d.). Harvard T.H. Chan School of Public Health. Retrieved May 1, 2013, from <http://www.hsph.harvard.edu/nutritionsource/vitamin-e/>

Reents, S. (2008, February 4). *Don't Underestimate the Power of Suggestion*. AthleteInMe.com. Retrieved May 1, 2013, from <http://www.athleteinme.com/ArticleView.aspx?id=1053>

The Data and Story Library (n.d.). Retrieved May 1, 2013, from <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html>

U.S. Department of Health and Human Services. (2019). Code of Federal Regulations: Title 45 Public Welfare Department of Health and Human Services, Part 46 Protection of Human Subjects, Section 46.111: Criteria for IRB Approval of Research.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

Births Time Series Data. (2013). General Register Office For Scotland. Retrieved April 3, 2013, from, <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html>

CO2 emissions (kt). (2013). The World Bank. Retrieved April 3, 2013, from, <http://databank.worldbank.org/data/home.aspx>

Consumer Price Index. (n.d.). United States Department of Labor: Bureau of Labor Statistics. Retrieved April 3, 2013, from, <http://data.bls.gov/pdq/SurveyOutputServlet>

Demographics: Children under the age of 5 years underweight. (n.d.). Indexmundi. Retrieved April 3, 2013, from, <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en>

Food Security Statistics. (n.d.). Food and Agriculture Organization of the United Nations. Retrieved April 3, 2013, from, <http://www.fao.org/economic/ess/ess-fs/en/>

Gunst, R., & Mason, R. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press.

Overweight and Obesity: Adult Obesity Facts. (n.d.). Centers for Disease Control and Prevention. Retrieved April 3, 2013, from, <http://www.cdc.gov/obesity/data/adult.html>

Presidents. (2007). Fact Monster. Retrieved April 3, 2013, from, <http://www.factmonster.com/ipka/A0194030.html>

Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more. (2013). Scholastic. Retrieved April 3, 2013, from, <http://www.scholastic.com/teachers/article/timeline-guide-us-presidents>

2.3 Measures of Central Tendency

Data. (n.d.). The World Bank. Retrieved April 3, 2013, from, <http://www.worldbank.org>

Demographics: Obesity – adult prevalence rate. (n.d.). Indexmundi. Retrieved April 3, 2013, from, <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en>

2.5 Measures of Location

1990 Census. (n.d.). United States Department of Commerce: United States Census Bureau. Retrieved April 3, 2013, from, <http://www.census.gov/main/www/cen1990.html>

Cauchon, D., & Overberg, P. (2012). Census data shows minorities now a majority of U.S. births. *USA Today*. Retrieved April 3, 2013, from, <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1>

Data. (n.d.). *San Jose Mercury News*.

Data. (n.d.). The United States Department of Commerce: United States Census Bureau. Retrieved April 3, 2013, from, <http://www.census.gov/>

Yankelovich Partners. (n.d.). Survey. *Time Magazine*.

2.6 Measures of Dispersion

Data. (n.d.). In *Microsoft Bookshelf*.

King, B. (n.d.). *Graphically Speaking*. Institutional Research, Lake Tahoe Community College. Retrieved April 3, 2013, from, <http://www.ltcc.edu/web/about/institutional-research>

3.2 The Terminology of Probability

Worldatlas. (2013). Countries List by Continent. In *Worldatlas.com*. Retrieved May 2, 2013, from, <http://www.worldatlas.com/cntycont.htm>

3.3 Contingency Tables

Blood Types. (n.d.). American Red Cross. Retrieved May 3, 2013, from, <http://www.redcrossblood.org/learn-about-blood/blood-types>

Data. (n.d.). National Center for Health Statistics, The United States Department of Health and Human Services.

