

Introduction to Applied Statistics for Psychology
Students

Introduction to Applied
Statistics for Psychology
Students

GORDON E. SARTY

OSAMA BATAINEH



Introduction to Applied Statistics for Psychology Students by Gordon E. Sarty is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.

See notes on specific copyright for screenshots from IBM® SPSS® Statistics software (“SPSS”) and R Project for Statistical Computing (“RStudio”).

Contents

About This Book	1
Licensing and Copyright	2
Acknowledgements	5
Statistical Software Used in this Book	6
Accessing SPSS and RStudio Through Your School	6
Downloading SPSS and RStudio	6
Why does this book cover both SPSS and RStudio?	7
University of Saskatchewan: Software Access	9
On-Campus Lab Access	9
Remote / Off-Campus Access	9
USask ICT Help	11
Data Sets	12
1. Background and Motivation	
1.1 Overview	15
1.1.1 Textbook Layout, * and ** Symbols Explained	15
1.1.2 Intro to Univariate Statistics	16
1.2 Basic Definitions	21
1.2.1 Types of Data (important!)	22
1.2.2 Measurement Scales (avoid this!)	23
1.2.3 Kinds of Sampling and Studies	24
1.3 Summation Convention	26

2. Descriptive Statistics: Frequency Data (Counting)

<u>2.1 Frequency Tables</u>	29
<u>2.2 Plotting Frequency Data</u>	36
<u>2.2.1 Stem and Leaf Plots</u>	43
<u>2.3 SPSS Lesson 1: Getting Started with SPSS</u>	45
<u>2.4 RStudio Lesson 1: Getting Started with RStudio</u>	61
<u>Osama Bataineh</u>	

3. Descriptive Statistics: Central Tendency and Dispersion

<u>3.1 Central Tendency: Mean, Median, Mode</u>	73
<u>3.1.1 Mean</u>	73
<u>3.1.2 Median</u>	77
<u>3.1.3 Mode</u>	78
<u>3.1.4 Midrange</u>	79
<u>3.1.5 Mean, Median and Mode in Histograms: Skewness</u>	80
<u>3.1.6 Mean, Median and Mode in Distributions: Geometric Aspects</u>	82
<u>3.2 Dispersion: Variance and Standard Deviation</u>	88
<u>3.3 z-score / z-transformation</u>	96
<u>3.4 SPSS Lesson 2: Combining variables and recoding</u>	98
<u>3.5 RStudio Lesson 2: Combining variables and recoding</u>	110

4. Probability and the Binomial Distributions

<u>4.1 Probability</u>	113
<u>4.2 Binomial Distribution</u>	118
<u>4.2.1 Practical Binomial Distribution Examples</u>	123

4.3 SPSS Lesson 3: Combining variables - advanced	126
4.4 RStudio Lesson 3: Combining variables - advanced	133
5. <u>The Normal Distributions</u>	
5.1 Discrete versus Continuous Distributions	137
5.2 **The Normal Distribution as a Limit of Binomial Distributions	141
5.3 Normal Distribution	151
5.3.1 <i>Computing Areas (Probabilities) under the standard normal curve</i>	153
6. <u>Percentiles and Quartiles</u>	
6.1 Discrete Data Percentiles and Quartiles	171
6.2 Finding Outliers Using Quartiles	175
6.3 Box Plots	176
6.4 Robust Statistics	178
6.5 SPSS Lesson 4: Percentiles	180
6.6 RStudio Lesson 4: Percentiles	185
7. <u>The Central Limit Theorem</u>	
7.1 <u>Using the Normal Distribution to Approximate the Binomial Distribution</u>	189
7.2 <u>The Central Limit Theorem</u>	191
8. <u>Confidence Intervals</u>	
8.1 <u>Confidence Intervals Using the z-Distribution</u>	199
8.2 **Bayesian Statistics	204

8.3 The t-Distributions	206
8.4 Proportions and Confidence Intervals for Proportions	209
8.5 Chi Squared Distribution	217
9. Hypothesis Testing	
9.1 Hypothesis Testing Problem Solving Steps	233
9.2 z-Test for a Mean	235
9.2.1 What p-value is significant?	242
9.3 t-Test for Means	244
9.4 z-Test for Proportions	248
9.5 Chi Squared Test for Variance or Standard Deviation	251
9.6 SPSS Lesson 5: Single Sample t-Test	260
9.7 RStudio Lesson 5: Single Sample t-Test	267
Osama Bataineh	
10. Comparing Two Population Means	
10.1 Unpaired z-Test	271
10.2 Confidence Interval for Difference of Means (Large Samples)	275
10.3 Difference between Two Variances - the F Distributions	279
10.4 Unpaired or Independent Sample t-Test	286
10.4.1 General form of the t test statistic	288
10.4.2 Two step procedure for the independent samples t test	288
10.5 Confidence Intervals for the Difference of Two Means	295
10.6 SPSS Lesson 6: Independent Sample t-Test	297
10.7 RStudio Lesson 6: Independent Sample t-Test	303

<u>10.8 Paired t-Test</u>	304
<u>10.9 Confidence Intervals for Paired t-Tests</u>	308
<u>10.10 SPSS Lesson 7: Paired Sample t-Test</u>	309
<u>10.11 RStudio Lesson 7: Paired Sample t-Test</u>	312
<u>11. Comparing Proportions</u>	
<u>11.1 z-Test for Comparing Proportions</u>	315
<u>11.2 Confidence Interval for the Difference between Two Proportions</u>	316
<u>12. ANOVA</u>	
<u>12.1 One-way ANOVA</u>	319
<u>12.2 Post hoc Comparisons</u>	320
<u>12.3 SPSS Lesson 9: One-way ANOVA</u>	321
<u>12.4 R Lesson 9: One-way ANOVA</u>	322
<u>12.5 Two-way ANOVA</u>	323
<u>12.6 SPSS Lesson 9: Two-way ANOVA</u>	324
<u>12.7 R Lesson 9: Two-way ANOVA</u>	325
<u>12.8 Higher Factorial ANOVA</u>	326
<u>12.9 Between and Within Factors</u>	327
<u>12.10 *Contrasts</u>	328
<u>13. Power</u>	
<u>13.1 Power</u>	331
<u>14. Correlation and Regression</u>	
<u>14.1 Scatter Plots</u>	335

<u>14.2 Correlation</u>	336
<u>14.3 SPSS Lesson 10: Scatterplots and Correlation</u>	337
<u>14.4 R Lesson 10: Scatterplots and Correlation</u>	338
<u>14.5 Linear Regression</u>	339
<u>14.6 r-squared and the Standard Error of the Estimate of y-prime</u>	340
<u>14.7 Confidence Interval for y-prime at a Given x</u>	341
<u>14.8 SPSS Lesson 11: Linear Regression</u>	342
<u>14.9 R Lesson 11: Linear Regression</u>	343
<u>14.10 Multiple Regression</u>	344
<u>14.11 SPSS Lesson 12: Multiple Regression</u>	345
<u>14.12 R Lesson 12: Multiple Regression</u>	346
<u>15. Chi Squared: Goodness of Fit and Contingency Tables</u>	
<u>15.1 Goodness of Fit</u>	349
<u>15.2 Contingency Tables</u>	350
<u>15.3 SPSS Lesson 13: Proportions, Goodness of Fit, and Contingency Tables</u>	351
<u>15.4 R Lesson 13: Proportions, Goodness of Fit, and Contingency Tables</u>	352
<u>16. Non-parametric Tests</u>	
<u>16.1 How to Rank Data</u>	355
<u>16.2 Median Sign Test</u>	356
<u>16.3 Paired Sample Sign Test</u>	357
<u>16.4 Two Sample Wilcoxon Rank Sum Test (Mann-Whitney U Test)</u>	358
<u>16.5 Paired Wilcoxon Signed Rank Test</u>	359

16.6 Kruskal-Wallis Test (H Test)	360
16.7 Spearman Rank Correlation Coefficient	361
16.8 SPSS Lesson 14: Non-parametric Tests	362
16.9 R Lesson 14: Non-parametric Tests	363
16.10 Runs Test	364
17. Overview of the General Linear Model	
17.1 Linear Algebra Basics	367
17.2 The General Linear Model (GLM) for Univariate Statistics	368
Appendix: Tables	369

About This Book

Introduction to Applied Statistics for Psychology Students, by Gordon E. Sarty (Professor, Department of Psychology, University of Saskatchewan) began as a textbook published in PDF format, in various editions between 2014-2017. The book was written to meet the needs of University of Saskatchewan psychology students at the undergraduate (PSY 233, PSY 234) and graduate (PSY 807) levels.

In 2019-2020, funding was provided through the Gwenna Moss Centre for Teaching and Learning, along with technical assistance from the Distance Education Unit, to update and adapt this book, making it more widely available in an easy-to-use and more adaptable digital (Pressbooks) format. This update included an expansion to add chapters on using RStudio, as an alternative to SPSS. The update also made revisions so that the book could be published with a license appropriate for **open educational resources (OER)**.

OERs are defined as “teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use and repurposing by others” ([Hewlett Foundation](#)). This textbook and other OERs like it are openly licensed using a [Creative Commons license](#), and are offered in various digital and e-book formats free of charge.

Printed editions of this book can be obtained for a nominal fee through the University of Saskatchewan bookstore.

Licensing and Copyright

Licensing

Except where otherwise noted (see notes below on the copyright for SPSS and R screenshots), the content of this book is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Under the terms of the CC BY-NC-SA license, you are free to copy, redistribute, modify or adapt this book as long as you provide attribution. You may not use the material for commercial purposes. If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. Additionally, if you redistribute this textbook, in whole or in part, in either a print or digital format, then you must retain on every physical and/or electronic page an attribution to the original author(s).

Copyright: SPSS Screenshots

SPSS Inc. was acquired by IBM in October, 2009. Reprints of images (i.e., screenshots) from IBM® SPSS® Statistics software (“SPSS”) appear courtesy of International Business Machines Corporation, © International Business Machines Corporation. IBM, the IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “IBM Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml. This consolidated credit paragraph and corresponding copyright notices must be listed on a title page or other conveniently viewable

location where any reprints of this material appear. Any repurposing of the material in this book should also follow these same requirements.

The University of Saskatchewan Open Press obtained specific permissions from IBM to reprint IBM SPSS Statistics screen images for the purposes of publishing this book, according to the conditions outlined here. Individuals who wish to use, duplicate, or redistribute any of these images are advised to do so in compliance with copyright law or to contact IBM directly for permissions: <http://www.ibm.com/contact/submissions/extsub.nsf/copyright>. If any derivative version of this book (i.e., remixed, transformed, modified, or built-upon version) is created, additional copyright permission from IBM should be acquired for including any of their images in the derivative version before it is released.

Copyright: RStudio Screenshots

Reprints of images (i.e., screenshots) from RStudio are © the R Foundation, from <http://www.r-project.org>, and may be reproduced for any purpose provided they are credited to the R statistical software using an attribution such as this.

Cover Image

Cover image by Ron Borowsky and Gordon Sarty, used for public talks and released with a CC BY-NC-SA license. The statistical methods that you will learn in this course were necessary to produce the functional MRI (fMRI) brain maps illustrated on the cover. In particular, a one-way ANOVA technique was used to detect

the brain activations shown in the images¹. The study shown was designed to reveal ventral and dorsal stream processing for ‘what’, ‘where’ and ‘how’ interpretations of words and pictures presented to the experimental subjects while they were in the Magnetic Resonance Imager (MRI)².

1. Sarty GE, Borowsky R. “Functional MRI Activation Maps from Empirically Defined Curve Fitting”, *Concepts in Magnetic Resonance Part B (Magnetic Resonance Engineering)*, 24B, 46-55, 2005.
2. Borowsky R, Loehr J, Friesen CK, Kraushaar G, Kingstone A, Sarty GE, “Modularity and Intersection of ‘What’, ‘Where’, and ‘How’ Processing of Visual Stimuli: A New Method of fMRI Localization”, *Brain Topography*, 18, 67-75, 2005.

Acknowledgements

A sincere thank you must go out to the following University of Saskatchewan personnel, in acknowledgement of their support and contributions to this updated open textbook:

- Julie Maier (Instructional Designer, Distance Education Unit), for technical assistance with Pressbooks, OER and licensing guidance, editing and formatting assistance, developing resources for statistical software access, and project coordination.
- Heather Ross (Educational Development Specialist, Gwenna Moss Centre for Teaching and Learning), for support with obtaining the funding that allowed this project to move forward.
- Kate Langrell (Copyright Coordinator, University of Saskatchewan Library), for answering copyright questions, particularly regarding the use of software screenshots and data files.
- Naveed Ahmed (Research Associate, Department of Agriculture and Resource Economics), for content updates and updated data sets for SPSS Lessons 1 to 7.
- Osama Bataineh (Lab Coordinator & Sessional Lecturer, Department of Mathematics & Statistics), for major content editing, editing the SPSS Lessons and writing the new RStudio Lessons throughout this book, devising and compiling the complete collection of finished data sets, porting material into Pressbooks, LaTeX refinement, and – last but not least – piloting this new version of the textbook with his PSY 233 students for the first time in Spring 2020.

Statistical Software Used in this Book

Throughout this book you will find **Lessons** that will take you through procedures to manipulate and analyze given data using two statistical software applications:

1. IBM® SPSS® Statistics software (referred to more simply as “SPSS”)
2. RStudio, from The R Project for Statistical Computing.

Accessing SPSS and RStudio Through Your School

See the page [University of Saskatchewan: Software Access](#) for more details on how to do this.

Downloading SPSS and RStudio

SPSS Statistics

SPSS Statistics is **not** a free program.

A trial version of SPSS can be downloaded at: <https://www.ibm.com/analytics/spss-trials>

If you really want to download the program (not in a trial version), see some information on student rates at: <https://www.ibm.com/>

[analytics/academic-statistical-software](#); however, consider carefully how necessary this is before you spend any of your own money, and look carefully at any terms of licensing (i.e., some licenses may only give you access for a set number of months). Unless you are in a position where you can get an employer or research supervisor to pay for it, you may want to stick with the cost-free options available to you.

RStudio

A free, open-source, non-commercial desktop version of RStudio can be downloaded at: <https://rstudio.com/products/rstudio/download/>

Why does this book cover both SPSS and RStudio?

While both SPSS and RStudio are powerful analytical tools, they operate differently and each have their pros and cons.

The history of SPSS Statistics goes back to the 1960s, and for many years it has been a standard for students and researchers working in the social sciences (SPSS, in fact, originally stood for *Statistical Package for the Social Sciences*, but was later changed to *Statistical Product and Service Solutions*). It is still an extremely popular and commonly-used package, and one that you are likely to find is used in labs and workplaces when you start to search for research and employment positions. For this reason, it is still essential for psychology graduates to have a solid grasp of how to use this program.

The more-recent R Project for Statistical Computing (which put

together RStudio) offers a free, open-source option. This means that anybody can access the software, and its community of users and developers can contribute to improving and updating the software. While it is increasing in popularity, it has yet to reach the ubiquitousness of SPSS.

So which Lessons should you do? SPSS or RStudio?

Well, first ask your instructor – if they want you to submit assignments that utilize a particular program, or if they are likely to ask you to interpret outputs from a particular program on your examinations, then you'd better complete the **Lessons** for that program!

Second, consider what skills you want to develop. Employers and/or supervisors are still very likely to expect psychology graduates to have SPSS skills. Would it benefit you, or set you apart from other research/employment candidates, to have a good grasp of both SPSS and RStudio? Do both sets of **Lessons** and practice using each program.

If you are a graduate student or somebody with a bit more freedom, who is just looking for some help analyzing data to support their research, then you could choose either program. Trying both for a little while might give you a good sense of which you prefer and why, or which might work better for your particular research situations.

University of Saskatchewan: Software Access

On-Campus Lab Access

If you are a University of Saskatchewan student working on-campus, all computers in the Arts & Science computer labs should have both SPSS and RStudio installed. See <https://artsandscience.usask.ca/it/labs/> for a list of lab locations for the Saskatoon campus.

Remote / Off-Campus Access

Virtual Lab

If you are a University of Saskatchewan student working remotely (off-campus), you can access both SPSS and RStudio via the **Virtual Lab** at <http://vlab.usask.ca/>.

- Log in with your NSID.
- Click “All” to expand the menu, then click on “Common U of S”.
- Select “SPSS 26” (for SPSS) or “RStudio” (for R) to launch either program within the Virtual Lab.

More information on the Virtual Computer Lab (VLab) can be found here: <https://wiki.usask.ca/x/lozDTg>

IMPORTANT NOTE: In order to open any of the given [Data Sets](#)

(.sav files) in the Virtual Lab, they first need to be added to your **Cabinet** drive. See the next sections for details on how to upload them.

Accessing your Cabinet Drive

The following links will guide you through gaining access your **Cabinet** drive so that you can then add files to it. Choose from the following options depending on if you are using Windows or Mac.

Ensure you follow the steps for connecting to **Cabinet**, specifically.

Try the steps without a VPN first. If you have issues, set up the **VPN (Virtual Private Network)** and try that way. The steps for setting up the VPN can be found here: <https://wiki.usask.ca/x/0YnDTg>

For Windows

- How do I map a network drive like Cabinet, Jade or Datastore on Windows?: https://wiki.usask.ca/x/_4nDTg

For Mac

- How do I map a network drive like Cabinet, Jade or Datastore on Windows?: https://wiki.usask.ca/x/_4nDTg

Adding Files to your Cabinet Drive

First, download all of the .sav files from the [Data Sets](#) page onto your computer.

Once you have access to your **Cabinet** drive, choose a designated folder within this drive where you will add the .sav files you want to work with; you may wish to create a new folder for this purpose, with a title like, e.g., **PSY 233 files**.

From there you can copy or move the .sav files from your computer into your designated **Cabinet** folder.

Then, they will be available for you to access them within the **Virtual Lab**.

USask ICT Help

Still stuck? Visit <https://www.usask.ca/ict/help-support/it-support-services.php> for more one-on-one assistance.

Data Sets

The dataset files listed here, which are used in the **SPSS Lessons** and **RStudio Lessons** of this book, were created by Osama Bataineh.

They are released with a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license.

[HyperactiveChildren.sav](#)

[Caregiver.sav](#)

[HeightLatency.sav](#)

[AgeSmoker.sav](#)

[HeadCircum.sav](#)

[pHLevel.sav](#)

[Methadone.sav](#)

..

[ActivityValue.sav](#)

I. BACKGROUND AND MOTIVATION

1.1 Overview

1.1.1 Textbook Layout, * and ** Symbols Explained

This textbook has been designed for use in the statistics classes for psychology I teach at the University of Saskatchewan. It is designed to replace the expensive, and inadequate, texts that have traditionally been used for these classes.

The courses covered by this text are:

1. Univariate Statistics I: Chapters 1 to 10 (Psy 233, undergraduate course)
2. Univariate Statistics II: Chapters 11 to 17 (Psy 234, undergraduate course)
3. Multivariate Statistics: Future project (Psy 807, graduate course)

Since these courses are applied statistics courses, students do not need to understand the derivations of the formulae and procedures. So these aspects, the “cookbook” approach, is what you need to learn to pass the applied statistics courses.

Sections Marked with ** : But, in the sections marked with a ** there are detailed derivations for those who don’t want to believe in magic. Most psychology students will want to skip the ** sections.

Sections Marked with * : Other sections are marked with a *; those sections contain applied statistics material that is not part of the course but is material that an experimental psychology student has a good chance of needing in experimental courses and research projects. (The graduate course Psy 805 is a review of Psy 233/234

with the additional * sections covered – so this text might also be used for Psy 805.)

Psychology students at the University of Saskatchewan are required to learn how to use the statistics program SPSS. So “Lessons” for learning SPSS are included throughout the text, with RStudio Lessons as an alternative using a different program.

For Univariate Statistics I, the class material is organized in 3 blocks:

- Block 1 is an introduction to the basic tools of statistics and probability – Chapters 1 to 6.
- Block 2 gets you into the ideas of hypothesis testing – Chapter 9.
- Block 3 is material on one- and two-sample t -tests – Chapters 9 and 10.

1.1.2 Intro to Univariate Statistics

So, to begin the course material proper, we may identify two “kinds” of statistics:

1. **Descriptive Statistics:** The presentation, organization and description of data. (Graphs, means, standard deviations, etc.) Block 1 material is primarily about descriptive statistics. Descriptive statistics lead to ideas about *probability* – we will cover probabilities as given by functions known as the *binomial distribution* and the *normal distribution*.
2. **Inferential Statistics:** The use of *probability* to infer things about a *population* from a *sample* through the use of *hypothesis testing*. Why do we need inferential statistics? Because it is usually impossible to measure (poll) an entire population.

The goal of Univariate Statistics I is to understand inferential statistics as embodied in the t -tests. With blocks 2 and 3 we will build up the background for, and then learn 3 kinds of “ t -tests” to infer means in populations. To foreshadow, let’s take a look at a simple example. Say we are interested in people’s heights. Let’s look at three situations, corresponding to the three types of t -tests we will learn.

i. *One sample t -test.* The situation is as illustrated in Figure 1.1.

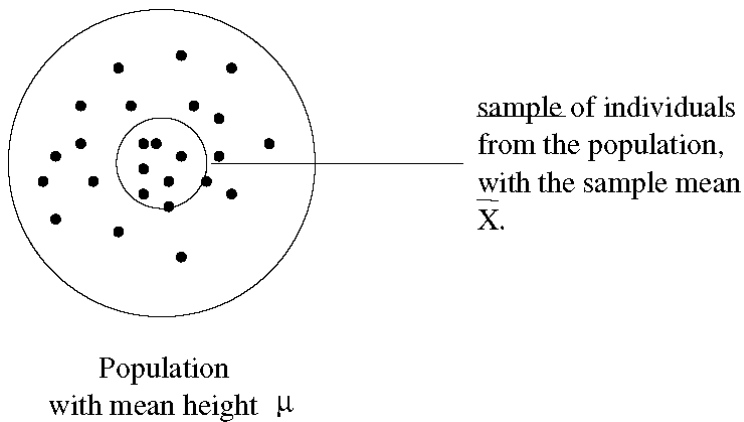


Figure 1.1: One sample t -test

The t -test will tell you when you may conclude that:

$$\mu = x_0 = \bar{x}$$

\uparrow \uparrow \uparrow
 pop. A priori sample
 mean guess mean
 about μ

Here the population could be the height of 10 year old children in Saskatchewan. The quantity μ is the actual average height of 10 year old kids in Saskatchewan. You could, in principle, measure all the 10 year olds in

Saskatchewan but, in practice you can't. Even if you spent the time finding them all and measuring their heights with a tape measure, they will be growing while you measure them all. It's generally impossible to measure a population in practice for some reason. Practically, we can only measure a small *sample* of children from the population. That sample will have a mean that we denote with \bar{x} . The *t*-test is a hypothesis test in which we compare the sample mean \bar{x} to a hypothetical mean x_0 and conclude with a probabilistic inference about μ .

ii. *Two sample t*-test. The situation is as illustrated in Figure 1.2.

The *t*-test will tell you when you can believe that $\mu_1 = \mu_2$ on the basis that $\bar{x}_1 \cong \bar{x}_2$. (The symbol \cong means "approximately equal to".)

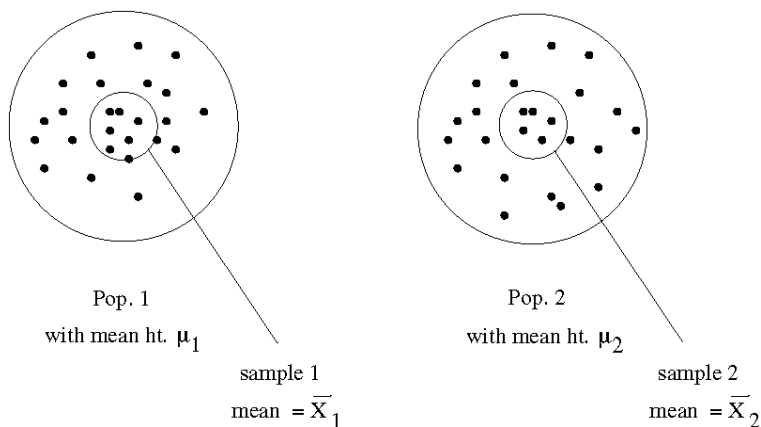


Figure 1.2: Two sample *t*-test

Here the two populations could be 10 year olds (population 1) and 11 year olds (population 2) in Saskatchewan. You might measure the two populations to get some idea about how

much 10 year old kids in Saskatchewan grow in one year. The two sample t -test will give you information on the difference of the average heights in the population, $\mu_1 - \mu_2$ on the basis of the difference of the means of small samples that you take from each population, $\bar{x}_1 - \bar{x}_2$.

iii. *Paired t -test.* The situation is as illustrated in Figure 1.3.

Say we want to know how fast a population grows in 1-year (e.g. pop = 10 year old kids). You can do the two-sample test with two separate populations but if you want to know how the environment affected the growth of the children (maybe you are concerned that they don't get enough to eat) then the two-sample test is only an approximation. The genetic composition, the natural ability to grow, may be different in the two separate populations. To get at the effect of the environment, without the measurements being confounded by individual differences, we would take a sample of 10 year old kids from the population now and measure their heights. Then we wait a year and measure the height of the same sample of now 11 year old kids. Then we combine the two samples of data into one data sample of differences. The Paired t -test will tell you if the average of differences (in heights) is zero or not.

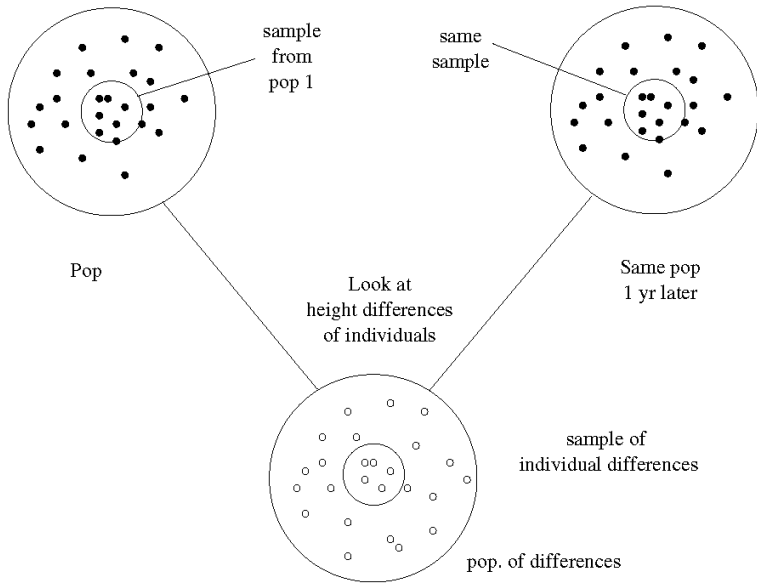


Figure 1.3: Paired t -test

1.2 Basic Definitions

Data : The numbers we collect. (Note the word data is plural. Datum is singular.) Data may be grouped into sets, hence *data set*.

Variable : A mathematical term used to denote something that can take on a range of *values*. There are important two types of variables :

- i. **Independent variable (IV)** : You set the value, a.k.a. *explanatory variable*.
- ii. **Dependent variable (DV)** : Value set (generally caused) by the independent variable, a.k.a. *outcome variable*. See Figure 1.4.

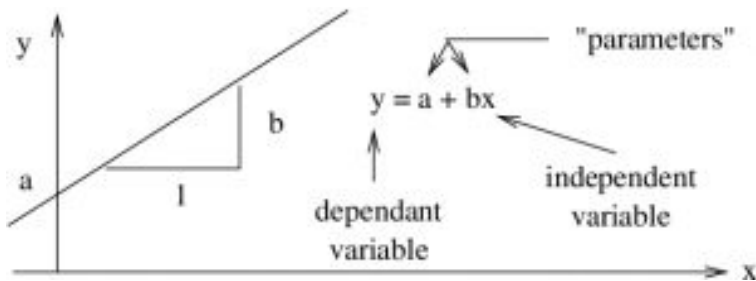


Figure 1.4: In the equation of a line, y is the dependent variable, x is the independent variable.

Random Variable : A dependent variable with random noise added. Value given by a *stochastic process*. We will only refer to random variables when discussing the theoretical relationship between probability distributions. Random variables, which we will denote with capital letters like X , are defined by their probability distribution. A stochastic process produces values that form a probability distribution if you allow the process that generates their values run for long enough.

Note : Data are frequently called “variables” in anticipation of how they will be used. The software program SPSS uses that convention.

1.2.1 Types of Data (important!)

Qualitative variable : described by a word, e.g. gender with “values” male or female. Qualitative variables are converted to *discrete quantitative variables* before analysis (e.g. male = 1, female = 2). In SPSS, you need to assign discrete numbers to qualitative variables in the “Values” column in the “Variable View” screen.

Quantitative variable : two types :

- i. **Discrete variable** : integer valued. In mathematical symbols $x \in \mathbb{Z}$ (read “the variable x belongs to [the set symbol \in means “belongs to”] the set of integers \mathbb{Z} ”). e.g. -2, -1, 0, 1, 2, 3, etc.
- ii. **Continuous variable** : real valued (essentially any number). In mathematical symbols $x \in \mathbb{R}$ (read “the variable x belongs to the set of real numbers \mathbb{R} ”). Geometrically, \mathbb{R} is the number line.

Note : Continuous variables can be converted to discrete variables by *grouping* :

heights \leq 5 ft = “short” (group value = 1)
heights $>$ 5 ft = “tall” (group value = 2)

Groups are also known as *classes*. We will be spending time defining classes in Chapter 2. Identifying what type of variable your data is will be the best way for you to decide what statistical test you need after you have learned and understood a number of different tests.

1.2.2 Measurement Scales (avoid this!)

Some texts, and the SPSS helper program (although I have never tried it), attempt to classify data into “scales” that try to go somewhat beyond the integers and real numbers. I don’t think such classification is particularly useful and recommend that you avoid such classification. Nevertheless, it exists, so we will take a very quick look at such scales. (There is no agreement about their definitions from source to source.)

One textbook that I used for a Univariate Statistics class for many years¹ lists 4 types of scales :

- i. nominal : discrete categories with no order (e.g. profession or gender) – qualitative.
- ii. ordinal : discrete categories with order (e.g. grades, A, B, C . . .) – qualitative.
- iii. interval : quantitative measure but no zero: ratios make no sense (e.g. temperature – makes no sense to say that one day was twice as hot as another day).
- iv. ratio : has zero, and hence ratios have meaning – quantitative.

SPSS uses :

- i. nominal.
- ii. ordinal.
- iii. scale : this scale is equivalent to the ordinal and ration scales listed above combined – as best as I can make out.

SPSS lets you specify a measurement scale under the “Measure”

1. Bluman AG, Elementary Statistics: A Step-by-Step Approach, numerous editions, McGraw-Hill Ryerson, circa 2005.

column in the “Variable View” screen. My recommendation is to leave it at “Unknown” or set it to “Scale”, otherwise it will try to restrict the statistical tests you can do when you don’t want it to. Measurement scales were invented to guide you to an appropriate statistical test but it doesn’t work that well. Instead, consider if your variable is continuous or discrete and then think about your situation.

1.2.3 Kinds of Sampling and Studies

This material properly belongs to a course on research methods and experimental design, but we will take a very quick look here. Ultimately your data need to be selected from the population at random. All mathematical statistical tests assume *random sampling*. The probability distributions that are used are *defined* by random sampling (the randomness – probability distribution relationship is pretty much a tautology). The real world is not ideal, however, and you may be forced to deal with bias introduced by the following sampling schemes :

1. Random Sampling : Samples selected from the population at random.
2. Systematic Sampling : The population is ordered somehow (e.g. by house address or by phone number) and there is a rule for selecting samples (e.g. every 4th house or every 10th phone number).
3. Stratified Sampling : The population is, or can be, ordered into groups and sampling is done at random from the groups.
4. Cluster Sampling : Restrict sampling to a few groups of the population (a few strata).

And, depending on the control you have over your independent variable, studies may be classified as :

1. **Observational Study** : Just watch. You have no control over the independent variables.
2. **Experimental Study** : Control some variables to isolate other variables. The object is to manipulate the independent variable.

Astronomy is a passion of mine; observing stars and planets through a telescope is an example of an observational study. Experimental studies can be affected (knowingly or unknowingly) by *confound variables*. These are causes (independent variables) that you are not interested in but which affect the outcome (dependent variables) and can lead to data *bias* that you need to account for. Such issues are beyond the scope of an introductory statistics course.

1.3 Summation Convention

For those of you who were ripped off in your high school education, a brief review of an important symbolic convention is given here. This convention will be used in the formulae that you will need to use.

The capital Greek Sigma, \sum , means sum or add. For example, suppose that you have 5 data sample values, represented abstractly by d_1, d_2, d_3, d_4 and d_5 , or more abstractly (using set notation) by:

$$d_i, i \in \{1, 2, 3, 4, 5\} \text{ (or } i = 1, 2, 3, 4, 5)$$

If you want to add the 5 values you would write:

$$d_1 + d_2 + d_3 + d_4 + d_5$$

or

$$\sum_{i=1}^5 d_i$$

Sometimes people get lazy and leave off the *limits* on the summation sign \sum and write

$$\sum d_i$$

where it is hopefully clear that i is the *summation index*. We can also leave off the summation index and write

$$\sum d$$

just to remind us that we need to add up a bunch of numbers generically represented by d . This last convention is useful for us because whenever we need to deal with a sum in a formula, we will get that sum from adding up numbers in a table that we have constructed.

2. DESCRIPTIVE STATISTICS: FREQUENCY DATA (COUNTING)

Statistical inference is based on probability and probability is based on counting (at least the “frequentist” definition of probability – more about that in Chapter 4). So let’s start counting!

2.1 Frequency Tables

Most material in this text is introduced first at an abstract level, then generally a step-by-step recipe is given and finally example problems are solved. This general to specific approach to learning statistics is the opposite of how many introductory statistics tests for the social sciences teach. For our first topic of frequency tables, the abstract concept is counting so let's dive into the recipe with the expectation that you won't get the complete picture until an example or two is worked.

The construction of a frequency table proceeds in two steps :

Step 1 : Determine the classes. There are two possibilities here, either the classes are given to you (pre-defined) or you have to define the classes based on the number of groups you want. So either

- i. Classes are given – nothing to do.
- ii. Define classes based on the number of groups you want. There are a number of different ways to group data into classes. We will cover a method here, different from Bluman's, that works for whole number data only. Here are the steps for that method :

(a) determine high data limit, H and the low data limit, L .

(b) compute the range $R = H - L$

(c) compute the class width :

$$W = \frac{R + 1}{G}$$

where G is the number of groups (or classes) you want.

(d) Begin the frequency table's first two columns :

Class	Class Boundaries
L to $(L + W - 1)$	$(L - 0.5)$ to $(L - 0.5 + W)$
$(L + W)$ to $(L + 2W - 1)$	$(L - 0.5 + W)$ to $(L - 0.5 + 2W)$
⋮	⋮
	$(H + 0.5 - W)$ to $(H + 0.5)$

Note : If the classes are given, you won't have, or need, the second column.

In the class column above a specific way of labelling classes is given. (We will see how this works exactly in the upcoming example.) This is to make the class names useful for seeing that the classes are uniquely defined – there will be no data points on the boundaries of the classes. The numbers in the labels will be whole numbers, since we are assuming that the data are whole numbers¹. In general we can label the classes any way we like.

Also we need to note that this procedure of defining classes using the formula given in step (2)(c) will only work for whole number data. In general the process of defining classes is a lot looser; there are few rules beyond thinking about what kind of information you hope to capture by defining the classes. Since I want to keep you focused on learning the basic ideas and not worry about stuff that is not really statistics all assignment and exam questions that ask for the construction of classes from quantitative data will be for whole number data only. The procedure given here does work in general but some data points may end up on class boundaries and will have

1. Whole numbers are 0 and the positive integers.

to make up an arbitrary rule about which class the data point should go in.

Step 2 : Construct the frequency table and fill it in :

Class	Class Boundaries	Tally	Frequency	Cumulative Freq.	Relative Freq.
			a	a	a/n
			b	$a + b$	b/n
			c	$a + b + c$	c/n
			\vdots	\vdots	\vdots
				n	

The last number in the cumulative frequency column, n , should equal number of data points as a check since it is the sum of the frequencies. And the sum of the relative frequencies will be 1 – we will see that this is an essential feature of probabilities. The tally column is optional.

Example 2.1 : 25 army inductees were tested for blood type. The data are :

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency table.

Solution :

Step 1 : Classes are given : A B O AB

Step 2 : Construct frequency table :

Class	Tally	Frequency	Cumulative Freq.	Relative Freq.
A		5	5	$5/25 = 0.20$
B		7	12	$7/25 = 0.28$
O		9	21	$9/25 = 0.36$
AB		4	25	$4/25 = 0.16$

The tally is actually silly in this case because you count² all the instances of A for the class A, etc., and you're done. The tally column will be more useful for the next example.

Example 2.2 : Given the high temperature data for each of 50 states for the month of July :

- The frequency of A is the number of times A is in the dataset, etc. ← **the take-home concept here.**

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Construct a frequency table using 7 classes.

Solution :

Step 1 :

(a) High limit, $H = 134$

Low limit, $L = 100$

(b) Range: $R = H - L = 134 - 100 = 34$

(c) Class width: $W = \frac{R+1}{G} = \frac{34+1}{7} = 5$

(d) (and continue to Step 2) :

Step 2 :

Class	Class Boundaries	Tally	Frequency	Cumulative Freq.	Relative Freq.
100 – 104	99.5 to 104.5		2	2	0.04
105 – 109	104.5 to 109.5		8	10	0.16
110 – 114	109.5 to 114.5	etc.	18	28	0.36
115 – 119	114.5 to 119.5		13	41	0.26
120 – 124	119.5 to 124.5		7	48	0.14
125 – 129	124.5 to 129.5		1	49	0.02
130 – 134	129.5 to 134.5		1	50	0.02
					= 1

Note how we can now use the tally column to keep track of our

counting. For example, for the class 100 – 104, we first count all the instances of 100 (there is 1), then 101 (none), 102 (none), 103 (none) and 104 (one). The sum of the frequencies is $n = 50$ and the sum of the relative frequencies is 1. Imagine that this data set represented the whole population and not just a sample. Then if you picked a random state there would be a 0.16 probability that the temperature would be between 105 and 109 inclusive. On other words relative frequency = probability for a population. Hence the term *frequentist* definition of probability. \square

You can also compute cumulative relative frequency in a frequency table. When you use SPSS to make a frequency table you will run up against the limitations of using black box canned software. SPSS produces only one style of frequency table and it doesn't match what we've been doing. In fact SPSS won't compute relative frequency; instead it computes "percentage". You need to convert percentage to relative frequency in your brain by dividing by 100.

2.2 Plotting Frequency Data

In general you may present your data, say in a report or paper, in tabular form or graphical form. Personally, I prefer graphical form – “a picture is worth a thousand words”. For frequency data, the frequency table is the tabular form. There are several ways of presenting the same data graphically, the primary way being the histogram:

1. Histogram – plot of frequency data using steps (mathematically: “step functions”).
2. Frequency polygon – plot of frequency data using straight lines (mathematically: “piece-wise linear functions”).
3. Cumulative frequency graph.
4. Pie charts, Pareto charts, Stem & Leaf plots – alternate ways of plotting frequency data

As a first step to plotting frequency data, you will need to construct a frequency table.

Example 2.3 : Continuing with the frequency table produced from the data given in Example 2.1 :

Class	Frequency	Cumulative Freq.	Relative Freq
A	5	5	0.20
B	7	12	0.28
O	9	21	0.36
AB	4	25	0.16

We will demonstrate most of the graph types using these data.

1. Histograms. First, the straight forward *histogram* is as shown in

Figure 2.1. This is a plot of the data in the frequency column of the frequency table.

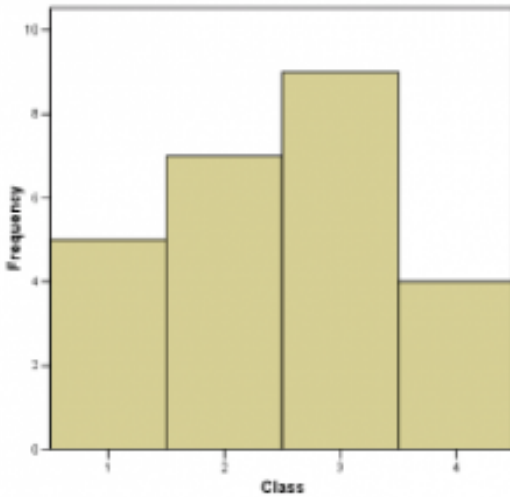


Figure 2.1 : Straight Forward histogram. A box or “step function” is used to show the frequency of each class. In this image, generated with SPSS, the classes are labelled with 1, 2, 3, and 4 which correspond to the classes A, B, O and AB. If we take these discrete quantitative class values literally, the class width is one. Keep that in mind when you look at Figure 2.2.

Next, still under the category histograms, is the *relative frequency histogram*. The relative frequency histogram for the blood type data is shown in Figure 2.2. It is a plot of the data in the relative frequency column of the frequency table.

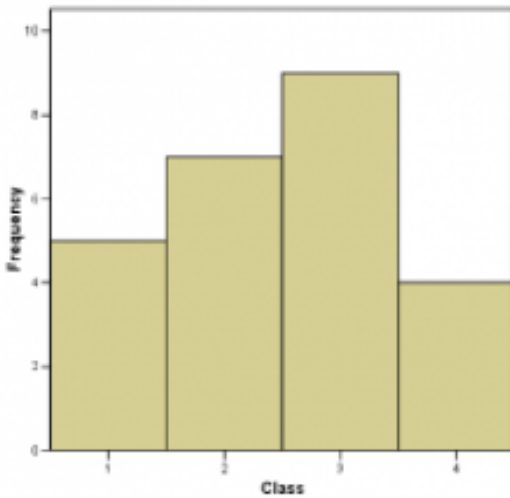


Figure 2.2 : Relative frequency histogram for the blood type data.

Very Important Concept : Look at Figure 2.2 and define the width of each class to be 1. Then the area under the histogram “curve” is $(0.2) \times 1 + (0.28) \times 1 + (0.36) \times 1 + (0.16) \times 1 = 1.00$. So, if we image that our data sample of the 25 army inductees is a whole population, then the relative frequency histogram may be interpreted as giving the following *probabilities* for getting a particular blood type for someone selected randomly from the population:

The probability of having type A blood is 0.20 (or 20%).

The probability of having type B blood is 0.28 (or 28%).

The probability of having type O blood is 0.36 (or 36%).

The probability of having type AB blood is 0.16 (or 16%).

2. Frequency Polygons. Frequency polygons are just another form of histogram. We have been talking about “area under the curve” to represent probability. The curve of a frequency polygon is a little bit smoother than the curve of a traditional histogram. Frequency

polygons can, of course be made for either straight frequency or relative frequency data. A frequency polygon for the relative frequency blood type data is shown in Figure 2.3.

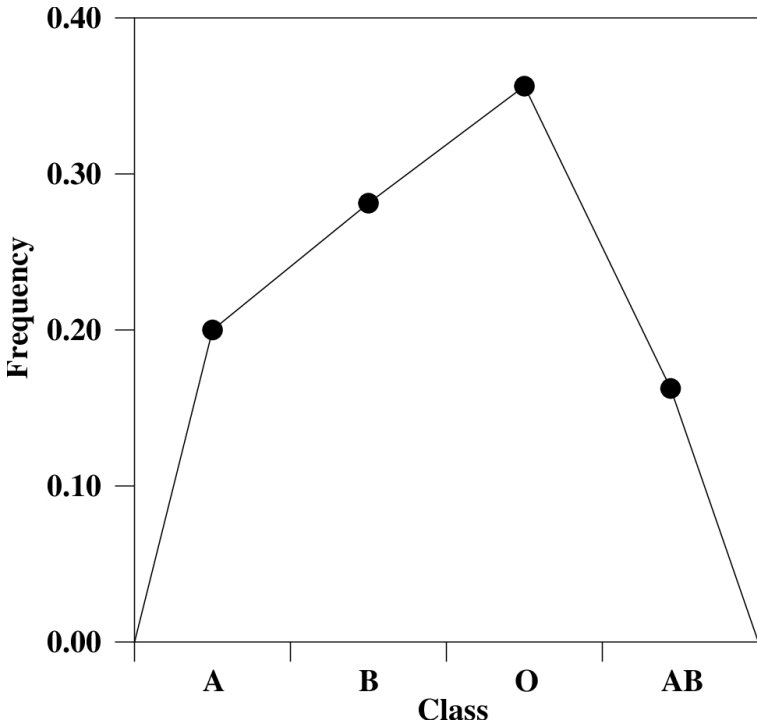


Figure 2.3 : Relative frequency polygon for the blood type data. Plot a dot at the center of each class at the y -value of the relative frequency then connect the dots as shown.

3. Cumulative Frequency Graph. Plotting the cumulative frequencies from the frequency table results in a cumulative frequency graph as shown in Figure 2.4. Cumulative relative frequencies can also be computed (add up relative frequencies as you move down the column) and plotted.

The cumulative frequency graph shows the “area under the curve” (of the traditional histogram) from the beginning of the first class

up to the given point. Cumulative frequencies or cumulative relative frequencies will therefore show up later as areas under probability distribution curves up to a given point (it represents the probability of having a value equal to or less than the given value if that quantity is pulled at random from the population.)

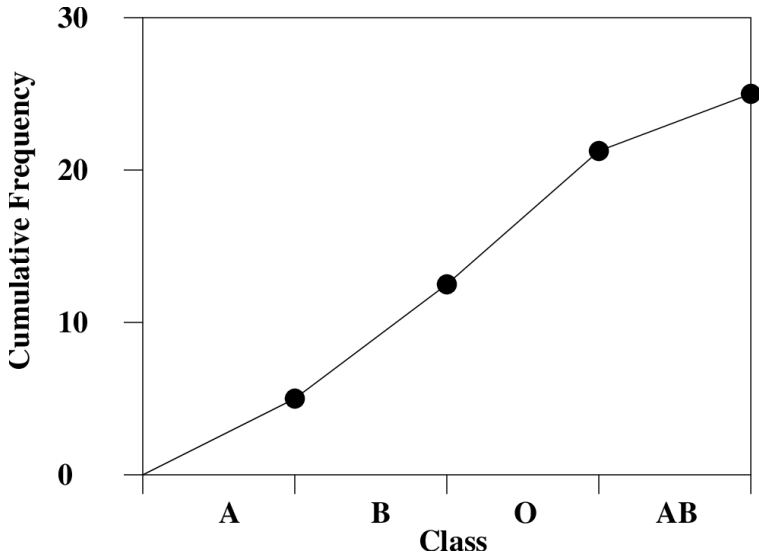


Figure 2.4 : Cumulative frequency graph for the blood sample data. Plot a dot at the end of the relevant class at a y -value equal to the cumulative frequency. Then connect the dots as shown.

4. Pie Chart. A pie chart is a round histogram. Everyone has seen a pie chart, it is intuitive. The angles in the pie chart are computed using:

$$\text{Angle} = \text{Relative Frequency} \times 360^\circ.$$

For the blood type data, the explicit angle calculations are :

Class	Angle
A	$0.20 \times 360^\circ = 72^\circ$
B	$0.28 \times 360^\circ = 100.8^\circ$
O	$0.36 \times 360^\circ = 129.6^\circ$
AB	$0.16 \times 360^\circ = 57.6^\circ$
	Check Sum = 360°

The pie chart for the blood type data is shown in Figure 2.5.

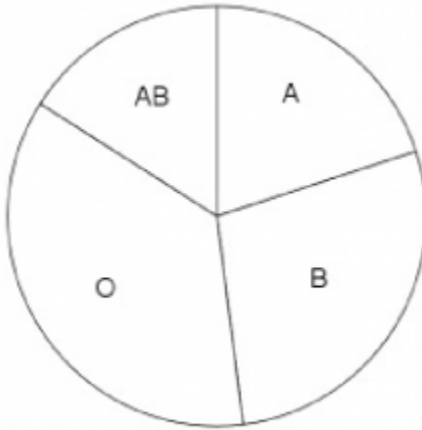


Figure 2.5 : Pie chart for the blood type data. It is a very good representation of the probability aspect of relative frequency. If you made the pie chart into a dart board and threw darts at it in a random fashion, then the probability of the dart landing in each class is equal to that class's relative frequency.

5. Pareto Chart. The Pareto chart is just an ordered histogram with

classes ordered from highest to lowest frequency. The classes need to be qualitative for this reordering to make sense of course. To construct a Pareto chart, writing an ordered frequency table down first will help :

Class	Frequency
A	5
B	7
O	9
AB	4

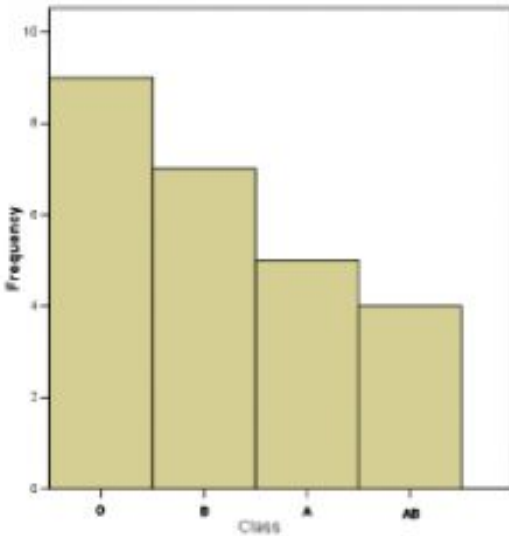


Figure 2.6: Pareto chart for the blood type data.

The Parato chart is plotted in Figure 2.6. The frequencies as ordered in a Parato chart can be given statistical meaning but that is a

subject beyond the scope of this course. Here you just have to be aware that such a chart exists and know how it is made.

2.2.1 Stem and Leaf Plots

A stem and leaf plot is a fancy kind of histogram that lets you see all your data instead of just class frequency information.

The steps for making a stem and leaf plot are :

1. Order the data (this is a frequently used, tedious, step for many procedures as we'll see).
2. Divide into classes of 10's or 5's (low decade and high decade).
3. Use "leading" and "trailing" digits of the data values to make the plot.

For step 3 you need to know what "leading" and "trailing" digits are. Let's illustrate that with an example.

Example 2.4 : Given classes: 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or equivalently, divide the classes into 5's and the data *in order* (i.e. with the tedious ordering step 1 already done) :

|50,51,51,52,53,53,|55,55,56,57,57,58,59,|62,63,|65,65,66,66,6
7,68,69,69|72,73,|75,75,77,78,79|

where the bars illustrate the division of the data into low and high decades, step 2. The first number of each data point is the leading digit (stem), the last, the trailing digit (leaf). So with this, step 3 leads to :

Stem	Leaf
5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

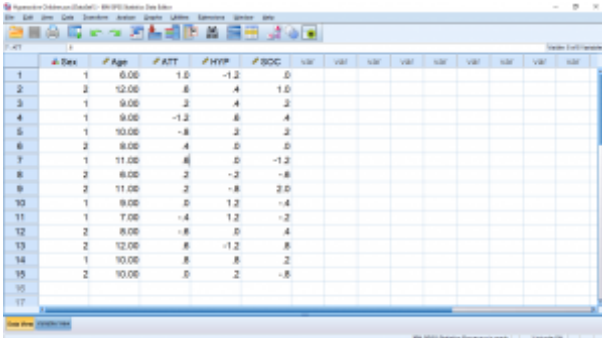
Notice how, since the numbers are all nicely lined up, that the stem and leaf plot is a histogram on its side. So you can visualize frequency information *and* see the values of the individual data points as well. One could use that information to compute accurate means from stem and leaf plots whereas, as we'll see, "class centers" need to be used with histogram (frequency table) data to estimate means with grouped data formulae.

2.3 SPSS Lesson I: Getting Started with SPSS

The following lesson will take you through an introduction to IBM® SPSS® Statistics software (referred to hereafter as “SPSS”).

First, you need to open SPSS. Ways to do that are detailed in the Front Matter of this book, in the section “[Statistical Software Used in this Book](#)“. Also in the Front Matter you will find the collection of provided [Data Sets](#); download the file “HyperactiveChildren.sav” and open it in SPSS.

You should see:



The screenshot shows the SPSS Data View window with a data matrix. The columns are labeled #Sex, #Age, #ATT, #HYP, #SOC, #VAR, #VAR, #VAR, #VAR, #VAR, #VAR, #VAR, and #VAR. The rows contain numerical data for each variable.

	#Sex	#Age	#ATT	#HYP	#SOC	#VAR	#VAR	#VAR	#VAR	#VAR	#VAR	#VAR
1	1	6.00	1.0	-1.2	.0							
2	2	12.00	.8	.4	1.0							
3	1	9.00	.2	.4	.2							
4	1	9.00	-1.2	.8	.4							
5	1	10.00	-.8	.2	.2							
6	2	8.00	.4	.0	.0							
7	1	11.00	.8	.0	-1.2							
8	2	8.00	.2	-.2	-.8							
9	2	11.00	.2	-.8	2.0							
10	1	8.00	.0	1.2	-.4							
11	1	7.00	-.4	1.2	-.2							
12	2	8.00	-.8	.0	.4							
13	2	12.00	.8	-1.2	.8							
14	1	10.00	.8	.8	.2							
15	2	10.00	.0	.2	-.8							
16												
17												

SPSS
screenshot ©
International
Business
Machines
Corporation.

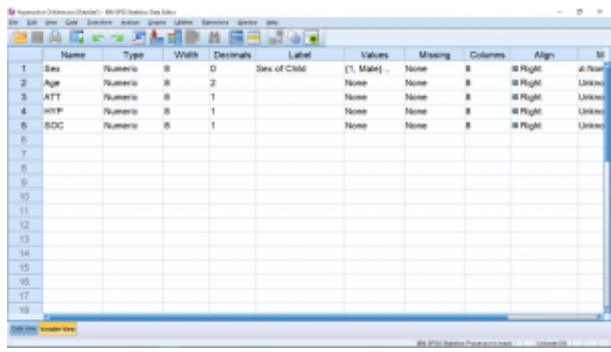
This is the “Data View” window. It is one of the three windows you will see when you use SPSS. The other two windows are the “Variable View” window and the “Output” window. You can get to the Variable View window either by clicking on the Variable View tab at the bottom of the window, or by double clicking one of the column headings (the “variable name”). But let’s talk about what’s on the Data View window before we look at the other two windows.

The Data View window is arranged in the form of a “data matrix”, which is an essential structure for multivariate statistics. This is the

first trap that people who try to use SPSS fall into – they collect data, put the data into SPSS and then go looking for an appropriate statistical test using help or the built-in “statistics coach”. Multivariate statistics is advanced. We need to learn a whole lot of basics before we can competently use multivariate statistics. This textbook covers *univariate* statistics. We are only going to learn how to deal with *one* dependent variable at a time. So many of the first SPSS lessons will be about how to combine multiple variables into one variable for analysis.

Back to the Data View window and the data matrix. *The rows represent individual subjects in the study.* In Psychology, the subjects (“participants”) are generally people but they could also be rats or schools or cities or whatever. To fix ideas, suppose the subjects are people. One line for each person in the study. *The columns represent variables.* SPSS doesn’t care what kind of variables you define (e.g. independent or dependent) so you need to keep track of their meaning yourself. As we said, we only need one independent variable for univariate tests.

The variables need to be defined. This is done by either double clicking on the variable name at the top of a column or by clicking the “Variable View” button at the bottom. Either way, you’ll end up in the Variable View window that looks like :



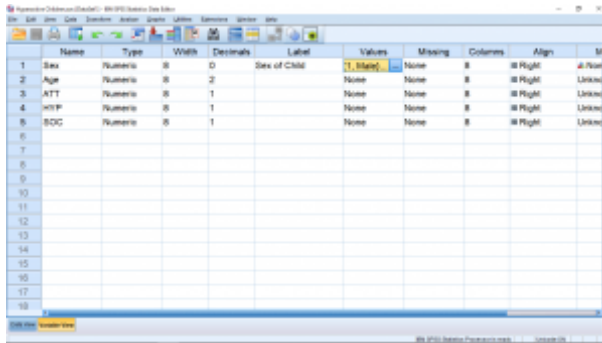
SPSS screenshot © International Business Machines Corporation.

Each line in the Variable View window lists the attributes of the

variables listed in the Data View window. You can usually leave most of the attributes as they come by default. The big exception is the Values attribute – it's important and we'll come back to that after a quick look at the other attributes.

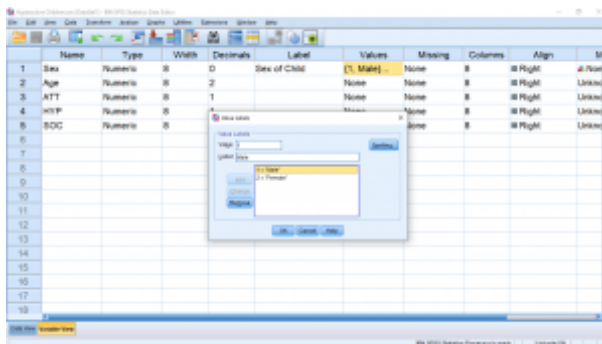
The Name attribute gives the name of the variable as it appears at the top of the columns in the Data View window. Type should be Numeric if you want to use the variable in any kind of statistical calculation. Having this set to String will cause errors if you are trying to use the variable as a qualitative variable (selection is via a pull down menu that appears when you click on a cell). Qualitative variables need to be Numeric and they are handled with the Values attribute – as we'll see shortly! The Width and Decimals attributes are just to format the appearance of the numbers in the Data View sheet; totally not critical. The Label is left over from early FORTRAN days. SPSS's heart is written in FORTRAN and variable names in FORTRAN used to be limited to eight characters which frequently makes it awkward to have good name for the variable. With Label you can give the variable a good name. If there is a value for Label then that value will be used on table and graph outputs that SPSS makes. If Label is blank then SPSS will use Name on table and graph outputs. We will largely ignore missing value issues in this course so leave the Missing attribute at None. Columns and Align are again used to make the Data View presentation look a little better; totally not critical. Leave Measure at Unknown or Scale, otherwise SPSS will try to interpret your data for you. SPSS is not very good at that and will tend to give strange errors that will make no sense to you, so leave Measure at Unknown or Scale. Leave Role at Input; this is a relatively new feature of SPSS and I don't know what it does, so don't muck with it.

Finally – the Values attribute! Here is where you make the link between a qualitative variable and the discrete values it needs to work in a computer setting. Let's take a look at the gender variable. Clicking in the cell brings up a thing with three dots :



SPSS screenshot © International Business Machines Corporation.

Clicking on the thing with three dots brings up a menu where you can define the connection between the qualitative description and your discrete number assignments :

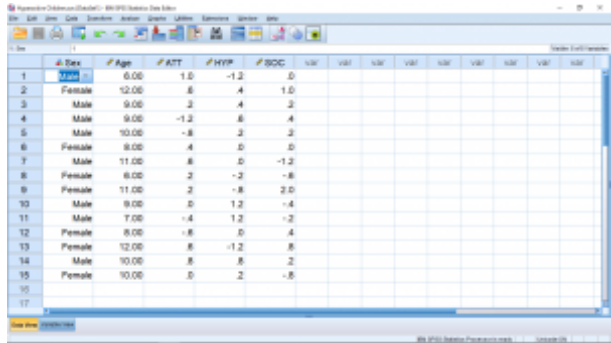


SPSS screenshot © International Business Machines Corporation.

Here I have clicked on the 1.00 = “Male” line to show that the Value is 1 (arbitrary discrete quantitative) and the Label is Male (qualitative). To enter new values, type them in the Value and Label box and then click Add to add them to the list.

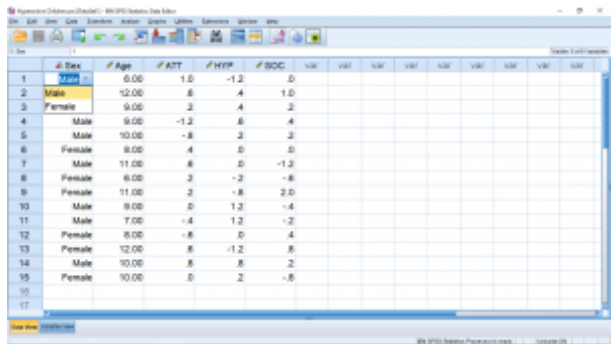
Let’s go back to the Variable View window to see how quantitative variables with discrete number assignments are handled. Look at the values in the sex variable column in the first image. The numbers 1 and 2 are shown which represent Male and Female. To see that

representation explicitly, click on the 1-A icon at the top of the window. You will then see:



SPSS screenshot © International Business Machines Corporation.

There's more. If you click on a cell in the gender variable, you will get a thing on the side of the cell and if you click on that thing, you will see:



SPSS screenshot © International Business Machines Corporation.

This pop-up allows you to change the value by clicking on the appropriate value. In one of your assignments you will get practice with entering qualitative data this way. In general, to enter data into SPSS from scratch, you can start by typing data into the Data View window and then fix up the attributes later in the Variable View window. For qualitative variables the best approach is to define the

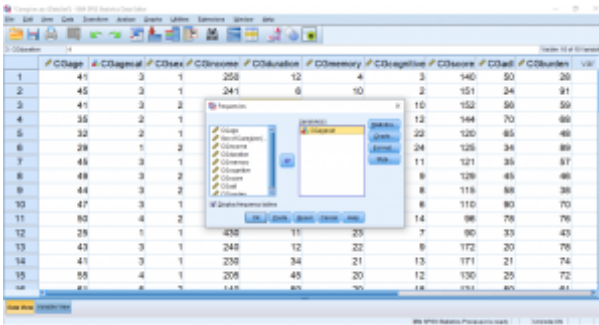
variable first in Variable View, getting the proper values into the Values attribute. Then you can go back to the Data View window and enter the qualitative data either by pulling down the menu when the mode of the 1-A icon is to show the labels or by remembering the number assignment and entering the numbers when the 1-A icon is set to show values.

Let's move on to do some descriptive statistics and see what results will look like in the Output window. For this load in the "Caregiver.sav" file from the [Data Sets](#):

	CChange	CChange1	CChange2	CChange3	CChange4	CChange5	CChange6	CChange7	CChange8	CChange9
1	41	3	1	350	12	4	3	140	50	20
2	45	3	1	341	6	10	2	151	34	81
3	41	3	2	320	4	11	10	152	58	59
4	35	2	1	290	23	19	12	144	70	88
5	33	2	1	140	22	20	22	120	65	48
6	29	1	2	132	23	31	24	125	34	89
7	45	3	1	76	32	33	11	121	35	97
8	49	3	2	322	31	15	9	129	45	46
9	44	3	2	348	25	19	6	118	88	38
10	47	3	1	810	18	19	6	110	80	70
11	50	4	2	440	18	49	14	96	78	76
12	29	1	1	430	11	23	7	80	33	43
13	43	3	1	240	12	22	9	172	20	78
14	41	3	1	230	34	21	13	171	21	74
15	55	4	1	205	45	20	12	130	25	72
16	41	4	1	105	25	20	18	131	40	81

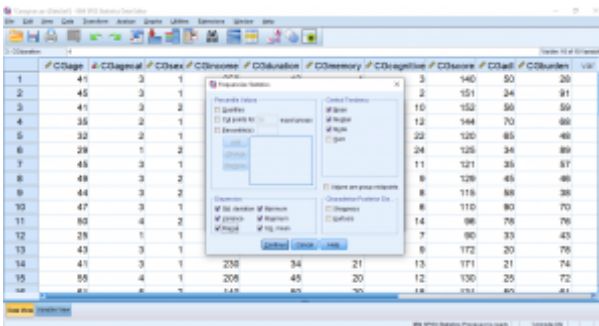
SPSS screenshot © International Business Machines Corporation.

There are 50 subjects in this file and 10 variables. One of the things we'll be learning, in later SPSS Lessons, is how to combine more than one variable into one variable. This is because we are studying univariate statistics which means we only want to deal with one dependent variable at a time. For now, let's pick on the variable CGDUR and see how we can generate descriptive statistics output. There are three ways to do this and they all begin in the Analyze → Descriptive Frequencies menu which looks like this (on a PC; very similar on a Mac):



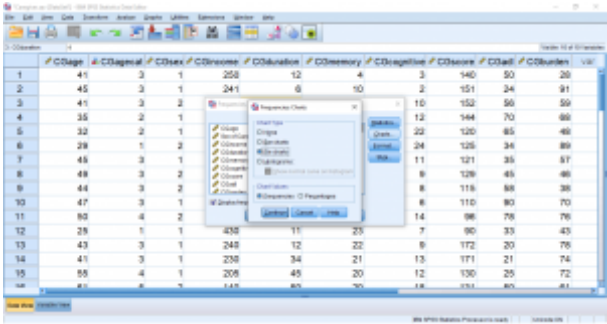
SPSS screenshot © International Business Machines Corporation.

Let's take a look at the submenus and set them up before we hit OK. First the Statistics... submenu. In that menu check off Mean (\bar{x}), Median (MD), Mode, Skewness, Kurtosis, Std. deviation (s), Variance (s^2), Range (R), Minimum (L) and, Maximum (H). We we look at all of those descriptive statistics in Chapter 3.



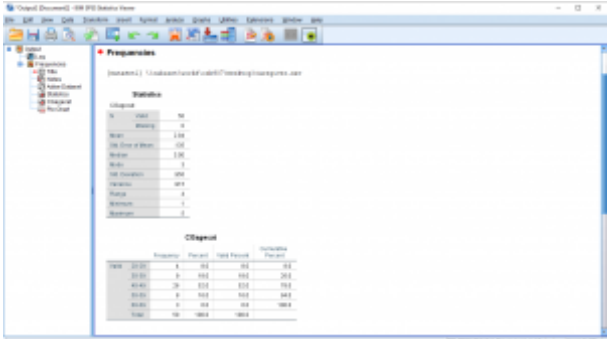
SPSS screenshot © International Business Machines Corporation.

Hit Continue, look at the Charts... menu and check off pie charts, just for fun:



SPSS screenshot © International Business Machines Corporation.

Hit Continue. You can look at the Format... and Style... menus if you want, they are not particularly interesting. Make sure “Display frequency tables” is checked (this will be important when you do the assignments), then hit OK. The Output window will pop up and in that window you will see:

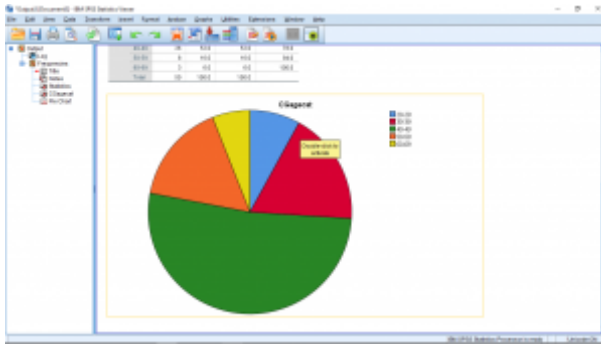


SPSS screenshot © International Business Machines Corporation.

The first table, Statistics, shows the descriptive statistics you asked for. Note, especially, for future reference (when we hit skewness in Chapter 3), the value of the skewness. It is 0.411. More to the point it is > 0 , or positive, meaning that the data set (CGagecatn) is right skewed or positively skewed. The

second table, labeled “highestQualification” is the frequency table (note how the variable Name and not the Label was used because the Label attribute for the highestQualification variable was blank). The structure of the frequency table is slightly different from how we will learn to construct one by hand. There is nothing you can do to make SPSS produce a frequency table that matches exactly like what you might want. There are limitations to using canned statistics software.

Scrolling down the Output window you will see the pie chart:

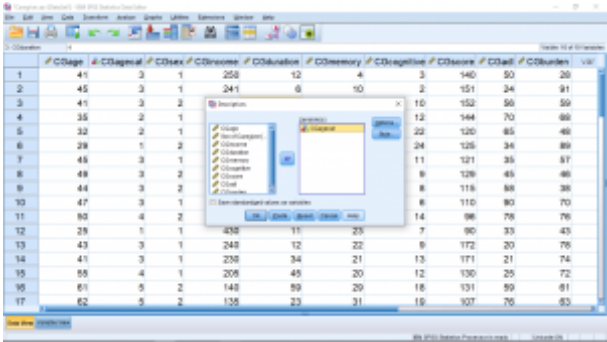


SPSS screenshot © International Business Machines Corporation.

Lets look at the Descriptives... menu next:

CCollege	N	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis	Missing	Valid
1	45	4.5	1.12	1	5	.00	-.00	0	45
2	45	4.5	1.12	1	5	.00	-.00	0	45
3	45	4.5	1.12	1	5	.00	-.00	0	45
4	35	4.5	1.12	1	5	.00	-.00	0	35
5	32	4.5	1.12	1	5	.00	-.00	0	32
6	28	4.5	1.12	1	5	.00	-.00	0	28
7	48	4.5	1.12	1	5	.00	-.00	0	48
8	48	4.5	1.12	1	5	.00	-.00	0	48
9	44	4.5	1.12	1	5	.00	-.00	0	44
10	47	4.5	1.12	1	5	.00	-.00	0	47
11	80	4.5	1.12	1	5	.00	-.00	0	80
12	25	4.5	1.12	1	5	.00	-.00	0	25
13	43	3	1	248	12	22	8	9	172
14	41	3	1	238	34	21	13	171	21
15	55	4	1	205	45	20	12	130	25
16	61	5	2	148	99	20	18	131	91
17	62	5	2	135	23	31	19	107	70

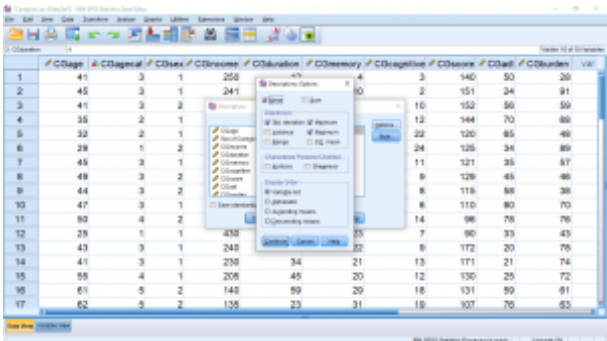
SPSS screenshot © International Business Machines Corporation.



SPSS screenshot © International Business Machines Corporation.

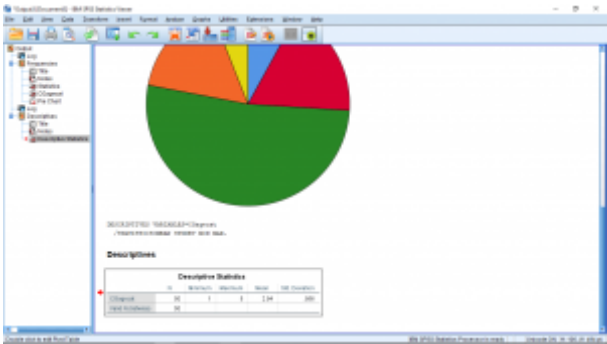
Move the CGagecat variable over as before and make sure to check off the “Save standardized values as variables”. We’ll learn about standardized values (z -values) in Chapter 3. Take note, this is the only way to get SPSS to compute z -values :

Click the options menu and check off descriptive statistics to compute, as before (S.E. mean is Standard Error of the mean which we’ll get to eventually also, we’ll just leave it off for now):



SPSS screenshot © International Business Machines Corporation.

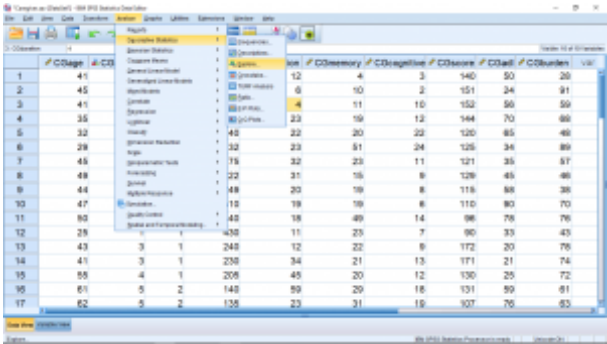
Hit Continue then OK and look at the results in the Output window. The output is straightforward:



SPSS
screenshot ©
International
Business
Machines
Corporation.

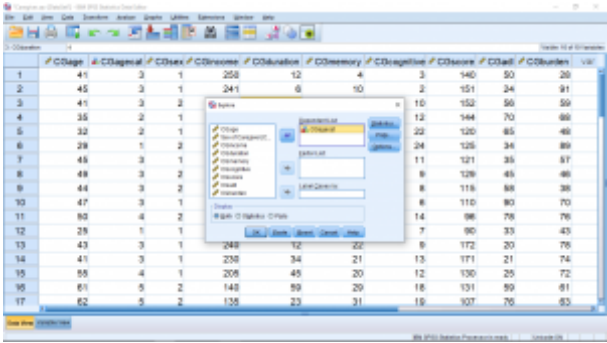
In Chapter 3 we will learn that the mean of a z -transformed variable is zero and the standard deviation is one. That is confirmed here. If you left the “Save standardized values as variables” box checked when you ran this, you’ll get another variable added in the Data View window – the z -transform of the z -transform. It’s the same, the z -transform of a z -transform give back the same numbers. But note that the skewness (0.411) of the z -transformed variable is the same as the skewness of the original variable. This means that z -transforming a variable doesn’t change anything about the variable except its mean and standard deviation. This is important when it comes to using and interpreting any analyses based on the z -transformed variable.

Finally, let’s look at the Explore... menu:



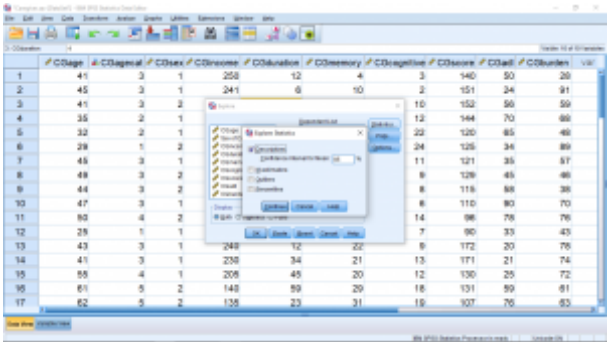
SPSS
screenshot ©
International
Business
Machines
Corporation.

Move CGagecat into the “Dependent List”. Don’t worry about “Factor List”, you should leave it blank (for future reference, “factor” is synonymous with “independent variable”):



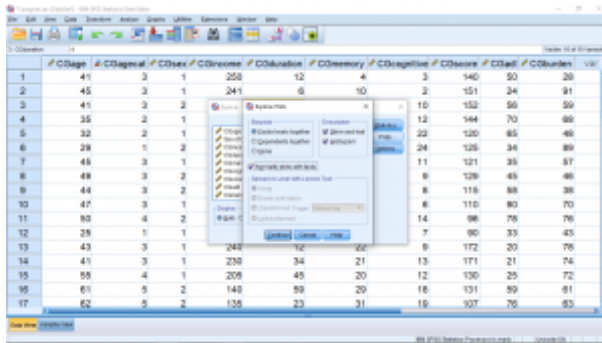
SPSS screenshot © International Business Machines Corporation.

Take a look at the Statistics... menu. You can leave it as it is (we’ll be learning about Confidence Intervals later):



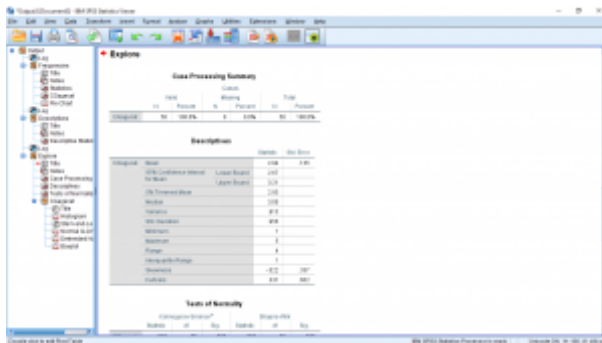
SPSS screenshot © International Business Machines Corporation.

Hit Continue and open the Plots... menu and check off the items as shown:



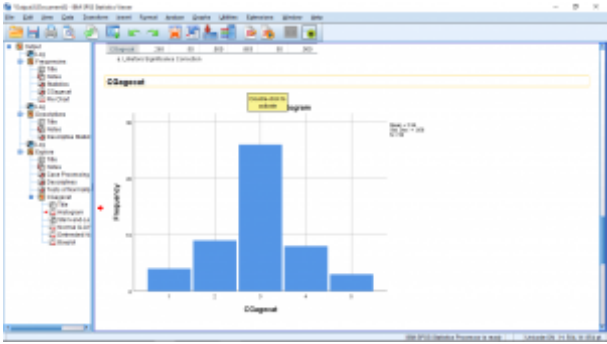
SPSS
screenshot ©
International
Business
Machines
Corporation.

We will talk about these different plots soon. For now, hit Continue, the OK and look at the output. First the tables:



SPSS
screenshot ©
International
Business
Machines
Corporation.

The first table is a “missing data report” that many SPSS procedures will output as a matter of course. You can ignore the missing data reports. Pay attention to the “Descriptive” table (it is something you could be asked about on exams!). You can ignore the “Tests of Normality” table. Next the plots. The first one is a histogram:



SPSS
screenshot ©
International
Business
Machines
Corporation.

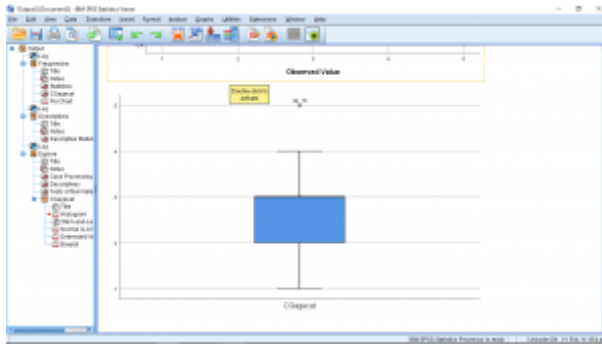
After we cover skewedness in Chapter 3, come back to this picture and note how the histogram is right skewed.

Next is the stem and leaf plot. Remember that the way to a stem and leaf plot in SPSS is through the Explore menu:



SPSS
screenshot ©
International
Business
Machines
Corporation.

You can ignore the Q-Q plots but note that a boxplot is produced:



SPSS
screenshot ©
International
Business
Machines
Corporation.

This is not a very good boxplot. Again, we'll be learning about boxplots later.

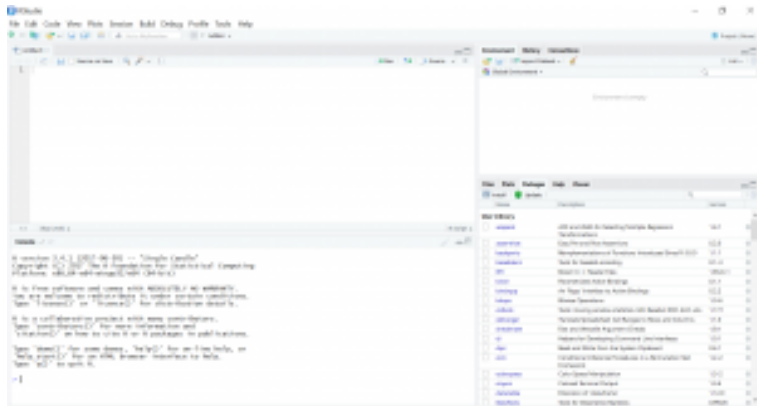
Looking at stuff here in SPSS before covering the concepts in class is a very real situation that people face in real life. They will go to a program like SPSS in the hopes that it is all they need for data analysis. But it will likely produce output that you don't understand if you don't have a basic education in statistics. If provided with output from SPSS (e.g., on an exam) you should be able to explain what the output means. For example, if given one of the tables shown above you should be able to determine what the standard deviation of a data set is and be able to use that number in a further calculation. It is also a good idea to do some calculations by hand when you first use SPSS for a procedure. If you can produce the same numbers as SPSS then you are sure you know what it is doing.

2.4 RStudio Lesson 1: Getting Started with RStudio

OSAMA BATAINEH

R is a free open source programming language widely used in statistical data analysis. It is very user friendly and it can be used in a variety of platforms such as MAC, Windows, LINUX, etc. R can be downloaded from <https://www.r-project.org/>. In most of the cases, R is being used with RStudio. RStudio is basically an additional interface which makes R more user friendly with a lot of additional features. A desktop version of RStudio can be downloaded (for free) at <https://rstudio.com/products/rstudio/download/>. More details are in the Front Matter of this book, in the section “[Statistical Software Used in this Book](#)”.

When you open RStudio, you will see this :



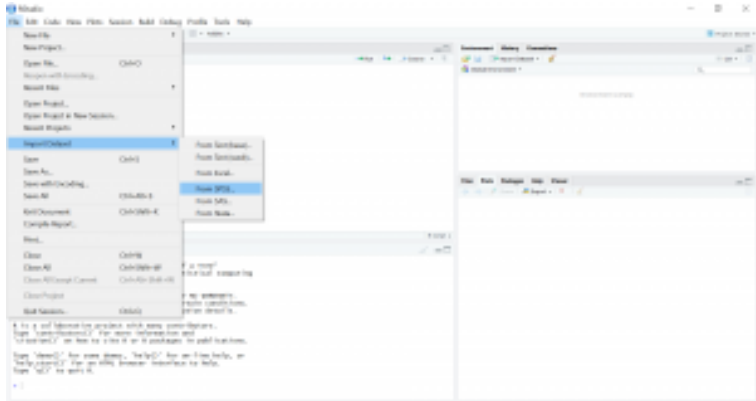
RStudio screenshot © the R Foundation.

Here you can see four windows. In the upper left window, you will write your R commands. The output is shown in the console in the

lower left. There are also ways to write commands in the console and subsequently get output in the following lines. The upper right window shows description of the dataset that you are working on such as number of variables, number of observations etc. Finally, the lower right window is showing different packages of R at the moment. We will talk more about the packages little later. The lower right window also has other purposes. For example, if you write commands to produce graphs, it will show here. It also shows the results of help command when you seek help regarding anything in R.

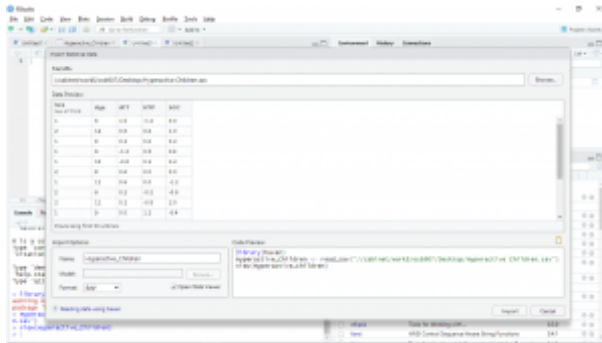
Now let's discuss a bit about how R functions. As the name suggests, R is a programming language. In other statistical packages such Stata, SPSS, EViews, Tableau etc., you have to choose the right options based on what results you want. However in R, rather than picking the right options and clicking on them, you have to write commands to get the desired output. To aid this process, there are different packages in R. These R packages were built for various purposes based on what analysis you want to do. Some of these packages are already built-in and installed. However, to use an existing installed package, you have to load it in every work secession before starting to use it. There are other packages of R which are not installed but available online. If you think these packages serve your purpose, then you have to install and load it before using it. Next time when you use these newly installed packages, you only have to load it in each work secession before starting to use it.

Now let's get started working with our dataset. First, download the dataset "HyperactiveChildren.sav" from the textbook [Data Sets](#). Then open RStudio and go to File > " title="Rendered by QuickLaTeX.com" height="11" width="12" style="vertical-align: 0px;"> Import Dataset > " title="Rendered by QuickLaTeX.com" height="11" width="12" style="vertical-align: 0px;"> From SPSS.



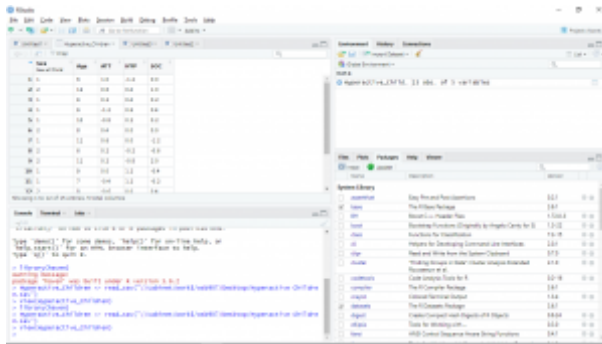
RStudio screenshot © the R Foundation.

After that, click the browse button in the pop-up window that will appear and select the dataset from the directory in which you have saved the data in your computer. Then click Import to insert the dataset in R. You can also do the same thing by manually executing the commands written in the Code Preview section by yourself.



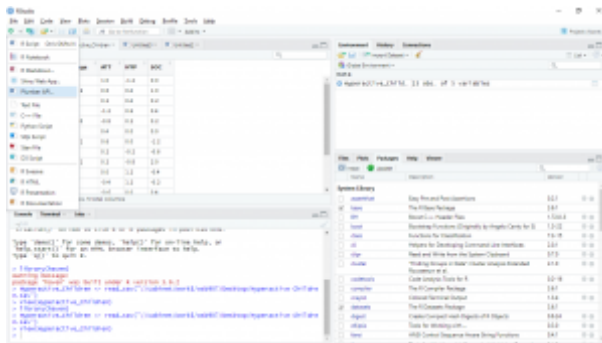
RStudio screenshot © the R Foundation.

After inserting the data, you will see this in RStudio.



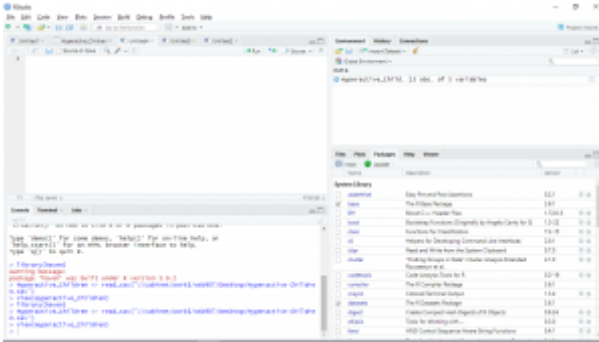
RStudio screenshot © the R Foundation.

Now we will write our commands in a new window in the upper left interface and save it in our desired folder in the computer to avoid rewriting the commands. This is known as R Script. To open a new R Script, click the arrow located just below and in between File and Edit and then select the first option R Script.



RStudio screenshot © the R Foundation.

After selecting the R script, a new R Script will open in a separate window named Untitled1. Save this script with a suitable name in your desired directory on your computer by clicking the old fashioned floppy disk icon. I have saved the script with the name Lesson 1_RScript in my computer.



RStudio
screenshot ©
the R
Foundation.

Before starting to work, let's get to know couple of things. Among these things, some are obligatory to know if anybody wants to work in R while others can make our life a lot easier and efficient. You can run the commands a lot quickly through the keyboard by clicking Ctrl and Enter. Other than the commands, you can also write notes in the R Script for your future references. To write anything other than the commands, just give a Hash (\#) sign in the beginning. Another thing you must know is that to work with any dataset in R after inserting it in the beginning, you have to attach it in the current work session to work with it further. To attach the Smoking dataset for the current work session, run the following command.

```
> attach(Hyperactive_Children)
```

Now let's start working with our dataset. Similar to other software packages, a variable has to be numeric to use it for statistical analysis. Thus any qualitative string variable needs to be transformed into numeric quantitative variable if we want to conduct analysis with it. In our dataset, the variable sex is a qualitative string variable with two categories male and female. Let's create a new variable sex_new which will take the value 1 for male and 2 for female. We can do it with the help of the function ifelse.

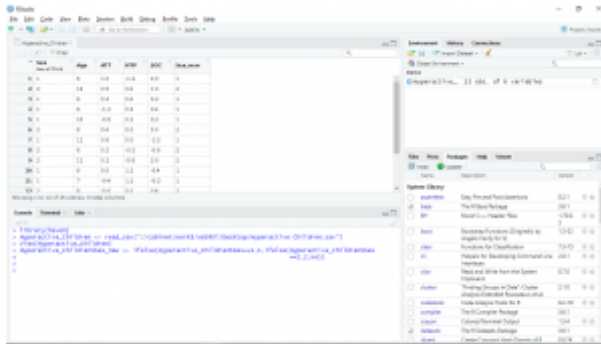
```
> Hyperactive_Children$Sex_new <-  
ifelse(Hyperactive_Children$Sex=="M",1,ifelse(Hyperactive_Children$
```

Sex

==2,2,NA))

A new variable Hyperactive_Children_new is being created. If we view the data by running the following command, we will see that at the end an additional column is being created named Hyperactive_Children which has numeric values of 1 and 2 only.

> View(Hyperactive_Children)



RStudio screenshot © the R Foundation.

Here a thing to be noted is the use of dollar sign (\$). The dollar sign basically calls the variable mentioned after it from the dataset mentioned prior to it or generates the new one mentioned after the sign to the dataset mentioned prior to it.

To get the frequency distribution of any variable in R similar in the way it's shown in many statistics textbook, you have to write codes for each of the columns separately. For example- to get the frequencies, cumulative frequencies, relative frequencies etc. – you have to write separate commands unless you are a program wizard and able to create an R package which will produce such table. First to get the frequency of a variable for its different categories, we can use the function `table`.

> Age.freq <- table(Hyperactive_Children\$Age)

> Age.freq

6 7 8 9 10 11 12

2 1 2 3 3 2 2

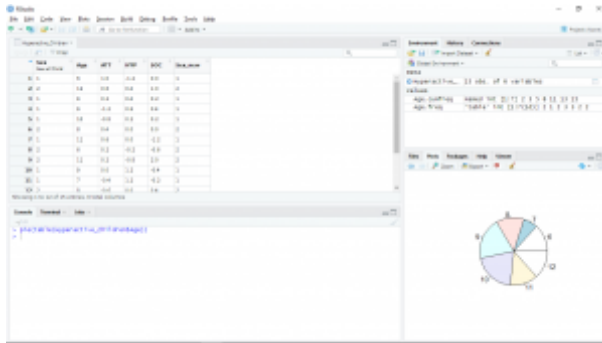
```
> cbind(Age.freq)
Age.freq
6  2
7  1
8  2
9  3
10 3
11 2
12 2
```

Here for a better presentation purpose, we have used another function *cbind* which basically shows the values of the variable and the number of observation it contains in a column. Then to get the cumulative frequencies, we can utilize the function *cumsum*.

```
> Age.cumfreq <- cumsum(table(Hyperactive_Children$Age))
> cbind(Age.freq, Age.cumfreq)
Age.freq Age.cumfreq
6  2      2
7  1      3
8  2      5
9  3      8
10 3     11
11 2     13
12 2     15
```

To produce pie chart, there is a specific function *pie* in R. However, since we need the pie chart of the frequency, we have to input `table(Hyperactive_Children$Age)` inside *pie*. The output will be shown in the lower right window.

```
> pie(table(Hyperactive_Children$Age))
```



RStudio
screenshot ©
the R
Foundation.

Now let's have a look at the descriptive statistics of this variable Educ. To calculate descriptive statistics, we need to install a specific package named *psych*. After installing and loading this package, we have to use the function *describe* which is a part of this new package. If any warning signs come, please ignore it.

```
> library(psych)
> describe(Hyperactive_Children$Age)
vars n mean sd median trimmed mad min max range skew
kurtosis se
X11 15 9.2 1.93 9 9.23 1.48 6 12 6 -0.21 -1.18
0.5
```

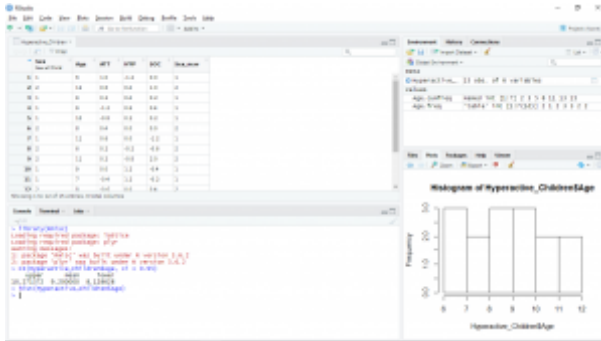
Similarly to get the confidence interval, you need another new package named *Rmisc*. While installing this package you have to keep the dependencies to true. After installing and loading this package, we have to use one of it's function *CI* to get the confidence interval. Also remember that in addition to inserting *Hyperactive_Children\$Age* in the domain of the function, you have to also specify the level of confidence interval. Ignore the warning signs here too.

```
> install.packages('Rmisc', dependencies = TRUE)
> library(Rmisc)
> CI(Hyperactive_Children$Age, ci = 0.95)
```


upper mean lower
 10.271372 9.200000 8.128628

Finally, to produce the histogram, stem and leaf display and boxplot, there are specific functions in R with their names. These functions are *hist*, *stem* and *boxplot* respectively.

```
> hist(Hyperactive_Children$Age)
```



RStudio screenshot © the R Foundation.

```
> stem(Hyperactive_Children$Age)
```

The decimal point is at the |

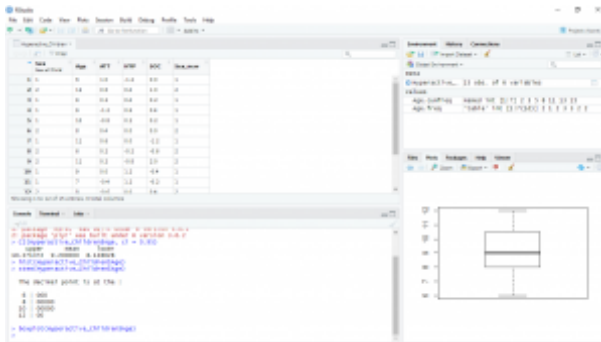
6 | 000

8 | 00000

10 | 00000

12 | 00

```
> boxplot(Hyperactive_Children$Age)
```



RStudio screenshot © the R Foundation.

3. DESCRIPTIVE STATISTICS: CENTRAL TENDENCY AND DISPERSION

3.1 Central Tendency: Mean, Median, Mode

Mean, median and mode are measures of the central tendency of the data. That is, as data are collected while sampling from a population, their values will tend to cluster around these measures. Let's define them one by one.

3.1.1 Mean

The mean is the average of the data. We distinguish between a sample mean and a population mean with the following symbols :

$$\bar{x} = \textit{sample mean}$$

$$\mu = \textit{population mean}$$

The formula for a sample mean is :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where n is the number of data points in the sample, the *sample size*. For a population, the formula is

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

where N is the size of the population.

Example 3.1 : Find the mean of the following data set :

84	12	27	15	40	18	33	33	14	4
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

To illustrate how the indexed symbols that represent the data in the formula work, they have been written below the data values. To get in the habit, let's organize our data as a table. We will need to do that for more complicated formulae and also that's how you need to enter data into SPSS, as a column of numbers :

x	label
84	x_1
12	x_2
27	x_3
15	x_4
40	x_5
18	x_6
33	x_7
33	x_8
14	x_9
4	x_{10}
Total = 280	

Since $n = 10$ we have $\bar{x} = \frac{\sum x_i}{n} = \frac{280}{10} = 28$.

□

Mean for grouped data : If you have a frequency table for a dataset but not the actual data, you can still compute the (approximate) mean of the dataset. This somewhat artificial situation for datasets will be a fundamental situation when we consider probability distributions. The formula for the mean of grouped data is

$$(3.1) \quad \bar{x} = \frac{\sum_{i=1}^G f_i x_{m_i}}{n}$$

where f_i is the frequency of group i , x_{m_i} is the class center of group i and n is the number of data points in the original dataset. Recall that $n = \sum f_i$ so we can write this formula as

$$\bar{x} = \frac{\sum_{i=1}^G f_i x_{m_i}}{\sum_{i=1}^G f_i}$$

which is a form that more closely matches with a generic weighted mean formula; the formula for the mean of grouped data is a special case of a more general weighted mean that we will look at next. The *class center* is literally the center of the class – the next example shows how to find it.

Example 3.2 : Find the mean of the dataset summarized in the following frequency table.

Class	Class Boundaries	Frequency, f_i	Midpoint, $x_{\{m_{\{i\}}\}}$	$f_i x_{m_i}$
1	5.5 - 10.5	1	8	8
2	10.5 - 15.5	2	13	26
3	15.5 - 20.5	3	18	54
4	20.5 - 25.5	5	23	115
5	25.5 - 30.5	4	28	112
6	30.5 - 35.5	3	33	99
7	35.5 - 40.5	2	38	76
sums		$n = \sum f_i = 20$		$\sum f_i x_{m_i} = 490$

Solution : The first step is to write down the formula to cue you to what quantities you need to compute :

$$\bar{x} = \frac{\sum_i f_i x_{m_i}}{n}$$

We need the sum in the numerator and the value for n in the

denominator. Get the numbers from the sums of the columns as shown in the frequency table :

$$\bar{x} = \frac{\sum_i f_i x_{m_i}}{n} = \frac{490}{20} = 24.5$$

□

Note that the grouped data formula gives an approximation of the mean of the original dataset in the following way. The exact mean is given by

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^G (\sum_{k=1}^{f_i} x_k)}{n}.$$

So the approximation is that

$$\sum_{k=1}^{f_i} x_k = f_i x_{m_i}$$

which would be exact only if all x_k in group i were equal to the class center x_{m_i} .

Generic Weighted Mean : The general formula for weighted mean is

$$(3.2) \quad \bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

where w_i is the *weight* for data point i . Weights can be assigned to data points for a variety of reasons. In the formula for grouped data, as a weighted mean, treats the class centers as data points and the group frequencies as weights. The next example weights grades.

Example 3.3 : In this example grades are weighted by credit units. The weights are as given in the table :

Course	Credit Units, w_i	Grade, x_i	$w_i x_i$
English	3	80	240
Psych	3	75	225
Biology	4	60	240
PhysEd	2	82	164
	$\sum w_i = 12$	$\sum x_i = 297$	$\sum w_i x_i = 869$

The formula for weighted mean is

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

so we need two sums. The double bars in the table above separate given data from columns added for calculation purposes. We will be using this convention with the double bars in other procedures to come. Using the sums for the table we get

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{869}{12} = 72.4$$

Note, that the unweighted mean for these data is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{297}{4} = 74.3$$

which is, of course, different from the weighted sum.



3.1.2 Median

The symbol we use for median is MD and it is the midpoint of the data set with the data put in order. We illustrate this with a couple of examples :

- If there are an odd number of data points, MD is the middle number.

Given data in order: 180 186 191 201 209 219 220

$$MD = 201$$



- If there are an even number of data points, MD is the average of the two middle points :

Given data in order: 656 684 702 764 856 1132 1133 1303

$$MD = \frac{764+856}{2} = 810$$



In these examples, the tedious work of putting the data in order from smallest to largest was done for us. With a random bunch of numbers, the work of finding the median is mostly putting the data in order.