Data. (n.d.). United States Senate. Retrieved May 2, 2013, from, <https://www.senate.gov/>

Haiman, C. A., Stram, D. O., Wilkens, L. R., Pike, M. C., Kolonel, L. N., Henderson, B. E., & le Marchand, L. (2006, January 26). Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer. *The New England Journal of Medicine*. <http://www.nejm.org/doi/full/10.1056/NEJMoa033250>

Human Blood Types. (2011). Unite Blood Services. Retrieved May 2, 2013, from, <http://www.unitedbloodservices.org/learnMore.aspx>

Samuel, T. M. (2013). *Strange Facts about RH Negative Blood*. eHow Health. Retrieved May 2, 2013, from, http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html

United States: Uniform Crime Report – State Statistics from 1960–2011. (n.d.). The Disaster Center. Retrieved May 2, 2013, from, <http://www.disastercenter.com/crime/>

3.7 Joint Probabilities

Data. (n.d.). *Baseball-Almanac*. Retrieved May 2, 2013, from, <https://www.baseball-almanac.com/>

Data. (n.d.). Field Research Corporation.

Data. (n.d.). The Roper Center: Public Opinion Archives at the University of Connecticut. Retrieved May 2, 2013, from, <http://www.ropercenter.uconn.edu/>

Data. (n.d.). The Wall Street Journal. <https://www.wsj.com/>

Data. (n.d.). U.S. Census Bureau. <https://www.census.gov/>

DiCamillo, M., & Field, M. (n.d.). *The File Poll*. Field Research Corporation. Retrieved May 2, 2013, from, <http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf>

Mayor's Approval Down. (n.d.). Forum Research. Retrieved May 2, 2013, from, http://www.forumresearch.com/forms/News_Archives/News_Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf

Rider, D. (2011, September 14). Ford support plummeting, poll suggests. *The Star*. Retrieved May 2, 2013, from, http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html

Roulette. (n.d.). In *Wikipedia*. <http://en.wikipedia.org/wiki/Roulette>

Shin, H. B., & Kominski, R. A. (2010, April 1). *Language Use in the United States: 2007*. United States Census Bureau. <https://www.census.gov/library/publications/2010/acs/acs-12.html>

4.3 Expected Value and Standard Deviation of a Discrete Random Variable

Course Catalog. (n.d.). Florida State University. Retrieved May 15, 2013, from, https://m.fsu.edu/default/course_catalog/index

World Earthquakes: Live Earthquake News and Highlights. (2012). World Earthquakes. Retrieved May 15, 2013, from, http://www.world-earthquakes.com/index.php?option=ethq_prediction

4.4 The Binomial Distribution

Access to electricity (% of population). (2013). The World Bank. Retrieved May 15, 2015, from, http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc

Distance Education. (n.d.). In *Wikipedia*. Retrieved May 15, 2013, from, http://en.wikipedia.org/wiki/Distance_education

NBA Statistics – 2013. ESPN. Retrieved May 15, 2013, from, http://espn.go.com/nba/statistics/_/seasontype/2

Newport, F. (2013, May 9). *Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income*. Gallup. <http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx>

Pryor, J. H., DeAngelo, L., Palucki Blake, L., Hurtado, S., & Tran, S. (2011). *The American Freshman: National Norms Fall 2011*. Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA. <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf>

The World FactBook. (n.d.). Central Intelligence Agency. Retrieved May 15, 2013, from, <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html>

What are the key statistics about pancreatic cancer? (2013). American Cancer Society. Retrieved

May 15, 2013, from, <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics>

4.5 The Poisson Distribution

ATL Fact Sheet. (2013). Department of Aviation at the Hartsfield-Jackson Atlanta International Airport. Retrieved February 18, 2019, from, <http://www.atl.com/about-atl/atl-factsheet/>

Children and Childrearing. (n.d.). Ministry of Health, Labour, and Welfare. Retrieved May 15, 2013, from, <http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html>

Daily Mail Reporter. (2011, June 9). One born every minute: The maternity unit where mothers are THREE to a bed. *Daily Mail*. Retrieved May 15, 2013, from, <http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers-bed.html>

Eating Disorder Statistics. (2006). South Carolina Department of Mental Health. Retrieved May 15, 2013, from, <http://www.state.sc.us/dmh/anorexia/statistics.htm>