3.1.3 Mode

In a given dataset the mode is the data value that occurs the most. Note that :

- it may be there is no mode.
- there may be more than one mode.

Example 3.4 : In the dataset

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

8 occurs 5 times, more than any other number. So the *mode* is 8.



Example 3.5 : The dataset

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 72

has no mode. Do not say that the mode is zero. Zero is not in the dataset.



Example 3.6 : The dataset

15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26

has two modes: 18 and 24. This data set is *bimodal*.

The concept of mode really makes more sense for frequency table/histogram data.



Example 3.7 : The mode of the following frequency table data is the class with the highest frequency.

Class	Class Boundaries	Freq
1	5.5 - 10.5	1
2	10.5 - 15.5	2
3	15.5 - 20.5	3
4	20.5 - 25.5	5 (Modal Class)
5	25.5 - 30.5	4
6	30.5 - 35.5	3
7	35.5 - 40.5	2



3.1.4

Midrange

The midrange, which we'll denote symbolically by MR, is defined simply by

$$MR = \frac{H + L}{2}$$

where H and L are the high and low data values.

Example 3.8 : Given the following data : 2, 3, 6, 8, 4, 1. We have

$$MR = \frac{8 + 1}{2} = 4.5$$



3.1.5 Mean, Median and Mode in Histograms: Skewness

If the shape of the histogram of a dataset is not too bizarre¹ (e.g. unimodal) then we may determine the *skewness* of the dataset's histogram (which would be a probability distribution of the data represented a population and not a sample) by comparing the mean or median to the mode. (Always compare something to the mode, no reliable information comes from comparing the median and mean.) If you have SPSS output with the skewness number calculated (we will see the formula for skewness later) then a left skewed distribution will have a negative skewness value, a symmetric distribution will have a skewness of 0 and, a right skewed distribution will have a positive skewness value.

1. For the purposes of deciding the skewness of a dataset in assignments and exams, you can assume that the histogram shape is not too bizarre.

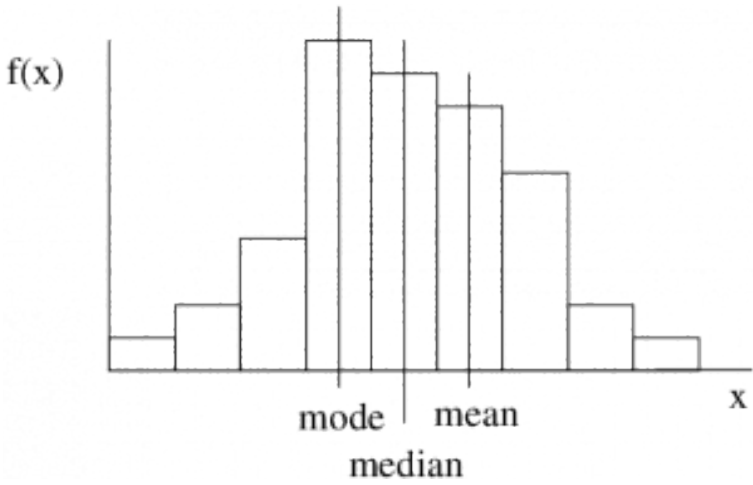


Figure 3.1: A right skewed histogram (or distribution) generally has the mean and median to the right, or positive side of the mode. The tail of the histogram stretches to the right or positive side.

Symmetric distribution

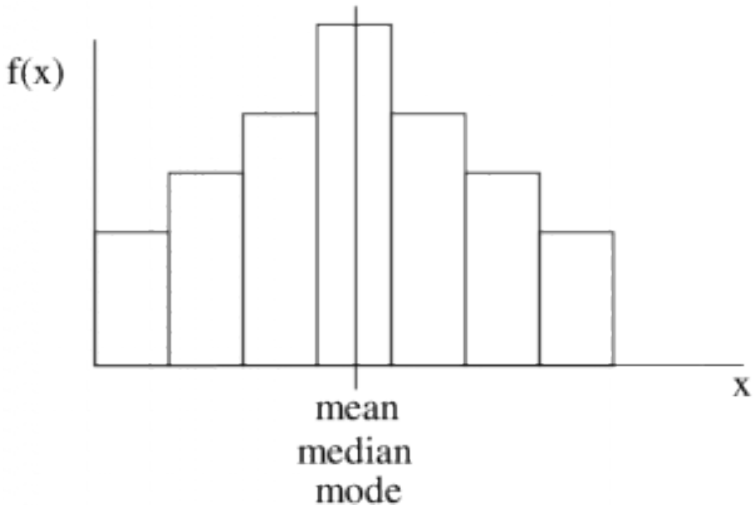


Figure 3.2: A symmetric distribution (histogram) has the mean, median and mode all in the same place. Its shape is symmetric.

Negatively skewed or left skewed histograms

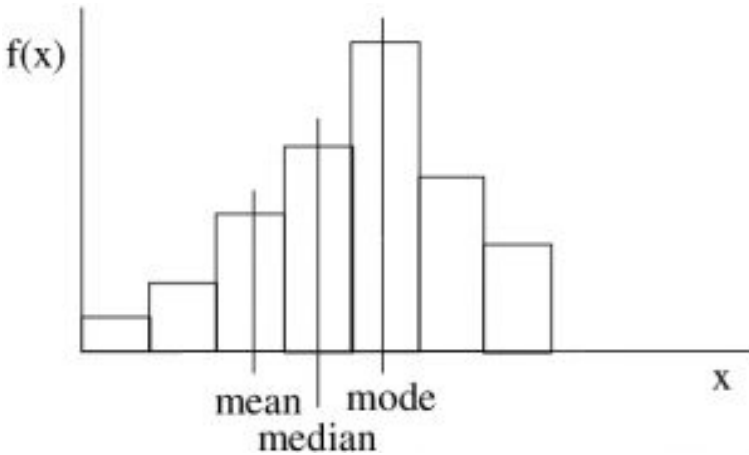


Figure 3.3: A left skewed histogram (or distribution) generally has the mean and median to the left, or negative side of the mode. The tail of the histogram stretches to the left or negative side.

3.1.6 Mean, Median and Mode in Distributions: Geometric Aspects

To understand the geometrical aspects of histograms we make the abstraction of letting the class widths shrink to zero so that the histogram curve becomes smooth. So let's consider the mode, median and mean in turn.

Mode

The mode is the x value where the frequency $f(x)$ is maximum, see Figure 3.4. More accurately the mode is a “local maximum” of

the histogram² (so if there are multiple modes, they don't all have to have the same maximum value).

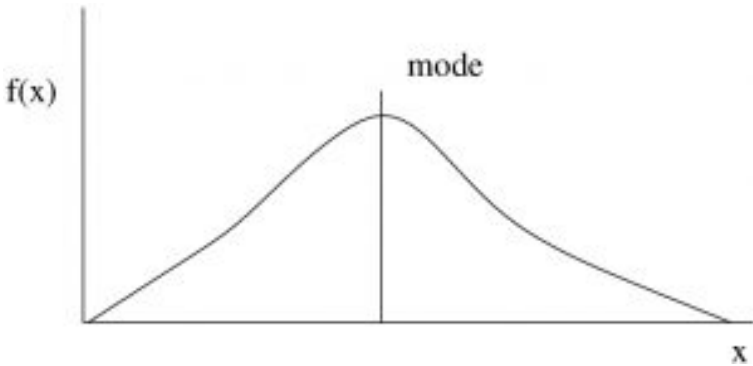


Figure 3.4: The mode is the maximum of the histogram (distribution).

Median

The area under the curve is equal on either side of the median. In Figure 3.5 each area A is the same. For relative frequencies (and so for probabilities) the total area under the curve is one. So the area on each side of the median is half. The median represents the 50/50 probability point; it is equally probable that x is below the median as above it.

2. **In calculus terms, local maximums and minimums (and inflexion points) are where the derivative equals zero, $\frac{df}{dx} = 0$.

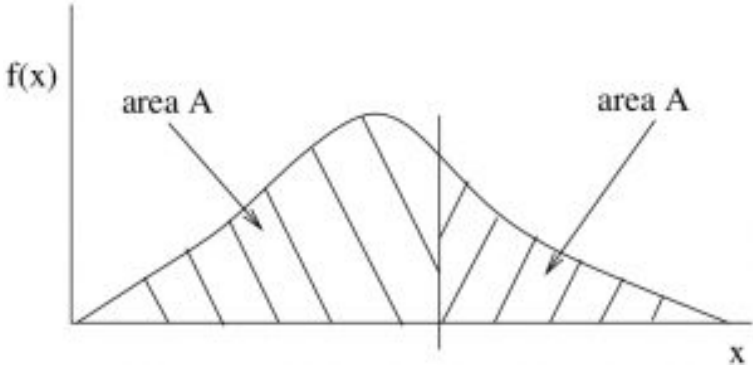


Figure 3.5: The median divides the area under the histogram into two equal areas A .

Mean

The mean is the balance point of the histogram/distribution as shown in Figure 3.6.

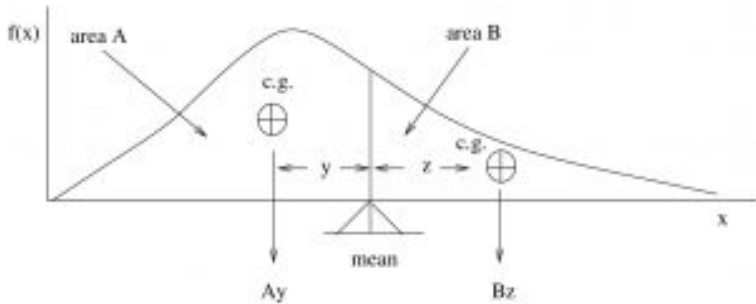


Figure 3.6: The mean is the balance point of the histogram. It is where the “first moments” of the area of the histogram balance. Here the moments are Ay and Bz balance. $Ay = Bz$.

****A proof that the mean is the center of gravity of a histogram:**

In physics, a moment is weight \times moment arm :

$$M = Wx$$

where M is moment, W is weight and x is the moment arm (a distance).

Say we have two kids, kid1 and kid2 on a teeter-totter (Figure 3.7).

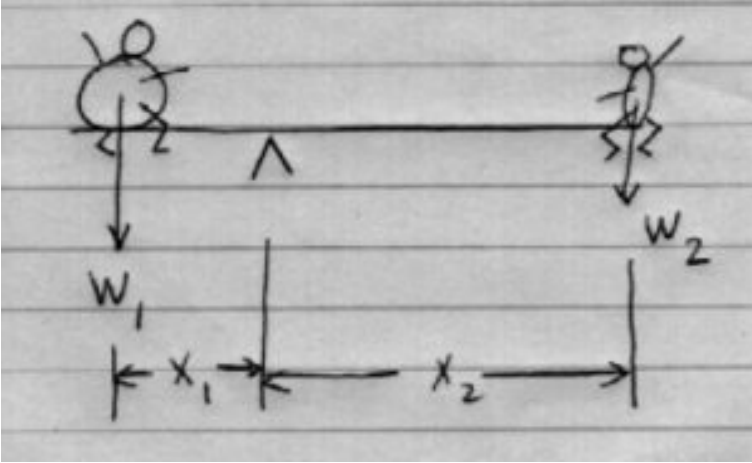


Figure 3.7

Kid1 with weight W_1 is heavy, kid2 with weight W_2 is light.

To balance the teeter-totter we must have

$$W_1x_1 = W_2x_2.$$

The moment arm, x_1 , of the heavier kid must be smaller than the moment arm, x_2 , of the lighter kid if they are to balance.

So now let's define the center of gravity. If you have a bunch of weights W_i with corresponding moment arms x_i then the center of gravity (c of g) is the moment arm x_g (distance) that satisfies :

$$\sum W_i x_i = W_t x_g$$

where $W_t = \sum W_i$ is the total weight.

With histograms, instead of weight W we have area A . You can think of area as having a weight. (Think of cutting out a piece of the

blackboard with a jigsaw after you draw a histogram on it.) So for a histogram (see Figure 3.8):

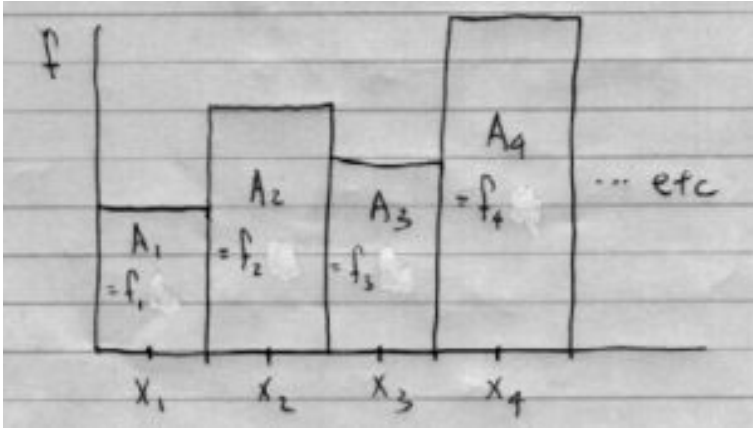


Figure 3.8

(We assume, for simplicity but “without loss of generality”, that x_i are integers and also the classes. This is the case for discrete probability distributions as we’ll see.) So, for the c of g,

$$\sum W_i x_i = W_t x_g$$

translates to

$$\begin{aligned} \sum A_i x_i &= A_t x_g \\ \sum f_i x_i &= (\sum f_i) x_g \\ \sum f_i x_i &= n x_g \\ x_g &= \frac{\sum f_i x_i}{n} \end{aligned}$$

where we have used $A_i = f_i$ because the class widths are one, so

$$x_g = \bar{x} = \frac{\sum f_i x_i}{n}.$$

Because our “weight” is area, \bar{x} is technically called the “1st

moment of area". (Variance, covered next, is the "2nd moment of area about the mean".)

□

3.2 Dispersion: Variance and Standard Deviation

Variance, and its square root standard deviation, measure how “wide” or “spread out” a data distribution is. We begin by using the formula definitions; they are slightly different for populations and samples.

1. Population Formulae :

Variance :

$$(3.3) \quad \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

where N is the size of the population, μ is the mean of the population and x_i is an individual value from the population.

Standard Deviation :

$$\sigma = \sqrt{\sigma^2}$$

The standard deviation, σ , is a population parameter, we will learn about how to make inferences about population parameters using statistics from samples.

2. Sample Formulae :

Variance :

$$(3.4) \quad s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

where n = sample size (number of data points), $n - 1$ = degrees of freedom for the given sample, \bar{x} and x_i is a data value.

Standard Deviation :

$$s = \sqrt{s^2}$$

Equations (3.3) and (3.4) are the definitions of variance as the second moment about the mean; you need to determine the means (μ or \bar{x}) before you can compute variance with those formulae. They

are algebraically equivalent to a “short cut” formula that allow you to compute the variance directly from sums and sums of squares of the data without computing the mean first. For the sample standard deviation (the useful one) the short cut formula is

$$(3.5) \quad s^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2}{n - 1}$$

At this point you should figure out how to compute \bar{x} , s and σ on your calculator for a given set of data.

Fact (not proved here) : The sample standard deviation s is the “optimal unbiased estimate” of the population standard deviation σ . s is a statistic”, the best statistic it turns out, that is used to estimate the population parameter σ . It is the $n - 1$ in the denominator that makes s the optimal unbiased estimator of σ . We won’t prove that here but we will try and build up a little intuition about what that should be so – why dividing by $n - 1$ should be better than dividing by n . ($n - 1$ is known as the degrees of freedom of the estimator s). First notice that you can’t guess or estimate a value for σ (i.e. compute s) with only one data point. There is no spread of values in a data set of one point! This is part of the reason why the degrees of freedom is $n - 1$ and not n . A more direct reason is that you need to remove one piece of information (the mean) from your sample before you can guess σ (compute s).

Coefficient of Variation

The coefficient of variation, CVar, is a “normalized” measure of data spread. It will not be useful for any inferential statistics that we will be doing. It is a pure descriptive statistic. As such it can be useful as a dependent variable but we treat it here as a descriptive statistic that combines the mean and standard deviation. The definition is :

$$\begin{aligned} \text{CVar} &= \frac{s}{\bar{x}} \times 100\% && \text{samples} \\ \text{CVar} &= \frac{\sigma}{\mu} \times 100\% && \text{population} \end{aligned}$$

Example 3.9 : In this example we take the data given in the following table as representing the whole population of size $N = 6$. So we use the formula of Equation (3.3) which requires us to sum $(x_i - \mu)^2$.

x_i	$(x_i - \mu)^2$
10	$(10 - 35)^2$
60	$(60 - 35)^2$
50	$(50 - 35)^2$
30	$(30 - 35)^2$
40	$(40 - 35)^2$
20	$(20 - 35)^2$
$\sum x_i = 210$	$\sum (x_i - \mu)^2 = 1750$

Using the sum in the first column we compute the mean :

$$\mu = \frac{\sum x_i}{N} = \frac{210}{6} = 35.$$

Then with that mean we compute the quantities in the second (calculation) column above and sum them. And then we may compute the variance :

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{1750}{6} = 291.7$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{291.7} = 17.1.$$

Finally, because we can, we compute the coefficient of variation:

$$\text{CVar} = \frac{\sigma}{\mu} \times 100\% = \frac{17.1}{35} \times 100\% = 48.9\%.$$

□

Example 3.10 : In this example, we have a *sample*. This is the usual circumstance under which we would compute variance and sample standard deviation. We can use either Equation (3.4) or (3.5). Using Equation (3.4) follows the sample procedure that is given in Example 3.9 and we'll leave that as an exercise. Below we'll apply the short-cut formula and see how s may be computed without knowing \bar{x} . The dataset is given in the table below in the column to the left of the double line. The columns to the right of the double line are, as usual, our calculation columns. The size of the sample is $n = 6$.

x_i	$(x_i - \bar{x})^2$	x_i^2
11.2		$11.2^2 = 125.44$
11.9		$11.9^2 = 141.61$
12.0	exercise	$12.0^2 = 144$
12.8		$12.8^2 = 163.84$
13.4		$13.4^2 = 179.56$
14.3		$14.3^2 = 204.49$
$\sum x_i = 75.6$		$\sum x_i^2 = 958.94$

To find s compute

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n - 1} = \frac{958.94 - \frac{(75.6)^2}{6}}{6 - 1} = \frac{958.94 - 952.56}{5} = 1.28$$

So

$$s = \sqrt{s^2} = \sqrt{1.28} = 1.13.$$

Note that s^2 is never negative! If it were then you couldn't take the square root to find s . Also not that we have not yet determined the mean. We can do that now:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{75.6}{6} = 12.60.$$

And with the mean we can then compute

$$\text{CVar} = \frac{s}{\bar{x}} = \frac{1.13}{12.6} \times 100\% = 9.0\%$$

□

Grouped Sample Formula for Variance

As with the mean, we can compute an approximation of the data variance from frequency table, histogram, data. And again this computation is precise for probability distributions with class widths of one. The grouped sample formula for variance is

$$(3.6) \quad s^2 = \frac{\sum_{i=1}^G (f_i \cdot x_{m_i}^2) - \left[\frac{(\sum_{i=1}^G f_i \cdot x_{m_i})^2}{n} \right]}{n - 1}$$

where G is the number of groups or classes, x_{m_i} is the class center of group i , f_i is the frequency of group i and

$$n = \sum_{i=1}^G f_i$$

is the sample size. Equation (3.6) the short-cut version of the formula. We can also write

$$s^2 = \frac{\sum_{i=1}^G f_i (x_{m_i} - \mu)^2}{n - 1}$$

or if we are dealing with a population, and the class width is one so that the class center $X_{m_i} = X_i$,

$$\sigma^2 = \frac{\sum_{i=1}^G f_i (X_{m_i} - \mu)^2}{N}$$

which will be useful when we talk about probability distributions. In fact, let's look ahead a bit and make the frequentist definition for the probability for X_i as $P(X_i) = f_i/N$ (which is the relative frequency of class i) so that

$$(3.7) \quad \sigma^2 = \sum_{i=1}^G P(X_i)(X_i - \mu)^2.$$

If we make the same substitution $P(X_i) = f_i/N$ in the grouped mean formula, Equation (3.1) with population items X and N in place of the sample items x and n , then it becomes

$$(3.8) \quad \mu = \sum_{i=1}^G P(X_i)X_i.$$

More on probability distributions later, for now let's see how we use Equation (3.6) for frequency table data.

Example 3.11 : Given the frequency table data to the left of the double dividing line in the table below, compute the variance and standard deviation of the data using the grouped data formula.

Class	Class Boundaries	Freq. f_i	Class Centre x_{m_i}	$f_i \cdot x_{m_i}$	$x_{m_i}^2$	$f_i \cdot x_{m_i}^2$
1	5.5 - 10.5	1	8	$1 \cdot 8 = 8$	$8^2 = 64$	$1 \cdot 64 = 64$
2	10.5 - 15.5	2	13	$2 \cdot 13 = 26$	$13^2 = 169$	$2 \cdot 169 = 338$
3	15.5 - 20.5	3	18	$3 \cdot 18 = 54$	$18^2 = 324$	$3 \cdot 324 = 972$
4	20.5 - 25.5	5	23	$5 \cdot 23 = 115$	$23^2 = 529$	$5 \cdot 529 = 2645$
5	25.5 - 30.5	4	28	$4 \cdot 28 = 112$	$28^2 = 784$	$4 \cdot 784 = 3136$
6	30.5 - 35.5	3	33	$3 \cdot 33 = 99$	$33^2 = 1089$	$3 \cdot 1089 = 3267$
7	35.5 - 40.5	2	38	$2 \cdot 38 = 76$	$38^2 = 1444$	$2 \cdot 1444 = 2888$
		$\sum f = 20$		$\sum fx_m = 490$		$\sum fx_m^2 = 13310$

The formula

$$s^2 = \frac{\sum (fx_m^2) - \left[\frac{(\sum fx_m)^2}{n} \right]}{n - 1}$$

tells us that we need the sums of fx_m^2 and fx_m after we compute the class centres x_m and their squares x_m^2 - these calculations we do in the columns added to the right of the double bar in the table above. With the sums we compute

$$s^2 = \frac{\sum f_i x_{m_i}^2 - \left[\frac{(\sum f_i x_{m_i})^2}{n} \right]}{n - 1} = \frac{13310 - \left[\frac{490^2}{20} \right]}{20 - 1} = \frac{13310 - 12005}{19} = 68.7.$$

So

$$s = \sqrt{s^2} = \sqrt{68.7} = 8.3.$$

The mean, from one of the sums already finished is

$$\bar{x} = \frac{\sum f_i x_{m_i}}{n} = \frac{490}{20} = 24.5$$

and the coefficient of variation is

$$\text{CVar} = \frac{s}{\bar{x}} \times 100\% = \frac{8.3}{24.5} \times 100\% = 33.9\%$$

□

Now is a good time to figure out how to compute \bar{x} and s (and σ) on your calculators.

3.3 z-score / z-transformation

The z -score is the result of transformation of data that converts a dataset of x values, $\{x_i\}$, that has a mean of \bar{x} and standard deviation s to a set of z values $\{z_i\}$ that has a mean of $\bar{z} = 0$ and a standard deviation of $s_z = 1$. It will be very useful when we need to compute probabilities associated with normal distributions. The z -transformation is defined by

$$z = \frac{x - \bar{x}}{s} \quad (\text{sample})$$

$$z = \frac{x - \mu}{\sigma} \quad (\text{population})$$

Example 3.12 : Find the z -scores of the data given in the left column of the table below.

Data x_i	x_i^2	z -score, z_i
18	324	$(18-9.9)/6.2 = 1.3$
15	225	$(15-9.9)/6.2 = 0.8$
12	144	$(12-9.9)/6.2 = 0.3$
6	36	$(6-9.9)/6.2 = -0.6$
8	64	$(8-9.9)/6.2 = -0.3$
2	4	$(2-9.9)/6.2 = -1.3$
3	9	$(3-9.9)/6.2 = -1.1$
5	25	$(5-9.5)/6.2 = -0.8$
20	400	$(20-9.5)/6.2 = -1.7$
10	100	$(10-9.5)/6.2 = 0.1$
$\sum x_i = 99$	$\sum x_i^2 = 1331$	

The dataset size is $n = 10$. You need to compute the z -score for

each data value separately. To do the calculation, both \bar{x} and s are needed. So in addition to the sum of the data, $\sum x$, we also need the sum of the x^2 values. The work of getting those sums is shown in the table above. With the x and x^2 sums we get

$$\bar{x} = \frac{\sum x_i}{n} = \frac{99}{10} = 9.9$$

and

$$s^2 = \frac{\sum x_i^2 - \left[\frac{(\sum x_i)^2}{n}\right]}{n-1} = \frac{1331 - \left[\frac{99^2}{10}\right]}{9} = \frac{1331 - 980.1}{9} = 39.0$$

$$\text{and } s = \sqrt{39} = 6.2.$$

Using these values for \bar{x} and s in the third column of the table above, compute the z -scores as shown. If we had computed the z -scores more accurately, they would add up to zero, $\sum z_i = 0$ (the mean of the z -scores is zero.)

□

3.4 SPSS Lesson 2: Combining variables and recoding

Frequently data collection results in a collection of many variables. This happens, for example, with tests or surveys where people answer questions on a 5 or 7 point *Lickert scale* where questions range from, say, “strongly agree” to “somewhat agree” to . . . to “strongly disagree”. A bunch of those questions may refer to, say, happiness and adding up the scores, perhaps averaging them, will lead to a single variable, one dependent variable, that becomes our measurement of happiness. This gives us not only a univariate variable that we can subject to a statistical test but likely gives us a stronger and more reliable measurement of happiness. A problem with combining variables in this way arises if the response “1” for “strongly agree” means happiness for one question (e.g. “I wake up happy”) and sadness in another question (e.g. “I go to bed sad”). In such a situation some of the variables will need to be reverse-scaled or *recoded* before they can be added. Let’s see how to combine and recode variables in SPSS.

Open the file “Caregiver.sav” from the textbook [Data Sets](#). This dataset is about the different attributes of diamonds such as its color, price, carat, cutting quality etc. Here one of the variables is `cut_new` which basically represents the cutting quality of diamond and takes values from 1 to 5 depending on the cutting quality with 5 being the best quality. Now let’s assume that we need to reverse scale this variable to use it in other calculations in a meaningful manner. To recode `cut_new` first open the Transform → Recode in Same Variables... menu :

The screenshot shows the SPSS Data Editor with a recoding dialog box open for the variable 'cut'. The dialog box has 'cut' selected in the 'Variable' list. The 'Old and New Values' section is empty. The 'Recoding into Different Variables...' option is selected. The 'New Variable Name' field is empty. The 'OK' button is highlighted.

	cut	Coloration	Clarity	Clarity2	Clarity3	Clarity4	Clarity5
1	41	250	12	4	3	140	50
2	45	241	6	10	2	151	24
3	41	320	4	11	10	152	56
4	35	290	23	19	12	144	70
5	32	140	22	20	22	120	65
6	29	132	23	51	24	125	34
7	45	76	32	23	11	121	35
8	48	222	31	15	9	129	45
9	44	249	25	19	8	115	58
10	47	810	18	19	8	110	80
11	50	440	18	49	14	96	76
12	28	430	11	23	7	80	33
13	43	240	12	22	9	172	20
14	41	230	34	21	13	171	21
15	55	205	48	20	12	130	25
16	61	140	58	29	18	131	59
17	62	135	23	31	19	137	76

SPSS screenshot © International Business Machines Corporation.

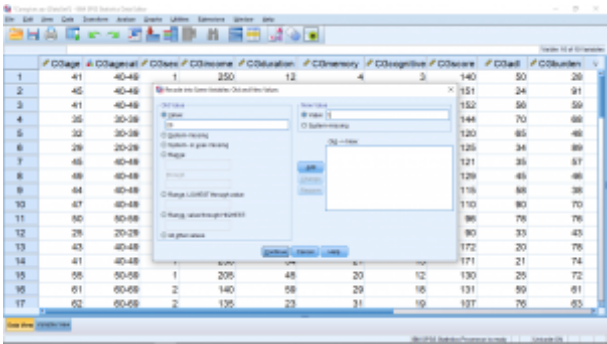
You can choose the Recode into Different Variables... if you want to, instead. That choice will lead to the creation of a new variable that you would use in place of cut_new for your analysis. With our choice of Recode in Same Variables... we will overwrite the old values of cut_new with new ones. (This is a danger if you make a mistake.) Our job is now to map 1 to 5, 2 to 4, 3 to 3, 4 to 2 and 5 to 1, recoding the variable. First move the cut_new variable over in the pop up menu :

The screenshot shows the SPSS Data Editor with a recoding dialog box open for the variable 'cut'. The dialog box has 'cut' selected in the 'Variable' list. The 'Old and New Values' section is empty. The 'Recoding into Different Variables...' option is selected. The 'New Variable Name' field is empty. The 'OK' button is highlighted.

	cut	Coloration	Clarity	Clarity2	Clarity3	Clarity4	Clarity5
1	41	250	12	4	3	140	50
2	45	241	6	10	2	151	24
3	41	320	4	11	10	152	56
4	35	290	23	19	12	144	70
5	32	140	22	20	22	120	65
6	29	132	23	51	24	125	34
7	45	76	32	23	11	121	35
8	48	222	31	15	9	129	45
9	44	249	25	19	8	115	58
10	47	810	18	19	8	110	80
11	50	440	18	49	14	96	76
12	28	430	11	23	7	80	33
13	43	240	12	22	9	172	20
14	41	230	34	21	13	171	21
15	55	205	48	20	12	130	25
16	61	140	58	29	18	131	59
17	62	135	23	31	19	137	76

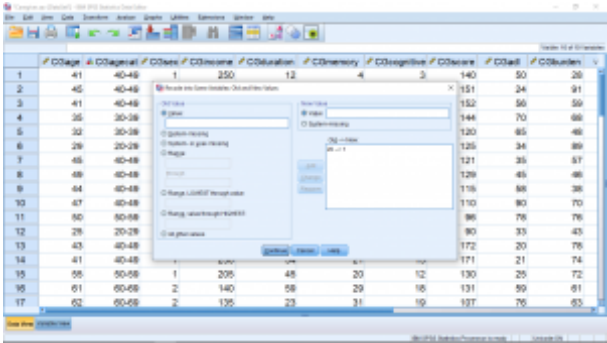
SPSS screenshot © International Business Machines Corporation.

then hit the Old and New Values.. button that will bring up a new pop up menu. Next enter 1 under Old Value and 5 in New Value :



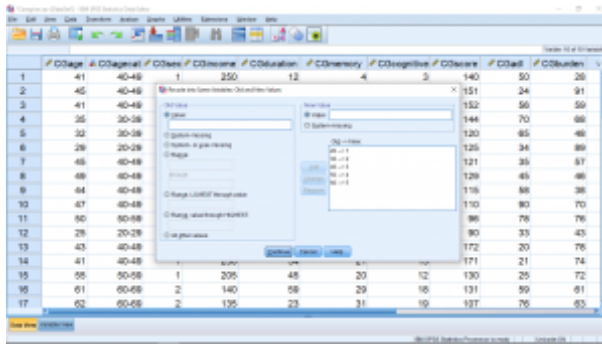
SPSS screenshot © International Business Machines Corporation.

then hit Add :



SPSS screenshot © International Business Machines Corporation.

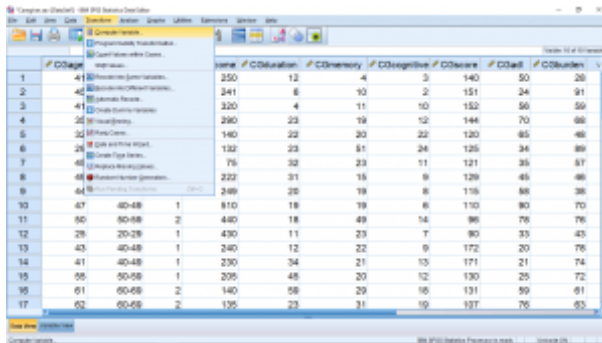
Continue this way to complete the recoding list :



SPSS screenshot © International Business Machines Corporation.

Hit Continue, then OK. The variable cut_new will now have the new values in the Data View window.

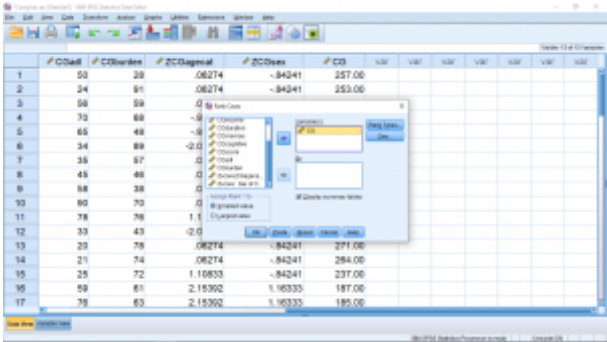
Now suppose we want to add multiple variables to create a new variable. Let's open the dataset Caregiver from the course website. This dataset is regarding the test scores of students from diverse background in UK. Here we will add the test scores of read, write, math and science to create a new variable totalscore. Pick the Transform → Compute Variable... menu :



SPSS screenshot © International Business Machines Corporation.

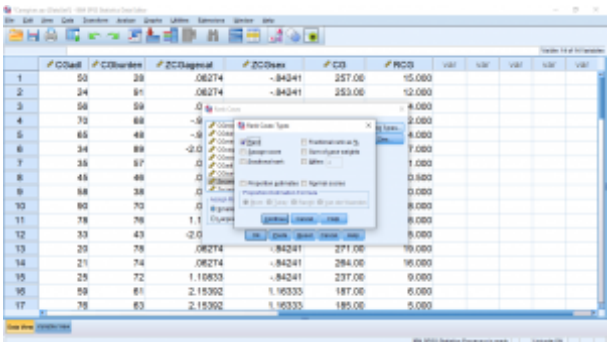
This will bring up a menu which is essentially the calculator feature of SPSS :

data in order from smallest to largest. This is tedious but SPSS can do it with a couple of mouse clicks (yes, yes SPSS can compute the median directly but whatever). There are a couple of approaches in SPSS to ordering, or ranking, data. One is to compute the rank, that is, give rank 1 to the lowest value, 2 to the next lowest up to n for the highest value. Pick Transform → Rank Cases and move totalscore into the Variable(s) box :



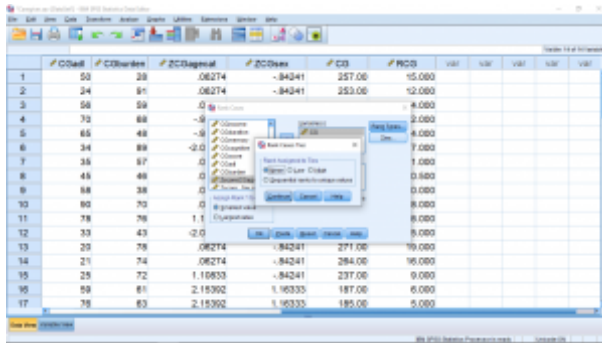
SPSS screenshot © International Business Machines Corporation.

This is a new menu for us, so let's take a look at the submenus. First, the Rank Types menu :



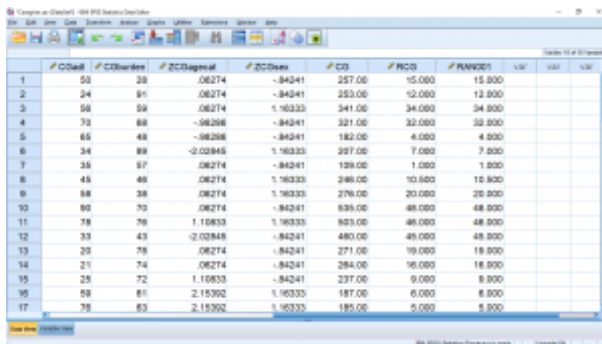
SPSS screenshot © International Business Machines Corporation.

Pretty fancy. Much too advanced for our use, so let's leave that one be, hit Continue. Next look at Ties...



SPSS screenshot © International Business Machines Corporation.

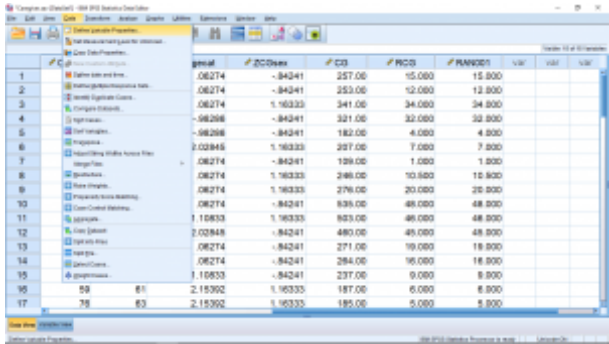
We will assign the average (mean) rank to ties in our classes. To understand the ties options, think of two people in a race who cross the finish line at exactly the same time, a tie. With the mean rank, they both come in 1.5 place. With lowest, they both come in 1st place, with highest, they both come in 2nd place. Hit Continue, the OK and a new variable Rtotal score will be formed in the Data View menu :



SPSS screenshot © International Business Machines Corporation.

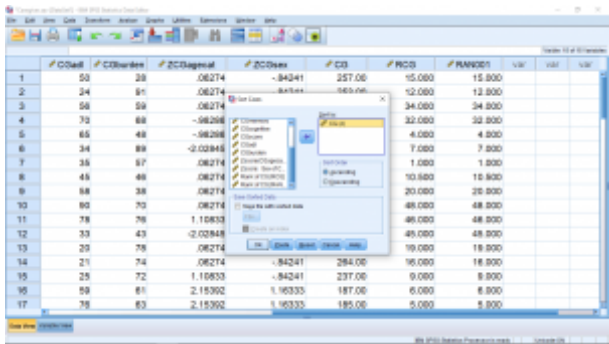
Here the variable RCG ranks the total score of the students. But it's very difficult from this data view to identify which students' rank the highest or lowest, let alone who falls in the middle to find the median. This is not quite what we are after to easily get the median.

Ranking will become useful on Psy 234 (in Chapter 16), but it's not that useful for us now. What we need, is to shuffle the numbers around from lowest to highest (of course we can do that directly). To shuffle pick Data → Sort Cases :



SPSS screenshot © International Business Machines Corporation.

which brings up, after moving over the RCG total score variable :



SPSS screenshot © International Business Machines Corporation.

Keep the ascending button selected (sort from lowest to highest), then hit OK to sort the file :

	CQual	CQualities	ZCQual	ZCQuals	CQ	PCQ	RANKED	v.1	v.2	v.3
1	35	57	.08274	-.84241	159.00	1.000	1.000			
2	44	82	1.15833	-.84241	142.00	2.000	2.000			
3	82	65	.08274	1.16333	150.00	3.000	3.000			
4	65	48	-.98266	-.84241	182.00	4.000	4.000			
5	78	63	2.15900	1.16333	185.00	5.000	5.000			
6	58	81	2.15900	1.16333	187.00	6.000	6.000			
7	34	89	-2.02845	1.16333	207.00	7.000	7.000			
8	45	103	-.98266	1.16333	236.00	8.000	8.000			
9	25	72	1.15833	-.84241	237.00	9.000	9.000			
10	45	48	.08274	1.16333	246.00	10.000	10.000			
11	81	71	.08274	-.84241	246.00	10.000	10.000			
12	24	81	.08274	-.84241	253.00	12.000	12.000			
13	80	57	.08274	1.16333	254.00	13.000	13.000			
14	47	64	.08274	-.84241	254.00	13.000	13.000			
15	50	28	.08274	-.84241	257.00	15.000	15.000			
16	21	74	.08274	-.84241	254.00	16.000	16.000			
17	65	54	-.98266	1.16333	295.00	17.000	17.000			

SPSS
screenshot ©
International
Business
Machines
Corporation.

Everything is sorted now. (Note how useful the id variable is now. If that wasn't there, we'd lose track of who's data was what.) Now if we scroll down, we will find that the middle two total test scores are both 210. Thus the median of total test score is 210.

As a final analysis of the Caregiver data, suppose we wanted some descriptive statistics for the male students separate from the female students. To do this we use the "split file" feature of SPSS. Select Data → Split File to get

SPSS
screenshot ©
International
Business
Machines
Corporation.

where the gender variable has been moved into the "Groups Based on" box – you will need to click on the "Organize output by groups" button also. We'll also leave the "Sort the file by grouping variables"

(gender in this case), this will shuffle the file yet again, putting all the males and females together. So, when you hit OK the result is

	CSAge	CSAgecat	CSGen	CSGenrec	CSEducation	CSMemory	CSIntelligence	CSStress	CSSkill	CSBundling
19	35	30-39	1	290	23	19	12	144	70	68
20	33	30-39	1	320	26	21	7	121	39	68
21	37	30-39	1	320	28	41	26	140	50	47
22	40	40-49	1	390	26	55	9	120	30	37
23	42	40-49	1	390	28	37	25	179	70	52
24	29	20-29	1	370	28	69	30	152	57	37
25	25	20-29	1	430	11	23	7	80	33	43
26	31	30-39	1	490	38	24	9	101	41	71
27	47	40-49	1	810	18	19	6	110	80	70
28	39	30-39	1	890	27	40	26	144	81	80
29	44	40-49	1	810	28	27	1	134	49	52
30	40	40-49	2	95	28	32	23	199	82	69
31	42	40-49	2	135	23	31	19	107	76	63
32	51	50-59	2	140	59	29	16	131	59	61
33	29	20-29	2	132	23	81	24	129	34	89
34	30	30-39	2	175	51	41	20	127	45	133
35	49	40-49	2	222	31	15	9	129	45	46

SPSS screenshot © International Business Machines Corporation.

Now the file is sorted into Male and Female (the 1-A button at the top has been pressed). Also note that “Split by gender” appears on the lower right corner of the Data View window. Now let’s do a simple descriptive statistics analysis of the total score variable. The output looks like :

Size of Caregiver # 1

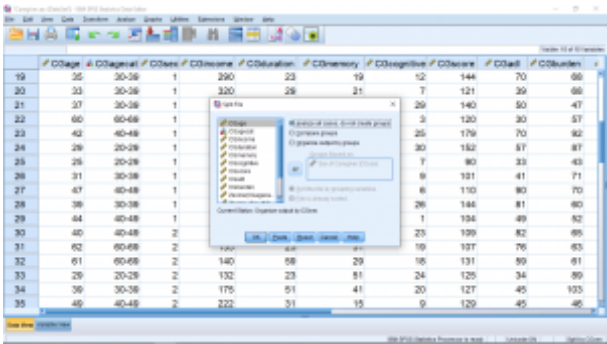
Statistic	Mean	Std. Deviation
Descriptive Statistics	28	1.39
Minimum	1	
Maximum	2	

Size of Caregiver # 2

Statistic	Mean	Std. Deviation
Descriptive Statistics	24	1.66
Minimum	1	
Maximum	2	

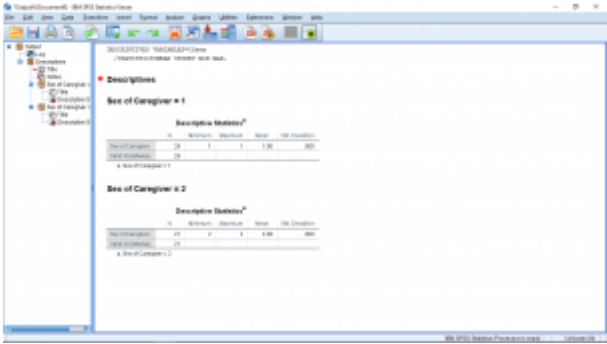
SPSS screenshot © International Business Machines Corporation.

To unsplit the file, go back to Data → Split File and hit the “Analyze all cases, do not create groups” button. This will remove the “Split” message from the lower right corner and when the descriptive statistics is run again, you will get :



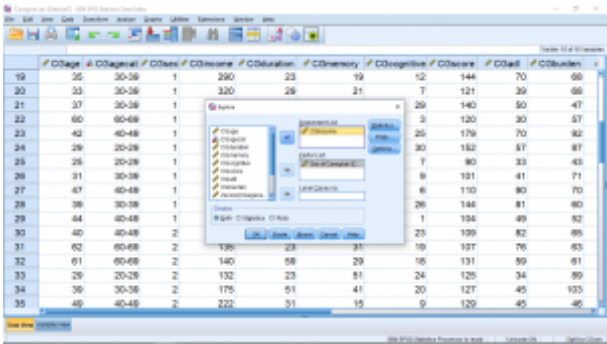
SPSS screenshot © International Business Machines Corporation.

From here, with the file unsplit, we can use gender as a factor to get separate descriptive statistics for males and female. Select Analyze → Explore and use gender as the factor, which results in :



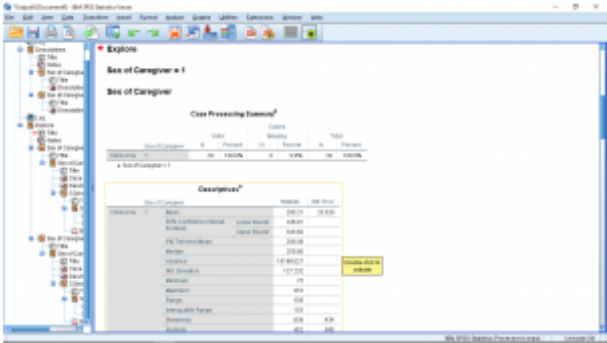
SPSS screenshot © International Business Machines Corporation.

From here, with the file unsplit, we can use gender as a factor to get separate descriptive statistics for males and female. Select Analyze → Explore and use gender as the factor :



SPSS screenshot © International Business Machines Corporation.

The result is :



SPSS screenshot © International Business Machines Corporation.

3.5 RStudio Lesson 2: Combining variables and recoding

[Coming soon]

4. PROBABILITY AND THE BINOMIAL DISTRIBUTIONS

4.1 Probability

The basic definition of probability is a ratio of things you can count (a ratio of their frequencies):

$$(4.1) \quad P(E) = \frac{n(E)}{n(S)}$$

where

$P(E)$ is the probability that event E happens,
 $n(E)$ is the number of ways E can happen and
 $n(S)$ is the total number of outcomes (all possibilities).

Example 4.1 : What is the probability of drawing a queen from a deck of cards :

$$P(E) = \frac{4}{52} = 0.077 \quad (7.7\% \text{ if we were to express the result in percentages})$$

□

To use $P(E)$ mathematically we set

$$0 \leq P(E) \leq 1$$

Where, probability-wise:

0 means E definitely will not occur, and

1 means E definitely will occur.

This is a method we can use instead of using percent. To compute probabilities, we first need to know how to count.

Fundamental Counting Rule

Say you have n events in order, and for event i there are k_i ways for it to happen. Then the number of ways for the n events to play out is :

$$k_1 \cdot k_2 \cdot k_3 \dots k_n = \prod_{i=1}^n k_i$$

(The giant pi symbolizes a multiplication convention in the same way that a giant sigma symbolizes a summation convention as described in Section 1.3.)

Example 4.2 How many combinations are there on a lock with 3 numbers?

Lay out the events as : $k_1 = 10$, $k_2 = 10$, and $k_3 = 10$. Note that each number can be anything from 0 to 9 giving 10 possibilities ($k_i = 10$) for each event. So the number of possible lock combinations is

$$k_1 k_2 k_3 = 10 \cdot 10 \cdot 10 = 10^3 = 1000$$

Note that you could have guessed this because the combination range from 000 to 999 – counting in base 10.

□

Example 4.3 Suppose that a hardware store can produce paints with the following qualities :

Colour : red, blue, white, black, green, brown, yellow (7 colours)

Type : latex, oil (2 types)

Texture : flat, semigloss, high-gloss (3 textures)

Use : indoor, outdoor (2 uses)

How many ways are there to combine these qualities to produce a can of paint?

Answer : From the above list $k_1 = 7$, $k_2 = 2$, $k_3 = 3$, $k_4 = 2$ and the number of possible paint kinds is:

$$7 \cdot 2 \cdot 3 \cdot 2 = 84$$

□

Applications of the Fundamental Counting Rule

We are interested in applying the fundamental counting rule to two special, important cases :

1. Permutations.
2. Combinations.

Let's define each one.

1. Permutations.

The number of ways, or permutations, of selecting r objects from a collection of n objects, while keeping track of the order of selection is

$${}_n P_r = \frac{n!}{(n-r)!}$$

This formula follows from the fundamental counting rule. With n objects there are $k_1 = n$ ways to select the first object. After selecting the first object there are $n - 1$ ways to choose the second object so $k_2 = n - 1$, etc. up to $k_r = n - r + 1$:

$$\begin{aligned} {}_n P_r &= (n)(n-1)(n-2)\dots(n-r+1) \\ &= \frac{(n)(n-1)\dots(2)(1)}{(n-r)(n-r-1)\dots(2)(1)} \end{aligned}$$

Example 4.4 : How many ways are there to choose 5 numbered balls from a bucket of 25 to make a lottery number?

Answer : $25 \cdot 24 \cdot 23 \cdot 22 \cdot 21 = 6,375,600$ possibilities. □

2. Combinations.

The number of ways of selecting x objects from a collection of n objects *without* caring about the order is :

$${}_n C_x = \frac{n!}{(n-x)!x!} = \frac{{}_n P_x}{x!} = \binom{n}{x}$$

That last symbol $\binom{n}{x}$ is colloquially called “ n choose x ”.

The second last expression demonstrates the application of the fundamental counting principal, it says

1. Recall that the definition of factorial follows

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \text{ etc.}$$

$$\binom{n}{x} = \frac{(n)(n-1)\dots(n-x-1)}{x!}$$

where $x!$ is just the number of ways of arranging x objects while caring about the order, $x! = {}_xP_x$.

As a practical matter, never try to compute $n!$. It will usually be unimaginably big. Use the formula that directly shows the fundamental counting rule as shown in the following example.

Example 4.5 : How many ways are there to select 10 balls from a bucket of 100?

Answer :

$$\binom{100}{10} = \frac{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \cdot 95 \cdot 94 \cdot 93 \cdot 92 \cdot 91}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{6.2815651 \times 10^{19}}{3,628,800} = \underline{17.3 \times 10^{12}}$$

□

The symbol $\binom{n}{x}$ is also known as the *binomial coefficient* because it shows up in algebra when you expand expressions of the form $(x + y)^n$. For example²

$$(x + y)^n = x^2 + 2xy + y^2$$

$$\begin{aligned} (x + y)^3 &= \binom{3}{0}x^3 + \binom{3}{1}x^2y + \binom{3}{2}xy^2 + \binom{3}{3}y^3 \\ &= x^3 + 3x^2y + 3xy^2 + y^3 \end{aligned}$$

2. You don't need this algebra for this statistics course. It's just interesting.

The binomial coefficients can be quickly computed using Pascal's triangle :

										$n =$		
				1						0		
			1		1					1		
		1		2		1				2		
		1	3		3		1			3		
	1		4		6		4		1	4		
1		1	5	10		10		5		1	5	
	1	6	15		20		15		6		1	6
					etc.							

Referring to Pascal's triangle we can quickly write

$$(x + y)^6 = x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6$$

for example.

4.2 Binomial Distribution

Given a success/failure situation (or yes/no, black/white, any 2 outcome, dichotomous situation) and a probability of success $P(S) = p$ (and so a probability of failure $P(F) = q = 1 - p$), what is the probability of achieving x successes in n trials? In symbols¹ what is $P(x \text{ successes} \mid n \text{ trials})$? Or with simpler notation, what is $P(x \mid n)$? The answer is :

$$(4.2) \quad P(x \mid n) = \binom{n}{x} p^x q^{n-x}.$$

****Proof of the $P(x \mid n)$ formula**

Use the boxes we used in defining the fundamental counting rule to represent each trial.

Consider $n = 1$.

The probability that a success occurs is the definition of p . So

$$P(1 \mid 1) = p = \binom{1}{1} p^1 q^0.$$

Consider $n = 2$. What is $P(0 \mid 2)$? This is all failures :

The probability of each failure is q so the probability of getting FF is $q \cdot q = q^2$. So

$$P(0 \mid 2) = q^2 = \binom{2}{0} p^0 q^2.$$

(Note that $\binom{2}{0} = 1$ by *definition*. There is exactly one way to draw no things from a collection of 2.)

1. Here the \mid is read as "given".

What is $P(1 | 2)$? Each probability of $p \cdot q$ ($p \cdot q$ for the first one, $q \cdot p$ for the second one). So

$$P(1 | 2) = 2 \cdot p \cdot q = \binom{2}{1} p^1 q^1.$$

For $x = 2$ we have

$$P(2 | 2) = \binom{2}{2} p^2 q^0.$$

We can continue this way for $n = 3, 4, \dots$ but this is clearly tedious. The way of “mathematical induction” is the formal way to proceed but let’s try a more intuitive approach.

For x successes in n trials, consider our n boxes, then any given sequence with x successes will have $n - x$ failures and so that given sequence will have a probability of $p^x q^{n-x}$. But how many specific sequences with x successes are there? Think of it this way. Of the n boxes, how many ways are there to write x S’s in the n boxes? There are n possibilities (n boxes are available) to write the first S, $n - 1$ ways after that to write the second S, etc. But we don’t care which order we wrote the S’s into the boxes so divide by $n!$. In other words there are $\binom{n}{x}$ specific sequences with x successes. Putting it all together :

$$P(x | n) = \binom{n}{x} p^x q^{n-x}.$$

□

Example 4.6 : In bucket of 100 toys with 20 dinosaurs and 80 bugs, consider drawing a dinosaur a success. So $P(S) = p = 0.2$ and $P(F) = q = 1 - p = 0.8$. Let us make an approximation and assume that p does not change with each draw²

2. **By assuming that p does not change, we will be lead to

the binomial distribution. If we more accurately assume that $P(S)$ changes with each draw we will be lead to the hypergeometric distribution. For fun, let's consider the case where $P(S)$ changes with each draw. It's just another application of the fundamental counting rule. To

begin, there are $\binom{100}{10} = 17.3 \times 10^{12}$ ways of

drawing 10 toys from the bucket without caring if it is a dinosaur or a bug. This is the size of the *sample space*; it is how many ways there are to make a sample of size 10 from the bucket of 100 choices; it is $n(S)$ in Equation (4.1). There are 17.3×10^{12} samples of 10 in the bucket.

If we want 3 dinosaurs in our sample, as in the example in text then of the 20 dinosaurs in the bucket, there are

$\binom{20}{3} = 1140$ ways to get 3 dinosaurs and

$\binom{80}{7} = 3.18 \times 10^9$ ways to get 7 bugs from the 80

in the bucket. So there are

$\binom{20}{3} \cdot \binom{80}{7} = 3.62 \times 10^{12}$ ways to draw 3

dinosaurs and 7 bugs from the bucket. This number is $n(E)$ in Equation (4.1). And so

$$P(3 \text{ dinosaurs} \mid 10 \text{ toys}) = \frac{\binom{20}{3} \binom{80}{7}}{\binom{100}{10}} = \frac{3.62 \times 10^{12}}{17.3 \times 10^{12}} = 0.209$$

Note how close this is to the answer from the binomial distribution of 0.201.

Say we want to know $P(3 \text{ successes} \mid 10 \text{ trials})$. In other words, what is the probability that if I take 10 toys out of the bucket that exactly 3 of them are dinosaurs? Using Equation (4.2) we find

$$P(3 \mid 10) = \binom{10}{3} 0.2^3 0.8^7 = 0.201.$$

The actual process of doing this calculation is somewhat tedious and therefore error prone. So in a test, for example, you will want to use the **Binomial Distribution Table** included in this text in the [Appendix](#). In the **Binomial Distribution Table**, you simply find the appropriate n and then x in the column on the left and then look under the appropriate p column to find $P(x \mid n)$ for the given p .

□

The complete binomial distribution specifies the probabilities of all x successes from 0 to n , and can be plotted as a histogram. Note that there is a binomial distribution for each x and p . Let's plot the binomial distribution for getting x successes (dinosaurs) in forming a sample of $n = 10$ toys with $p = 0.2$. The **Binomial Distribution Table** contains the relative frequency table for the histogram that represents the binomial distribution shown in Figure 4.1.

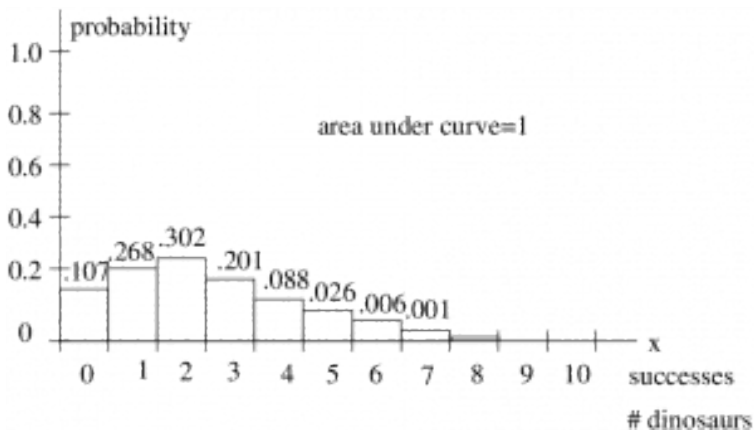


Figure 4.1: The binomial distribution for the example of forming samples of $n = 10$ toys with x representing the number of dinosaurs in the sample and $p = 0.2$ being the probability of selecting a dinosaur in forming the sample. Note that the probability of $x = 8, 9$ or 10 is not zero, just less than 0.001 .

The binomial distribution is an example of a *discrete probability distribution*. It is a histogram of relative frequencies obtained by counting possibilities in sample space.³

The mean and variance of any discrete distribution are given by

$$\mu = \sum_x x \cdot P(x)$$

$$\sigma^2 = \sum_x (x - \mu)^2 \cdot P(x) = \left[\sum_x x^2 \cdot P(x) \right] - \mu^2$$

3. Sample space is the set of all possible samples.

These two formulae come from the grouped data expressions $\mu = \sum f(x)x/n$ and $\sigma^2 = \sum f(x)(x - \mu)^2/n$, by substituting $P(x) = f(x)/n$. If we substitute Equation 4.2 for $P(x)$ in these general equations we get

$$\begin{aligned}\mu &= np \\ \sigma^2 &= npq\end{aligned}$$

which are the mean and variance for a binomial distribution with parameters n and p . The mean is the *expected value*.

Example 4.7 : For the bucket of toys example:

$$\mu = n \cdot p = 10 \cdot 0.20 = 2$$

So given any random sample of 10 toys we *expect* that 2 of them will be red.



4.2.1 Practical Binomial Distribution Examples

The examples given here illustrate the *sampling theory* for forming samples from a dichotomous (with success/fail items; items of interest and no interest) population. In this situation we know exactly what is in the population and ask questions about what kind of samples can be formed and what is their probability. The sampling theory is completely described by the binomial distribution. Later, we will have a sampling theory based on the Central Limit Theorem which will lead us to the normal distribution.

In practically solving these kinds of problems keep in mind that you need to identify: n , p and x .

Example 4.8 : It was reported that 5% of Americans are afraid of being alone in a house at night. In a random sample of 20 Americans, what are the probabilities that the sample contains

1. exactly 5 afraid people?

2. at most 3 afraid people?
3. at least 3 afraid people?

Solution : First identify: $n = 20$, $p = 0.05$ and the x as specific to each question :

1. For this case, $x = 5$, so from the **Binomial Distribution Table** get $P(x = 5) = 0.002$.
2. For this case $x = 0, 1, 2$ and 3 and we have to add up the probabilities

From the **Binomial Distribution Table**:

$$P(x = 0) = 0.358$$

$$P(x = 1) = 0.377$$

$$P(x = 2) = 0.189$$

$$P(x = 3) = 0.060$$

So

$$P(x \text{ is at most } 3) = 0.358 + 0.377 + 0.189 + 0.060 = 0.989$$

3. $x = 3, 4, 5, 6, 7, \dots, 20$

From the **Binomial Distribution Table**:

$$P(x = 3) = 0.060$$

$$P(x = 4) = 0.013$$

$$P(x = 5) = 0.002$$

$$P(x = 6 \text{ or more}) = \text{approximately zero}$$

Since the probabilities of high x are too small to appear in the **Binomial Distribution Table** (and there would be many terms to consider if they weren't) we should use the following trick :

$$\begin{aligned}P(x = 3 \text{ or more}) &= 1 - P(x \text{ is less than } 3) \\&= 1 - [P(0) + P(1) + P(2)] \\&= 1 - [0.358 + 0.377 + 0.189] = 0.076\end{aligned}$$

□

4.3 SPSS Lesson 3: Combining variables - advanced

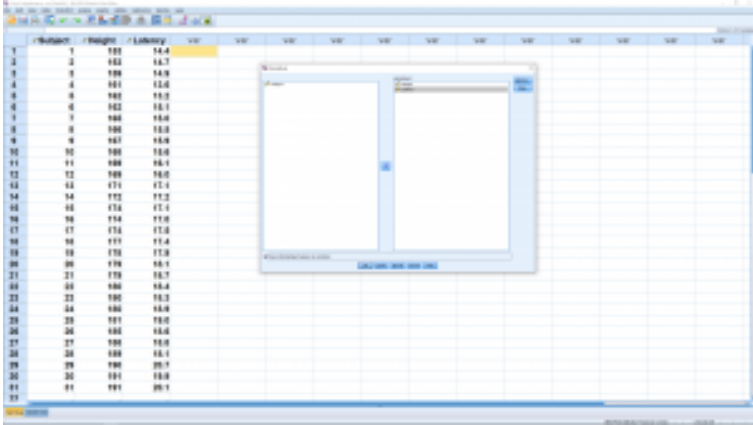
In SPSS Lesson 2 we saw how we can take variables defined on a Lickert scale and add them together, reverse scaling if necessary, to produce a single, better, variable for analysis. This works because the Lickert scale variables all have the same “units” (number of answer choices). You can combine any variables that have the same units, like feet or years or whatever. But if the units are different, but the variables still measure the same thing, like, for example, number of diet days per week and calories eaten per meal both measure levels of healthy eating habits but it makes no sense to simply add two such variables. It is literally like adding apples and oranges. The solution is to z -transform the variables you want to add first. The z -transform converts whatever units the original variable has to the z -transformed variable’s units of standard deviation distance from the mean. So when you add two z -transformed variables you end up with another variable whose units are standard deviation distance from the mean.

Let’s start by opening the file “HeightLatency.sav” from the [Data Sets](#). There are two variables in this file that we will combine into fewer variables. We begin by combining the variables Height and Latency into a new variable.

	Height	Latency
1	1	100
2	2	102
3	3	100
4	4	101
5	5	102
6	6	102
7	7	100
8	8	100
9	9	101
10	10	100
11	11	100
12	12	100
13	13	101
14	14	102
15	15	101
16	16	101
17	17	101
18	18	101
19	19	101
20	20	100
21	21	100
22	22	100
23	23	100
24	24	100
25	25	101
26	26	100
27	27	100
28	28	100
29	29	100
30	30	101
31	31	101
32	32	101

SPSS screenshot © International Business Machines Corporation.

Since Height and Latency have different units, we need to z -transform them first by running a descriptive analysis, making sure you have the “Save standardized values as variables” box checked :



SPSS screenshot © International Business Machines Corporation.

Hit Ok. This will produce two new variables, visible in the Data

View window, called ZHeight and ZLatency. We don't care about the actual descriptive statistics output here. Now you can simply add the Z-transforms to produce the required new variable :

The screenshot shows a data editor window with the following columns: 'Substact', 'Height', 'Latency', 'ZHeight', 'ZLatency', and several empty columns. The data rows are as follows:

	Substact	Height	Latency	ZHeight	ZLatency
1	1	180	14.4	-1.30288	-1.37228
2	2	182	14.7	-1.30288	-1.37228
3	3	188	14.9	-1.30288	-1.37228
4	4	181	14.6	-1.30288	-1.37228
5	5	180	14.5	-1.30288	-1.37228
6	6	182	14.1	-1.30288	-1.37228
7	7	188	14.8	-1.30288	-1.37228
8	8	186	14.9	-1.30288	-1.37228
9	9	187	14.6	-1.30288	-1.37228
10	10	188	14.8	-1.30288	-1.37228
11	11	188	14.1	-1.30288	-1.37228
12	12	188	14.6	-1.30288	-1.37228
13	13	171	17.1	.32832	.32832
14	14	182	17.2	-.32832	.32832
15	15	174	17.1	.32832	.32832
16	16	174	17.0	.32832	.32832
17	17	174	17.8	.32832	.32832
18	18	177	17.4	.32832	.32832
19	19	178	17.8	.32832	.32832
20	20	178	18.1	.32832	.32832
21	21	179	18.7	.32832	.32832
22	22	180	18.4	.32832	.32832
23	23	180	18.3	.32832	.32832
24	24	180	18.8	.32832	.32832
25	25	181	18.0	.32832	.32832
26	26	180	18.6	.32832	.32832
27	27	180	18.8	.32832	.32832
28	28	188	18.1	.32832	.32832
29	29	188	18.7	.32832	.32832
30	30	191	18.8	.32832	.32832
31	31	191	19.1	.32832	.32832

SPSS screenshot © International Business Machines Corporation.

Now let's combine a couple of sets of variables that have compatible units. First add ZHeight to ZLatency (note the fancy new way to add) to produce a new variable Sub :

The screenshot shows the same data editor window as above, but with a dialog box open in the center. The dialog box has a 'Variables' list on the left containing 'ZHeight' and 'ZLatency'. The 'Add' button is highlighted. The background data is partially visible through the dialog box.

SPSS screenshot © International Business Machines Corporation.

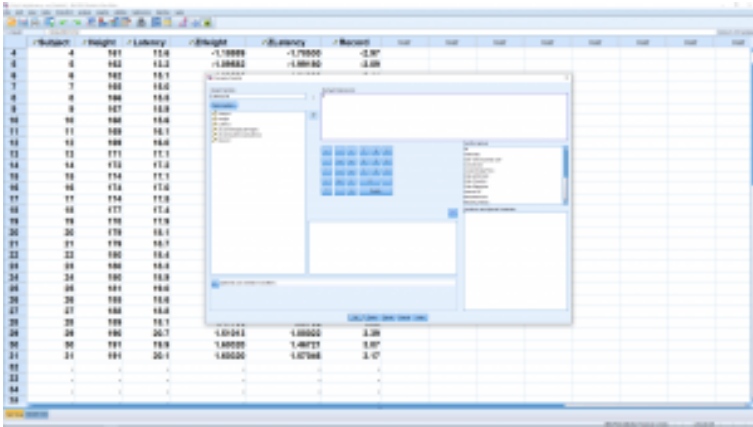
The new variable shows clearly on SPSS sheet :

Subject	Height	Latency	ZHeight	ZLatency	Residual
1	180	14.4	1.8044488701000	-1.57229	-0.771
2	182	14.7	1.85588	-1.21740	-0.36
3	186	14.9	1.97969	-1.17495	-0.409
4	181	14.8	1.83889	-1.23822	-0.407
5	180	14.9	1.89989	-1.09989	-0.409
6	182	14.1	1.89982	-1.81289	-0.911
7	186	14.6	1.99981	-1.69981	-0.409
8	180	14.9	1.72984	-1.59989	-0.873
9	187	14.9	1.89987	-1.59987	-0.359
10	182	14.6	1.82740	-1.72719	-0.299
11	188	14.1	1.99982	-1.99982	-0.4
12	189	14.9	1.99989	-1.99989	-0.99
13	171	17.1	1.28989	1.28989	0.24
14	172	17.2	1.38982	1.37989	0.99
15	174	17.1	1.47989	1.37989	0.4
16	174	17.0	1.47989	1.37989	-0.71
17	174	17.0	1.47989	1.37989	0.99
18	177	17.4	1.69989	1.57989	0.6
19	179	17.9	1.78989	1.67987	0.23
20	176	18.1	1.68984	1.67989	0.469
21	179	18.7	1.78989	1.67989	1.223
22	186	18.4	1.97989	1.67989	1.07
23	190	18.2	2.07989	1.67111	1.223
24	186	18.6	1.97989	1.67989	0.469
25	191	18.0	2.07989	1.66989	1.669
26	190	18.0	2.06979	1.66989	1.664
27	190	18.0	2.06989	1.66989	1.223
28	188	18.1	1.97794	1.67989	1.669
29	190	18.7	2.07989	1.66989	0.99
30	191	18.9	2.08229	1.66779	1.07
31	191	18.1	2.08989	1.67989	0.717

SPSS screenshot © International Business Machines Corporation.

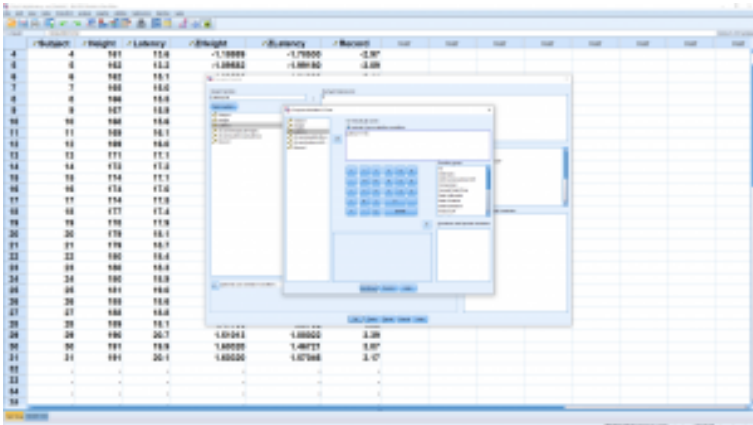
Next we will make a conversion from a quantitative variable to a qualitative variable essentially by dividing the data into classes. First a simple case. Create the new variable Life from the variable Latency as the following :

$$\text{Latency} = \begin{cases} 1 & \text{if Latency} < 17.5 \\ 2 & \text{if Latency} \geq 17.5 \end{cases}$$



SPSS screenshot © International Business Machines Corporation.

We'll need to do this in two steps. First pull up Transform → Compute Variable and set it up so that 1 is in the Numeric Expression box. Then hit the If... button at the bottom left hand of the menu window to bring up :



SPSS screenshot © International Business Machines Corporation.

Then click Continue, then hit OK. That will create the new LatencyCat variable, with missing values. Those values will be filled in the next step.

Subject	Weight	Latency	Weight	Abundance	Percent	LatencyCat	
4	4	101	11.0	-1.0000	-1.7000	-2.30*	1.00
6	6	102	11.2	-1.0000	-1.8000	-2.40	1.00
8	8	103	11.1	-1.0000	-1.9000	-2.50	1.00
7	7	104	11.0	-1.0000	-1.8000	-2.40	1.00
9	9	105	11.0	-1.0000	-1.9000	-2.50	1.00
5	5	107	11.0	-1.0000	-1.8000	-2.40	1.00
10	10	106	11.0	-1.0000	-1.9000	-2.50	1.00
11	11	109	11.1	-1.0000	-1.8000	-2.40	1.00
12	12	108	11.0	-1.0000	-1.9000	-2.50	1.00
13	13	111	11.1	-1.0000	-2.0000	-2.60	1.00
14	14	110	11.0	-1.0000	-2.1000	-2.70	1.00
15	15	114	11.1	-1.0000	-2.2000	-2.80	1.00
16	16	113	11.0	-1.0000	-2.3000	-2.90	1.00
17	17	116	11.0	-1.0000	-2.4000	-3.00	1.00
18	18	117	11.0	-1.0000	-2.5000	-3.10	1.00
19	19	118	11.0	-1.0000	-2.6000	-3.20	1.00
20	20	119	11.1	-1.0000	-2.7000	-3.30	1.00
21	21	120	11.1	-1.0000	-2.8000	-3.40	1.00
22	22	121	11.0	-1.0000	-2.9000	-3.50	1.00
23	23	122	11.0	-1.0000	-3.0000	-3.60	1.00
24	24	123	11.0	-1.0000	-3.1000	-3.70	1.00
25	25	124	11.0	-1.0000	-3.2000	-3.80	1.00
26	26	125	11.0	-1.0000	-3.3000	-3.90	1.00
27	27	126	11.0	-1.0000	-3.4000	-4.00	1.00
28	28	127	11.1	-1.0000	-3.5000	-4.10	1.00
29	29	128	11.0	-1.0000	-3.6000	-4.20	1.00
30	30	129	11.0	-1.0000	-3.7000	-4.30	1.00
31	31	130	11.0	-1.0000	-3.8000	-4.40	1.00
32	32	131	11.0	-1.0000	-3.9000	-4.50	1.00
33
34
35

SPSS screenshot © International Business Machines Corporation.

Pull up Transform → Compute Variable again and, leaving LatencyCat where it is, put 2 in the Numeric Expression box, then hit the “If” button again and change the expression in the condition box, then hit Continue, then OK. Now LatencyCat is either 1 or 2 with no missing values :

	*Number	*Height	*Latitude	*Elight	*Bussess	*Market	*Latencolour							
1	1	100	14.4		-1.07229	-0.371	1.000							
2	2	102	14.7	-1.05088	-1.21740	-0.366	1.000							
3	3	106	14.9	-1.07500	-1.19419	-0.409	1.000							
4	4	101	14.0	-1.08888	-1.70420	-0.267	1.000							
5	5	102	14.0	-1.08888	-1.59990	-0.289	1.000							
6	6	102	14.1	-1.09402	-1.51049	-0.171	1.000							
7	7	100	14.0	-1.09402	-1.60409	-0.168	1.000							
8	8	100	14.0	-1.22088	-1.60409	-0.163	1.000							
9	9	101	14.0	-1.09402	-1.60409	-0.164	1.000							
10	10	100	14.0	-1.07140	-1.70219	-0.159	1.000							
11	11	100	14.1	-1.08888	-1.60409	-0.164	1.000							
12	12	100	14.0	-1.09402	-1.54020	-0.169	1.000							
13	13	111	17.1	-1.08888	-1.60409	-0.164	1.000							
14	14	112	17.2	-1.08888	-1.60409	-0.169	1.000							
15	15	114	17.1	-1.08888	-1.60409	-0.164	1.000							
16	16	114	17.0	-1.07140	-1.60409	-0.161	1.000							
17	17	114	17.0	-1.07140	-1.60409	-0.161	1.000							
18	18	111	17.4	-1.08888	-1.60409	-0.169	1.000							
19	19	110	17.0	-1.08888	-1.60409	-0.161	1.000							
20	20	116	18.1	-1.08888	-1.60409	-0.166	1.000							
21	21	119	18.7	-1.08888	-1.60409	-0.163	1.000							
22	22	100	14.4	-1.07140	-1.60409	-0.167	1.000							
23	23	100	14.2	-1.07140	-1.60409	-0.162	1.000							
24	24	100	14.0	-1.07140	-1.60409	-0.162	1.000							
25	25	101	14.0	-1.07140	-1.60409	-0.160	1.000							
26	26	100	14.0	-1.08888	-1.60409	-0.164	1.000							
27	27	100	14.0	-1.07140	-1.60409	-0.161	1.000							
28	28	100	14.1	-1.07140	-1.60409	-0.160	1.000							
29	29	100	14.7	-1.07140	-1.60409	-0.169	1.000							
30	30	101	14.0	-1.08888	-1.60409	-0.167	1.000							
31	31	101	14.1	-1.08888	-1.60409	-0.167	1.000							

SPSS screenshot © International Business Machines Corporation.

4.4 RStudio Lesson 3: Combining variables - advanced

[Coming soon]

5. THE NORMAL DISTRIBUTIONS

5.1 Discrete versus Continuous Distributions

We can describe populations in terms of discrete variables ($x \in \mathbb{Z}$) or continuous variables ($x \in \mathbb{R}$). In the last chapter we saw how to describe discrete probability distributions with the example of the binomial distributions. Discrete probabilities need to be added in inferential statistics and this can lead to complicated formulae. Calculus turns sums into integrals¹ which generally lead to simpler formulae. In the following table we compare, and show the relationship between, discrete and continuous variables and their associated probability distributions.

1. If you have no calculus background, an integral is a way of calculating areas under curves.

Discrete	Continuous
<ul style="list-style-type: none"> • We have a finite number of values between the high and low values • A histogram plot of the random variables x may be interpreted as a probability distribution. 	<ul style="list-style-type: none"> • We have an infinite number of values between the high and low values. • With continuous random variables we have a probability density.
<p>By increasing the number of values in an appropriate limiting way you make \longrightarrow the discrete probability distribution \longrightarrow approach a probability density.</p>	
<ul style="list-style-type: none"> • The units of $P(x)$ are probability. 	<ul style="list-style-type: none"> • The units of $P(x)$ are probability density. Probabilities are given by areas under the curve only.

We will be slurring our language and call a probability density, a probability distribution. So we'll say normal distribution instead of normal density. Continuing the comparison, probability distributions and densities have means, moments, skewness, etc. :

- Means and variances of a discrete probability distribution, $P(x)$, are given by the application of the grouped data formulae we saw in Chapter 4 :

$$\mu = \sum_x x \cdot P(x) \quad \sigma^2 = \sum_x [(x - \mu)^2 \cdot P(x)]$$

- Means and variances of a continuous probability density, $P(x)$ are given by the integrals :

$$\mu = \int x \cdot P(x) dx \quad \sigma^2 = \int (x - \mu)^2 \cdot P(x) dx$$

Recall that the variance is the second moment of x about the mean μ .

We don't have to stop at the second moment about the mean. The third and fourth moments about the mean are called skewness and kurtosis respectively :

	Discrete	Continuous
Skewness	$\mu_3 = \frac{1}{\sigma^3} \sum (x - \mu)^3 P(x)$	$\mu_3 = \frac{1}{\sigma^3} \int (x - \mu)^3 P(x) dx$
Kurtosis	$\mu_4 = \frac{1}{\sigma^4} \sum (x - \mu)^4 P(x)$	$\mu_4 = \frac{1}{\sigma^4} \int (x - \mu)^4 P(x) dx$

SPSS will easily compute skewness and kurtosis. μ_3 is positive for a positively skewed distribution, negative for a negative skewed distribution. The σ^3 and σ^4 are “normalization” factors; they make the moments of the normal distribution simple.

The moments of a probability distribution are important. In fact, if you specify all the moments of a distribution then you have completely specified the distribution. Let's say that in another way. The specify a probability distribution you can either give its formula (as generally derived from counting) or you can give all its moments. The normal distribution with a mean of μ and a variance of σ^2 is specified by the formula

$$(5.1) \quad P(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

or by its moments. The normal distribution with a mean of μ and a variance of σ^2 is the only continuous probability distribution with moments (from first to second and on up) of: $\mu, \sigma^2, 0, 1, 0, 1, 0,$

. . . The normal distribution is special that way among probability distributions.

5.2 **The Normal Distribution as a Limit of Binomial Distributions

The results of the derivation given here may be used to understand the origin of the Normal Distribution as a limit of Binomial Distributions¹. A mathematical “trick” using logarithmic differentiation will be used.

First, recall the definition of the Binomial Distribution² as

$$(5.2) \quad w_n(x) = \binom{n}{x} p^x q^{n-x}$$

where p is the probability of success, $q = 1 - p$ is probability of failure and

1. The formula for the Binomial Distribution was apparently derived by Newton according to: Lindsay RB, Margenau. Foundations of Physics. Dover, New York, 1957 (originally published 1936). For that claim, Lindsay & Margenau quote: von Mises R. Probability, Statistics, and Truth. Macmillan, New York, 1939 (originally published 1928). The derivation of the Normal Distribution presented here largely follows that given in Lindsay & Margenau's book.
2. In class we denoted the Binomial distribution as $P(x | n)$. Here we use $w_n(x) = P(x | n)$ to avoid using too many P's and p's.

$$(5.3) \quad \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the binomial coefficient that counts the number of ways to select x items from n items without caring about the order of selection. Here x is a discrete variable, $x \in \mathbb{Z}$, with $0 \leq x \leq n$.

The trick is to find a way to deal with the fact that $x \in \mathbb{Z}$ (x is a discrete variable) for the Binomial Distribution and $x \in \mathbb{R}$ (x is a continuous variable) for the Normal Distribution³ In other words as we let $n \rightarrow \infty$ we need to come up with a way to let Δx shrink⁴ so that a probability density limit (the Normal Distribution) is reached from a sequence of probability distributions (modified Binomial Distributions). So let $w(x)$ represent the Normal Distribution with mean $\bar{x} = np$ and variance $\sigma^2 = npq$. We will show how $\lim_{n \rightarrow \infty} w_n(x) = w(x)$ where each Binomial Distribution $w_n(x)$ also has mean $\bar{x} = np$ and variance $\sigma^2 = npq$.

The heart of the trick is to notice⁵ that

$$(5.4) \quad \frac{d}{dx} \ln w(x) = \lim_{\Delta x \rightarrow 0} \frac{w(x + \Delta x) - w(x)}{w(x)\Delta x}.$$

This is perfectly true for the density $w(x)$. The trick is to

3. Remember that the Normal Distribution is technically a probability density but we slur the use of the word distribution between probability distribution (discrete x) and probability density (continuous x) like everyone else.
4. $\Delta x = 1$ for the Binomial Distribution.
5. Remember that $\frac{d}{dx} \ln(x) = \frac{1}{x}$ and use the chain rule to notice this.

substitute the distribution $w_n(x)$ for the density $w(x)$ in the RHS of Equation (5.4) to get :

$$(5.5) \quad \frac{w_n(x + \Delta x) - w_n(x)}{w_n(x)\Delta x} = \frac{w_n(x + 1) - w_n(x)}{w_n(x)}$$

because $\Delta x = 1$. The trick is to now pretend that $w_n(x)$ is a continuous function defined at all $x \in \mathbb{R}$; we just don't know what its values should be for non-integer x . With such a "continuation" of $w_n(x)$ we can write⁶

6. You can probably imagine many ways to continue the Binomial Distribution from $x \in \mathbb{Z}$ to $x \in \mathbb{R}$. It doesn't matter which one you pick as long as the behaviour of your new function is not too crazy between the integers; that is, $\lim_{n \rightarrow \infty} w_n(x)$ should exist at all $x \in \mathbb{R}$.

(..)

$$\frac{d}{dx} \ln w(x) = \lim_{n \rightarrow \infty} \frac{w_n(x+1) - w_n(x)}{w_n(x)} \quad (5.6)$$

$$= \lim_{n \rightarrow \infty} \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^x q^{n-x}} - 1 \quad (5.7)$$

$$= \lim_{n \rightarrow \infty} \frac{n-x}{x+1} \frac{p}{q} - 1. \quad (5.8)$$

Equation (5.8) has no limit; it blows up as $n \rightarrow \infty$. We need to transform x in such a way to gain control on Δx (getting it to shrink as $n \rightarrow \infty$) and to get something that converges. To do that we introduce $h = \frac{1}{\sqrt{n}}$ and a new variable $u = h(x - \bar{x}) = h(x - np)$. With this transformation of variables, the chain rule gives

$$(5.9) \quad \frac{d}{dx} \ln w(x) = \frac{du}{dx} \frac{d}{du} \ln w(u) = h \frac{d}{du} \ln w(u)$$

and the RHS of Equation (5.8) becomes, using $x = \frac{u}{h} + np$

(..)

$$\frac{n - x}{x + 1} \frac{p}{q} - 1 = \frac{\left(n - \frac{u}{h} - np\right) p}{\left(\frac{u}{h} + np + 1\right) q} \quad (5.10)$$

$$= \frac{\left(n(1 - p) - \frac{u}{h}\right)}{\left(\frac{u}{h} + np + 1\right) \frac{q}{p}} \quad (5.11)$$

$$= \frac{\left(nq - \frac{u}{h}\right)}{\left(\frac{uq}{hp} + nq + \frac{q}{p}\right)} \quad (5.12)$$

$$= \frac{\left(1 - \frac{u}{nhq}\right)}{\left(\frac{u}{nhp} + 1 + \frac{1}{np}\right)} \quad (5.13)$$

$$= \frac{\left(1 - \frac{u}{nhq}\right)}{\left(1 + \frac{u+h}{nhp}\right)} \quad (5.14)$$

Using Equation (5.9), for the LHS, and Equation (5.14), for the RHS, Equation (5.8) becomes

(..)

$$h \frac{d}{du} \ln w(u) = \lim_{n \rightarrow \infty} \frac{1 - \frac{u}{nhq}}{1 + \frac{u+h}{nhp}} - 1 \quad (5.15)$$

$$= \lim_{n \rightarrow \infty} \left(1 - \frac{u}{nhq} \right) \left[1 - \frac{u+h}{nhp} + \left(\frac{u+h}{nhp} \right)^2 - \dots \right] - 1 \quad (5.16)$$

$$= \lim_{n \rightarrow \infty} -\frac{1}{np} - \frac{u}{nhq} - \frac{u}{nhp} + O\left(\frac{1}{n}\right) \quad (5.17)$$

$$= \lim_{n \rightarrow \infty} -\frac{1}{np} - \frac{u}{nhpq} + O\left(\frac{1}{n}\right) \quad (5.18)$$

$$= \lim_{n \rightarrow \infty} -\frac{u}{nhpq}. \quad (5.19)$$

where $O\left(\frac{1}{n}\right)$ means terms that will go to zero as $n \rightarrow \infty$, and we have used the relation $\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots$ to get Equation (5.16) and $p + q = 1$ to go from Equation (5.17) to Equation (5.18). Dividing both sides of Equation (5.19) by h leaves

$$(5.20) \quad \frac{d}{du} \ln w(u) = \lim_{n \rightarrow \infty} -\frac{u}{nh^2pq} = -\frac{u}{pq}.$$

Our transformation, with its \sqrt{n} , has given us the exact control we need to keep the limit from disappearing or blowing up. Integrating Equation (5.20) gives

$$(5.21) \quad w(u) = Ce^{-\frac{u^2}{2pq}}$$

where C is the a constant of integration. Switching back to the x variable

(.)

$$w(x) = Ce^{-\frac{(h[x-\bar{x}])^2}{2pq}} \tag{5.22}$$

$$= Ce^{-\frac{(x-\bar{x})^2}{2npq}} \tag{5.23}$$

$$= Ce^{-\frac{(x-\bar{x})^2}{2\sigma^2}}. \tag{5.24}$$

To evaluate the constant of integration, C , we impose $\int_{-\infty}^{\infty} w(x) dx = 1$ because we want $w(x)$ to be a probability distribution. So

$$(5.25) \quad C \int_{-\infty}^{\infty} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx = C\sqrt{2\pi\sigma^2} = 1$$

so

$$(5.26) \quad C = \frac{1}{\sqrt{2\pi\sigma^2}}$$

and

$$(5.27) \quad w(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

which is the Normal Distribution that approximates Binomial Distributions with the same mean and variance as n gets large.

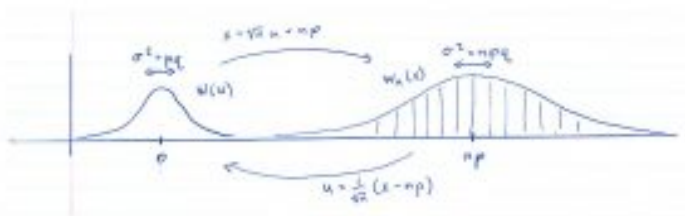


Figure 5.1: The transformation $u = \frac{(x-np)}{\sqrt{n}}$ effectively shrinks the Δx of the Binomial Distribution with mean $\bar{x} = np$ and variance $\sigma^2 = npq$ by pulling a continuous version $w_n(x)$ back to the constant Normal Distribution $w(u)$. Another way of thinking about it is that the transformation $x = \sqrt{n}u + np$ takes the fixed Normal Distribution $w(u)$ to the Normal Distribution $w(x)$ that provides a better and better approximation of $w_n(x)$ as $n \rightarrow \infty$.

You may be wondering why that transformation $u = \frac{1}{\sqrt{n}}(x - np)$ worked because it seems to have been pulled from the air. According to Lindsay & Margenau, it was Laplace who first used this transformation and derivation in 1812. What this transformation does is pull the Binomial Distribution $w_n(x)$ back to have a mean of zero (by subtracting $\bar{x} = np$) which keeps x from running off to infinity and, more importantly, allows us to define a function $w(u)$ with $u \in \mathbb{R}$ that has a constant variance of pq that we can match to npq when we transform back to x at each n , see Figure 5.1. Looking at it the other way around,

the Normal Distribution⁷ $w(x)$ with $x = \sqrt{nu} + np$ is an approximation for Binomial Distribution $w_n(x)$ that “asymptotically” approaches $w_n(x)$ as $n \rightarrow \infty$.

This is not the only way to form a probability density limit from a sequence of Binomial distributions. It is one that gives a good approximation of the Binomial Distribution when n is fairly small if the term $\frac{1}{np}$ in Equation (5.18) becomes small quickly. If p is very small, this does not happen and another limit of Binomial Distributions that leads to the Poisson Distribution is more appropriate. When p and q are close to 0.5 or more generally when $np \geq 5$ and $nq \geq 5$ then the Normal approximation is a good one. Either way, the density limit is a mathematical idealization, a convenience really, that is based on a discrete probability distribution that just summarizes the result of counting outcomes. Counting gives the foundation for probability theory.

7. Our symbols here are not mathematically clean; we should write something like $w(u(x))$ instead of $w(x)$ or w composed with u at x , $w \circ u_n(x)$, instead of $w(x)$. But to emphasize the intuition we use $w(x)$. In clean symbols, the function $w \circ u_n(x)$ asymptotically approaches $w_n(x)$ where $u_n(x) = \frac{(x-np)}{\sqrt{n}}$.

5.3 Normal Distribution

Let us now take a detailed look at the normal distribution and learn how to apply it to probability problems (in sampling theory) and statistical problems. Its formula (which you will never have to use because we have tables and SPSS) is again:

$$(5.28) \quad P(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

The factor $\sigma\sqrt{2\pi}$ is a normalization factor that ensures that the area under the whole curve is one:

$$\int P(x) dx = 1.$$

Without that factor we just have a bell-shaped curve¹ with the area under the curve equal to one we have a probability function since the total probability is one. For those with a bad math background, the letters in Equation (5.28) are: $e = 2.718\dots$ ², $\pi = 3.1415\dots$ ³, $\mu =$ mean and $\sigma =$ standard deviation of the normal distribution. The normal distribution's shape is as shown in Figure 5.2.

1. **Whose shape is determined essentially by the shape of $y = e^{-x^2}$. Plot $y = e^{-x}$ and think about the square preventing any negative values for the argument.
2. ** The number e is the natural base implied by functions whose values match how fast it changes, i.e. the derivative of the function is the same as the function.
3. ** Of course, π comes from circles: $\pi =$ circumference/diameter.

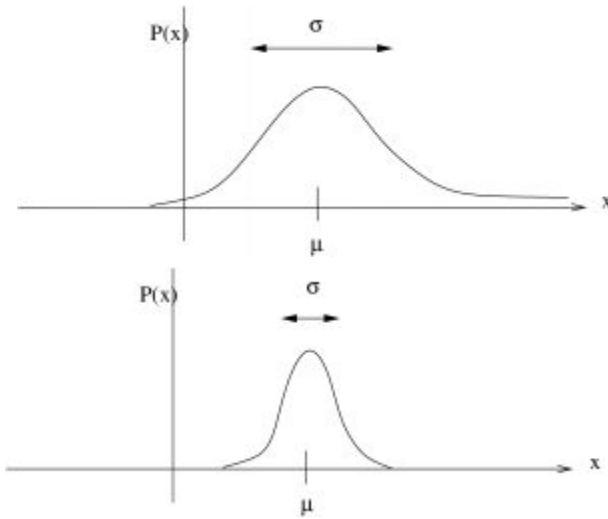


Figure 5.2: The normal distribution. It is a bell-shaped curve with its mode (= mean and median because it's symmetric, $\mu_3 = 0$) centred on its mean μ . On the left is a distribution with a large σ^2 and on the right one with a smaller σ^2 .

To work with normal distribution, in particular so we can use the **Standard Normal Distribution Table** and the **t Distribution Table** in the [Appendix](#), we need to transform it to the *standard normal distribution* using the *z*-transform. We need to transform $P(x)$, which has a mean μ and standard deviation σ to $P(z)$ which has a mean of 0 and a standard deviation of 1. Recall the definition of the *z*-transform:

$$z = \frac{x - \mu}{\sigma}$$

applying this to $P(x)$ gives

$$(5.29) \quad P(z) = \frac{P(x) - \mu}{\sigma}.$$

If we substitute Equation (5.28) into Equation (5.29) and do the algebra we get :

$$(5.30) \quad P(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}.$$

Equation (5.30) defines the standard normal distribution, or as we'll call it, the z -distribution.

Areas under $P(z)$ are given in the **Standard Normal Distribution Table** in the [Appendix](#).

5.3.1 Computing Areas (Probabilities) under the standard normal curve

Here we learn how to use the **Standard Normal Distribution Table** to get probabilities associated with any old area under the normal curve that we can dream up. The general layout of areas under the z -distribution is shown in Figures 5.3 and 5.4.

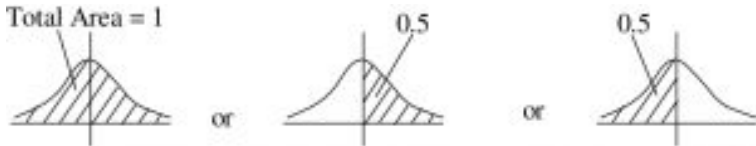


Figure 5.3 : The Z -distribution is a probability distribution (total area = 1) and symmetric, so the area on either side of the mean (which is 0) is a half. You will need to remember this information as you calculate areas using the Standard Normal Distribution Table.

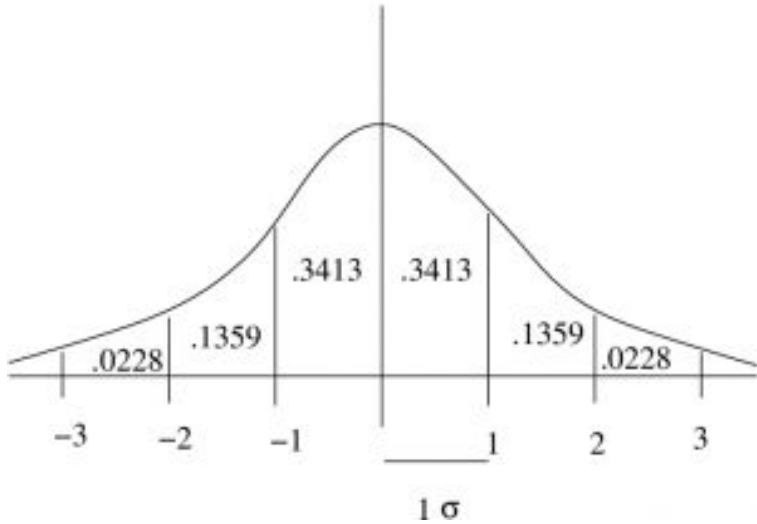


Figure 5.4 : The units of z in $P(z)$ are standard deviations. No matter what the measurement units of x were before the z -transformation, the units of z are “standardized” to be standard deviation units. With SPSS you will learn how to standardize (z -transform) variables so that you can sensibly combine multiple dependent variables into one dependent variable for univariate statistical analysis. The areas, probabilities, associated with each increment in σ are shown here.

Let’s divide the types of areas we want to compute into cases, following Bluman⁴. For all these cases we’ll use the notation $A(z)$ to represent the area we look up in the **Standard Normal Distribution Table** associated with z .

Case 1 : Areas on one side of the mean. This is the case of finding an area between 0 (which corresponds to the mean before any z -transformations) and a given z . For this case we simply use the

4. Bluman AG, *Elementary Statistics: A Step-by-Step Approach*, numerous editions, McGraw-Hill Ryerson, circa 2005.

tabulated values, $P(0 \leq x \leq z) = A(z)$, see Figure 5.5. This case also covers when z is a negative number: $P(-z \leq x \leq 0) = A(z)$.

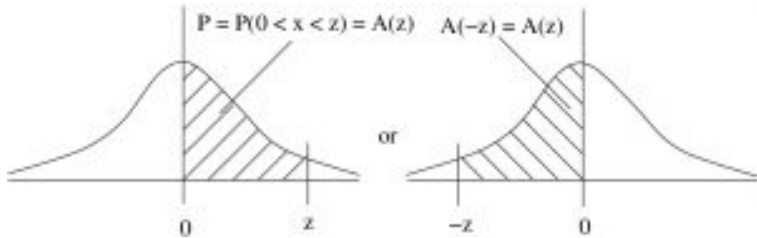


Figure 5.5 : Case 1: Areas on one side of the mean.

Example 5.1 : Find the probability that z is between 0 and 2.34.

Solution : Look up $A(2.34)$ in the **Standard Normal Distribution Table**, see Figure 5.6. $P = P(0 < z < 2.34) = A(2.34) = 0.4904$. (Note that it makes no difference whether we use $<$ or \leq because the probability of a single value is 0. That's why we need to use areas.)

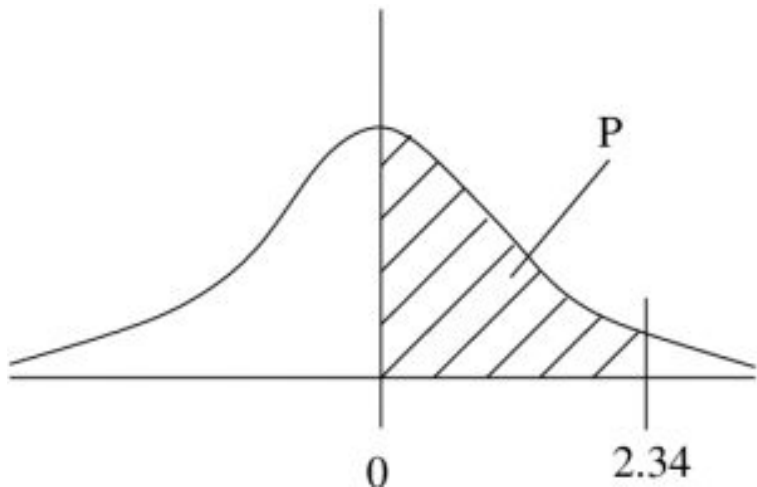


Figure 5.6 : The situation for Example 5.1.

□

Example 5.2 : Find the probability that z is between -1.75 and 0 .

Solution : $P(-1.75 < z < 0) = A(1.75) = 0.4599$, see Figure 5.7.

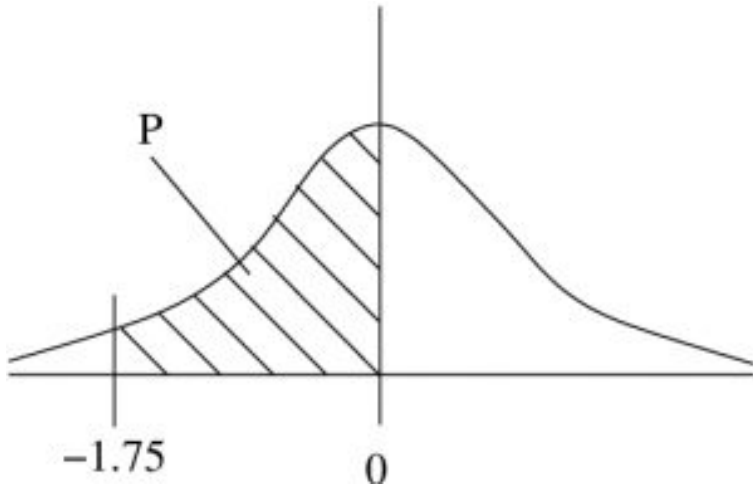


Figure 5.7 : The situation for Example 5.2.