Giving Birth in Manila. (2011, June 8). *The Guardian*. Retrieved May 15, 2013, from, <http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2>

Lenhart, A. (2012, March 19). *Teens, Smartphones & Texting*. Pew Research Center. Retrieved May 15, 2013, from, http://www.pewinternet.org/~media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf

Smith, A. (2011, September 19). *How Americans Use Text Messaging*. Pew Research Center. Retrieved May 15, 2013, from, <http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx>

Teen Drivers: Fact Sheet, Injury Prevention & Control: Motor Vehicle Safety. (2012, October 2). Center for Disease Control and Prevention. Retrieved May 15, 2013, from, http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html

Vanderkam, L. (2012, October 8). *Stop Checking Your Email, Now*. Fortune. Retrieved May 15, 2013, from, <http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/>

World Earthquakes: Live Earthquake News and Highlights. (n.d.). World Earthquakes Live. Retrieved May 15, 2013, from, http://www.world-earthquakes.com/index.php?option=ethq_prediction

5.4 The Standard Normal Distribution

2012 College-Bound Seniors Total Group Profile Report. (2012). CollegeBoard. Retrieved May 14, 2013, from, <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf>

Blood Pressure of Males and Females. (n.d.). StatCrunch. Retrieved May 14, 2013, from, <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960>

Data. (n.d.). National Basketball Association. Retrieved May 14, 2013, from, www.nba.com

Data. (n.d.). *San Jose Mercury News.*

Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009. (2009). National Center for Education Statistics. Retrieved May 14, 2013, from, http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp

Janssen, S. (Ed.). (n.d.). *The World Almanac and Book of Facts.* World Almanac Books.

List of stadiums by capacity. (n.d.). In *Wikipedia.* Retrieved May 14, 2013, from, https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity

The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores. (2009). London School of Hygiene and Tropical Medicine. Retrieved May 14, 2013, from, http://conflict.lshtm.ac.uk/page_125.htm

5.5 Calculating Probabilities for a Normal Distribution

Facebook Statistics. (n.d.). Statistics Brain. Retrieved May 14, 2013, from, <http://www.statisticbrain.com/facebook-statistics/>

Naegele's rule. (n.d.). In *Wikipedia.* Retrieved May 14, 2013, from, http://en.wikipedia.org/wiki/Naegele's_rule

NUMMI. (2010, March 26). This American Life. Retrieved May 14, 2013, from, <http://www.thisamericanlife.org/radio-archives/episode/403/nummi>

Scratch-Off Lottery Ticket Playing Tips. (n.d.). WinAtTheLottery.com. Retrieved May 14, 2013, from, <http://www.winatthelottery.com/public/department40.cfm>

Smart Phone Users, By The Numbers. (n.d.). Visual.ly. Retrieved May 14, 2013, from, <http://visual.ly/smart-phone-users-numbers>

6.2 Sampling Distribution of the Sample Mean

Baran, D. (n.d.). *20 Percent of Americans Have Never Used Email.* WebGuild. Retrieved May 14, 2013, from, <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email>

Data. (n.d.). The Flurry Blog. Retrieved May 17, 2013, from, <http://blog.flurry.com>

Data. (n.d.). The United States Department of Agriculture.

7.2 Confidence Intervals for a Single Population Mean with

Known Population Standard Deviation

American Fact Finder. (n.d.). U.S. Census Bureau. Retrieved July 2, 2013, from, <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

Disclosure Data Catalog: Candidate Summary Report 2012. (n.d.). U.S. Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/data/index.jsp>

Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall. (n.d.). Foothill De Anza Community College District. Retrieved September 30, 2013, from, http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm

Kuczumski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z. , Wei, R., Curtin, L. R., Roche, A. F., & Johnson, C. L. (2002, May). Vital Health Statistics: 2000 CDC Growth Charts for the United States: Methods and Development. *Centers for Disease Control and Prevention*, 11(246). Retrieved July 2, 2013, from, <http://www.cdc.gov/growthcharts/2000growthchart-us.pdf>

Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates. (n.d.). American Fact Finder, U.S. Census Bureau. Retrieved July 2, 2013, from, http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table

Metadata Description of Candidate Summary File. (n.d.). U.S. Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml>

National Health and Nutrition Examination Survey. (n.d.). Centers for Disease Control and Prevention. Retrieved July 2, 2013, from, <http://www.cdc.gov/nchs/nhanes.htm>

Ralph, N., & German, K. (2011, June 1). *Cell phones with the highest radiation levels (pictures)*. CNET. Retrieved July 2, 2013, from, <http://reviews.cnet.com/cell-phone-radiation-levels/>

7.3 Confidence Intervals for a Single Population Mean with Unknown Population Standard Deviation

America's Best Small Companies. (2013). Forbes. Retrieved July 2, 2013, from, <http://www.forbes.com/best-small-companies/list/>

Data. (n.d.). Businessweek. <http://www.businessweek.com/>.

Data. (n.d.). Forbes. <http://www.forbes.com/>.

Data. (n.d.). In *Microsoft Bookshelf*.

Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012. (n.d.). Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/data/index.jsp>

Human Toxome Project: Mapping the Pollution in People. (n.d.). Environmental Working Group. Retrieved July 2, 2013, from, <http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn>

Metadata Description of Leadership PAC List. (n.d.). Federal Election Commission. Retrieved July 2, 2013, from, <http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml>

7.4 Confidence Intervals for Population Proportions

2013 Teen and Privacy Management Survey. (n.d.). Pew Research Center: Internet and American Life Project. Retrieved July 2, 2013, from, http://www.pewinternet.org/~-/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf

52% Say Big-Time College Athletics Corrupt Education Process. (2013, May 16). Rasmussen Reports. Retrieved July 2, 2013, from, http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process

Jensen, T. (2013, May 10). *Democrats, Republicans Divided on Opinion of Music Icons.* Public Policy Polling. Retrieved July 2, 2013, from, <https://www.publicpolicypolling.com/polls/democrats-republicans-divided-on-opinion-of-music-icons/>

Madden, M., Lenhart, A., Coresi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013, May 21). *Teens, Social Media, and Privacy.* Pew Research Center. Retrieved July 2, 2013, from, <https://www.pewresearch.org/internet/2013/05/21/teens-social-media-and-privacy/>

Saad, L. (2013, May 23). *Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity.* Gallup. Retrieved July 2, 2013, from, <http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx>

The Field Poll. (n.d.). Field. Retrieved July 2, 2013, from, <http://field.com/fieldpollonline/subscribers/>

Zogby. (2013, May 16). *New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security.* Zogby Analytics. Retrieved July 2, 2013, from, <http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll>

8.2 Null and Alternative Hypotheses

Data. (n.d.). The National Institute of Mental Health. <http://www.nimh.nih.gov/publicat/depression.cfm>

8.6 Hypothesis Tests for a Population Mean with Known Population Standard Deviation, 8.7 Hypothesis Tests for a Population Mean with Unknown Population Standard Deviation, 8.8 Hypothesis Tests for a Population Proportion

Allen, E. I., & Seaman, J. (2005). *Growing by Degrees: Online Education in the United States, 2005*. The Sloan Consortium.

Amit Schitai, A. (n.d.). Data.

Data. (n.d.). American Automobile Association. Retrieved June 27, 2013, from, www.aaa.com

Data. (n.d.). American Library Association. Retrieved June 27, 2013, from, <https://www.ala.org/>

Data. (n.d.). Bureau of Labor Statistics. <http://www.bls.gov/oes/current/oes291111.htm>.

Data. (n.d.). Centers for Disease Control and Prevention. Retrieved June 27, 2013, from, www.cdc.gov

Data. (n.d.). Energy.Gov. Retrieved June 27, 2013, from, <http://energy.gov>

Data. (n.d.). Gallup. Retrieved June 27, 2013, from <https://www.gallup.com/home.aspx>

Data. (n.d.). La Leche League International. <http://www.lalecheleague.org/Law/BAFeb01.html>

Data. (n.d.). Toastmasters International. <http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1>.