□

Case 2 : Tail areas. A tail area is the opposite of the area given in the **Standard Normal Distribution Table** on one half of the normal distribution, see Figure 5.8. The tail area after a given positive z is $P = P(x > z) = 0.5 - A(z)$ or before a given negative value $-z$ is $P = P(x < -z) = 0.5 - A(z)$.

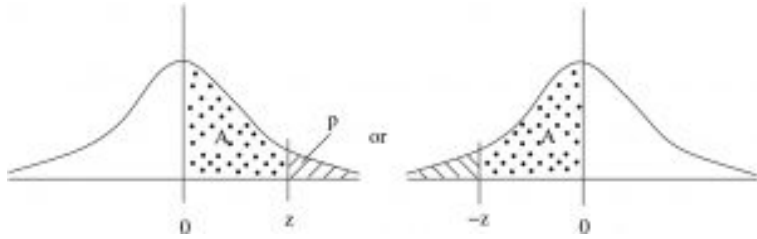


Figure 5.8 : Case 2 : Tail areas.

Example 5.3: What is the probability that $z > 1.11$?

Solution

$$P(z > 1.11) = 0.5 - A(1.11) = 0.5 - 0.3665 = 0.1335$$

where $A(1.11) = 0.3665$ is the area under the standard normal curve between 0 and 1.11. See Figure 5.9.

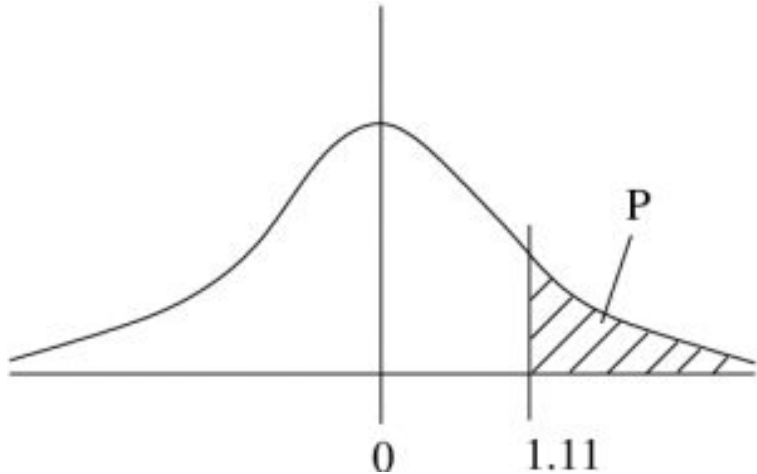


Figure 5.9 : The situation for Example 5.3.

□

Example 5.4 : What is the probability that $z < -1.93$?

Solution

:

$P = P(z < -1.93) = 0.5 - A(1.93) = 0.5 - 0.4732 = 0.0268$
 , see Figure 5.10.

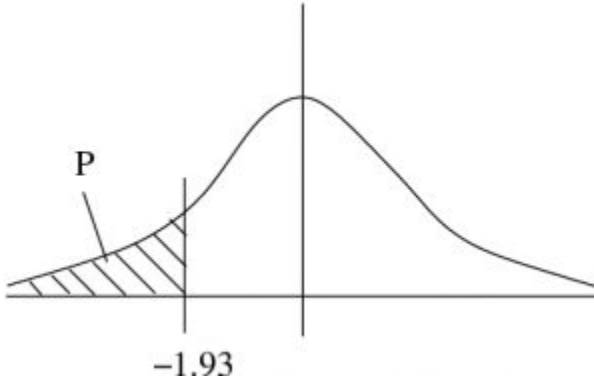


Figure 5.10 : The situation for Example 5.2.

□

Case 3 : An interval on one side of the mean. Recall that $\mu = 0$ for the z -distribution. So we are looking for the probabilities $P = P(z_1 < x < z_2)$ for an interval to the right of the mean or $P = P(-z_2 < x < -z_1)$ for an interval to the left of the mean. In either case $P = A(z_2) - A(z_1)$, see Figure 5.11.

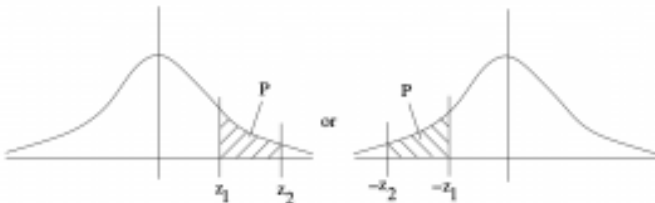


Figure 5.11: Case 3: An interval on one side of the mean.

Example 5.5 : What is the probability that z is between 2.00 and 2.97?

Solution

$P(2.00 < z < 2.97) = A(2.47) - A(2.00) = 0.4932 - 0.4772 = 0.0160$
, see Figure 5.12.

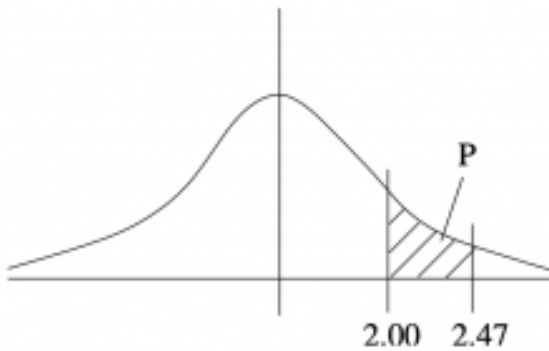


Figure 5.12: The situation for Example 5.5.

□

Example 5.6 : What is the probability that z is between -2.48 and -0.83?

Solution

$P(-2.48 < z < -0.83) = A(2.48) - A(0.83) = 4.934 - 0.2967 = 0.1967$
, see Figure 5.13.

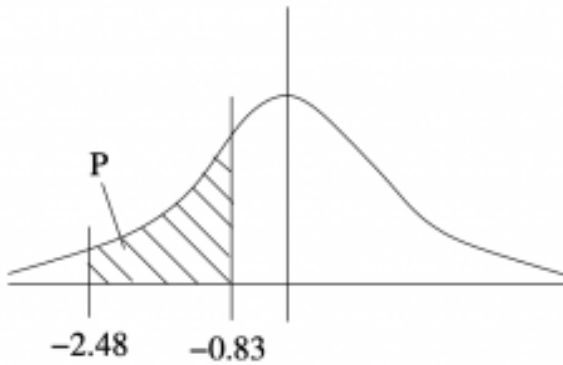


Figure 5.13: The situation of Example 5.6.

□

Case 4 : An interval containing the mean. The situation is as shown in Figure 5.14 with the interval being between a negative and a positive number. In that case $P(-z_1 < x < z_2) = A(z_1) + A(z_2)$.

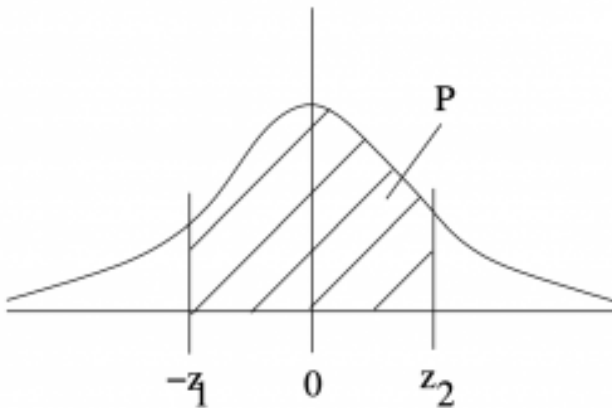


Figure 5.14: Case 4: An interval containing the mean.

Example 5.7 : What is the probability that z is between -1.37 and 1.68 ?

Solution :

$P(-1.37 < z < 1.68) = A(1.37) + A(1.68) = 0.4147 + 0.4535 = 0.8682$
, see Figure 5.15.

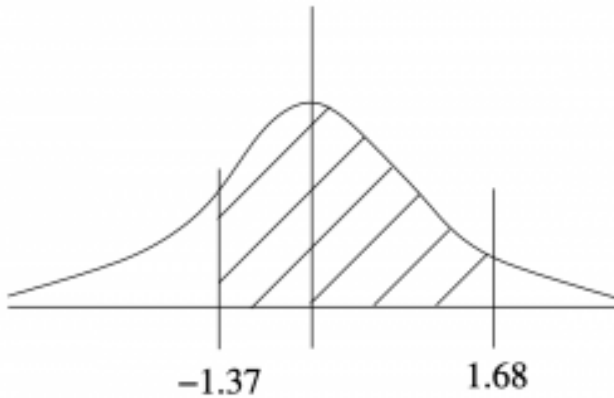


Figure 5.15: The situation for Example 5.7.

□

Cases 5 & 6 : Excluding tails. Case 5 is excluding the right tail, $P(x < z)$. Case 6 is excluding the left tail, $P(x > -z)$ $-z$ " title="Rendered by QuickLaTeX.com" height="18" width="83" style="vertical-align: -4px;". See Figure 5.16. Case 5 is the situation which gives the percentile position of z if you multiply the area by 100. More about percentiles in Chapter 6. In either case, $P = 0.5 + A(z)$.

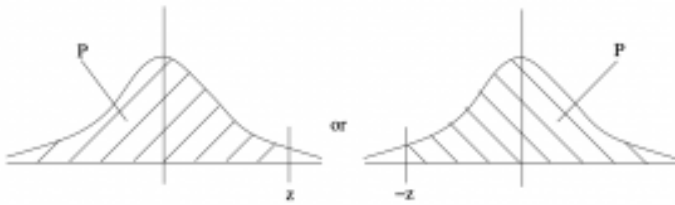


Figure 5.16: Left: Case 5. Right: Case 6.

Case 7 : Two unequal tails. In this case we add the areas of the left and right tails, see Figure 5.17. The special case where the tails have equal areas (i.e. when $z_1 = z_2$ in the notation we have been using) is the case we will encounter for two-tail hypothesis testing. $P = P(x < -z_1) + P(x < z_2) = (0.5 - A(z_1)) + (0.5 - A(z_2))$

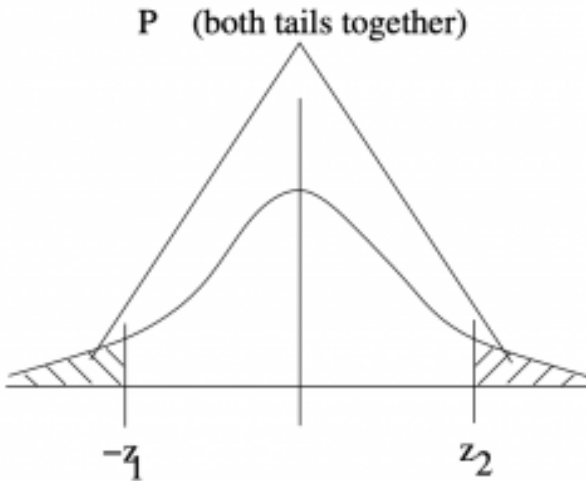


Figure 5.17: Case 7: Two unequal tails.

Example 5.8 : Find the areas of the tails shown in Figure 5.18.

Solution :

$$\begin{aligned} P(z < -3.01 \text{ or } z > 2.43) &= 2A(3.01) + 2A(2.43) \\ &= (0.5 - A(3.01)) + (0.5 - A(2.43)) \\ &= (0.5 - 0.4987) + (0.5 - 0.4925) \\ &= 0.0013 + 0.0075 \\ &= 0.0088. \end{aligned}$$

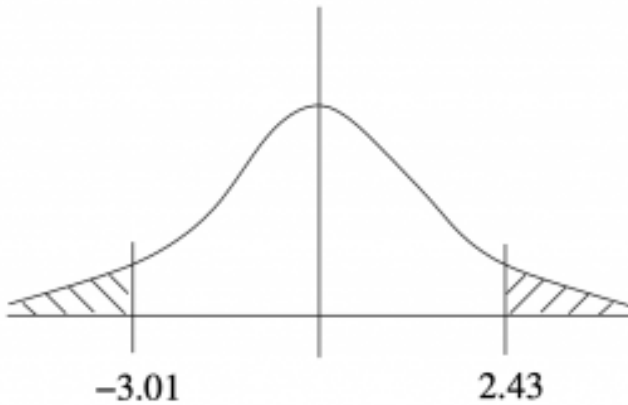


Figure 5.18: The situation for Example 5.8.

□

Using the Standard Normal Distribution Table backwards

Up until now we've used the **Standard Normal Distribution Table** directly. For a given z , we look up the area $A(z)$. Now we look at how to use it backwards: We have a number that represents the area between 0 and z , what is z ? Let's illustrate this process with an example.

Example 5.9 : We are given an area $P = 0.2123$ as shown in Figure 5.19. What is $\{z\}$?

Solution : Look in the **Standard Normal Distribution Table** for the

closest value to the given P . In this case 0.2123 corresponds exactly to $z = 0.56$.

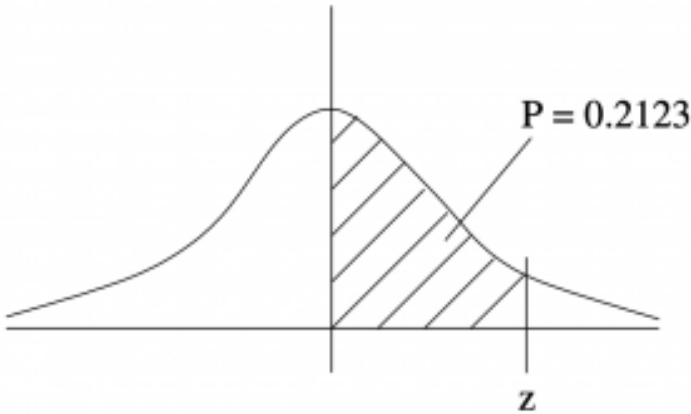


Figure 5.19: The situation for Example 5.9.

□

Example 5.9 was artificial in that the given area appeared exactly in the **Standard Normal Distribution Table**. Usually it doesn't. In that case pick the nearest area in the table to the given number and use the z associated with the nearest area. This, of course, is an approximation. For those who know how, linear interpolation can be used to get a better approximation for z .

The z -transformation preserves areas

In a given situation of sampling a normal population, the mean and standard deviation of the population are not necessarily 0 and 1. We have just learned how to compute areas under a standard normal curve. How do we compute areas under an arbitrary normal curve? We use the z -transformation. If we denote the original normal distribution by $P(x)$ and the z -transformed distribution by $P(z)$ then areas under $P(x)$ will be transformed to areas under $P(z)$ that are the same. *The z -transformation preserves areas.* So we can compute areas, or probabilities under $P(z)$ using

the **Standard Normal Distribution Table** and instantly have the probabilities we need for the original $P(x)$. Let's follow an example.

Example 5.10 : Suppose we know that the amount of garbage produced by households follows a normal distribution with a mean of $\mu = 28$ pounds/month and a standard deviation of $\sigma = 2$ pounds/month. What is the probability of selecting a household that produces between 27 and 31 pounds of trash/month?

Solution : First convert $x = 27$ and $x = 31$ to their z -scores:

$$z_1 = z(27) = \frac{27 - 28}{2} = \frac{-1}{2} = -0.5$$

$$z_2 = z(31) = \frac{31 - 28}{2} = \frac{3}{2} = 1.5$$

Then, referring to Figure 5.20, we see that the probability is $P = A(0.5) + A(1.5) = 0.1915 + 0.4332 = 0.6247$

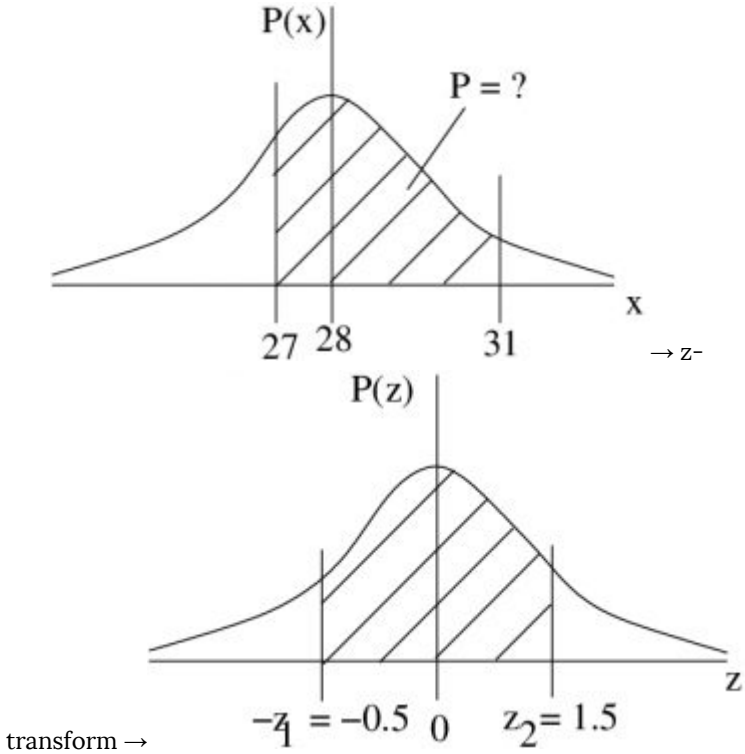


Figure 5.20 : The situation of Example 5.10. Left is the given population, $P(x)$. On the right is the z -transformed version of the population $P(z)$. The value 27 is z -transformed to -0.5 and 31 is z -transformed to 1.5.

□

In Example 5.10 we used the **Standard Normal Distribution Table** directly. You will also need to know how to solve problems in which you use this table backwards. The next example shows how that is done. For this kind of problem you will find the z first and then you will need to find x using the inverse z -transformation :

$$x = z \cdot \sigma + \mu.$$

which is derived by solving the z -transformation, $z = \frac{x-\mu}{\sigma}$ for x .

Example 5.11 : In this example we work from given P . To be a police person you need to be in the top 10\% on a test that has results that follow a normal distribution with an average of $\mu = 200$ and $\sigma = 20$.

What score do you need to pass?

Solution : First, find the z such that $P = P(y > z) = 0.10$ $z) = 0.10$ title="Rendered by QuickLaTeX.com" height="18" width="163" style="vertical-align: -4px;". That P is a right tail area (Case 2), so we need $A(z) = 0.4$, look at Figure 5.21 to see that. Then, going to the **Standard Normal Distribution Table**, look for 0.4 in the middle of the table then read off z backwards. The closest area is 0.3997 which corresponds to $z = 1.28$. Using the inverse z -transformation, convert that z to an x :

to get

$$x = 1.28 \times 20 + 200 = 25.60 + 200 = 225.60$$

or, rounding, use $x = 226$. There are frequently consequences to our calculations and in this case we want to make sure that we have a score that guarantees a pass. So we round the raw calculation up to ensure that.

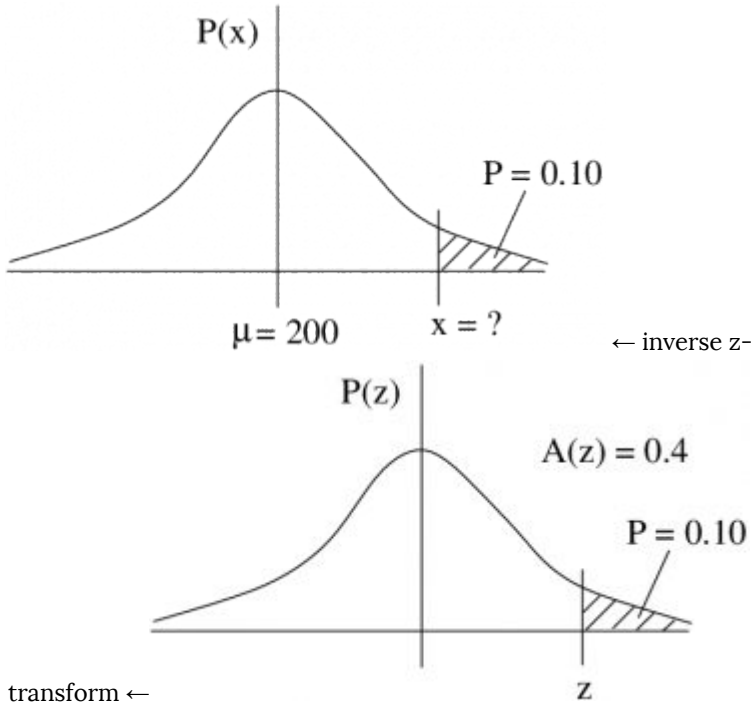


Figure 5.21 : The situation of Example 5.11



6. PERCENTILES AND QUARTILES

The concept of percentile¹ applies to either a data set (sample, as represented by a histogram – a discrete distribution) or to a continuous distribution (which represents a population) as shown in Figure 6.1.

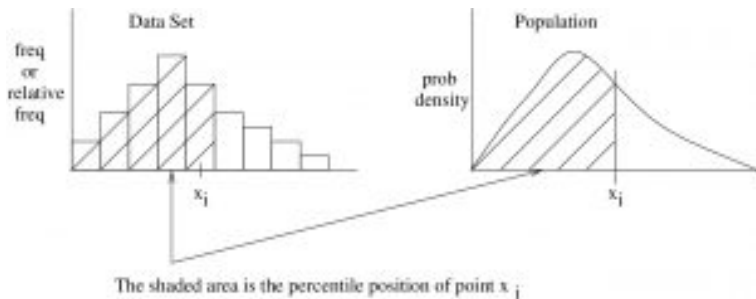


Figure 6.1: The concept of percentile applies to either a data set or to a continuous distribution.

The *percentile* position of the data point x_i , denoted here by $P(x_i)$, is the percentage of the area under the curve up to the

1. This percentile stuff is all about cumulative frequency or (thinking about probabilities) cumulative relative frequencies. The corresponding probability functions are called Cumulative Distribution Functions or CDFs. You will encounter CDFs in SPSS; they are mentioned later in this chapter.

point x_i . *Notation warning:* Do not confuse percentile and probability, we use P to denote both!! (They are related though.)

To determine the percentile position for x_i from a normal distribution of values, convert x_i to z_i via the z -transformation, determine the area under the standard normal curve up to z_i and multiply by 100. We have, therefore, already seen how to compute $P(x_i)$ given x_i or how to compute x_i for a given percentile P . See Case 5 in Section 5.3 and remember how to use the **Standard Normal Distribution Table** forward and backwards.

6.1 Discrete Data Percentiles and Quartiles

Before we get into how to calculate percentile in a data set, note that we can see percentiles directly on a cumulative frequency plot, see Figure 6.2.

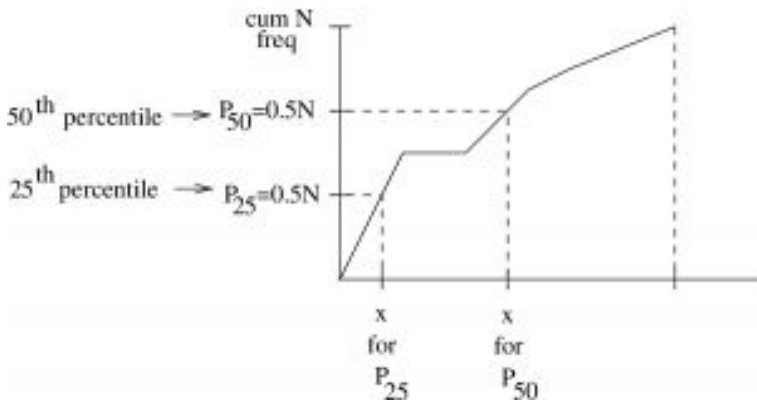


Figure 6.2 : With a cumulative frequency plot, we can read percentiles off the y axis. If you have a newborn baby and take it to the doctor for their first check up, they will measure the baby's head circumference and tell you the baby's head size percentile by looking at such a chart. The doctor's chart will be based on an accumulation of a very large number of essentially population data. Cumulative frequency graphs, or more exactly cumulative probability graphs, can be made for continuous distributions like the normal distribution. The resulting function is the Cumulative Distribution Function, or CDF, and is, for example, $P(z)$ represents the z -distribution then CDF $(x) = \int_0^x P(z)dz$. We will see this CDF in SPSS.

Computing percentile positions of discrete data. Let i be the ordered position of a data set of n data points, then we define the percentile position of x_i to be

$$(6.1) \quad P(x_i) = \frac{(i - 1)}{(n - 1)} \times 100.$$

This formula has the property that $P(x_1 = L) = 0$ and $P(x_n = H) = 100$. It is what we will use as a percentile formula but it is not the only one. Look at Figure 6.1. The way the histogram there is shaded the formula would be $P(x_i) = \frac{i}{n} \times 100$ which would have the property that $P(L) = \frac{100}{n}$ and $P(H) = 100$. There are other, not necessarily wrong, ways to define the percentile position of discrete data but we will use Equation 6.1.

If you want to find the position, i , of the data point corresponding to a given percentile P then compute

$$(6.2) \quad i = \left[\frac{P \times (n - 1)}{100} \right] + 1.$$

Equation (6.2) is derived by solving Equation (6.1) for i . Note that Equation (6.2) gives the *position* of the data point x_i , not its value. To clarify that, let's look at an example.

Example 6.1 : Consider the dataset given below. Data would originally be given as the numbers in the first line. So the first step in answering any question about percentiles is to order the data, the same as what you need to do to determine the median of a dataset. Once the data are ordered, then you may assign a position number to each data point as shown in the third line.

original data	18	15	12	6	8	2	3	5	20	10
ordered data	2	3	5	6	8	10	12	15	18	20
i	1	2	3	4	5	6	7	8	9	10
$n = 10$										

Q : What is the percentile rank of $x_i = 12$?

A : $i = 7$ so

$$P(12) = P(x_7) = \frac{(7-1)}{10-1} \times 100 = \frac{6}{4} \times 100 = 67^{\text{th}}$$

percentile.

Q : What is the value corresponding to the 25th percentile, P_{25} ?

A :

$$i = \left[\frac{P \times (n-1)}{100} \right] + 1 = \left[\frac{(25) \times (10-1)}{100} \right] + 1 = \left[\frac{25 \times 9}{100} \right] + 1 = 2.25 + 1 = 3.25$$

The closest i is 3 and $x_3 = 5$. We can write $P_{25} = 5$.

□

Decile :

$D(x_i) \equiv$ The decile of data value x_i in the *ordered* position i is defined as

$$D(x_i) = \frac{P(x_i)}{10} \qquad 0 \leq D(x_i) \leq 10$$

We will not make much use of decile except to see that quartile is defined in the same way.

Quartile :

$Q(x_i) \equiv$ The quartile of data value x_i in the ordered position 1.

(6.3)

$$Q(x_i) = \frac{P(x_i)}{25} \qquad 0 \leq Q(x_i) \leq 4$$

Notation : (This notation also applies to P and D .) We write :

$Q_0 \equiv 0^{\text{th}}$ quartile

$Q_1 \equiv 1^{\text{st}}$ quartile

$Q_2 \equiv 2^{\text{nd}}$ quartile

$Q_3 \equiv 3^{\text{rd}}$ quartile

$Q_4 \equiv 4^{\text{th}}$ quartile

Quartiles are useful because we do not have to compute percentile first and then divide by 25 as given by Equation (6.3). Instead, we can use the following handy tricks after ordering our data:

$$Q_2 = \text{MD (median)}$$

$$Q_1 = \text{MD of values less than } Q_2$$

$$Q_3 = \text{MD of values greater than } Q_2$$

$$Q_0 = L$$

$$Q_4 = H$$

Example 6.2 : Example with an even number of data points. With the data *in order*, first find the median, then the medians of the two halves of the dataset :

$$5 \quad 6 \quad 12 \quad 13 \quad 15 \quad 18 \quad 22 \quad 50$$

$$Q_1 = \frac{6+12}{2} = 9$$

$$MD = \frac{13+15}{2} = 14 = Q_2$$

$$Q_3 = \frac{18+22}{2} = 20$$

$$Q_0 = L = 5$$

$$Q_4 = H = 50$$

□

Example 6.3 : Example with an even number of data points. With the data *in order*, first find the median, then the medians of the two halves of the dataset :

$$2 \quad 5 \quad 11 \quad 14 \quad 18 \quad 25 \quad 35$$

$$Q_1 = 5$$

$$MD = 14 = Q_2$$

$$Q_3 = 25$$

$$Q_0 = L = 2$$

$$Q_4 = H = 35$$

□

6.2 Finding Outliers Using Quartiles

We can use quartiles to identify *outliers* or data points that are wildly discrepant with the rest of the data. For this application, we need another definition of data dispersion :

$$\text{Interquartile Range} = IQR = Q_3 - Q_1$$

With the IQR any data value that satisfies:

(a) less than $Q_1 - (1.5 \times IQR)$

or

(b) greater than $Q_3 + (1.5 \times IQR)$

...is considered an outlier. This is one of many ways one can define an outlier. As we will discuss below, it is a robust way of identifying outliers.

Example 6.4 : Consider the data of Example 6.2. We found

$$Q_1 = 9 \quad Q_2 = 14 \quad Q_3 = 20$$

so,

$$IQR = Q_3 - Q_1 = 20 - 9 = 11.$$

Following our rules for finding outliers, we compute:

(a) lower acceptable value limit

$$= Q_1 - (1.5 \times IQR)$$

$$= 9 - (1.5 \times 11)$$

$$= 9 - 16.5 = 7.5$$

(b) upper acceptable value limit

$$= Q_3 + (1.5 \times IQR)$$

$$= 20 + (1.5 \times 11)$$

$$= 20 + 16.5 = 36.5$$

and $50 > 36.5$ so 50 is considered an outlier.

□

6.3 Box Plots

A box plot is a plot that shows Q_1 , Q_3 and MD ($= Q_2$) along with H and L ($= Q_0$ and Q_4) as shown in Figure 6.3. It especially emphasizes the IQR.

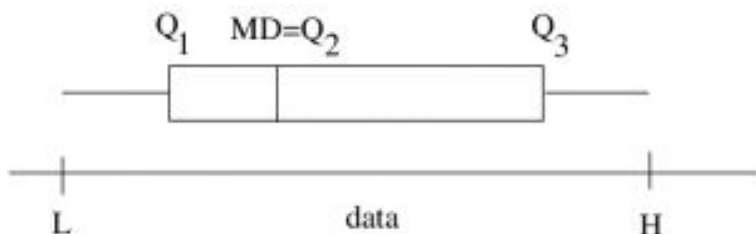


Figure 6.3: The features of a box plot, also known as a box-and-whiskers plot. When one of the whiskers is more than 1.5 times the length of the box (the IQR) then there are outliers by our definition in Section 6.2. The data line shown below the box plot is a construction line and not part of the box plot.

Example 6.5 : Construct a box plot for the data shown in Figure 6.4. Again, someone has done the first, tedious, step of ordering the data for us.

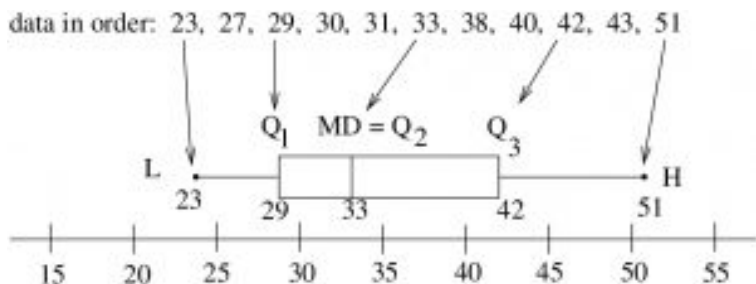


Figure 6.4: Construction of a box plot from the given data.



Box plots can also be drawn vertically. SPSS draws box plots vertically; this is especially useful for comparing datasets.

6.4 Robust Statistics

A *robust statistic* or *resistant statistic* is one that is less affected by outliers than a non-robust or non-resistant statistic. If you look at the numbers in Example 6.2 you can see that the value of the MD (and IQR) is completely unaffected by the value of the outlier data point 50. The mean and the standard deviation will, however, be greatly affected by the value of the outlier. So while some people may identify outliers as those being (say) 3σ from the mean, we see that that is a non-robust way of identifying outliers. In summary:

Measures of central tendency and dispersion	
Robust	Non-robust
MD IQR	\bar{x} s

It would seem that inferential statistics based on robust statistics would be better than statistics based on non-robust values. Maybe. But, traditionally, statistical analysis like the t -tests, ANOVA and regression, are based on the non-robust statistics of means and standard deviations (or variance). People tend to use robust statistics in “Exploratory Data Analysis” (EDA). With EDA one is not concerned so much with testing hypothesis as in trying to get an understanding of general trends in the data. The techniques, and statistics, the fall under the two categories are:

Traditional	Exploratory Data Analysis (EDA)
Frequency Tables Histogram Mean, \bar{x} Standard Deviation, s	Stem and Leaf Plot Box Plot Median, MD Interquartile Range, IQR

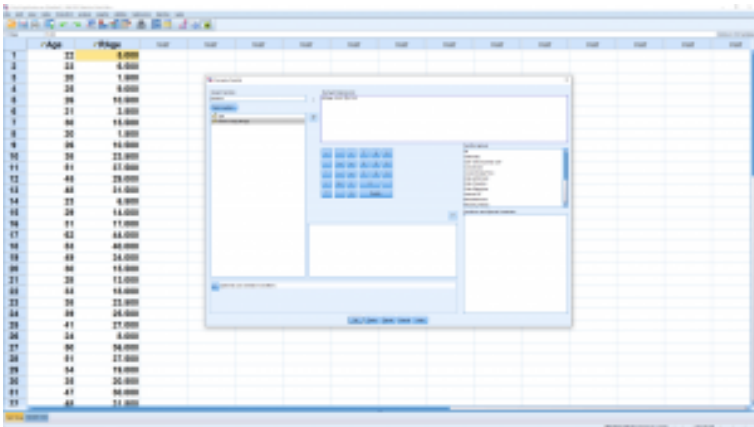
You will find an EDA menu under Analyze → Descriptives in SPSS.

This produces the ranking variable RAge, visible in the Data View window.

	Age	Wage
1	22	8,000
2	22	8,000
3	25	1,000
4	28	8,000
5	36	14,000
6	21	2,000
7	36	14,000
8	20	1,000
9	36	14,000
10	36	14,000
11	31	22,000
12	42	28,000
13	33	9,000
14	33	9,000
15	28	14,000
16	31	11,000
17	42	14,000
18	43	24,000
19	31	11,000
20	39	11,000
21	39	11,000
22	34	14,000
23	36	24,000
24	41	27,000
25	34	8,000
26	36	14,000
27	36	14,000
28	31	27,000
29	34	19,000
30	33	20,000
31	47	34,000
32	45	31,000

SPSS screenshot © International Business Machines Corporation.

Now use that ranking variable in Equation 6.1 by pulling up Transform → Compute Variable :



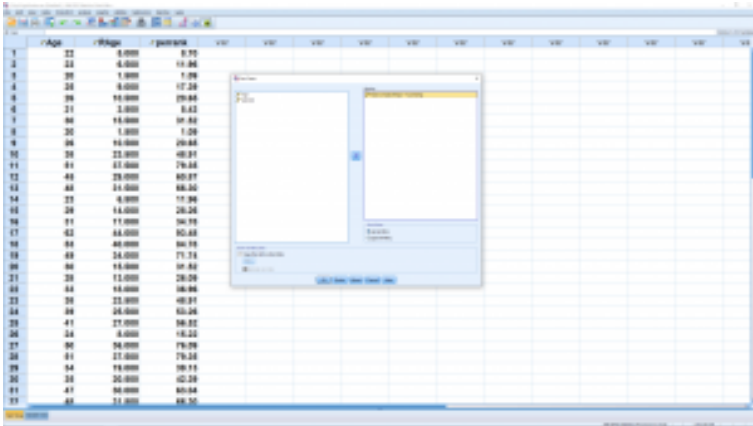
SPSS screenshot © International Business Machines Corporation.

The result, in the Data View window, looks like :

	Age	Wage	percentile															
1	22	0.0000	0.00															
2	22	0.0000	0.00															
3	20	1.0000	1.00															
4	20	0.0000	0.00															
5	00	00.0000	00.00															
6	21	2.0000	0.03															
7	00	00.0000	00.00															
8	20	1.0000	1.00															
9	00	00.0000	00.00															
10	20	22.0000	40.01															
11	01	01.0000	00.00															
12	00	20.0000	00.00															
13	00	24.0000	00.00															
14	22	0.0000	0.00															
15	20	0.0000	00.00															
16	21	11.0000	00.00															
17	02	00.0000	00.00															
18	00	00.0000	00.00															
19	00	20.0000	00.00															
20	20	12.0000	00.00															
21	00	00.0000	00.00															
22	20	20.0000	40.01															
23	00	20.0000	00.00															
24	00	20.0000	00.00															
25	01	21.0000	00.00															
26	20	0.0000	00.00															
27	00	00.0000	00.00															
28	01	21.0000	00.00															
29	00	00.0000	00.00															
30	20	20.0000	00.00															
31	01	00.0000	00.00															
32	00	01.0000	00.00															

SPSS screenshot © International Business Machines Corporation.

We can sort the data on RAGE using Data → Sort Cases :



SPSS screenshot © International Business Machines Corporation.

Note how the smallest value has percentile rank 0. If you scroll to the end of the list you will see that the largest value has percentile rank 100.

	Age	Wage	percentile															
1	20	1.0000	1.00															
2	20	1.0000	1.00															
3	21	1.0000	6.62															
4	21	2.0000	6.62															
5	22	6.0000	9.70															
6	22	6.0000	17.06															
7	24	6.0000	17.06															
8	24	8.0000	18.22															
9	26	9.0000	17.06															
10	26	10.0000	20.68															
11	26	10.0000	20.68															
12	27	12.0000	22.91															
13	28	12.0000	26.09															
14	28	14.0000	26.09															
15	30	16.0000	28.82															
16	30	16.0000	31.92															
17	31	17.0000	26.70															
18	32	18.0000	16.96															
19	34	18.0000	26.70															
20	34	20.0000	42.28															
21	35	20.0000	46.88															
22	35	22.0000	46.88															
23	36	22.0000	46.88															
24	36	24.0000	46.88															
25	36	26.0000	50.29															
26	38	26.0000	52.26															
27	41	27.0000	56.82															
28	42	28.0000	66.70															
29	46	29.0000	66.70															
30	47	28.0000	62.54															
31	48	31.0000	66.70															
32	49	31.0000	66.70															

SPSS screenshot © International Business Machines Corporation.

	Age	Wage	percentile															
1	20	1.0000	1.00															
2	20	1.0000	1.00															
3	21	1.0000	6.62															
4	21	2.0000	6.62															
5	22	6.0000	9.70															
6	22	6.0000	17.06															
7	24	6.0000	17.06															
8	24	8.0000	18.22															
9	26	9.0000	17.06															
10	26	10.0000	20.68															
11	26	10.0000	20.68															
12	27	12.0000	22.91															
13	28	12.0000	26.09															
14	28	14.0000	26.09															
15	30	16.0000	28.82															
16	30	16.0000	31.92															
17	31	17.0000	26.70															
18	32	18.0000	16.96															
19	34	18.0000	26.70															
20	34	20.0000	42.28															
21	35	20.0000	46.88															
22	35	22.0000	46.88															
23	36	22.0000	46.88															
24	36	24.0000	46.88															
25	36	26.0000	50.29															
26	38	26.0000	52.26															
27	41	27.0000	56.82															
28	42	28.0000	66.70															
29	46	29.0000	66.70															
30	47	28.0000	62.54															
31	48	31.0000	66.70															
32	49	31.0000	66.70															

SPSS screenshot © International Business Machines Corporation.

CDF stands for Cumulative Distribution Function. It is literally the cumulative area under a probability distribution function, in this case the normal distribution. So multiplying it by 100 give the percentile rank. The output, in the Data View window looks like :

	Age	gparank	perrank	z1	z2	z3	z4	z5	z6	z7	z8	z9	z10
1	20	1.000	1.00	-1.38000									
2	20	1.000	1.00	-1.38000									
3	21	0.000	0.00	-1.30750									
4	21	1.000	0.00	-1.23500									
5	22	0.000	0.00	-1.16250									
6	22	0.000	10.00	-1.09000									
7	23	0.000	10.00	-1.01750									
8	24	0.000	10.00	-0.94500									
9	24	0.000	10.00	-0.87250									
10	25	0.000	20.00	-0.80000									
11	25	0.000	20.00	-0.72750									
12	27	10.000	20.00	-0.65500									
13	28	1.000	20.00	-0.58250									
14	28	0.000	20.00	-0.51000									
15	28	0.000	20.00	-0.43750									
16	28	0.000	20.00	-0.36500									
17	31	10.000	20.00	-0.29250									
18	31	0.000	20.00	-0.22000									
19	31	0.000	20.00	-0.14750									
20	31	0.000	20.00	-0.07500									
21	33	20.000	20.00	-0.00250									
22	33	0.000	20.00	0.07000									
23	33	20.000	20.00	0.14250									
24	33	20.000	20.00	0.21500									
25	33	20.000	20.00	0.28750									
26	33	20.000	20.00	0.36000									
27	33	20.000	20.00	0.43250									
28	33	20.000	20.00	0.50500									
29	33	20.000	20.00	0.57750									
30	33	20.000	20.00	0.65000									
31	33	20.000	20.00	0.72250									
32	33	20.000	20.00	0.79500									
33	33	20.000	20.00	0.86750									
34	33	20.000	20.00	0.94000									
35	33	20.000	20.00	1.01250									
36	33	20.000	20.00	1.08500									
37	33	20.000	20.00	1.15750									
38	33	20.000	20.00	1.23000									
39	33	20.000	20.00	1.30250									
40	33	20.000	20.00	1.37500									
41	33	20.000	20.00	1.44750									
42	33	20.000	20.00	1.52000									
43	33	20.000	20.00	1.59250									
44	33	20.000	20.00	1.66500									
45	33	20.000	20.00	1.73750									
46	33	20.000	20.00	1.81000									
47	33	20.000	20.00	1.88250									
48	33	20.000	20.00	1.95500									
49	33	20.000	20.00	2.02750									
50	33	20.000	20.00	2.10000									
51	33	20.000	20.00	2.17250									
52	33	20.000	20.00	2.24500									
53	33	20.000	20.00	2.31750									
54	33	20.000	20.00	2.39000									
55	33	20.000	20.00	2.46250									
56	33	20.000	20.00	2.53500									
57	33	20.000	20.00	2.60750									
58	33	20.000	20.00	2.68000									
59	33	20.000	20.00	2.75250									
60	33	20.000	20.00	2.82500									
61	33	20.000	20.00	2.89750									
62	33	20.000	20.00	2.97000									
63	33	20.000	20.00	3.04250									
64	33	20.000	20.00	3.11500									
65	33	20.000	20.00	3.18750									
66	33	20.000	20.00	3.26000									
67	33	20.000	20.00	3.33250									
68	33	20.000	20.00	3.40500									
69	33	20.000	20.00	3.47750									
70	33	20.000	20.00	3.55000									
71	33	20.000	20.00	3.62250									
72	33	20.000	20.00	3.69500									
73	33	20.000	20.00	3.76750									
74	33	20.000	20.00	3.84000									
75	33	20.000	20.00	3.91250									
76	33	20.000	20.00	3.98500									
77	33	20.000	20.00	4.05750									
78	33	20.000	20.00	4.13000									
79	33	20.000	20.00	4.20250									
80	33	20.000	20.00	4.27500									
81	33	20.000	20.00	4.34750									
82	33	20.000	20.00	4.42000									
83	33	20.000	20.00	4.49250									
84	33	20.000	20.00	4.56500									
85	33	20.000	20.00	4.63750									
86	33	20.000	20.00	4.71000									
87	33	20.000	20.00	4.78250									
88	33	20.000	20.00	4.85500									
89	33	20.000	20.00	4.92750									
90	33	20.000	20.00	5.00000									

SPSS screenshot © International Business Machines Corporation.

Note how the percentile ranks of gparank are different from, but close to, the percentile ranks of perrank computed using the data's own distribution. This indicates that the data themselves follow an approximately normal distribution.

6.6 RStudio Lesson 4: Percentiles

[Coming soon]

7. THE CENTRAL LIMIT THEOREM

Before we can learn about confidence intervals in Chapter 8 and hypothesis testing in the Chapter 9, we need a couple of results that form the foundation of the usefulness of the normal distribution. We have mentioned that the normal distribution can be derived as a limit of binomial distributions. This fact can be used in reverse and we can use the normal distribution to approximate the binomial distribution. This approximation will be useful for inferences (confidence intervals and hypothesis testing) on proportions. The second result is the *very important* central limit theorem where the normal distribution pops out as the answer to the characterization of random sample means. The central limit theorem gives us the *sampling theory* for all statistical inference procedures involving means.

7.1 Using the Normal Distribution to Approximate the Binomial Distribution

Recall the definitions: p = probability of success, $q = 1 - p$ = probability of failure and n = sample size. When $np \geq 5$ and $nq \geq 5$ then the normal distribution is very close, numerically, to the binomial distribution.

Using the histogram way of drawing the binomial distribution, a good fit looks like that shown in Figure 7.1.

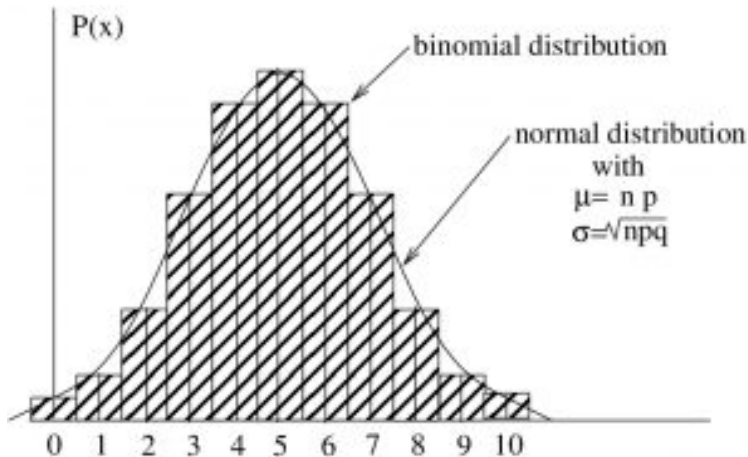


Figure 7.1: A normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$ is a good fit to the binomial distribution with the same mean and standard deviation as long as $np \geq 5$ and $nq \geq 5$.

A couple of things to note about this approximation:

1. Although the values of the normal and the binomial

distributions match well at x equal to integer values when $np \geq 5$ and $nq \geq 5$, the areas match not as well. A “correction for continuity” can be used to better make the areas match but we won’t be worrying about such fine details in our studies.

2. We will use the normal approximation to the binomial make inferences on proportions. In that case p , the probability of success will represent a proportion in a population.

7.2 The Central Limit Theorem

Now we come to the *very important* central limit theorem. First, let's introduce it intuitively as a process :

1. Suppose you have a large population (in theory infinite) with mean μ and standard deviation σ (and any old shape).
2. Suppose you have a large sample, size n , of values from that population. (In practise we will see that $n > 30$ is large.)
Take the mean, \bar{x}_1 , of that sample. Put the sample back into the population¹
3. Randomly pick another sample of size n . **Compute the mean of the new sample, \bar{x}_2 .** Return the sample to the population..
4. Repeat step 3 an infinite number of times and **build up your collection of sample means \bar{x}_i .**
5. Then² the distribution of the sample means will be *normal* will have a mean equal to the population mean, μ , and will have a standard deviation of

1. This is redundant since the population is infinite, but for conceptual purposes imagine that you return the items to the population.
2. More precisely, the distribution of sample means asymptotically approaches a normal distribution as $n \rightarrow \infty$. But 30 is close enough to infinity for most practical purposes and the statistical inferential tests that we will study will assume that the distribution of sample means will be normal.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population's standard deviation.
 $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is known as the *standard error of the mean*.

Now let's visualize this same process using pictures :

- Take a sample of size n from the population and compute the mean \bar{x} (see Figure 7.2a).

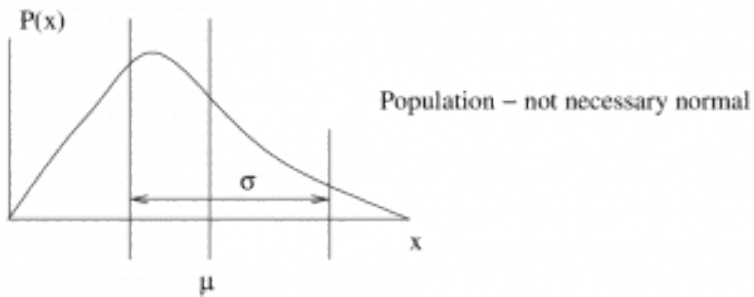


Figure 7.2a

- Put them back and take n more data points.
- Do this over and over to get a bunch of values for \bar{x} . Those values for \bar{x} will be distributed as shown in Figure 7.2b.