Data. (n.d.). United States Census Bureau. Retrieved June 27, 2013, from, <https://www.census.gov/programs-surveys/sis/resources/data-tools/quickfacts.html>

Data. (n.d.). United States Census Bureau. <http://www.census.gov/hhes/socdemo/language/>.

Data, (n.d.). Weather Underground. Retrieved June 27, 2013, from, <https://www.wunderground.com/>

Deprez, E. E. *NYC Smoking Rate Falls to Record Low of 14%, Bloomberg Says*. Businessweek. Retrieved June 27, 2013, from <https://www.bloomberg.com/news/articles/2011-09-15/new-york-city-adult-smoking-rate-falls-to-all-time-low-of-14-mayor-says#:~:text=New%20York's%20adult%20smoking%20rate,are%20smoking%2C%20the%20mayor%20said>

FBI. (n.d.). *Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005*. The Disaster Center. Retrieved June 27, 2013, from, <http://www.disastercenter.com/kentucky/crime/3868.htm>

Foothill-De Anza Community College District. (2006, Winter). De Anza College. http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf

Johansen, C., Boice, Jr., J., McLaughlin, J., Olsen, J. (2001). Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark. *Journal of National Cancer Institute*, 93(3), 203-207. <https://doi.org/10.1093/jnci/93.3.203>

How often does sexual assault occur? (n.d.). RAINN. Retrieved June 27, 2013, from, <http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault>

9.2 Statistical Inference for Two Population Means with Known Population Standard Deviations

Data. (n.d.). United States Census Bureau. <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>

FBI. (n.d.). *Texas Crime Rates 1960–1012*. Uniform Crime Reports, The Disaster Center. Retrieved June 17, 2013, from, <http://www.disastercenter.com/crime/txcrime.htm>

Hinduja, S. (2013). *Sexting Research and Gender Differences*. Cyberbullying Research Center. Retrieved June 17, 2013, from, <http://cyberbullying.us/blog/sexting-research-and-gender-differences/>

Smart Phone Users, By the Numbers. (2013). Visually. Retrieved June 17, 2013, from, <http://visual.ly/smart-phone-users-numbers>

Smith, A. (2013). *35% of American adults own a Smartphone*. Pew Research Center. Retrieved June 17, 2013, from, http://www.pewinternet.org/~media/Files/Reports/2011/PIP_Smartphones.pdf

State-Specific Prevalence of Obesity Among Adults—United States, 2007. Morbidity and Mortality Weekly Report, Centers for Disease Control and Prevention. Retrieved June 17, 2013, from, <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm>

9.3 Statistical Inference for Two Population Means with Unknown Population Standard Deviations

Baseball-Almanac. (2013). World Series History. In *Baseball-Almanac, 2013*. Retrieved June 17, 2013, from, <http://www.baseball-almanac.com/ws/wsmenu.shtml>

Data. (n.d.). Graduating Engineer + Computer Careers. <http://www.graduatingengineer.com>

Data. (n.d.). In *Microsoft Bookshelf*.

Data. (n.d.). United States Senate. Retrieved June 17, 2013, from <https://www.senate.gov/>

List of current United States Senators by Age. (n.d.). In *Wikipedia*. http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age

Sectoring by Industry Groups. (n.d.). Nasdaq. Retrieved June 17, 2013, from, <http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry>

Strip Clubs: Where Prostitution and Trafficking Happen. (2013). Prostitution Research & Education. Retrieved June 17, 2013, from <https://prostitutionresearch.com/strip-clubs-where-prostitution-and-trafficking-happen/>

9.5 Statistical Inference for Two Population Proportions

Data. (n.d.). American Cancer Society. Retrieved June 17, 2013, from, <http://www.cancer.org/index>

Data. (1994, November). Chancellor's Office, California Community Colleges.