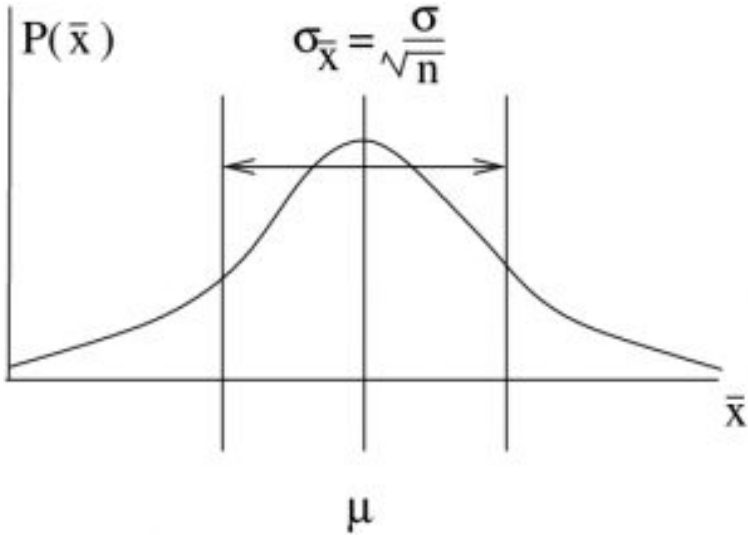


Figure 7.2b

The central limit theorem is our *fundamental sampling theory*. It tells us the *if* we know what the mean and standard deviation of a population³ are *then* we can assign the probabilities of getting a certain mean \bar{x} in a randomly selected sample from that population via a normal distribution of sample means that has the same mean as the population and a standard deviation equal to the standard error of the mean.

To apply this central limit theorem sampling theory we will need to compute areas P under the normal distribution of means. In order to do that, so we can use the **Standard Normal Distribution Table**, we need to convert the values (\bar{x}) to a standard normal

3. In hypothesis testing we know what the mean of the population in the null hypothesis is.

z using the z -transformation as usual: $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$. So, for the distribution of sample means the appropriate z -transformation is :

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{n}}$$

Example 7.1 : Assume that we know, say from SGI's database, that the mean age of registered cars is $\mu = 96$ months and that the population standard deviation of the cars is $\sigma = 16$ months. We make no assumption about the shape of the population distribution. Then, what is the answer to the following sampling theory question: What is the probability that the mean age is between 90 and 100 months in a sample of 36 cars?

Solution : The central limit theorem tells us that sample means will be distributed as shown in Figure 7.3.

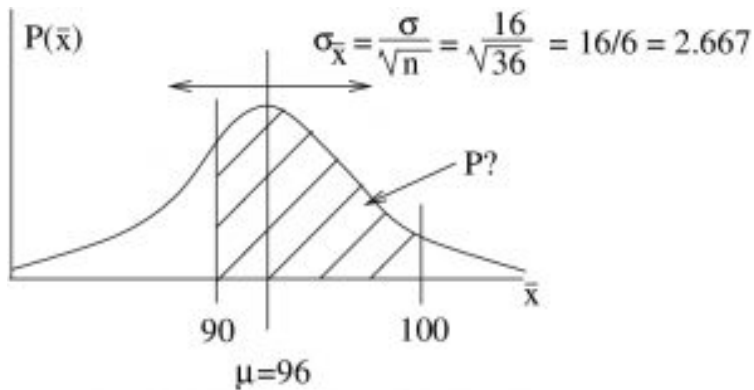


Figure 7.3 : Distribution of mean age from samples of 36 cars.

Convert 90 and 100 to z -scores as usual:

$$z(90) = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} = \frac{90 - 96}{2.667} = -2.25$$
$$z(100) = \frac{100 - 96}{2.667} = 1.50$$

Then, the required probability using the **Standard Normal Distribution Table** is

$$\begin{aligned} P &= A(2.25) + A(1.50) \\ &= 0.4878 + 0.4332 \\ &= 0.921 \quad (92.1\%) \end{aligned}$$

□

8. CONFIDENCE INTERVALS

8.1 Confidence Intervals Using the z-Distribution

With confidence intervals we will make our first statistical inference. Confidence intervals give us a direct inference about the population from a sample. The probability statement is one about hypotheses about the mean μ of the population based on the mean \bar{x} and standard deviation s of the sample. This is a fine point. The frequentist definition of probability gives no way to assign a probability to a hypothesis. How do you count hypotheses? The central limit theorem makes a statement about the sample means \bar{x} on the basis of a hypothesis about a population, about its mean μ and standard deviation σ . If the population is fixed then the central limit theorem gives the results of counting sample means, frequentist probabilities. If we let H represent a hypothesis about a population (i.e. that it is described by μ and σ) and let D represent data (with mean \bar{x}) then the central limit theorem gives the probability $P(D | H) = P(\bar{x} | \mu, \sigma)$. The confidence intervals that we'll look at first give $P(H | D) = P(\mu | \bar{x}, \sigma)$. We'll look at the recipe for computing confidence intervals for means first, then return to this discussion about probabilities for hypotheses.

Our goal is to define a symmetric interval about the population mean μ that will contain all potentially measured values of \bar{x} with a probability¹ of \mathcal{C} .

1. Because of this issue about probabilities of hypotheses, many prefer to say "confidence" and not probability. But we will learn enough about Bayesian probability to say "probability".

Typically C will be

$$C = 0.90 \quad (90\% \text{ confidence})$$

$$C = 0.95 \quad (95\% \text{ confidence})$$

$$C = 0.99 \quad (99\% \text{ confidence})$$

The assumptions that we need in order to use the z -distribution to compute confidence intervals for means are :

1. The population standard deviation, σ , is known (a somewhat artificial assumption since it is usually not known in an experimental situation) or
2. The sample size is greater than (or equal to) 30, $n \geq 30$ and we use $\sigma = s$, the sample standard deviation in our confidence interval formula.

Definition : Let $z_C = z_{\alpha/2}$ where $C = 1 - \alpha$ be the z -value, from the **Standard Normal Distribution Table** that corresponds to an area, between 0 and z_C of $C/2$ as shown in Figure 8.1.

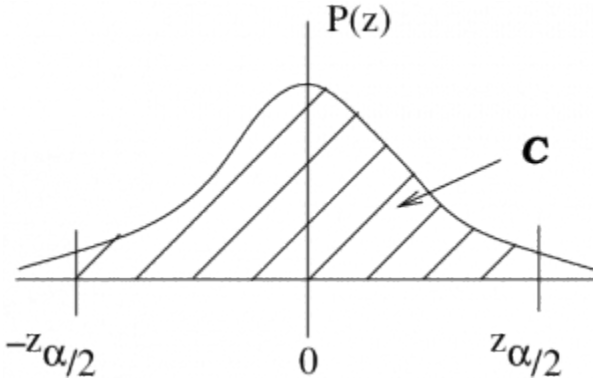


Figure 8.1: The z -distribution areas of interest associated with $z_C = z_{\alpha/2}$.

To get our confidence interval we simply inverse z -transform the picture of Figure 8.1, taking the mean of 0 to the sample mean \bar{x} and

the standard deviation of 1 to the standard error σ/\sqrt{n} as shown in Figure 8.2.

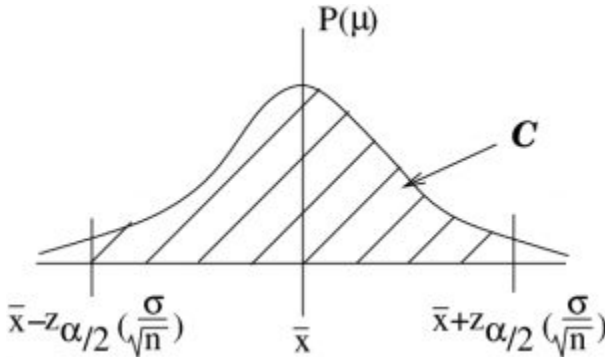


Figure 8.2 : The inverse Z -transformation of Figure 8.1 gives the confidence interval for μ .

So here is our recipe from Figure 8.2. The \mathcal{C} -confidence interval for the mean, under one of the two assumptions given above, is :

$$\bar{x} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{x} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

or using notation that we will use as a standard way of denoting symmetric confidence intervals

$$(8.1) \quad \bar{x} - E < \mu < \bar{x} + E$$

where

$$E = z_{\mathcal{C}} \left(\frac{\sigma}{\sqrt{n}} \right).$$

The notation $z_{\mathcal{C}}$ is more convenient for us than $z_{\alpha/2}$ because we will use the **t Distribution Table** in the [Appendix](#) to find $z_{\mathcal{C}}$ very quickly. We could equally well write

$$\mu = \bar{x} \pm E$$

but we will use Equation (8.1) because it explicitly gives the bounds for the confidence interval.

Notice how the confidence interval is *backwards* from the picture that the central limit theorem gives, the picture shown in Figure 8.3. We actually had no business using the inverse z -transformation $\mu = (z - \bar{x})/(\sigma/\sqrt{n})$ to arrive at Figure 8.2. It reverses the roles of μ and \bar{x} . We'll return to this point after we work through the mechanics of an example.

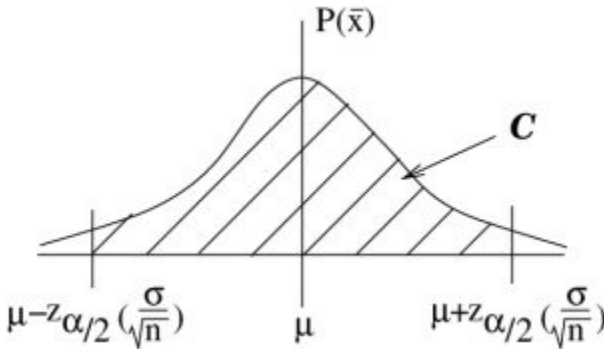


Figure 8.3 : The central limit theorem is about distributions of sample means.

Example 8.2 : What is the 95% confidence interval for student age if the population σ is 2 years, sample $n = 50$, $\bar{x} = 23.2$?

Solution : So $C = 0.95$. First write down the formula prescription so you can see with numbers you need:

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = z_{95\%} \frac{\sigma}{\sqrt{n}}.$$

First determine $z_C = z_{\alpha/2}$. With the tables in the Appendices, there are two ways to do this. The first way is to use the **Standard Normal Distribution Table** noting that we need the z associated with a table area of $0.95/2 = 0.475$. Using the table backwards we find $z_C = 1.96$. The **second way**, the recommended way

especially during exams, is to use the **t Distribution Table**. Simply find the column for the 95% confidence level and read the z from the last line of the table. We quickly find $z_{95\%} = 1.960$.

Either way we now find

$$E = 1.96\left(\frac{2}{\sqrt{50}}\right) = 0.6$$

so

$$\bar{x} - E < \mu < \bar{x} + E$$

$$23.2 - 0.6 < \mu < 23.2 + 0.6$$

$$22.6 < \mu < 23.8$$

with 95% confidence.



8.2 **Bayesian Statistics

Now that we've seen how easy it is to compute confidence intervals, let's give it a proper probabilistic meaning. To extend probability from the frequentist definition to the Bayesian definition, we need Bayes' rule. Bayes' rule is, for events A and B :

$$P(A | B)P(B) = P(B | A)P(A).$$

Study Figure 8.4 to convince yourself that Bayes' rule is true. Notice that

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(B | A) = \frac{P(A \cap B)}{P(A)}.$$

So, equating $P(A \cap B)$ from each of those two perspectives, we get Bayes' rule.

If we let $A = H$ (hypothesis) and $B = D$ (data), Bayes' rule gives us a way to define the probability of hypothesis through

$$(8.2) \quad P(H | D) = P(D | H) \left[\frac{P(H)}{P(D)} \right].$$

The quantity $[P(H)/P(D)]$ is known as the *prior probability* of the data relative to the hypothesis and is something that can be computed in theory if probabilities are assigned in a reasonable manner. The specification of prior probabilities is a contentious issue with the Bayesian approach. Really, it represents a *prior belief*. The quantity $P(D | H)$ is what sampling theory, like the central limit theorem, gives and is known as the *likelihood*. Finally the quantity $P(H | D)$ is known as the *posterior probability*. Equation (8.2) is an expression about probability

distributions as well as individual probabilities (just allow H and D to vary).

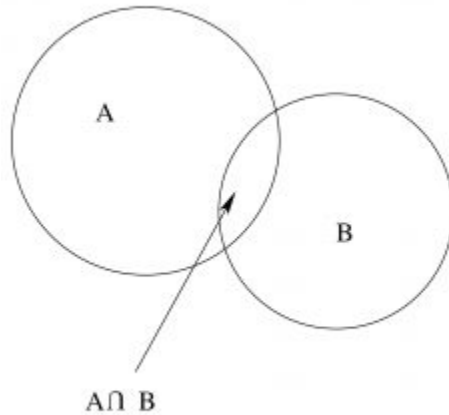


Figure 8.4 : Venn diagram illustration of Bayes rule.

If we assign $[P(H)/P(D)] = 1$ for the prior probability then $P(H | D) = P(D | H)$. We can switch the roles of D and H ! Of course $[P(H)/P(D)] = 1$ is not a probability distribution because the area under a function whose value is always 1 is infinite. The area under a probability distribution must be 1. So $[P(H)/P(D)] = 1$ is an *improper distribution* (as a function of either H or D). But note that an improper distribution times a proper distribution here gives rise to a proper distribution. With this slight of hand, we can give confidence intervals a probabilistic interpretation.

8.3 The t -Distributions

As a broad introduction, the t -distributions are family of distributions that give different approximations to the z -distribution as shown in Figure 8.5.

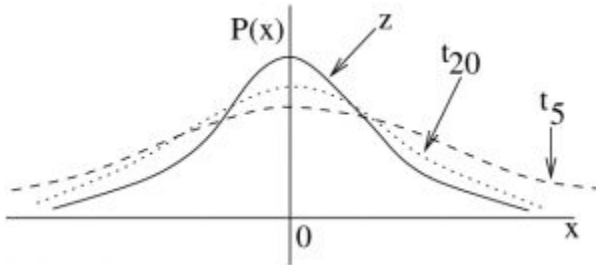


Figure 8.5: The t -distributions are a family of distributions, labeled here by their degrees of freedom ν as in t_ν .

As the degrees of freedom, ν , increases, t_ν become closer to z , $\lim_{\nu \rightarrow \infty} t_\nu = z$. In practice, as reflected in the **t Distribution Table**, t_{30} is very very close to z .

The t -distributions arise as a corollary to the central limit theorem; they give the distribution of sample means when knowledge of the population σ is replaced by using the sample mean s . When we encounter the χ^2 distribution later, we will give a more exact mathematical specification of the t -distributions.

Similar, to the z -distribution case, the C confidence interval for the mean μ for small n samples is given by

$$\bar{x} - E < \mu < \bar{x} + E$$

where, now

$$E = t_{\nu, C} \left(\frac{s}{\sqrt{n}} \right).$$

With this new formula for E we have replaced σ with s in

comparison with the formula we used in [Section 8.1: Confidence Intervals using the z-distribution](#) and, of course, replaced z_C with $t_{\nu,C}$. Some books use $t_{\nu,C} = t_{\nu,\alpha/2}$ like the z_C of Section 8.1. We use $t_{\nu,C}$ because we'll look up its value in the **t Distribution Table** in the column for C confidence intervals (just like we did with z) and with the degrees of freedom ν specifying the row. The formula for the degrees of freedom in this case is :

$$\nu = n - 1.$$

The $t_{\nu,C}$ specify a probability C as shown in Figure 8.6. As before, the inverse z -transform, in the form $x = t_{\nu,C}s + \bar{x}$ from the t -distribution on the left of Figure 8.6 to the distribution on the right of Figure 8.6 leads to our confidence interval formula for small means. And as before we should justify using that transform from a Bayesian perspective.

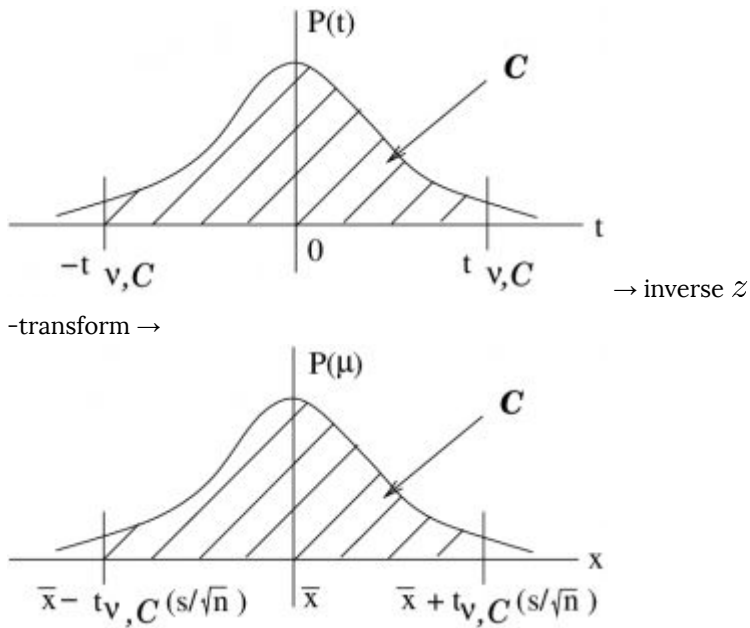


Figure 8.6 : Derivation of confidence intervals for means of small samples.

Example 8.2 : Given the following data:

5460 5900 6090 6310 7160 8440 9930

find the 99% confidence interval for the mean.

Solution : First count $n = 7$ and then, with your stats calculator compute

$$\bar{x} = 7041.4 \quad \text{and} \quad s = 1610.3.$$

Using the **t Distribution Table** with $\nu = n - 1 = 6$ in the 99% confidence interval column, find

$$t_{n-1, C} = t_{6, 99\%} = 3.707.$$

With these numbers, compute

$$E = t_{n-1, C} \left(\frac{s}{\sqrt{n}} \right) = 3.707 \left(\frac{1610.3}{\sqrt{7}} \right) = 2256.2$$

so

$$\bar{x} - E < \mu < \bar{x} + E$$

$$7041.4 - 2256.2 < \mu < 7041.4 + 2256.2$$

$$4785.2 < \mu < 9297.6$$

is the 99% confidence interval for μ .

□

8.4 Proportions and Confidence Intervals for Proportions

We will now make use of the approximation of the binomial distribution by the z -distribution given in [Section 7.1: Using the Normal Distribution to Approximate the Binomial Distribution](#). As usual, the confidence interval will switch the roles of population and sample quantities. The recipe will be laid out first, then we will connect it to what you know about the binomial distribution.

First some definitions. Let X be the number of items in a population of size N that have a given quality. (e.g. the number of females in a population; or the number of people at the U of S wearing yellow sweaters). Then the proportion of the population having the given quality is

$$p = \frac{X}{N}$$

Given a sample from the population of size n , the best estimate for p is:

$$\hat{p} = \frac{x}{n}$$

where x is the number of items in the sample having the given quality. To go along with \hat{p} we also have

$$\hat{q} = 1 - \hat{p}$$

which is the proportion of items in the sample without the given quality.

To compute an \mathcal{C} confidence interval for a proportion p we need to compute

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

and it must be true that both $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$ (otherwise we need to use the binomial distribution directly).

With E , the \mathcal{C} confidence interval for a proportion is given by $\hat{p} - E < p < \hat{p} + E$.

To derive the proportions confidence interval formula we'll begin with the sampling theory given by the binomial distribution and the corresponding z -approximation. Then we'll switch the roles of p and \hat{p} . Let

$$x_{\text{pop}} = \frac{n}{N}X = np$$

be the mean, the expected value, of x that you expect to find in a sample of size n randomly selected from the population with a proportion p of items of interest. This is true because p is also the probability of randomly selecting an item of interest (the probability of success) from the population as per what we did in Chapter 4. The binomial distribution tells you the probability of getting different numbers x of items of interest in your sample given p . The binomial distribution that describes our situation is shown in Figure 8.7; it has a standard deviation of $\sigma = \sqrt{npq}$.

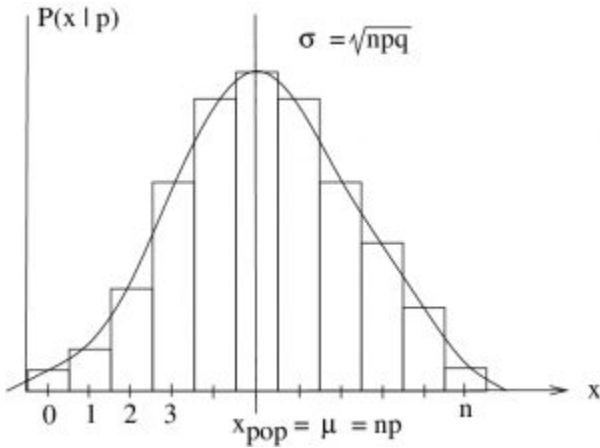


Figure 8.7: The binomial distribution relevant to forming a sample of size N with X items of interest from a population with a proportion P of items of interest. The normal distribution with the same μ and σ is shown.

Moving to the normal approximation, we have the picture of Figure 8.8.

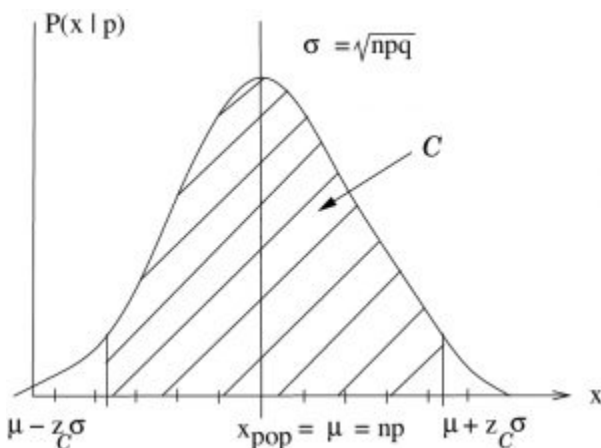


Figure 8.8 : The normal distribution relevant to forming a sample of size n with x items of interest from a population with a proportion p of items of interest. The boundaries of the area C follow from an inverse Z -transform of the Z -distribution to a normal distribution of mean μ and standard deviation σ , $x = z\sigma + \mu$.

Figure 8.8 says :

$$\mu - z_C \sigma < x < \mu + z_C \sigma$$

$$np - z_C \sqrt{npq} < x < np + z_C \sqrt{npq}$$

with a (frequentist) probability of C . This is our sampling theory.

Divide by n :

$$p - z_C \sqrt{\frac{pq}{n}} < \frac{x}{n} < p + z_C \sqrt{\frac{pq}{n}}$$

$$p - z_C \sqrt{\frac{pq}{n}} < \hat{p} < p + z_C \sqrt{\frac{pq}{n}}$$

Swapping the roles of the population and sample, we arrive at the confidence interval formula :

$$\hat{p} - z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Time for a worked example.

Example 8.3 : A sample of 500 nursing applications included 60 men. Find the 90% confidence interval of the true proportion of men who applied to the nursing program.

Solution : From the **t Distribution Table**, look up

$$z_c = 1.65$$

and compute

$$\hat{p} = \frac{x}{n} = \frac{60}{500} = 0.12$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.12 = 0.88$$

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.65 \sqrt{\frac{(0.12) \cdot (0.88)}{500}} = 0.024.$$

Then

$$\hat{p} + E < p < \hat{p} - E$$

$$0.12 + 0.024 < p < 0.12 - 0.024$$

$$0.096 < p < 0.144$$

is the confidence interval with 90% confidence. □

Sample size need for a poll

Measuring proportions is what pollsters do. For example in an election you might want to know how many people will vote for liberals (items of interest) and how many will vote for conservatives (items not of interest)¹ In a news paper you might see: “The poll

1. We assume here that there are only two parties. For the real life situation of more than two parties we need the

says that 72% of the voters will vote liberal. The poll is considered accurate to 2 percentage points 19 time out of 20." This means that the 95% confidence interval ($19/20 = 0.95$) of the proportion of liberal voters is 0.72 ± 0.02 (note how proportions are presented as percentages in the newspaper). The error here is $E = 0.02$. Before the pollster starts telephoning people, she must know how many people to phone to arrive at that goal error of 2%. She needs to know what the sample size n needed is. In general, the minimum sample size needed to attain a goal error E on a confidence interval of C is

$$n = \hat{p}\hat{q} \left(\frac{z_C}{E} \right)^2 .$$

Here \hat{p} and \hat{q} could come from a previous survey if available. If there is no such survey or if you want to be sure of ending up with an error equal to or less than a goal E , then use $\hat{p} = \hat{q} = 0.5$, see Figure 8.9.

multinomial distribution and to approximate it with a multivariate normal distribution. That is a topic for multivariate statistics but the principles are the same as what we cover here.

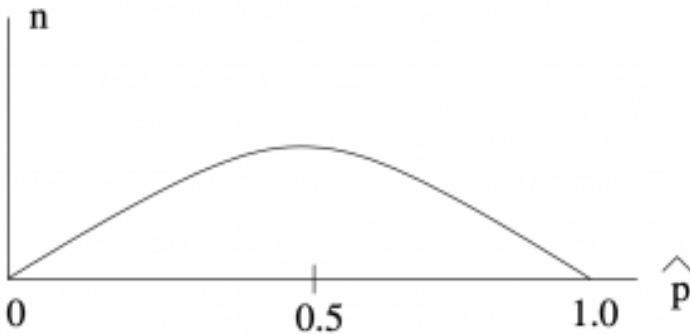


Figure 8.9 : The formula $n = \hat{p}\hat{q} \left(\frac{z_C}{E}\right)^2$ is a quadratic formula.

Substitute $\hat{q} = 1 - \hat{p}$ to get $n = \hat{p}(1 - \hat{p}) \left(\frac{z_C}{E}\right)^2$ or $n = (\hat{p} - \hat{p}^2) \left(\frac{z_C}{E}\right)^2$. The maximum of $n_{\max} = \frac{1}{4} \left(\frac{z_C}{E}\right)^2$ is at $\hat{p} = 0.5$.

Example 8.4 : We want to estimate, with 95% confidence, the proportion of people who own a home computer. A previous study gave an answer of 40%. For a new study we want an error of 2%. How many people should we poll?

Solution : From the question we have :

$$\begin{aligned} \hat{p} &= 0.40, & \hat{q} &= 0.60 \\ E &= 0.02, & \alpha &= 0.95 \end{aligned}$$

From the **t Distribution Table** (or the **Standard Normal Distribution Table** if you think about the areas correctly) we find

$$z_C = z_{95\%} = 1.960.$$

Therefore

$$n = \hat{p}\hat{q} \left(\frac{z_{\alpha/2}}{E} \right)^2 = (0.40)(0.60) \left(\frac{1.96}{0.02} \right)^2 = 2304.96$$

Which we round up to a sample size of 2305 to ensure that $E < 0.02$.

□

8.5 Chi Squared Distribution

The χ^2 (chi squared) distribution is a consequence of a random process based on the normal distribution. It is derived from the normal distribution as the result of the following stochastic process :

1. Suppose you have a population that has variance σ^2 and is normally distributed.
2. Take a sample of size n from the population and compute $x_1 = \frac{(n-1)s_1^2}{\sigma^2}$ using the sample standard deviation s_1 from that sample.
3. Put the sample back into the population.
4. Take another sample of size n from the population and compute $x_2 = \frac{(n-1)s_2^2}{\sigma^2}$ using the sample standard deviation s_2 from that sample.
5. etc.
6. The distribution of the values of $x_i = \frac{(n-1)s_i^2}{\sigma^2}$ values will be a χ^2 distribution with $\nu = n - 1$ degrees of freedom.

Like the t -distributions, the χ^2 distributions are a family, see Figure 8.10.

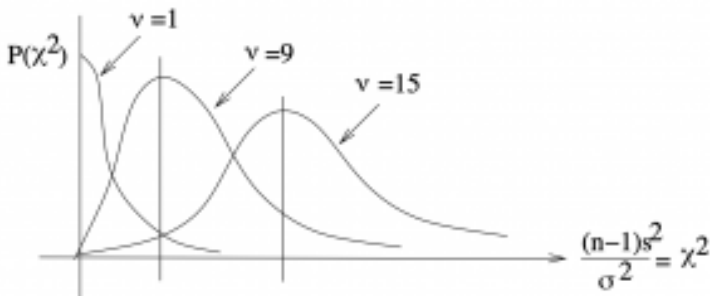


Figure 8.10 : The χ^2 distributions are enumerated by degrees of freedom.

The χ^2 distribution underlies why s is the best estimate for σ . Its mean, or expected value is $\nu = n - 1$ so the expected value of s is σ . The expected value of $\sum(x - \bar{x})/n$ in a random sample of size n is not σ .

Confidence Intervals on σ and σ^2

The χ^2 distribution is already normalized in its definition through including s in its definition. Therefore no z -transforms are needed and we can work directly with a table that gives right tail areas under the χ^2 distribution. That table is the **Chi-squared Distribution Table**, in the [Appendix](#), and it gives values of χ^2 for given values of area to the right of χ^2 , see Figure 8.11.

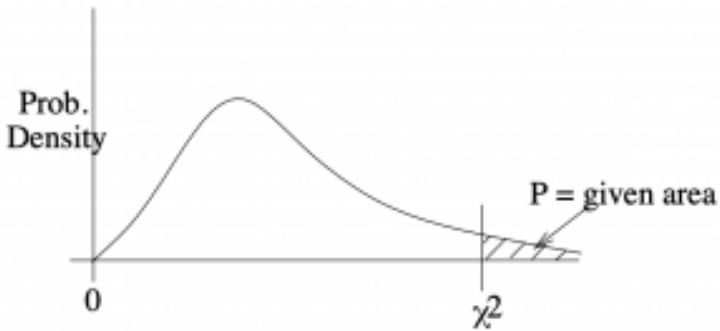


Figure 8.11 : The Chi-squared Distribution Table gives χ^2 associated with given right tail areas.

We'll need χ^2_{left} and χ^2_{right} such that the tail areas are equal and such that the area between them is C , see Figure 8.12.

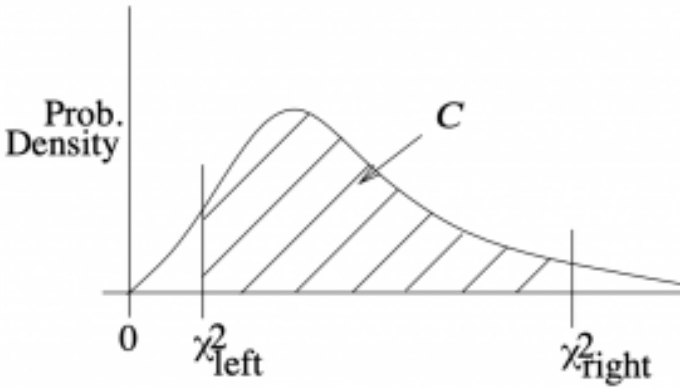


Figure 8.12 : The values χ^2_{left} and χ^2_{right} define the confidence region \mathcal{C} .

Notation : Let's call the α in the **Chi-squared Distribution Table** α_T and let $\chi^2(\alpha_T)$ be the table value that corresponds to α_T . In other words $\chi^2(\alpha_T)$ is the χ^2 value that corresponds to a right tail area of α_T .

So given \mathcal{C} , the appropriate χ^2_{left} and χ^2_{right} are the following values from the **Chi-squared Distribution Table**:

$$\chi^2_{\text{right}} = \chi^2 \left(\frac{1 - \mathcal{C}}{2} \right)$$

$$\chi^2_{\text{left}} = \chi^2 \left(1 - \left[\frac{1 - \mathcal{C}}{2} \right] \right).$$

Note the symmetry of the **Chi-squared Distribution Table**. If χ^2_{right} comes from the column 3 columns from the right edge of the table then χ^2_{left} comes from a column 3 columns from the left edge of the table. Only small and large areas appear in the table, there are no intermediate values.

Finally, the confidence interval for σ^2 is given by

$$\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^{-2}}$$

and for σ by:

$$\sqrt{\frac{(n-1)s^2}{\chi_{\text{right}}^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_{\text{left}}^{-2}}}$$

Where the χ^2 distribution with $\nu = n - 1$ degrees of freedom (giving the line to use in the **Chi-squared Distribution Table**) is used.

Example 8.5 : Find the 90% confidence interval on σ and σ^2 for the following data

59, 54, 53, 52, 51, 39, 49, 46, 49, 48

Solution : Compute, using your calculator :

$$s^2 = 28.2$$

$$\nu = n - 1 = 9.$$

From the **Chi-squared Distribution Table**, in the $\nu = 9$ line, find :

$$\chi_{\text{right}}^2 = \chi^2 \left(\frac{1 - 0.90}{2} \right) = \chi^2(0.05) = 16.919$$

and

$$\chi_{\text{left}}^2 = \chi^2(1 - 0.05) = \chi^2(0.95) = 3.325$$

So

$$\frac{(n-1)s^2}{\chi_{\text{right}}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\text{left}}^2}$$
$$\frac{9 \cdot 28.2}{16.919} < \sigma^2 < \frac{9 \cdot 28.2}{3.325}$$
$$15.0 < \sigma^2 < 76.3$$

with 90% confidence.

Taking square roots:

$$3.87 < \sigma < 8.73$$

with 90% confidence.

□

9. HYPOTHESIS TESTING

The process of hypothesis testing can be simplified into :

1. Transform (“reduce”) your given data into a test statistic that you can locate on probability distribution given by the sampling theory under a null hypothesis (H_0) about the population. (e.g. z , t or χ^2 test statistic).
2. See if your test statistic falls into a critical region of the distribution or not. The critical, or rejection region as we’ll call it, represents an area of low probability that the null hypothesis, H_0 is true. If the test statistic falls in the rejection region, then we make the decision to reject H_0 as the conclusion of the hypothesis test.

Before we define the critical region under the null hypothesis, we need to define what a null hypothesis is. We’ll define two hypotheses, actually, because the null hypothesis needs to be contrasted to its logical opposite :

H_0 : Null Hypothesis, the hypothesis that nothing is going on; no effect; no signal.

H_1 : Alternative Hypothesis, the hypothesis that H_0 is not true; there is an effect; there is a signal.

A good experimental design will be set up so that the effects of interest define H_1 . (Your “claim” will be H_1 .) Why? It’s about signal to noise ratios. A test statistic is literally signal/noise, a signal to noise ratio. When you do not reject H_0 you are saying that there is more noise than signal. When you reject H_0 (essentially accepting H_1) you are saying that there is more signal than noise. Usually you are interested in the signal (also known as an “effect”) so your claim would be H_1 . You perform your experiment to find evidence for H_1 . If you are interested in noise (can happen, for example to test assumptions on which tests are based) then your claim would be H_0 . The examples that follow here don’t follow

these experimentally correct rules for which of H_0 or H_1 should be the claim to emphasize the logical nature of the decision making process. But test statistics are signal to noise ratios and in real life you will be interested in signals.

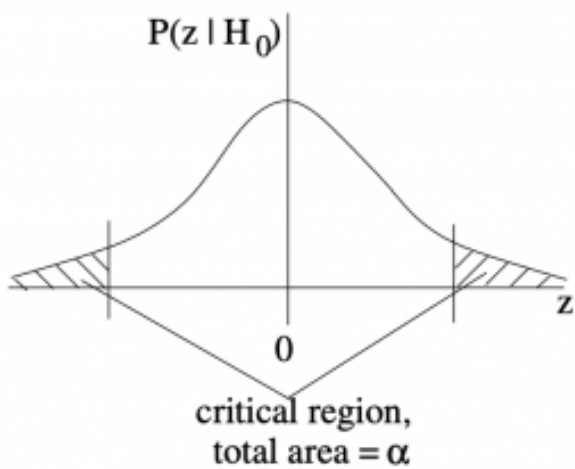
To fix ideas about hypothesis testing, we'll first look at hypotheses on the means of populations (μ). Later we'll consider hypotheses on σ and on p (proportions).

With means there are three combinations of H_0 and H_1 to consider :

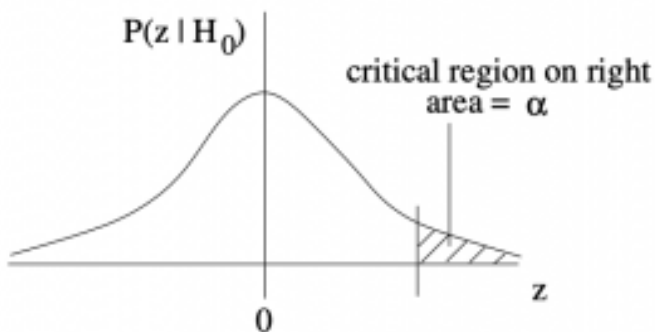
Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu = k$	$H_0: \mu \leq k$	$H_0: \mu \geq k$
$H_1: \mu \neq k$	$H_1: \mu > k$	$H_1: \mu < k$

Here k is a given number. Note that the rightness or the leftness of the one-tailed test is reflected in H_1 . H_1 is generally what people are interested in. Then the critical regions, which are on z distributions as we'll see, for each case look like :

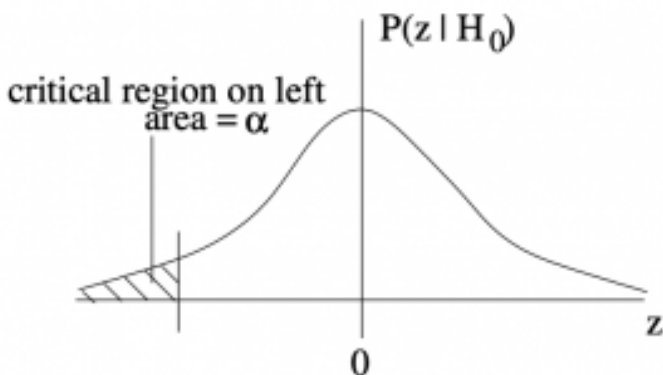
1. Two-tailed test:



2. Right-tailed test:



3. Left-tailed test:



The critical regions, or rejection regions, appear in the probability distributions $P(z | H_0)$, which is the probability distribution that the sample test statistic, z , that would occur if H_0 were true. These z -distributions are z -transforms of the distribution of sample means under H_0 given by the central limit theorem. More about this when we introduce the formula for the z distribution. For now, let's focus on the decision making process.

When your statistic ends up in the critical region, you conclude that H_0 is false. You reject H_0 . The *critical region* is the *rejection region*.

In the two tailed test, the critical region, with total area α is the opposite to the region $C = 1 - \alpha$ that we have been using for confidence intervals. Compare the two-tail critical region sketch above to Figure 8.1.

There are four possible outcomes to a statistical hypothesis test given by the so-called¹ "confusion matrix" :

1. So called not because it is confusing but because you are never 100% sure which decision is correct.

	H_0 true	H_1 true
Reject H_0 (believe H_1)	Type I error α	Correct decision 1- β
Do not reject H_0 (believe H_0)	Correct decision 1- α	Type II error β

The probabilities are relative to the realities. The probabilities in the columns add to 1. The probability of making a Type I error, α , is the area in the critical region. The diagram with the critical region on it assumes that H_0 is the reality. We will see how to compute β in Chapter 13. The quantity $1 - \beta$ is defined as the *power* of the statistical test.

We can view the confusion matrix from a medical test point of view. A medical test is a hypothesis test has the following hypotheses pairs :

H_0 : negative test result, healthy patient

H_1 : positive test result, sick patient

Then :

	Healthy	Sick
Positive Result (believe sick)	Type I error α	Correct decision 1- β
Negative Result (believe healthy)	Correct decision 1- α	Type II error β

In medical tests, the quantity $1 - \alpha$ is known as the test's *specificity*, the probability of finding true negatives. The quantity $1 - \beta$ is the test's *sensitivity*, the probability of finding true positives. Generally α and β are functions of some other decision parameter. In the hypothesis tests that we consider here, α is the decision parameter.

Back to understanding the meaning of hypothesis testing. As we said, a good experimental design will be set up so that H_1 is your favourite theory that there is an effect. In that case H_0 represents the case that there is *no effect* : the position of \bar{x} away from k , or

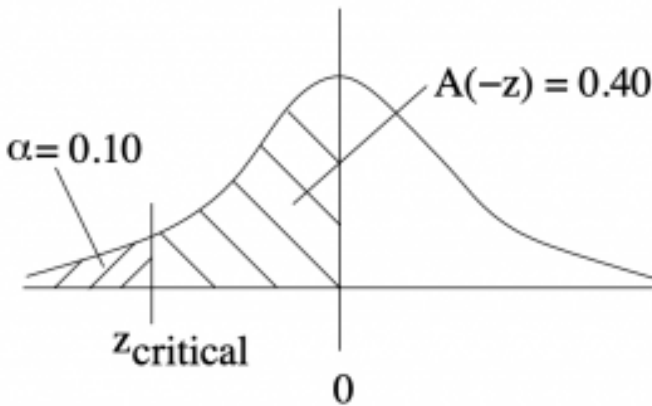
z away from 0 (in the case of hypothesis testing of μ) is just due to noise. If your experiment is then successful in proving your theory, i.e. you reject H_0 , then α represents the probability that you are wrong. The number α actually defines a decision point for rejecting H_0 . Later we will see how to compute a value, p , that is associated with the test statistic. This p -value is then a more refined value for the probability that you are wrong if you reject H_0 . From another point of view, p would be the probability that your measurement is entirely due to noise.

Let's do some examples to build our mechanical skills at defining critical regions for z distributions.

Example 9.1 : Critical Areas on z -distributions with hypothesis testing on the mean, μ .

(a) Left-tailed test with $\alpha = 0.10$. Find the critical value z_{critical} .

First step, draw a picture :



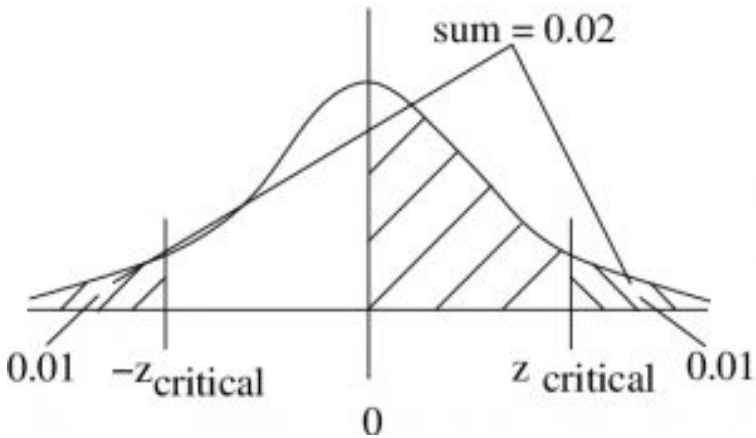
With the tables we have in the [Appendix](#), there are two ways to find z_{critical} :

- Method (a) : Look up area in the **Standard Normal Distribution Table** equal to 0.40 : Closest z is 1.28 so $z_{\text{critical}} = -1.28$.
- Method (b) : Use the last line in the **t Distribution Table** for the one tailed test column. Find a z of 1.282 and add a minus sign because we have a left tail test. So $z_{\text{critical}} = -1.282$.

Use Method (b) on tests and exams. It is faster, requires less thinking about areas (and so less chance for making a mistake) and gives a slightly more accurate result. The critical area or critical region or the rejection region is where $z < -1.282$. The critical value that defines the region in this case is $z = -1.282$.

(b) A two tailed test with $\alpha = 0.02$. Find the critical value z_{critical} .

Draw a picture :



- Method (a) : Look up area in the **Standard Normal Distribution Table** equal to 0.49. The closest z is 2.33. So, because we have a two-tailed test, $z_{\text{critical}} = \pm 2.33$.

- Method (b): Use the last line in the **t Distribution Table**, for two tailed test, $\alpha = 0.02$. Find $z = 2.326$, $z_{\text{critical}} = \pm 2.326$.

Again, Method (b) is the recommended approach.

So the critical *areas* are those where

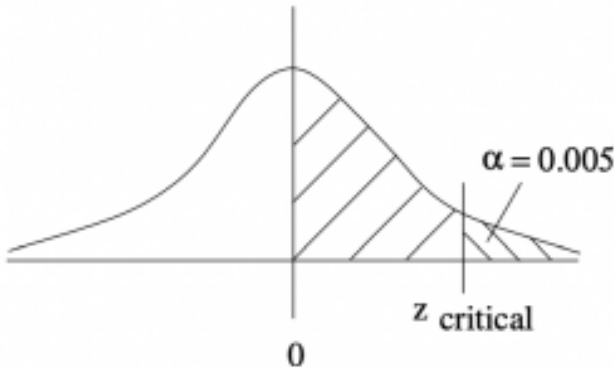
$$z > 2.326 \text{ and } z < -2.326$$

and the critical *values* are $z_{\text{critical}} = 2.326$ and

$$z_{\text{critical}} = -2.326.$$

- (c) A right tailed test with $\alpha = 0.005$. Find the critical value z_{critical} .

Draw a picture :



- Method (a) Look up area in the **Standard Normal Distribution Table** equal to 0.495, the Closest z is 2.58. So $z_{\text{critical}} = 2.58$
- Method (b) Use the last line in the **t Distribution Table** for one tailed test, $\alpha = 0.005$ and find $z_{\text{critical}} = 2.576$.

So the critical area is that where $z > 2.576$

title="Rendered by QuickLaTeX.com" height="13" width="74" style="vertical-align: 0px;"> and the critical value is $z_{\text{critical}} = 2.576$.



One final note on setting up the hypotheses. When setting up the hypotheses H_0 and H_1 , one of the two alternatives will be the *claim* (what the problem says you really want to test). As mentioned before, a good experimental design will have H_1 as the claim. But this may not always be possible to arrange (especially in tests of assumptions). So many of the exercises in the text and assignments will have H_0 as the claim.

9.1 Hypothesis Testing Problem Solving Steps

Now that we have some background on setting up hypotheses and finding critical regions, we introduce the steps needed for every hypothesis testing procedure. Hypothesis testing is based directly on sampling theory and the probabilities $P(\text{test statistic} \mid H_0)$ that the sampling theory gives. Here are the steps we will follow :

1. **Hypotheses** : Formulate H_0 and H_1 . State which is the claim
2. **Critical statistic** : Find the critical values and regions. (Use tables of z , t , χ^2 , etc. values).
3. **Test statistic** : Compute the test statistic from your data. It summarizes your data in one number. The p -value follows from the test statistic.
4. **Decision** : If the test statistic falls in the critical region (rejection region), reject H_0 . (This decision can also be made using the p -value.)
5. **Interpretation** : Summarize results in a sentence and/or present a graphic or table.

The definition of a p -value will be covered below. For now you should know that a computer program (SPSS) will give you a p -value but not a critical statistic. So there is no Step 2 if you use SPSS.

A generic test statistic may be defined by :

$$\text{test value} = \frac{(\text{observed value}) - (\text{expected } H_0 \text{ value})}{\text{standard error}}.$$

The numerator represents a signal or an effect. The denominator

represents noise. Not all test statistics will have this form (e.g. some χ^2 test statistics), but all test statistics represent a signal-to-noise ratio. Much of the tabular output of SPSS gives the numerator and denominator of this generic form with or without the corresponding test statistic.

9.2 z-Test for a Mean

This is our first hypothesis test. Use it to test a sample's mean when :

1. The population σ is known.
2. Or When $n \geq 30$, in which case use $\sigma = s$ in the test statistic formula.

The possible hypotheses are as given in the table you saw in the previous section (one- and two-tailed versions):

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu = k$	$H_0: \mu \leq k$	$H_0: \mu \geq k$
$H_1: \mu \neq k$	$H_1: \mu > k$	$H_1: \mu < k$

In all cases the test statistic is

$$(9.1) \quad z_{\text{test}} = \frac{\bar{x} - k}{(\sigma/\sqrt{n})}.$$

In real life, we will never know what the population σ is, so we will be in the second situation of having to set $\sigma = s$ in the test statistic formula. When you do that, the test statistic is actually a t test statistic as we'll see. So taking it to be a z is an approximation. It's a good approximation but SPSS never makes that approximation. SPSS will always do a t -test, no matter how large n is. So keep that in mind when solving a problem by hand versus using a computer.

Let's work through a hypothesis testing example to get the procedure down and then we'll look at the derivation of the test statistic of Equation (9.1).

Example 9.2 : A researcher claims that the average salary of assistant professors is more than \$42,000. A sample of 30 assistant professors has a mean salary of \$43,260. At $\alpha = 0.05$, test the claim that assistant professors earn more than \$42,000/year (on average). The standard deviation of the population is \$5230.

Solution :

1. Hypothesis :

$$H_0 : \mu \leq 42,000$$

$$H_1 : \mu > 42,000$$

(This is a right-tailed test.)

(This is a right-tailed test.)

2. Critical Statistic.

- Method (a) : Find z such that $A(z) = 0.45$ from the **Standard Normal Distribution Table**: $z_{\text{critical}} = 1.65$; or
- Method (b) : Look up z in the **t Distribution Table** corresponding to one tail $\alpha = 0.05$ (column), and read the last (z) line: $z_{\text{critical}} = 1.645$.

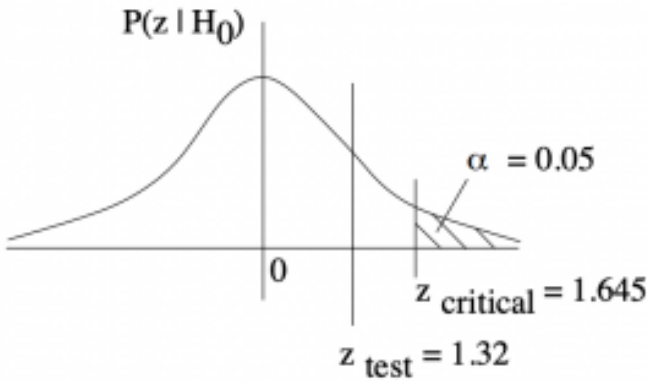
Method (b) is the recommended method not only because it is faster but also because the procedure for the upcoming t -test will be the same for the z -test.

3. Test Statistic.

$$z_{\text{test}} = \frac{\bar{x} - k}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{43260 - 42000}{\left(\frac{5230}{\sqrt{30}}\right)} = 1.32$$

4. Decision.

Draw a picture so you can see the critical region :



So z is in the non-critical region: Do not reject H_0 .

5. Interpretation.

There is not enough evidence, from a z -test at $\alpha = 0.05$, to support the claim that professors earn more than \$42,000/year on average.

□

So where does Equation (9.1) come from? It's an application of the central limit theorem! In Example 9.2, $\bar{x} = 43,260$, $n = 30$, $\sigma = 5230$ and $k = 42,000$ on the null hypothesis of a right-tailed test. The central limit theorem says that if H_0 is true then we can expect the sample means, \bar{x} to be distributed as shown in the top part of Figure 9.1. Setting $\alpha = 0.05$ means that if the actual sample mean, \bar{x} ends up in the tail of the expected (under H_0) distribution of sample means then we consider that either we picked an unlucky 5% sample or the null hypothesis, H_0 , is not true. In taking that second option, rejecting H_0 , we are willing to live with the 0.05 probability that we made a wrong choice – that we made a type I error.

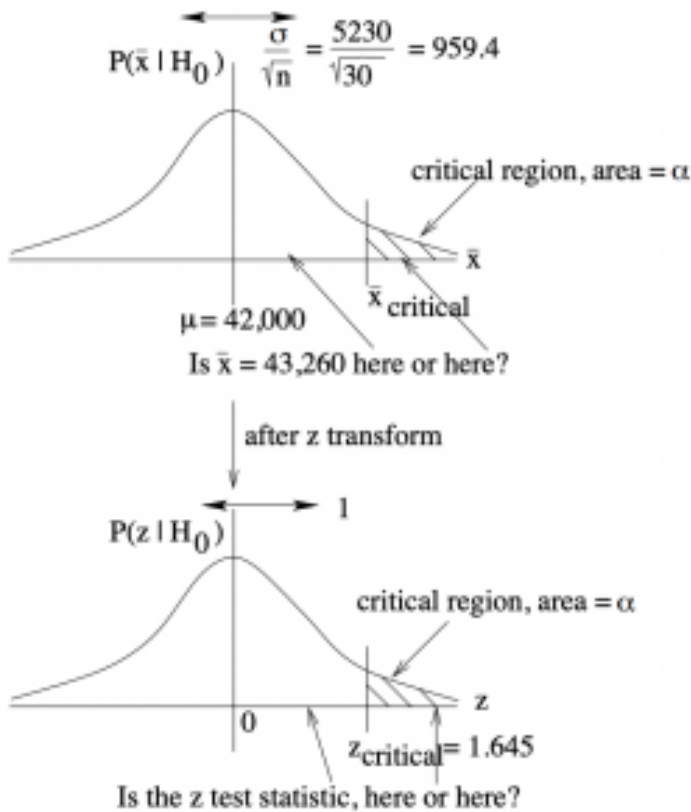


Figure 9.1: Derivation of the Z test statistic.

Referring to Figure 9.1 again, $z_{\text{critical}} = 1.645$ on the lower picture defines the critical region of area $\alpha = 0.05$ (in this case). It corresponds to a value $\bar{x}_{\text{critical}}$ on the upper picture which also defines a critical region of area $\alpha = 0.05$. So comparing \bar{x} to $\bar{x}_{\text{critical}}$ on the original distribution of sample means, as given by the sampling theory of the central limit theorem, is equivalent, after z -transformation, to comparing z_{test} with z_{critical} . That is, z_{test} is the z -transform of the data value \bar{x} , exactly as given by Equation (9.1).

One-tailed tests

From a frequentist point of view, a one-tailed test is a bit of a cheat. You use a one-tailed test when you know *for sure* that your test value or statistic is greater than (or less than) the null hypothesis value. That is, for the case of means here, you know *for sure* that the mean of the population, if it is different from the null hypothesis mean, is greater than (or less than) the null hypothesis mean. In other words, you need some *a priori* information (a Bayesian concept) *before* you do the formal hypothesis test.

In the examples that we will work through in this course, we will consider one-tailed tests when they make logical sense and will not require formal *a priori* information to justify the selection of a one-tailed test. For a one-tail test to make logical sense, the alternate hypothesis, H_1 , must be true on the face value of the data. That is, if we substitute the value of \bar{x} for μ into the statement of H_0 (for the test of means) then it should be a true statement. Otherwise, H_1 is blatantly false and there is no need to do any statistical testing. In any statistical test, H_1 must be true at face value and we do the test to see if H_1 is *statistically true*. Another way to think about this is to think of \bar{x} as a fuzzy number. As a sharp number a statement like " $\bar{x} > k$ " may be true, but \bar{x} is fuzzy because of s (think $\bar{x} = \bar{x} \pm s$ to get the fuzzy number idea). So " $\bar{x} > k$ " may not be true when \bar{x} is considered to be a fuzzy number¹

When we make our decision (step 4) we consider the equality part of the H_0 statement in one-tailed tests. This equality is the strict

1. Fuzzy numbers can be treated rigorously in a mathematical sense. See, e.g. Kaufmann A, Gupta MM, *Introduction to fuzzy arithmetic: theory and applications*, Van Nostrand Reinhold Co., 1991.

H_0 under all circumstances but we use \geq or \leq is H_0 statements simply because they are the logical opposite of $<$ or $>$ in the H_1 statements. So people may have an issue with this statement of H_0 but we will keep it because of the logical completeness of the H_0, H_1 pair and the fact that hypothesis testing is about choosing between two well-defined alternatives.

p-Value

The critical statistic defines an area, a probability, α that is the maximum probability that we are willing to live with for making a type I error of incorrectly rejecting H_0 . The test statistic also defines an analogous area, called p or the p -value or (by SPSS especially) the significance. The p -value represents the best guess from the data that you will make a type I error if you reject H_0 . Computer programs compute p -values using CDFs. So when you use a computer (like SPSS) you don't need (or usually have) the critical statistic and you will make your decision (step 4) using the p -value associated with the test statistic according to the rule:

If $p \leq \alpha$ reject H_0 .

If $p > \alpha$ do not reject H_0 .

The method of comparing test and critical statistics is the traditional approach, popular before computers because it is less work to compute the two statistics than it is to compute p . When we work problem by hand we will use the traditional approach. When we use SPSS we will look at the p -value to make our decision. To connect the two approaches pedagogically we will estimate the p -value by hand for a while.

Example 9.3 : Compute the p -value for $z_{test} = 1.32$ of Example 9.2.

Solution : This calculation can happen as soon as you have the test statistic in step 3. The first thing to do is to sketch a picture of the p -value so that you know what you are doing, see Figure 9.2.

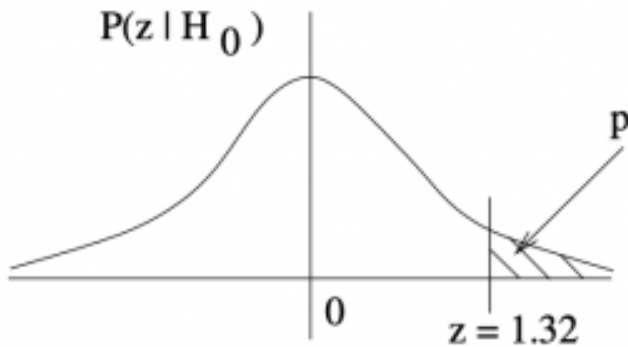


Figure 9.2 : The p -value associated with $z_{\text{test}} = 1.32$ in a one-tail test.

Using the **Standard Normal Distribution Table** to find the tail area associated with $z_{\text{test}} = 1.32$, we compute :

$$\begin{aligned} p(z_{\text{test}}) &= 0.5 - A(z_{\text{test}}) \\ &= 0.5 - 0.4066 = 0.0934 \end{aligned}$$

That is $p = 0.0934$. Since $(p = 0.0934) > (\alpha = 0.05)$ ($\alpha = 0.05$), we do not reject H_0 in our decision step (step 4).

□

When using the **Standard Normal Distribution Table** to find p -values for a given z you compute).

- For *two-tailed* tests: $p(z) = 2(0.5 - A(z))$. See Figure 9.3.
- For *one-tailed* tests: $p(z) = 0.5 - A(z)$ (as in Example 9.3)².

2. Of course substitute $-z$ in the formula for a left tail test.

Don't try to remember these formula, draw a picture to see what the situation is.

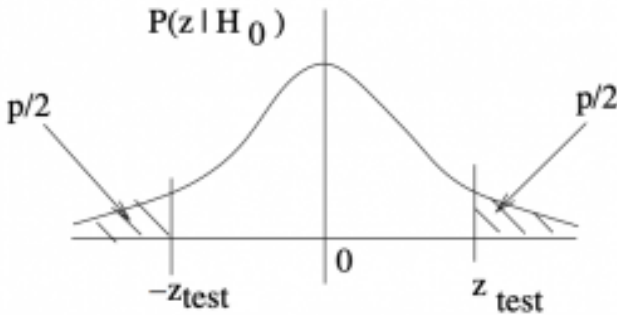


Figure 9.3 : The p -value associated with a two-tailed z_{test} . Since α is defined by, $\pm z_{critical}$, p is defined by $\pm z_{test}$.

9.2.1 What p -value is significant?

By culture, psychologists use $\alpha = 0.05$ to define the decision point for when to reject H_0 . In that case, if $p < 0.05$ then it means that the data (the test statistic) indicates there is less than a 5% chance that the result is a statistical fluke; that there is less than a 5% chance that the decision is a Type I error. So, in this course, we assume that $\alpha = 0.05$ unless α is otherwise given explicitly for pedagogical purposes. The choice of $\alpha = 0.05$ is actually fairly lax and has led to the inability to reproduce psychological experiments in many cases (about 5% of course). The standards in other scientific disciplines can be different. In particle physics experiments, for example, $p < 0.003$ is referred to as “evidence” for a discovery and they must have $p < 0.0000006$ before an actual discovery, like the discovery of the Higgs boson, is announced. With z test statistics, $\alpha = 0.003$ represents the area in the tails of the z distribution by 3 standard deviations, or 3σ , from the mean. The value

$\alpha = 0.0000006$ represents tail area 5σ , from the mean. So you may hear physicists saying that they have “5 sigma” evidence when they announce a discovery.

9.3 t-Test for Means

Hypothesis testing for means for sample set sizes in $n < 30$ where s is used as an estimate for σ is the same as for $n \geq 30$ except that t and not z is the test statistic¹. Specifically, the test statistic is

$$t_{\text{test}} = \frac{\bar{x} - k}{s/\sqrt{n}}$$

for k from any of the hypotheses listed in the table you saw in the previous section (one- and two-tailed versions):

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu = k$	$H_0: \mu \leq k$	$H_0: \mu \geq k$
$H_1: \mu \neq k$	$H_1: \mu > k$	$H_1: \mu < k$

The critical statistic is found in the **Distribution Table** with the degrees of freedom $\nu = n - 1$.

Example 9.4 : A physician claims that joggers, maximal volume oxygen uptake is greater than the average of all adults. A sample of 15 joggers has a mean of 40.6 ml/kg and a standard deviation of 6 ml/kg. If the average of all adults is 36.7 ml/kg, is there enough evidence to support the claim at $\alpha = 0.05$?

1. Hypothesis.

$$H_0 : \mu \leq 36.7$$

$$H_1 : \mu > 36.7 \text{ (claim)}$$

1. Again, SPSS applies the t -test, uses s directly, for any sample size.

2. Critical statistic.

In the **Distribution Table**, find the column for one-tailed test at $\alpha = 0.05$ and the line for degrees of freedom $\nu = n - 1 = 14$. With that find

$$t_{\text{critical}} = 1.761$$

3. Test statistic.

$$t_{\text{test}} = \frac{\bar{x} - k}{s/\sqrt{n}}$$

To compute this we need: $\bar{x} = 40.6$, $s = 6$ and $n = 15$ from the problem statement. From the hypothesis we have $k = 36.7$. So

$$t_{\text{test}} = \frac{40.6 - 36.7}{(6/\sqrt{15})} = 2.517$$

At this point we can estimate the p -value using the **Distribution Table**, which doesn't have as much information about the t -distribution as the **Standard Normal Distribution Table** has about the z -distribution, so we can only estimate. The procedure is: In the $\nu = 14$ row, look for t values that bracket $t_{\text{test}} = 2.517$. They are 2.145 (with $\alpha = 0.025$ in the column heading for one-tailed tests) and 2.624 (associated with a one-tail $\alpha = 0.01$).

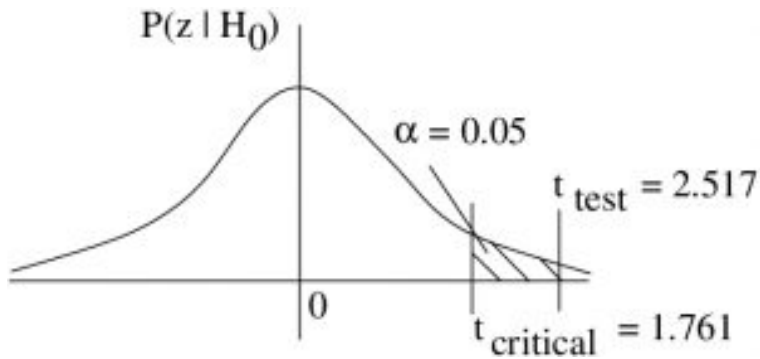
So,

$$0.010 < p < 0.025$$

is our estimate² for p .

4. Decision.

2. If you know how to interpolate then you can find a single value for p .



Reject H_0 . We can also base this decision on our p -value estimate since :

$$(0.010 < p < 0.025) < (\alpha = 0.05)$$

5. Interpretation.

There is enough evidence to support the claim that the joggers' maximal volume oxygen uptake is greater than 36.7 ml/kg using a t -test at $\alpha = 0.05$.

□

Fine point. When we use s in a t (or z test) as an estimate for σ , we are actually assuming that distribution of sample means is normal. The central limit theorem tells us that the distribution of sample means is approximately normal so generally we don't worry about this restriction. If the population is normal then the distribution of sample means will be exactly normal. Some stats texts state that we need to assume that the population is normal for a t -test to be valid. However, the central limit theorem's conclusion guarantees that the t -test is robust to violations of that assumption. If the population has a very wild distribution then s may be bad estimate for σ because the distribution of sample s values will not follow the χ^2 distribution. The chance if this happening becomes smaller the larger the n , again by the central limit theorem.

Origin of the t -distribution

We can easily define the t -distribution via random variables associated with the following stochastic processes. Let :

Z = a random variable with a z -distribution

X = a random variable with a χ^2 distribution with ν degrees of freedom.

Then the random variable

$$T = \frac{Z}{X}$$

is a random variable that follows a t -distribution with ν degrees of freedom.

9.4 z-Test for Proportions

The possible hypothesis pairs are :

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0 : p = k$	$H_0 : p \leq k$	$H_0 : p \geq k$
$H_1 : p \neq k$	$H_1 : p > k$	$H_1 : p < k$

The steps in hypothesis testing for proportions are the same as hypothesis testing for means. Even the generic test statistic formula is the similar :

$$\text{test value} = \frac{(\text{observed value}) - (\text{expected } H_0 \text{ value})}{\text{standard error}}.$$

but now the observed and expected values are proportions, \hat{p} and p respectively. The standard error in this case is

$$\sqrt{\frac{pq}{n}} = \frac{\sigma_{\text{binomial}}}{n} = \frac{\sqrt{npq}}{n}$$

Using this information with the generic form, which mimics a t test statistic, the proportions test statistic is

$$z_{\text{test}} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

where p is the number k which appears in the H_0 hypothesis statement (see table above). This test statistic is valid only if $np \geq 5$ and $nq \geq 5$ (so that the normal distribution provides a good approximation for the relevant binomial distribution). But, even though the test statistic can be moulded into the generic form,

the proportions test statistic comes from the sampling theory given by the binomial distributions and not from any distribution that has a standard error $\{\backslash\text{em per se}\}$. The normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$ (remember those binomial distribution formulae?) z -transformed to a z -distribution with mean 0 and standard deviation 1 gives the test statistic formula. See the discussion in Section 8.4.

Example 9.5 : An attorney claims that more than 25% of all lawyers advertise. A sample of 200 lawyers in a certain city showed that 63 had used some form of advertising. At $\alpha = 0.05$, is there enough evidence to support the attorney's claim?

Solution :

1. Hypotheses.

$$H_0 : p \leq 0.25 \quad , \quad H_1 : p > 0.25 \text{ (claim)}$$

2. Critical statistic.

Using the **Distribution Table** (last line) for a one tailed test at $\alpha = 0.05$ we find $z_{\text{critical}} = 1.645$

3. Test statistic.

$$z_{\text{test}} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

So using

$$\hat{p} = \frac{63}{200} = 0.315 \quad p = 0.25$$

$$q = 1 - 0.25 = 0.75 \quad n = 200$$

find

$$z_{\text{test}} = \frac{0.35 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{200}}} = 2.12.$$

We can also find the p value along with the critical statistic. (See the picture for the next step.) Use the **Standard Normal Distribution Table** to find

$$\begin{aligned}
 p(z) &= 0.5 - A(z) \\
 &= 0.5 - 0.4830 = 0.017 \\
 p &= 0.017
 \end{aligned}$$

4. Decision.

Refer to the diagram in Figure 9.4. It shows t_{test} in the rejection region. So we reject H_0 .

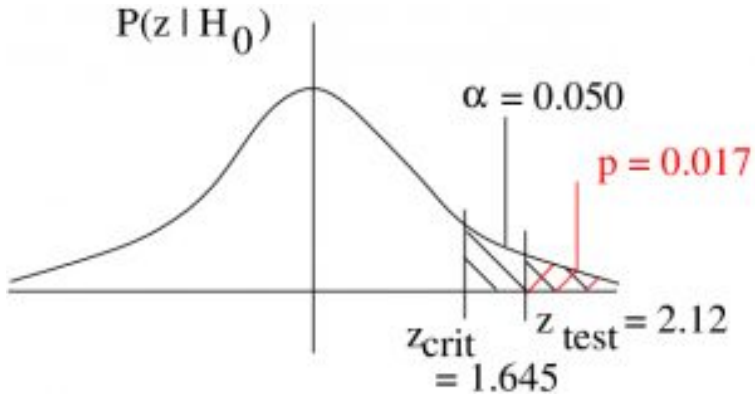


Figure 9.4 : The null hypothesis situation for Example 9.5

We come, of course, to the same decision by considering the p -value :

$$(p = 0.017) < (\alpha = 0.05)$$

5. Interpretation.

There is enough evidence, using a z -test at $\alpha = 0.05$, to support the claim that more than 25% of the lawyers use some form of advertising.

□

9.5 Chi Squared Test for Variance or Standard Deviation

The possible hypothesis pairs are, for variance :

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0 : \sigma^2 = k$	$H_0 : \sigma^2 \leq k$	$H_0 : \sigma^2 \geq k$
$H_1 : \sigma^2 \neq k$	$H_1 : \sigma^2 > k$	$H_1 : \sigma^2 < k$

For standard deviation we use the square roots of everything :

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0 : \sigma = k$	$H_0 : \sigma \leq k$	$H_0 : \sigma \geq k$
$H_1 : \sigma \neq k$	$H_1 : \sigma > k$	$H_1 : \sigma < k$

Note that we did not square root k . This is because we are using k to stand in for whatever number. That number from H_0 will appear in our formulae as either σ^2 or σ depending on the set up. Generally we will work with variance as we work through the problem and convert to standard deviation only in the last interpretation step if required by the wording of the question.

The new test statistic is :

$$\chi_{\text{test}}^2 = \frac{(n - 1)s^2}{\sigma^2}$$

where s comes from the sample and σ^2 comes from the number k in H_0 . The degrees of freedom associated with the test statistic (for finding the critical statistic) is $\nu = n - 1$. There is no

mystery where this test statistic came from – this is just how χ^2 as a probability distribution is defined. So, for this test to be valid, the population must be normally distributed. The χ^2 test here is not very robust to violations of that assumption because there is no normalizing intermediate central limit theorem here.

The critical regions on the χ^2 distribution will appear as shown in Figure 9.5.

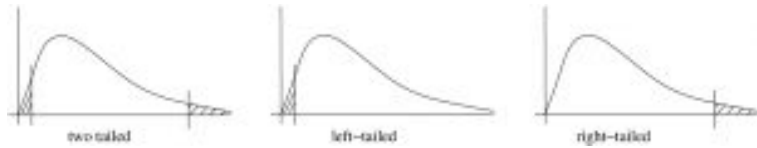


Figure 9.5 : Schematics of the critical regions for χ^2 tests of variance. In the two-tailed situation the tail areas are equal.

Let's work through an example of each hypotheses pair case. In all of the examples we assume that the population is normally distributed.

Example 9.6 : An instructor wishes to see whether the variance in scores of the 23 students in her class is less than the variance of the population. The variance of the class is 198. is there enough evidence to support the claim that the variation of the students is less than the population variance $\sigma^2 = 225$ at $\alpha = 0.05$?

Solution :

1. Hypotheses.

$$H_0 : \sigma^2 \geq 225 \quad H_1 : \sigma^2 < 225$$

2. Critical statistic.

Refer to Figure 9.6 as we get the critical statistic from the **Chi-squared Distribution Table**. As we see in that figure, we must look in the column that corresponds to a right tail area of 0.95. The row we need is for $\nu = n - 1 = 23 - 1 = 22$. With that information we find $\chi^2_{\text{crit}} = 12.338$.

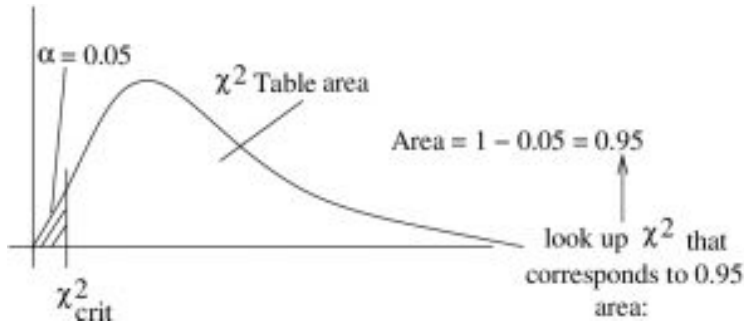


Figure 9.6 : Schematics of the critical regions for χ^2 tests of variance. In the two-tailed situation the tail areas are equal.

3. Test statistic.

The values we need for the test statistic are $\sigma^2 = 225$ (from H_0), $s^2 = 198$ and $n - 1 = 22$ from the information in the problem. So :

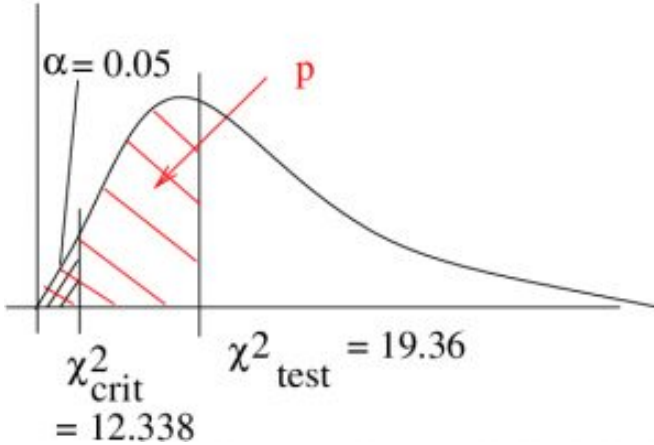
$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

$$\chi^2 = \frac{(22)(198)}{225} = 19.36$$

At this point we can also estimate the p value from the **Chi-squared Distribution Table**. The p value is the area under the χ^2 distribution with $\nu = 22$ to the left of $\chi^{2_{\text{test}}}$. In the $\nu = 22$ row of the **Chi-squared Distribution Table** (in general use the closest ν if your particular value is not in the **Chi-squared Distribution Table**) hunt down the test statistic value of 19.38. You won't find it but you can bracket it with values higher and lower than 19.38. Those numbers are 14.042 which has a right tail area of 0.90 (and so a left tail area of 0.10) and 30.813 which has a right tail area of 0.10 (and so a left tail area of 0.90). Recall that the α in

the column headings of the **Chi-squared Distribution Table** refers to right tail areas. So, considering the left tail areas we know that $0.10 < p < 0.90$ since $30.813 > 19.38 > 14.042$ for the relevant χ^2 values.

4. Decision.



Since χ^2_{test} doesn't fall in the rejection region, do not reject H_0 . We come to the same conclusion with our p -value estimate:

$$(0.10 < p < 0.90) > (\alpha = 0.05)$$

5. Interpretation.

There is not enough evidence, at $\alpha = 0.05$ with a χ^2 test, to support the claim that the variation in test scores of the class is less than 225.

□

Example 9.7 : A hospital administrator believes that the standard deviation of the number of people using out-patient surgery per day is greater than eight. A random sample of 15 days is selected. The data are shown below. At $\alpha = 0.10$ is there enough evidence to support the administrator's claim?

25 30 5 15 18 42 16 9 10 12 12 38 8 14 27

Solution :

0. Data reduction.

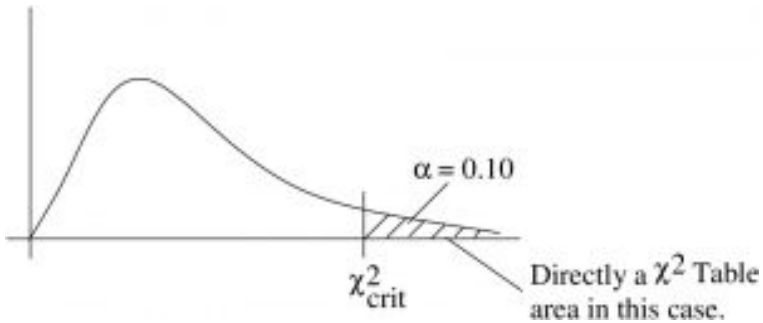
We'll introduce a step 0 when it looks like we should do some preliminary calculations with or data. In this case we should enter the dataset into our calculations and determine s . We find $s = 11.2$.

1. Hypotheses.

$$H_0 : \sigma^2 \leq 64 \quad H_1 : \sigma^2 > 64 \text{ (claim)}$$

Note conversion to σ^2 right away.

2. Critical statistic.



In the $\nu = 15 - 1 = 14$ line and $\alpha_T = 0.10$ column of the **Chi-squared Distribution Table**, look up $\chi^2_{crit} = 21.064$

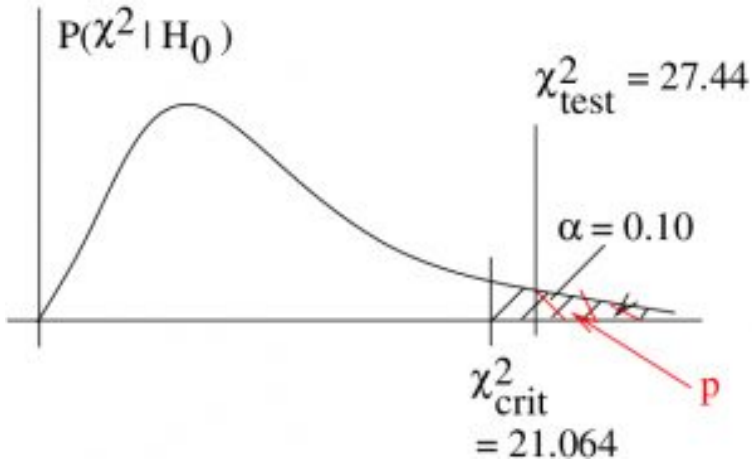
3. Test statistic.

$$\chi^2_{test} = \frac{(n - 1)s^2}{\sigma^2}$$
$$\chi^2_{test} = \frac{(14)(11.2)^2}{64} = 27.44$$

To estimate the p value, find the bracketing values of $\chi^2_{test} = 27.44$ in the $\nu = 14$ line of the **Chi-squared**

Distribution Table. They are : 26.119 ($\alpha = 0.025$) and 29.141 ($\alpha = 0.010$), so $0.010 < p < 0.025$.

4. Decision.



Reject H_0 since χ^2_{test} is in the rejection region. Our estimate of p leads to the same conclusion :
 $(0.010 < p < 0.025) < (\alpha = 0.10)$

5. Interpretation.

There is enough evidence, at $\alpha = 0.10$ with a χ^2 test, to support the claim that the standard deviation is greater than 8. (Note how we convert to a statement about standard deviation after working through the problem using variances.)

□

Example 9.8 : A cigarette manufacturer wishes to test the claim that the variance of the nicotine content of its cigarettes is 0.644. Nicotine content is measured in milligrams, assume that it is normally distributed. A sample of 20 cigarettes has a standard deviation of 1.00 kg. At $\alpha = 0.05$, is there enough evidence to reject the manufacturer's claim?

Solution :

1. Hypotheses.

$$H_0 : \sigma^2 = 0.644 \text{ (claim)} \quad H_1 : \sigma^2 \neq 0.64$$

2. Critical statistic.

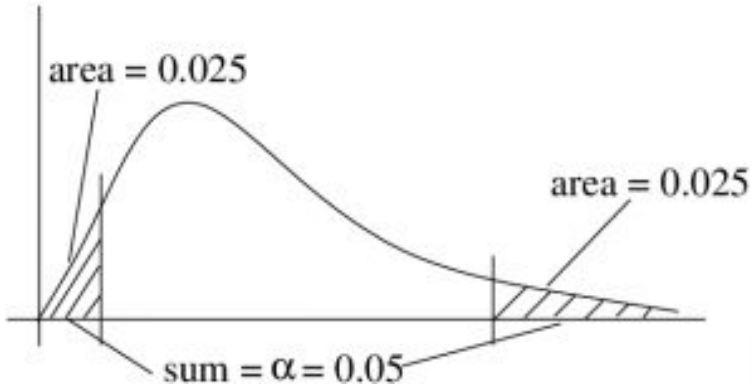


Figure 9.7 : Critical regions for a two tailed test.

Referring to Figure 9.7, we see that we need two χ_{crit}^2 values, one with a tail area of 0.025 and the other with a tail area of $1 - 0.025 = 0.975$. From the **Chi-squared Distribution Table** in the $\nu = n - 1 = 19$ line find $\chi_{\text{crit}}^2 = 8.907$ from the $\alpha_T = 0.975$ column and $\chi_{\text{crit}}^2 = 32.852$ from the $\alpha_T = 0.025$ column.

3. Test statistic.

$$\chi_{\text{test}}^2 = \frac{(n - 1)s^2}{\sigma^2}$$
$$\chi_{\text{test}}^2 = \frac{(19)(1^2)}{(0.644)} = 29.50$$

To estimate the p value find the bracketing value of $\chi_{\text{test}}^2 = 29.50$ in the $\nu = 19$ row, They are 27.204 (

$\alpha_T = 0.10$) and 30.144 ($\alpha_T = 0.05$). The α_T are right tail areas, which is ok, but we need to multiply them by 2 because those right tail areas represent $p/2$ as shown in Figure 9.8. So $0.10 < p < 0.20$.

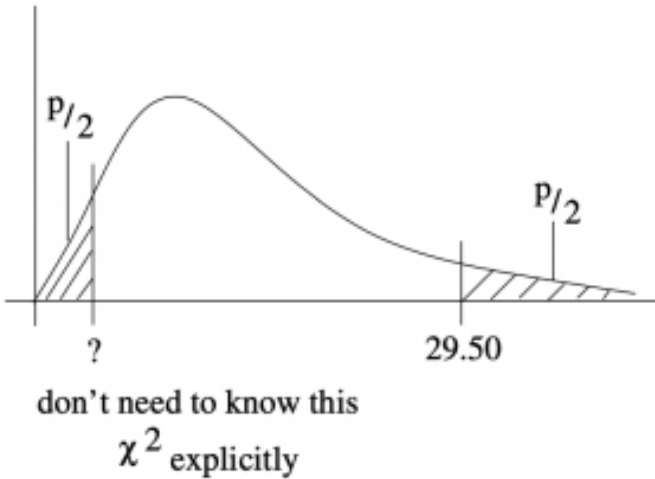
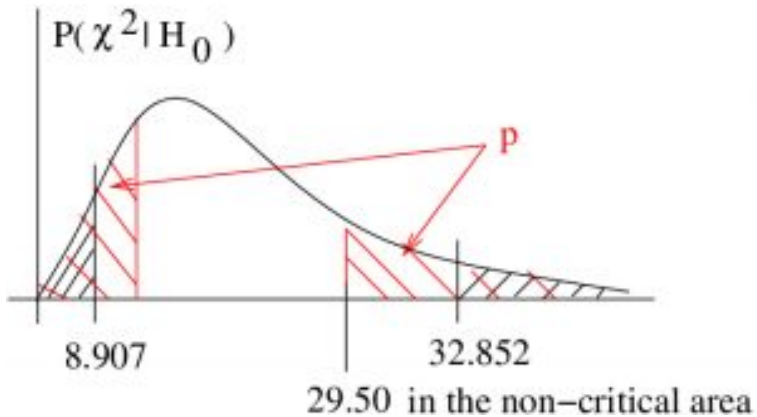


Figure 9.8 : Areas for p associated with the test statistic (29.50 here) in a two tail test.

4. Decision.



Do not reject H_0 . The estimate p value leads to the same conclusion :

$$(0.10 < p < 0.20) > (\alpha = 0.05)$$

5. Interpretation.

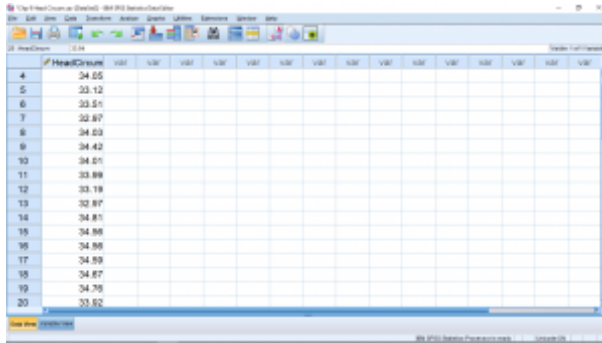
There is not enough evidence, at $\alpha = 0.05$ with a χ^2 test, to reject the manufacturer's claim that the variance of the nicotine content of the cigarettes is equal to 0.644.

Notice, with the claim on H_0 , that failing to reject H_0 does not provide any evidence that H_0 is true. We just have the weaker conclusion that we couldn't disprove it. Such is the double negative nature of the logic behind hypothesis testing that arises where we don't assign probabilities to hypothesis.



9.6 SPSS Lesson 5: Single Sample t-Test

Open “HeadCircum.sav” from the textbook [Data Sets](#):



The screenshot shows the SPSS Data Editor window with a single variable named 'HeadCircum'. The data is as follows:

Case	HeadCircum
4	34.65
5	33.12
6	33.51
7	32.87
8	34.53
9	34.42
10	34.51
11	33.89
12	33.19
13	32.87
14	34.81
15	34.96
16	34.96
17	34.59
18	34.87
19	34.76
20	33.82

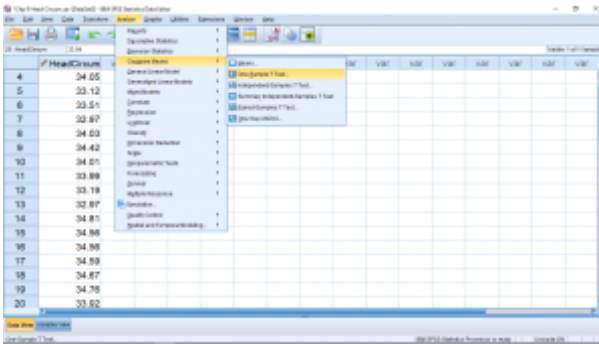
SPSS
screenshot ©
International
Business
Machines
Corporation.

Look at how simple it is! One variable. This is our single sample. Let’s do a t -test for the hypotheses:

$$H_0 : \mu = 33.8$$

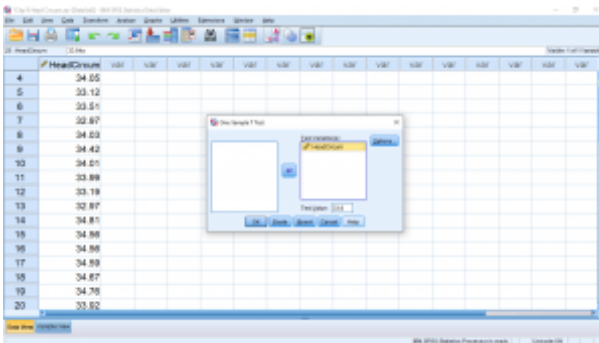
$$(9.2) \quad H_1 : \mu \neq 33.8$$

where we have used $k = 34.5$ as the potentially inferred population value. Selecting the value for k is something that you will need to think about when doing single sample t -tests. Some possibilities are: past values, data range midpoints or chance level values. To run the t -test in SPSS, pick Analyze → Compare Means → One-Sample T Test:



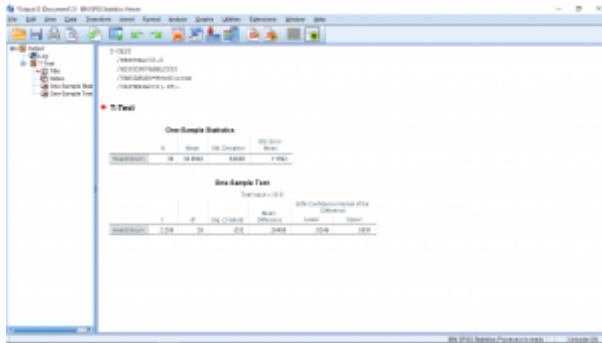
SPSS screenshot © International Business Machines Corporation.

The pop up menu is:



SPSS screenshot © International Business Machines Corporation.

where we have moved our variable into the Test Variable(s) box. If more than one variable is in this box then a separate t -test will be run for each variable. The value $k = 33.8$ has been entered into the Test Value box. That's how SPSS knows that the hypotheses to test is that of the statement (9.2) above. If you open the Options menus, you will have a chance to specify the associated confidence interval. Running the analysis gives the very simple output:



SPSS
screenshot ©
International
Business
Machines
Corporation.

The output is simple but it requires your knowledge of the t -test to interpret. As you get more experience with using SPSS, or any canned statistical software, you will get into the habit of looking for the p -value. In SPSS it is in the Sig. (for Significance) column. Here $p = 0.032$, which is less than $\alpha = 0.05$, so we reject the null hypothesis and conclude that there is evidence that the population mean is not 34.5. Note that this p -value is for a two-tailed test. What if you wanted to do a one-tailed test? Well, then you have to think because SPSS won't do that for you explicitly. For a one-tailed test, $p = 0.016$, half that of the two-tailed test. Remember that the two-tailed p has two tails, each with an area of 0.016 as defined by $\pm t_{\text{test}}$, so getting rid of one of those areas gives the p for the one-tailed test. Another way to remember to divide the two-tailed p by 2 to get the one-tailed value is to remember that people try to go for a one-tailed test when they can because it has more power – it is easier to reject the null hypothesis with a one-tailed test meaning the p -value will be smaller for a one-tailed test.

Let's look at the rest of the output. There is a lot of redundant information there. You can use that redundant information to check to make sure you know what SPSS is doing and I can use that redundant information to see if you understand what SPSS is doing by reducing the redundancy and asking you to calculate the missing pieces. In the first output table, "One-Sample Statistics" is the

information that you would get out of your calculator. The first three columns are n , \bar{x} and s . The last column is s/\sqrt{n} .

In the second output table “One-Sample Test”, notice that the test value of 33.8 is printed to remind you what the hypotheses being tested is. The columns give: t_{test} , ν , p and $\bar{x} - k$. Notice that the first column, t_{test} is the fourth column $\bar{x} - k$ divided by the last column of the first table, s/\sqrt{n} . The last two columns give the 95% confidence interval

$$(9.3) \quad 0.0249 < \mu - 33.80 < 0.5031$$

Note that zero is not in this confidence interval which is consistent with rejecting the null hypothesis. Simply add $k = 33.80$ to Equation (9.3) to get the form we go for when we do confidence intervals by hand:

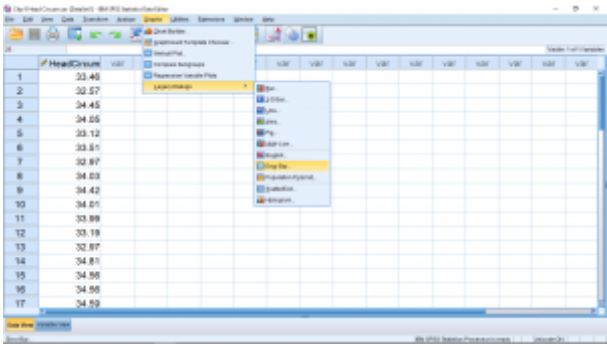
$$(9.4) \quad 33.8249 < \mu < 34.3031$$

You can use the output here to compute a further quantity, known as standardized effect size. You’ll get a little practice with doing that in the assignments. The standardized effect size, d , is a purely descriptive statistic (although it can be used in power calculations) and is defined by

$$(9.5) \quad d = \frac{\bar{x} - k}{s} = \frac{t}{\sqrt{n}}$$

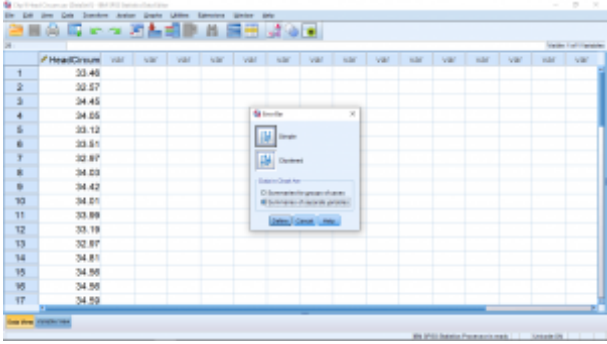
where, by t we mean t_{test} . Being a descriptive statistic, people use the following rule of thumb to describe d . If d is approximately 0.2 then d is considered “small”; if d is approximately 0.5 then d is considered “medium”; d is approximately 0.8 then d is considered “large”.

For the presentation of data graphically in reports and papers, an error bar plot is frequently used. To get such a plot for the data here, select Graphs \rightarrow Legacy Dialogs \rightarrow Error Bar:



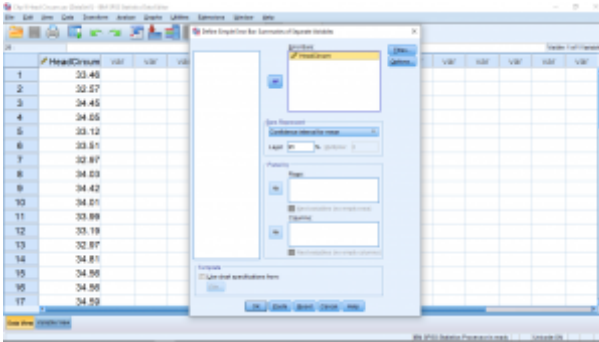
SPSS screenshot © International Business Machines Corporation.

Choose Simple and “Summaries of separate variables”:



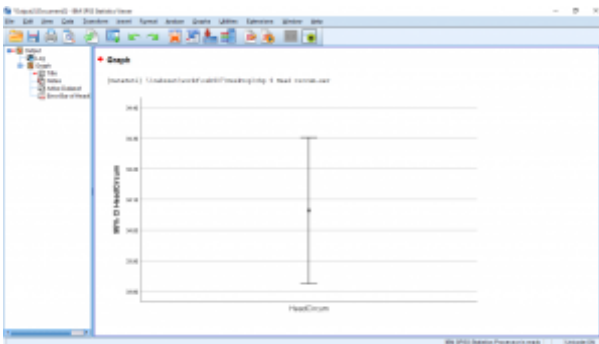
SPSS screenshot © International Business Machines Corporation.

and hit Define. Then set up the menu as follows:



SPSS screenshot © International Business Machines Corporation.

noting that we have chosen “Bars Represent” as “Standard error of the mean” so that the error bars will be $\bar{x} \pm \frac{s}{\sqrt{n}}$:

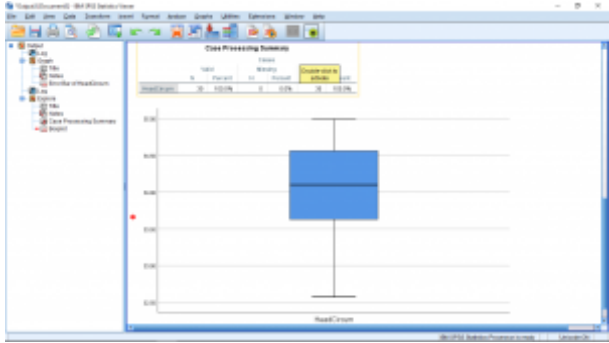


SPSS screenshot © International Business Machines Corporation.

With an error bar plot like this, you can intuitively check the meaning of rejecting H_0 from the formal t -test. Here the error bars do not include the value of 33.80 which is consistent with the conclusion that we reject 33.80 as a possible value for the population mean. We can see this more directly, and exactly, if we choose the value 95 confidence interval in the Bars Represent pull down of the plot menu.

This is a plot of Equation (9.4). The value $k = 33.80$ is not in the 95% confidence interval.

Finally, selecting Graphs → Legacy Dialogs → Boxplot gives a EDA type of data presentation:



SPSS
screenshot ©
International
Business
Machines
Corporation.

9.7 RStudio Lesson 5: Single Sample t-Test

OSAMA BATAINEH

Let's insert and attach the dataset "HeadCircum.sav" (from [Data Sets](#)). To get the descriptive statistics, first we need to install and load the package *pastecs*. I have already installed it before. So I just loaded this package using the *library* function. Then we utilize a function of this package *round* to get the descriptive statistics. Here we use our variable *age* as the argument and set *digits* equal to 2 to make it look nice.

```
> library(pastecs)
```

Warning message:

```
package 'pastecs' was built under R version 3.4.4
```

```
> descriptive_stat <- round(stat.desc(age), digits = 2)
```

```
> descriptive_stat
```

```
x
```

```
nbr.val 5534.00
```

```
nbr.null 0.00
```

```
nbr.na 0.00
```

```
min 10.00
```

```
max 43.00
```

```
range 33.00
```

```
sum 129718.00
```

```
median 23.00
```

```
mean 23.44
```

```
SE.mean 0.06
```

```
CI.mean.0.95 0.12
```

```
var 22.29
```

```
std.dev 4.72
```

```
coef.var 0.20
```

Now we will be doing a T-test for mean equal to 33.80. To do it,

we will take the help of the function `t.test`. Here we can see that in the argument, we have to mention the value of our null hypotheses (33.80 in the case), the type of T-test we are doing (one sided or two sided) and the confidence interval.

Now we will be doing a T-test for mean equal to 23.30. To do it, we will take the help of the function `t.test`. Here we can see that in the argument, we have to mention the value of our null hypotheses (23.30 in the case), the type of T-test we are doing (one sided or two sided) and the confidence interval.

```
> t.test(age, mu=23.30, alternative="two.sided", conf.level=0.95)
```

One Sample t-test

data: age

t = 2.2088, df = 5533, p-value = 0.02723

alternative hypothesis: true mean is not equal to 23.3

95 percent confidence interval:

23.31577 23.56461

sample estimates:

mean of x

23.44019

```
> t.test(age, mu=23.30, alternative="greater", conf.level=0.95)
```

One Sample t-test

data: age

t = 2.2088, df = 5533, p-value = 0.01361

alternative hypothesis: true mean is greater than 23.3

95 percent confidence interval:

23.33578 Inf

sample estimates:

mean of x

23.44019

10. COMPARING TWO POPULATION MEANS

There are two types of two-sample t -tests. (The test we covered in Chapter 9 that compared the mean of one sample to a fixed number k is known as a one-sample t -test.) These tests are:

Unpaired or independent sample t -test:

The two populations are “independent”. There is no relation between the x_1 and x_2 variables (as we’ll call them).

This is a “between subjects” test, the experimental subjects in each of the two populations are different.

Paired or dependent sample t -test:

There is a natural pairing between the two variables x_1 and x_2 , usually they are measured from the same subject.

A paired t -test is an example of a “repeated measures” or “within subject” test.

We will introduce the independent sample t -test with a z -test approximation first to build ideas. As before, note that SPSS doesn’t do these approximate z -tests. It does t -tests even for large samples.

10.1 Unpaired z-Test

We have two populations and two sample sets, one from each population :

	Sample Mean	Sample std. dev.
From population 1	\bar{x}_1	s_1
From population 2	\bar{x}_2	s_2

The population means are μ_1 and μ_2 and just as with the single population test, there are 3 possible hypothesis tests :

Two Tailed	Right Tailed	Left Tailed
$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \leq \mu_2$	$H_0 : \mu_1 \geq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 > \mu_2$	$H_1 : \mu_1 < \mu_2$
or	or	or
$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 \leq 0$	$H_0 : \mu_1 - \mu_2 \geq 0$
$H_1 : \mu_1 - \mu_2 \neq 0$	$H_1 : \mu_1 - \mu_2 > 0$	$H_1 : \mu_1 - \mu_2 < 0$

In the second row the hypotheses are written in terms of a difference. Irrespective of which way you write the hypotheses, give population 1 priority. Write population 1 first. That way you won't mess up your signs or your interpretation.