Data. (December). *Educational Resources*.

Data. (n.d.). Hilton Hotels. Retrieved June 17, 2013, from, <http://www.hilton.com>

Data. (n.d.). Hyatt Hotels. Retrieved June 17, 2013, from, <http://hyatt.com>

Data. (n.d.). Statistics. United States Department of Health and Human Services.

Data. (n.d.). Whitney Exhibit on loan to San Jose Museum of Art.

State of the States. (2013). Gallup. Retrieved June 17, 2013, from, <http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive>

West Nile Virus. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Vector-Borne Diseases (DVBD). Retrieved June 17, 2013, from, <http://www.cdc.gov/ncidod/dvbid/westnile/index.htm>

10.3 Statistical Inference for a Single Population Variance

AppleInsider Price Guides. (n.d.). Apple Insider. Retrieved June 17, 2013, from, http://appleinsider.com/mac_price_guide

Data. (n.d.). World Bank.

10.4 The Goodness-of-Fit Test

Current Population Reports. (n.d.). U.S. Census Bureau.

Data. (n.d.). The College Board. <http://www.collegeboard.com>.

Data. (n.d.). U.S. Census Bureau.

Ma, Y., Bertone, E. R., Stanek III, E. J., Reed, G. W., Hebert, J. R., Cohen, N. L., Merriam, P. A., & Ockene, I. S. (2003, July 1). Association between Eating Patterns and Obesity in a Free-living

US Adult Population. *American Journal of Epidemiology*, 158(1), 85-92. <https://doi.org/10.1093/aje/kwg117>

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2012, January). *Prevalence of Obesity in the United States, 2009–2010. NCHS Data Brief no. 82*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Retrieved May 24, 2013 from <http://www.cdc.gov/nchs/data/databriefs/db82.pdf>

Stevens, B. J. (n.d.). *Multi-family and Commercial Solid Waste and Recycling Survey*. Arlington County. Retrieved May 24, 2013, from, <http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf>

10.5 The Test of Independence

DiCamilo, M., & Field, M. (2013, February 14). *Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs*. The Field Poll. Retrieved May 24, 2013, from, <http://field.com/fieldpollonline/subscribers/Rls2436.pdf>

Favorite Flavor of Ice Cream. (2016, October 22). Statistic Brain Research Institute. <http://www.statisticbrain.com/favorite-flavor-of-ice-cream>

Youngest Online Entrepreneurs List. (2016, June 29). Statistic Brain Research Institute. <http://www.statisticbrain.com/youngest-online-entrepreneur-list>

11.3 Statistical Inference for Two Population Variances

MLB Vs. Division Standings – 2012. (n.d.). ESPN. http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true

11.4 One-Way ANOVA and Hypothesis Tests for Three or More Population Means

Data. (n.d.). Fourth-grade classroom in 1994 in a private K – 12 school, San Jose, CA.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A Handbook of Small Datasets: Data for Fruitfly Fecundity*. Chapman & Hall.

MLB Standings – 2012. ESPN. http://espn.go.com/mlb/standings/_/year/2012

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992). A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association*, 268, 1578-1580.

12.2 Linear Equations

Data. (n.d.). Centers for Disease Control and Prevention.

Data. (n.d.). National Center for Agency Reporting Flu Cases and TB Prevention.

12.5 The Regression Equation

Data. (n.d.). Centers for Disease Control and Prevention.

Data. (n.d.). National Center for Agency Reporting Flu Cases and TB Prevention.

Data. (n.d.). National Center for Health Statistics.

Data. (n.d.). United States Census Bureau. http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

VERSIONING HISTORY

This page provides a record of edits and changes made to this book since its initial publication. Whenever edits or updates are made in the text, we provide a record and description of those changes here. If the change is minor, the version number increases by 0.1. If the edits involve a number of changes, the version number increases to the next full number.

The files posted alongside this book always reflect the most recent version.

Version	Date	Change	Affected Web Page
1.0	August 2022	First Publication	N/A