The test statistic to use, in all cases¹ is

1. You could specify a non-zero null hypothesis, e.g.

$H_0 : \mu_1 - \mu_2 = k$, in which case you would have

$$(10.1) \quad z_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where n_1 = sample set size from population 1 and n_2 = sample set size from population 2. This test statistic is based on a distribution of sample means as shown in Figure 10.1.

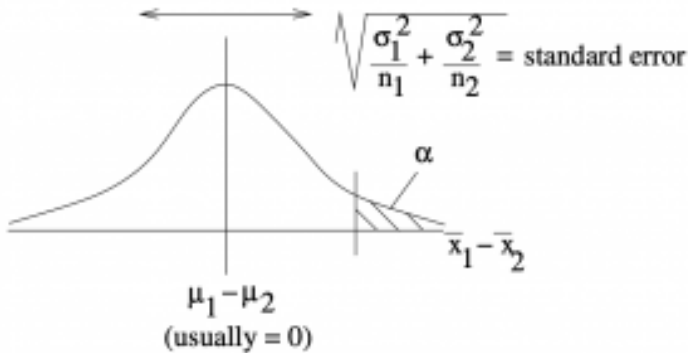


Figure 10.1: The distribution of the difference of sample means $\bar{x}_1 - \bar{x}_2$ under the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$. A one-tail example is shown here. The test statistic of Equation 10.1 follows from a Z -transformation of this picture.

Example 10.1 : A researcher hypothesizes that the average number of sports colleges offer for males is greater than the average number of sports offered for females. Samples of the number of sports offered to each sex by randomly selected colleges is given here :

$$z_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2) - k}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

course.

Males (pop. 1)	Females (pop. 2)
$n_1 = 50$	$n_2 = 50$
$\bar{x}_1 = 8.6$	$\bar{x}_2 = 7.9$
$s_1 = 3.3$	$s_2 = 3.3$

At $\alpha = 0.10$ is there enough evidence to support the claim?

Solution :

1. Hypotheses.

$$H_0 : \mu_1 \leq \mu_2 \qquad H_1 : \mu_1 > \mu_2 \text{ (claim)}$$

Note that $\bar{x}_1 > \bar{x}_2$ ($8.6 > 7.9$) so $H_1 : \mu_1 > \mu_2$ is true on the face of it. If H_1 is not true on the face of it then

H_1 is just plain false without the need for any statistical test.

With the hypotheses direction set correctly, the question becomes: Is \bar{x}_1 significantly greater than \bar{x}_2 ? The term “statistically significant” corresponds to “reject H_0 ”.

2. Critical statistic.

From the **t Distribution Table**, one-tailed test at $\alpha = 0.10$ we find

$$z_{\text{crit}} = 1.282$$

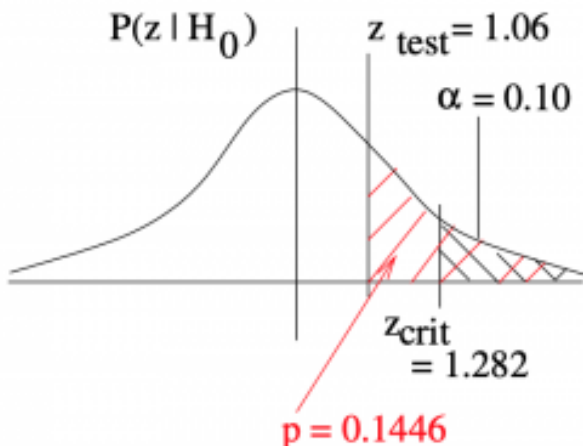
Note that z_{critical} is positive because this is a right-tailed test. For left tailed tests make z_{crit} negative. For two-tailed tests you have $\pm z_{\text{crit}}$.

3. Test statistic.

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(8.6 - 7.9)}{\sqrt{\frac{3.3^2}{50} + \frac{3.3^2}{50}}} \\ &= 1.06 \end{aligned}$$

Using the **Standard Normal Distribution Table**, we can find the p -value. Since $A(z) = A(1.06) = 0.3554$, $p = 0.05 - 0.3554 = 0.1446$.

4. Decision.



Do not reject H_0 since z_{test} is not in the rejection region.

The p -value reflects this :

$$(p = 0.1446) > (\alpha = 0.10)$$

5. Interpretation.

There is not enough evidence, at $\alpha = 0.10$ under a z -test, to support the claim that colleges offer more sports for males than females.

□

10.2 Confidence Interval for Difference of Means (Large Samples)

Swapping the roles of sample and population in the sampling theory, we have the confidence interval corresponding to the hypothesis test of Section 10.1

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

where

$$E = z_C \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 10.2 : Find the 95% confidence interval for the difference between the means for the data of Example 10.1.

Solution : First, recall our data :

$$\bar{x}_1 = 88.42, s_1 = 5.62, n_1 = 50$$

$$\bar{x}_2 = 80.61, s_2 = 4.83, n_2 = 50$$

From the **t Distribution Table**, look up the z for the 95% confidence interval: $z_{95\%} = 1.960$. Then compute:

$$\bar{x}_1 - \bar{x}_2 = 88.42 - 80.61 = 7.81$$

and

$$\begin{aligned} E &= z_{95\%} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= 1.960 \sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}} \\ &= 2.05 \end{aligned}$$

so

$$7.81 - 2.05 < (\mu_1 - \mu_2) < 7.81 + 2.05$$

or

$$5.76 < (\mu_1 - \mu_2) < 9.83$$

with 95% confidence. Notice that it is also correct to write

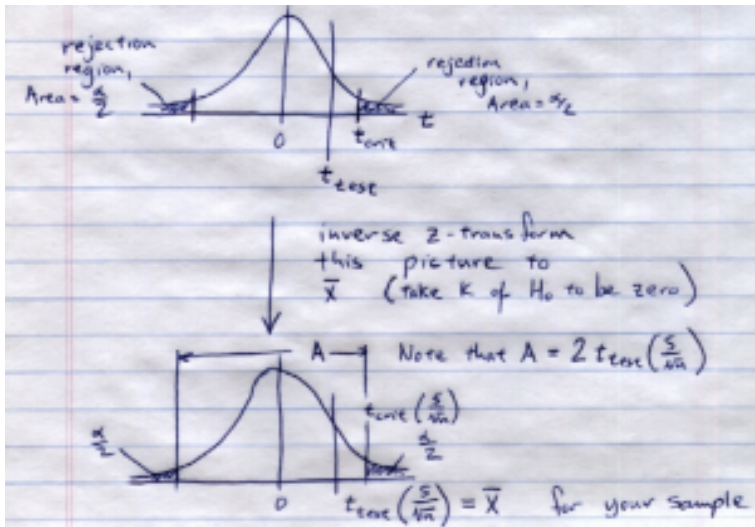
$$\mu_1 - \mu_2 = 7.81 \pm 2.05 \text{ with } 95\% \text{ confidence.}$$

□

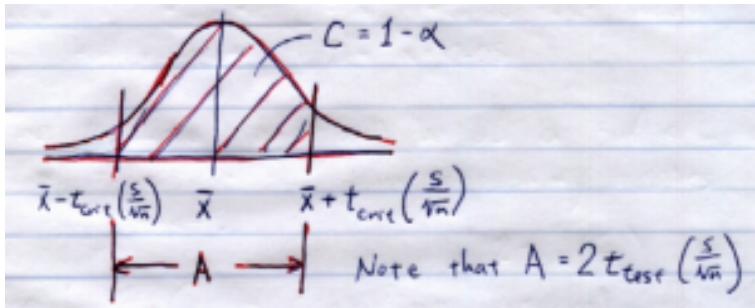
This is a good point to make an important observation. A two-tailed hypothesis test at a given α is complementary to a confidence interval of $C = 1 - \alpha$ in the sense that if 0 is in the confidence interval then the complementary hypothesis test will not reject H_0 .

Let's illustrate this principle with a one-sample t -test under $H_0 : \mu = 0$. (We need $k = 0$ for this principle to work.) Look at the two possible outcomes :

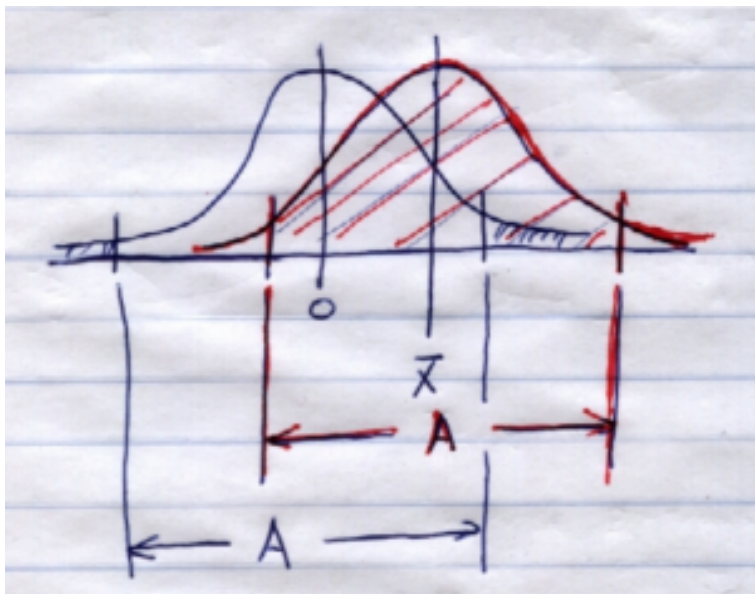
Case 1 : 0 in the confidence interval, fail to reject H_0 . In the hypothesis test you would find :



In the confidence interval calculation you would find:



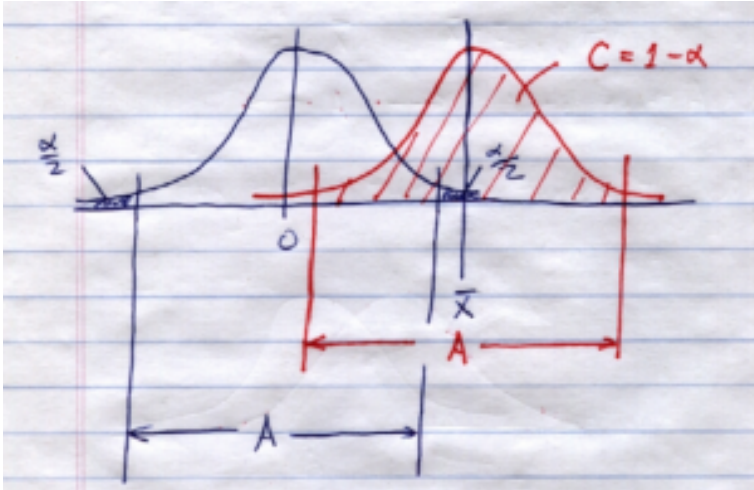
Putting the two pictures together gives:



See, 0 is in the confidence interval if \bar{x} is not in the rejection region. The red distribution that defines the confidence interval

is just the blue (identical) distribution slid over from 0 to \bar{x} . The distance A is the same because $C = 1 - \alpha$.

Case 2 : 0 not in the confidence interval, reject H_0 . In this case the combined picture looks like:



Before we can consider the independent sample t -test, we need a tool for checking what the variances of the populations are. The formula for the t test statistic will depend on whether the two variances are the same or not. So let's take a look at comparing population variances.

10.3 Difference between Two Variances - the F Distributions

Here we have to assume that the two *populations* (as opposed to sample mean distributions) have a distribution that is almost normal as shown in Figure 10.2.

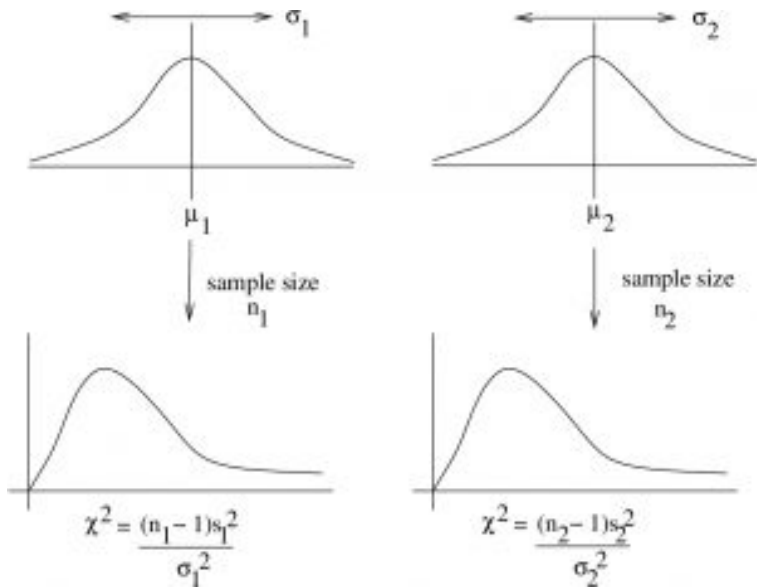


Figure 10.2: Two normal populations lead to two χ^2 distributions that represent distributions of sample variances. The F distribution results when you build up a distribution of the ratio of the two χ^2 sample values.

The ratio $\frac{s_1^2}{s_2^2}$ follows an F -distribution if $\sigma_1 = \sigma_2$. That F

distribution has two degrees of freedom: one for the numerator (d.f.N. or ν_1) and one for the denominator (d.f.D. or ν_2). So we denote the distribution more specifically as F_{ν_1, ν_2} . For the case of Figure 10.2, $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$. The F ratio, in general is the result of the following stochastic process. Let X_1 be random variable produced by a stochastic process with a $\chi^2_{\nu_1}$ distribution and let X_2 be random variable produced by a stochastic process with a $\chi^2_{\nu_2}$ distribution. Then the random variable $F = X_1/X_2$ will, by definition, have a F_{ν_1, ν_2} distribution.

The exact shape of the F_{ν_1, ν_2} distribution depends on the choice of ν_1 and ν_2 , But it roughly looks like a χ^2 distribution as shown in Figure 10.3.

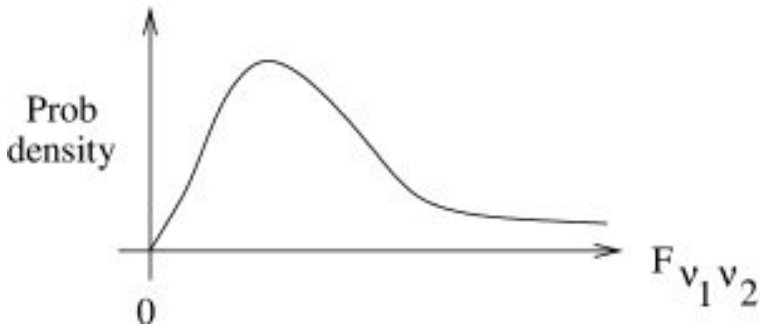


Figure 10.3: A generic F distribution.

F and t are related :

$$F_{1, \nu} = t_{\nu}^2$$

so the t statistic can be viewed as a special case of the F statistic.

For comparing variances, we are interested in the follow hypotheses pairs :

Right-tailed	Left-tailed	Two-tailed
$H_0 : \sigma_1^2 \leq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 > \sigma_2^2$ <small>\sigma^2_2</small> <small>title="Rendered by QuickLaTeX.com"</small> <small>height="21"</small> <small>width="97"</small> <small>style="vertical-align: -6px; "></small>	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 \neq \sigma_2^2$

We'll always compare variances (σ^2) and not standard deviations (σ) to keep life simple.

The test statistic is

$$F_{\text{test}} = F_{\nu_1, \nu_2} = \frac{s_1^2}{s_2^2}$$

where (for finding the critical statistic), $\mu_1 = n_1 - 1$ and $\mu_2 = n_2 - 1$.

Note that $F_{\nu_1, \nu_2} = 1$ when $s_1^2 = s_2^2$, a fact you can use to get a feel for the meaning of this test statistic.

Values for the various F critical values are given in the **F Distribution Table** in the [Appendix](#). We will denote a critical value of F with the notation :

$$F_{\text{crit}} = F_{\alpha, \nu_1, \nu_2}$$

Where:

α = Type I error rate

ν_1 = d.f.N.

ν_2 = d.f.D.

The **F Distribution Table** gives critical values for small right tail areas only. This means that they are useless for a left-tailed test. But that does not mean we cannot do a left-tail test. A left-tail test is easily converted into a right tail test by switching the assignments of populations 1 and 2. To get the assignments correct in the first place then, always define populations 1 and 2 so that $\sigma_1^2 > \sigma_2^2$

$\sigma^2_{-}\{2}\}$ title="Rendered by QuickLaTeX.com" height="21" width="60" style="vertical-align: -6px;". Assign population 1 so that it has the largest sample variance. Do this even for a two-tail test because we will have no idea what F_{crit} on the left side of the distribution is.

Example 10.3 : Given the following data for smokers and non-smokers (maybe its about some sort of disease occurrence, who cares, let's focus on dealing with the numbers), test if the population variances are equal or not at $\alpha = 0.05$.

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Note that $s_1^2 > s_2^2$ $s_{-}\{2}\}$ title="Rendered by QuickLaTeX.com" height="21" width="55" style="vertical-align: -6px; so we're good to go.

Solution :

1. Hypothesis.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

2. Critical statistic.

Use the **F Distribution Table**; it is a bunch of tables labeled by " α " that we will designate at α_T , the table values that signify right tail areas. Since this is a two-tail test, we need $\alpha_T = \alpha/2$. Next we need the degrees of freedom:

$$\text{d.f.N.} = \nu_1 = n_1 - 1 = 26 - 1 = 25$$

$$\text{d.f.D.} = \nu_2 = n_2 - 1 = 18 - 1 = 17$$

So the critical statistic is

$$F_{crit} = F_{\alpha/2, \nu_1, \nu_2} = F_{0.05/2, 25, 17} = F_{0.025, 25, 17} = 2.56.$$

3. Test statistic.

$$F_{\nu_1, \nu_2} = \frac{s_1^2}{s_2^2}$$

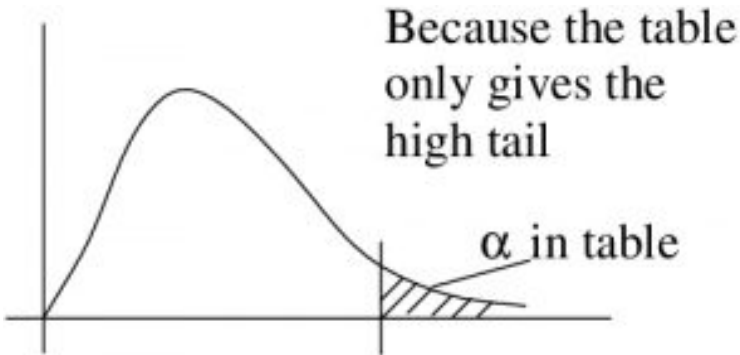
$$F_{\text{test}} = F_{25, 17} = \frac{36}{10} = 3.6$$

With this test statistic, we can estimate the p -value using the **F Distribution Table**. To find p , look up all the numbers with d.f.N = 25 and d.f.N = 17 (24 & 17 are the closest in the tables so use those) in all the the **F Distribution Table** and form your own table. For each column in your table record α_T and the F value corresponding to the degrees of freedom of interest. Again, α_T corresponds to $p/2$ for a two-tailed test. So make a row above the α_T row with $p = 2\alpha_T$. (For a one-tailed test, we would put $p = \alpha_T$.)

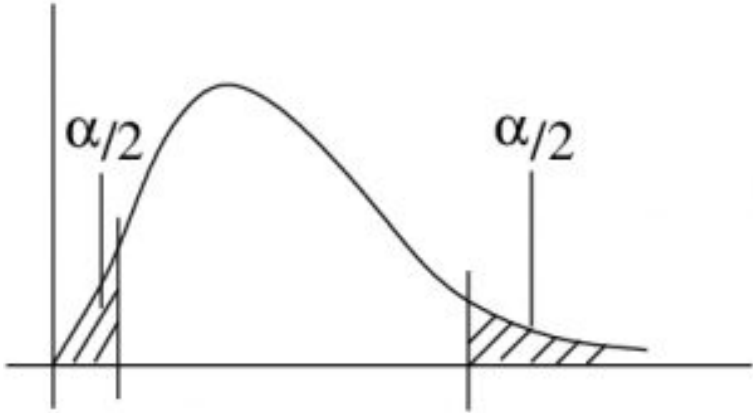
p	0.20	0.10	0.05	0.02	0.01	
α_T	0.10	0.05	0.025	0.01	0.005	
F	1.84	2.19	2.56	3.08	3.51	3.6 is over here somewhere so $p < 0.01$

Notice how we put an upper limit on p because F_{test} was larger than all the F values in our little table.

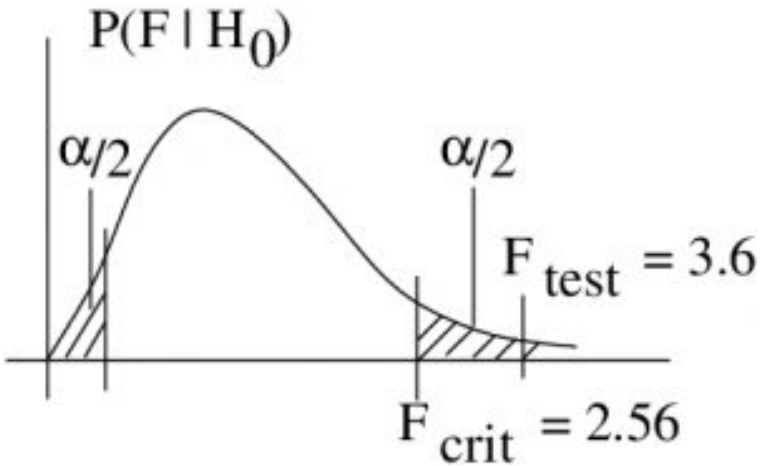
Let's take a graphical look at why we use $p = 2\alpha$ in the little table and $\alpha_T = \alpha/2$ for finding F_{crit} for two tailed tests :



But in a two-tailed test we want α split on both sides:



4. Decision.



Reject H_0 . The p -value estimate supports this :
 $(p < 0.01) < (\alpha = 0.05)$

5. Interpretation.

There is enough evidence to conclude, at $\alpha = 0.05$ with an F -test, that the variance of the smoker population is different from the non-smoker population.



10.4 Unpaired or Independent Sample t-Test

In comparing the variances of two populations we have one of two situations :

1. Homoscedasticity : $\sigma_1^2 = \sigma_2^2$
2. Heteroscedasticity : $\sigma_1^2 \neq \sigma_2^2$

These terms also apply when there are more than 2 populations. They either all have the same variance, or not. This affects how we do an independent sample t -test because we have two cases :

1. Variances of the two populations assumed unequal. $\sigma_1^2 \neq \sigma_2^2$

Then the test statistic is :

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This is the same formula as we used for the z -test. To find the critical statistic we will use, when solving problems by hand, degrees of freedom

$$(10.2) \quad \nu = \min(n_1 - 1, n_2 - 1).$$

This choice is a conservative approach (harder to reject H_0). SPSS uses a more accurate

$$(10.3) \quad \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{s_1^2}{n_1}\right)}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)}{n_2-1}\right]}$$

You will not need to use Equation (10.3), only Equation (10.2). Equation (10.3) gives fractional degrees of freedom. The t test statistic for this case and the degrees of freedom in Equation (10.3) is known as the Satterwaite approximation. The t -distributions are strictly only applicable if $\sigma_1 = \sigma_2$. The Satterwaite approximation is an adjustment to make the t -distributions fit this $\sigma_1 \neq \sigma_2$ case.

2. Variances of the two populations assumed equal.

$\sigma_1 = \sigma_2 = \sigma$.

In this case the test statistic is:

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

This test statistic formula can be made more intuitive by defining

$$(10.4) \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

as the *pooled estimate of the variance*. s_p is the data estimate for the common population σ . s_p^2 is the weighted mean of the sample variances s_1^2 and s_2^2 . Recall the generic weighted mean formula, Equation (3.2). The weights are $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$; their sum is $\nu_1 + \nu_2 = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$. In other words

$$s_p^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}$$

and we can write the test statistic as

$$(10.5) \quad t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

See that $s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is clearly a standard error of the mean.

10.4.1 General form of the t test statistic

All t statistics have the form :

$$t_{\text{test}} = \frac{\text{Difference of means}}{\text{Standard error of the mean}} = \frac{\text{Signal}}{\text{Noise}}.$$

Remember that! Memorizing complicated formulae is useless, but you should remember the basic form of a t test statistic.

10.4.2 Two step procedure for the independent samples t test

We will use the F test to decide whether to use case 1 or 2. SPSS uses a test called “Levine’s test” instead of the F test we developed to test $H_0 : \sigma_1^2 \neq \sigma_2^2$. Levine’s test also produces an F test statistic. It is a different F than our F but you interpret it in the same way. If the p -value of the F is high (larger than α) then assume $\sigma_1 = \sigma_2$, if the p -value is low (smaller than α) then assume $\sigma_1 \neq \sigma_2$.

In real life, homoscedasticity is almost always assumed because the t -test is robust to violations of homoscedasticity until one sample set contains twice as many, or more, data points as the other.

Example 10.4: Case 1 example.

Given the following data summary :

$s_1 = 38$	$\bar{x}_1 = 191$	$n_1 = 8$
$s_2 = 12$	$\bar{x}_2 = 199$	$n_2 = 10$

(Note that $(s_1 = 38) > (s_2 = 12)$) Is \bar{x}_1 significantly different from \bar{x}_2 ? That is, is μ_1 different from μ_2 ? Test at $\alpha = 0.05$.

Solution :

So the question is to decide between

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

a two-tailed test. But before we can test the question, we have to decide which t test statistic to use: case 1 or 2. So we need to do two hypotheses tests in a row. The first one to decide which t_{test} statistic to use, the second one to test the hypotheses of interest given above.

Test 1 : See if variances can be assumed equal or not.

1. Hypothesis.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

(Always use a two-tailed hypothesis when using the F test to decide between case 1 and 2 for the t test statistic.)

2. Critical statistic.

$$F_{\text{crit}} = F_{\alpha/2, \nu_1, \nu_2} = F_{0.05/2, 7, 9} = F_{0.025, 7, 9} = 4.20$$

(from the **F Distribution Table**)

(Here we used α given for the t -test question. But that is not necessary. You can use $\alpha = 0.05$ in general; the consequence of a type I error here is small because the t -test is robust to violations of the assumption of homoscedasticity.)

3. Test statistic.

$$F_{\text{test}} = F_{7,9} = \frac{s_1^2}{s_2^2} = \frac{38^2}{12^2} = 10.03$$

4. Decision.

$10.03 > 4.20$ 4.20" title="Rendered by QuickLaTeX.com" height="13" width="95" style="vertical-align: -1px;">($F_{\text{test}} > F_{\text{crit}}$)
 $F_{\text{test}} > F_{\text{crit}}$ title="Rendered by QuickLaTeX.com" height="16" width="89" style="vertical-align: -4px;">- drawing a picture would be a safe thing to do here as usual) so reject H_0 .

5. Interpretation.

Assume the variances are unequal, $\sigma_1^2 \neq \sigma_2^2$, and use the t test statistic of case 1.

Test 2 : The question of interest.

1. Hypothesis.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

2. Critical statistic.

From the **t Distribution Table**, with $\nu = \min(n_1 - 1, n_2 - 1) = \min(8 - 1, 10 - 1) = 7$, and a two-tailed test with $\alpha = 0.05$ we find

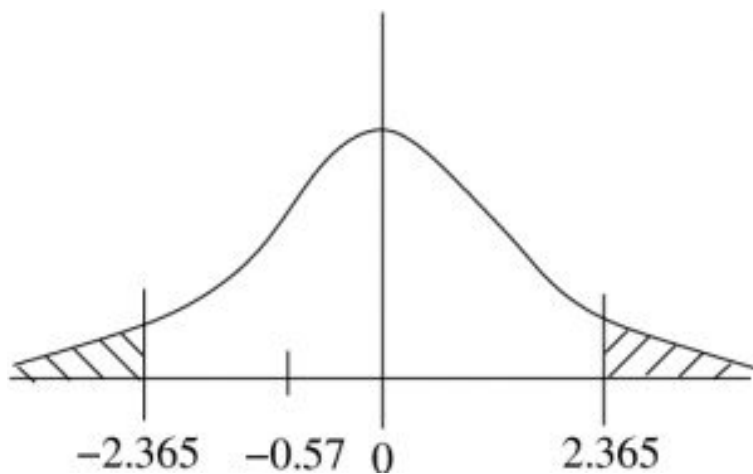
$$t_{\text{crit}} = \pm 2.365$$

3. Test Statistic.

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
 &= \frac{(191 - 199)}{\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}} = -0.57
 \end{aligned}$$

The p -value may be estimated from the **t Distribution Table** using the procedure given in Chapter 9: from the **t Distribution Table**, $\nu = 7$ line, find the values that bracket 0.57. There are none, the smallest value is 0.711 corresponding to $\alpha = 0.50$. So all we can say is $p > 0.50$.

4. Decision.



$t_{\text{test}} = -0.57$ is not in the rejection region so do not reject H_0 . The estimate for the p -value confirms this decision.

5. Interpretation.

There is not enough evidence, at $\alpha = 0.05$ with the independent sample t -test, to conclude that the means of the populations are different.

□

Example 10.5 (Case 2 example) :

The following data seem to show that private nurses earn more than government nurses :

Private Nurses Salary	Government Nurses Salary
$\bar{x}_1 = 26,800$	$\bar{x}_2 = 25,400$
$s_1 = 600$	$s_2 = 450$
$n_1 = 10$	$n_2 = 8$

Testing at $\alpha = 0.01$, do private nurses earn more than government nurses?

Solution :

First confirm, or change, the population definitions so that $s_1^2 > s_2^2$. This is already true so we are good to go.

Test 1 : See if variances can be assumed equal or not. This is a test of $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$. After the test we find that we believe that $\sigma_1^2 = \sigma_2^2$ at $\alpha = 0.05$. So we will use the case 2, equal variances, t -test formula for test 2, the test of interest.

Test 2 : The question of interest.

1. Hypothesis.

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

(Note how H_1 reflects the face value of the data, that private nurses appear to earn more than government nurses in the population – it is true in the samples.)

2. Critical statistic.

Use the **t Distribution Table**, one-tailed test, $\alpha = 0.01$ (column) and $\nu = n_1 + n_2 - 2 = 10 + 8 - 2 = 16$ to find

$$t_{\text{crit}} = 2.583$$

3. Test statistic.

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t_{\text{test}} = \frac{(26,800 - 25,400)}{\sqrt{\frac{(10-1)600^2 + (8-1)450^2}{10+8-2}} \sqrt{\frac{1}{10} + \frac{1}{8}}}$$

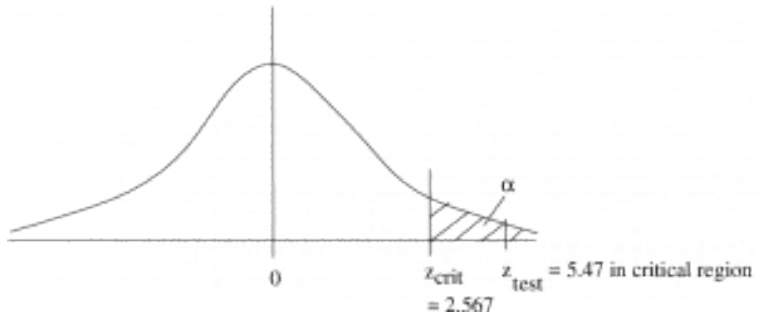
$$t_{\text{test}} = \frac{1400}{\sqrt{\frac{(9)(360000) + (7)(202500)}{16}} \sqrt{0.1 + 0.125}}$$

$$t_{\text{test}} = \frac{1400}{\sqrt{\frac{3240000 + 1417500}{16}} \sqrt{0.225}}$$

$$t_{\text{test}} = \frac{1400}{(\sqrt{291093.75})(\sqrt{0.225})} = 5.47$$

To estimate the p -value, look at the $\nu = 16$ line in the **t Distribution Table** to see if there are a pair of numbers that bracket $t_{\text{test}} = 5.47$. They are all smaller than 5.47 so p is less than the α associated with the largest number 2.921 whose α is 0.005 (one-tailed, remember). So $p < 0.005$.

4. Decision.



Reject H_0 since t_{test} is in the rejection region and $(p < 0.005) < (\alpha = 0.01)$.

$$t_{test} > t_{crit} \quad (5.47 > 2.583) \quad t_{crit} \quad \text{\hspace{.25in}} \\ (5.47 > 2.583) \quad \text{\hspace{.25in}} \text{" title="Rendered by QuickLaTeX.com" >}$$

5. Interpretation.

From a t -test at $\alpha = 0.01$, there is enough evidence to conclude that private nurses earn more than government nurses.

□

10.5 Confidence Intervals for the Difference of Two Means

The form of the confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

but, as with hypothesis testing, we have two cases to choose from to get the formula for E :

Case 1 : Variances of the 2 populations unequal}

$$E = t_C \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the degrees of freedom to use when looking up t_C in the **t Distribution Table** is

$$\nu = \min[(n_1 - 1), (n_2 - 1)]$$

Case 2 : Variances of the 2 populations equal

$$E = t_C \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where we use

$$\nu = n_1 + n_2 - 2$$

when looking up t_C .

To select the appropriate formula for E we need to do a preliminary hypothesis test on $H_0 : \sigma_1^2 = \sigma_2^2$. An odd combination of hypothesis test followed by confidence interval calculation.

Insight! By now you should have noticed that the formulae for E are just t times standard error of the mean. This whole z -transformation thing should be becoming somewhat transparent.

Example 10.6 : Find the 95% confidence interval for $\mu_1 - \mu_2$ for the data of Example 10.4 :

$s_1 = 38$	$\bar{x}_1 = 191$	$n_1 = 8$
$s_2 = 12$	$\bar{x}_2 = 199$	$n_2 = 100$

Solution :

First use F -test to see which formula to use. We did this already in Example 10.4 (the data come from that question) and found that we believed $\sigma_1^2 \neq \sigma_2^2$ with $\alpha = 0.05$.

Next, look up t_C in the **t Distribution Table** for 95% confidence interval for $\nu = 7$:

$$t_{95\%} = 2.365$$

Compute

$$E = t_{95\%} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

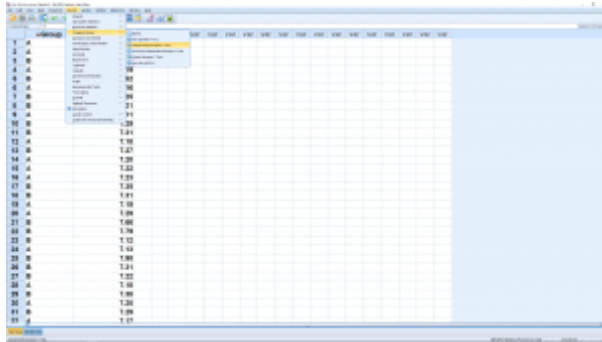
$$E = 2.365 \sqrt{\frac{38^2}{8} + \frac{12^2}{10}} = 33.01$$

So

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - E &< \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E \\ (191 - 199) - 33.02 &< \mu_1 - \mu_2 < (191 - 199) + 33.02 \\ -8 - 33.02 &< \mu_1 - \mu_2 < -8 + 33.02 \\ -41.02 &< \mu_1 - \mu_2 < 25.02 \end{aligned}$$

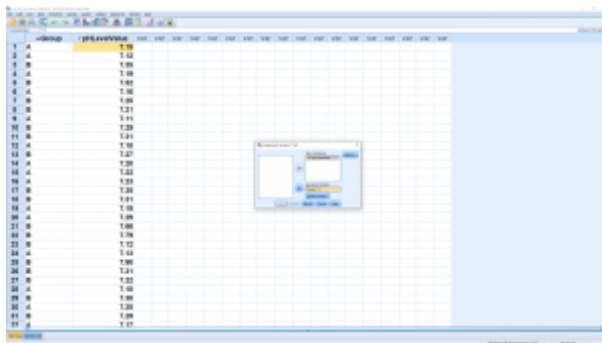
be careful of the order!

□



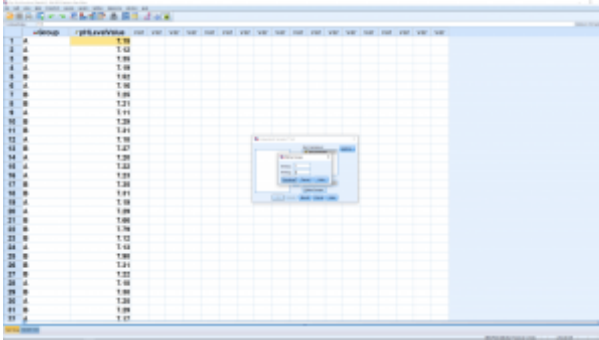
SPSS
screenshot ©
International
Business
Machines
Corporation.

Select Sepal.Length as the Test Variable (dependent variable) and Species as the group variable (independent variable) :



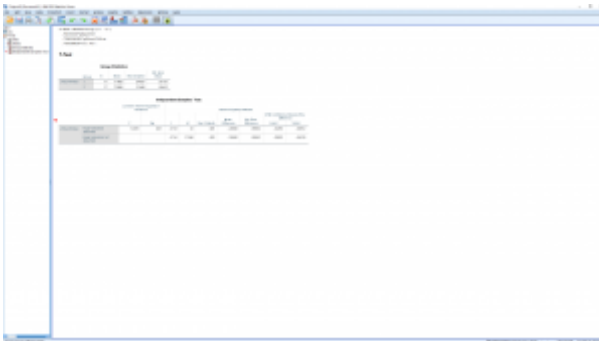
SPSS
screenshot ©
International
Business
Machines
Corporation.

You need to do some work to let SPSS know that the two levels of the “grouping variable” are 1 and 2 (as can be seen in the Variable View window). So hit Define Groups... and enter:



SPSS
screenshot ©
International
Business
Machines
Corporation.

Hit Continue, then OK (the Options menu will allow you to set the confidence level percent) to get:



SPSS
screenshot ©
International
Business
Machines
Corporation.

The first table shows descriptive statistics for the two groups independently. These numbers, excluding standard error numbers can be plugged into the t_{test} formulae for pencil and paper calculations.

The important table is the second table. First, what hypothesis are we testing? It is important to write it out explicitly:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$(10.6) \quad H_1 : \mu_1 - \mu_2 \neq 0$$

This, as you recall, is our test of interest. When we did this test

by hand, we had to do a preliminary F test to see if we could assume homoscedasticity or not :

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 \\ (10.7) \quad H_1 : \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

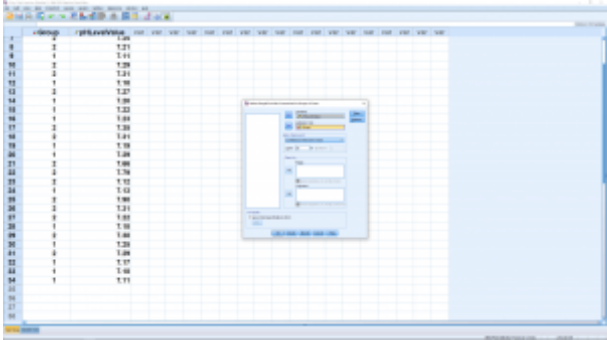
That preliminary test is given to us as Levine's test in the first two columns of the second table. Levine's test is similar to but not exactly the same as the F test we used but it also uses F as a test statistic. Here we see $F_{\text{test}} = 12.061$ with $p = 0.001$, so we reject H_0 and assume that population variances are unequal. That means we look at only the second line of the second table corresponding to "Equal variances not assumed". SPSS computes t and p using both t formulae but it does not decide for you which one is correct. You need to decide that yourself on the basis of the Levine's test.

Again the information is fairly redundant. Looking across the second row we have $t_{\text{test}} = -3.741$ (note that it is the same as the t in the first row - that's because the sample is large, making z a good approximation for both), $\nu = 32$ (notice the fractional ν here for the heteroscedastic case - recall Equation (10.3)), $p = 0.001$ (note that it is for a two-tailed hypothesis, if your hypothesis is one-tailed then divide p by 2), $\bar{x}_1 - \bar{x}_2 = -0.208$, and the standard error, the denominator of the t test statistic formula (t is mean over standard error). The p value is small, so we reject H_0 , the difference of the sample means is significant. The last two columns give the 95% confidence interval as

$$(10.8) \quad 0.75429 < \mu_1 - \mu_2 < 1.10571$$

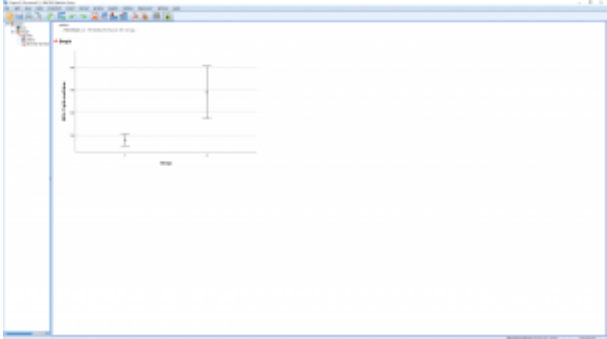
Notice that zero is not in the confidence interval, consistent with rejecting H_0 .

We can also make an error bar plot. Go through Graphs \rightarrow Legacy Dialogs \rightarrow Errorbar and pick Simple and "Summaries for groups of cases" in the next menu and:



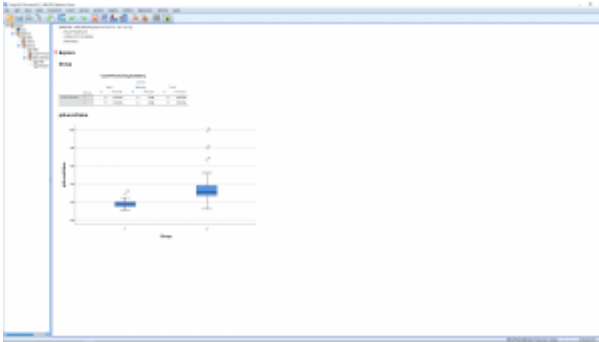
SPSS
screenshot ©
International
Business
Machines
Corporation.

which results in:



SPSS
screenshot ©
International
Business
Machines
Corporation.

or you could generate a boxplot comparison:



SPSS
screenshot ©
International
Business
Machines
Corporation.

Finally, we throw in a couple of effect size (descriptive) measures. One is the standardized effect size defined as:

$$(10.9) \quad d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where s_p is the pooled variance as given by Equation (10.4). Another measure is the strength of association

$$(10.10) \quad \eta^2 = \frac{t^2}{t^2 + (n_1 + n_2 - 2)}$$

which measures a kind of “correlation” between x_1 and x_2 . The larger t , the closer η^2 is to 1.

10.7 RStudio Lesson 6: Independent Sample t-Test

[Coming soon]

10.8 Paired t-Test

Here two measurements x_1 and x_2 are taken from every subject. We could say that we measure a vector $\vec{x} = [x_1 \ x_2]^T$ as the independent variable for every subject instead of just a number as the independent variable¹. This is a *within subject* design. Within subject designs tend to be more statistically powerful than independent or between subjects designs that have two completely different bunches of people for each variable. The extra power comes because we take the difference $D = x_1 - x_2$ for every subject. So any overall differences, or variances, in x_1 or x_2 due to individuals has been removed from the data.

The paired t -test is a *univariate test*. The difference between univariate and multivariate statistics is the the independent variables are numbers for univariate statistics and vectors for multivariate statistics. For the paired t -test, the vector is converted to a number by taking a difference. To convert vector data to difference data, make a table :

x_1	x_2	$D = x_1 - x_2$
1	2	-1
2	3	-1
3	5	-2
1	-2	3

Note here that the differences in individuals are gone after we take differences D .

The data from the D column are what you will work with.

1. An introduction to vectors will be given in Chapter 17.

Compute \bar{D} and s_D the mean and sample standard deviation of these data. With D the procedure becomes a single sample t -test of D against zero. Specifically we can test :

Two-tailed	Left-tailed	Right-tailed
$H_0 : \mu_D = 0$	$H_0 : \mu_D \geq 0$	$H_0 : \mu_D \leq 0$
$H_1 : \mu_D \neq 0$	$H_1 : \mu_D < 0$	$H_1 : \mu_D > 0$

The test statistic is

$$t_{\text{test}} = \frac{\bar{D}}{(s_D/\sqrt{n})}$$

with $\nu = n - 1$ (for finding t_{crit}).

Example 10.7 : A Physical Education director claims that a vitamin will increase a weight lifter's strength. Eight athletes are selected and tested on how much they can bench press. They are each tested once before taking the vitamin and again after taking the vitamin for two weeks. We want to test the director's claim at $\alpha = 0.05$

The data are :

Athlete	Before(x_1)	After(x_2)	$D = x_1 - x_2$
1	210	219	-9
2	230	236	-6
3	182	179	3
4	205	204	1
5	262	270	-8
6	253	250	3
7	219	222	-3
8	216	216	0

Here we have listed the differences which is actually part of step 0 of the solution. The x_1 and x_2 columns are what you enter into SPSS as your independent variables. With SPSS you never see the differences.

Solution :

0. Data reduction.

Compute $\bar{D} = -2.375$, $s_D = 4.84$ by entering the difference data into your calculator.

1. Hypothesis.

$$H_0 : \mu_D \geq 0$$

$$H_1 : \mu_D < 0$$

Note that a negative difference, based on $x_1 - x_2$ (always consistently give population 1 priority if you want to stay out of trouble without thinking), indicates an increase in strength. It is important to interpret positive or negative differences correctly by thinking about what they mean.

2. Critical statistic.

Using the **t Distribution Table** with the column for one-tailed tests, $\alpha = 0.05$, and row $\nu = n - 1 = 8 - 1 = 7$, find

$$t_{\text{crit}} = -1.895$$

(We added the negative sign because this is a left-tailed test.)

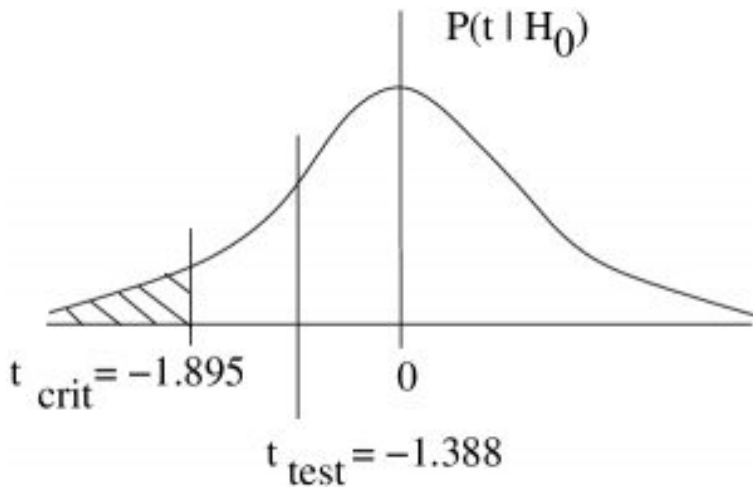
3. Test statistic.

$$t_{\text{test}} = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)} = \frac{-2.375 - 0}{\left(\frac{4.84}{\sqrt{8}}\right)}$$

$$t_{\text{test}} = -1.388$$

To estimate the p -value, from the **t Distribution Table**, $\nu = 7$ line, find $0.10 < p < 0.25$.

4. Decision.



Do not reject H_0 . ($0.10 < p < 0.25$) $>$ ($\alpha = 0.05$).

5. Interpretation.

Under a paired t -test, at $\alpha = 0.05$, there is not enough evidence to conclude that the vitamin increases strength.

□

10.9 Confidence Intervals for Paired t-Tests

The usual form applies :

$$\bar{D} - E < \mu_D < \bar{D} + E$$

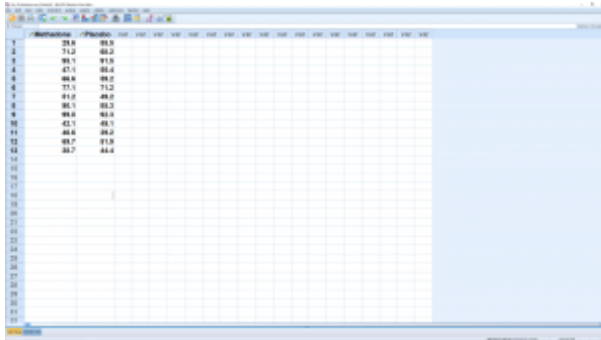
where now

$$E = t_C \left(\frac{s_D}{\sqrt{n}} \right)$$

and t_C can be found from the **t Distribution Table** in the $\nu = n - 1$ line using the “confidence intervals” heading.

10.10 SPSS Lesson 7: Paired Sample t-Test

To follow along, load in the [Data Set](#) “Methadone.sav”:

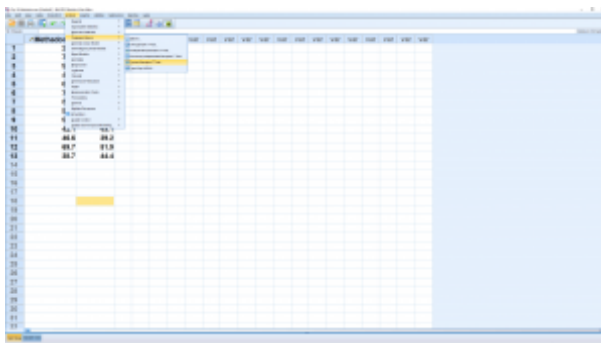


The screenshot shows the SPSS Data Editor window for the file 'Methadone.sav'. The data is organized into two columns. The first column contains values: 28.0, 71.2, 98.7, 27.1, 88.0, 77.1, 91.0, 88.7, 42.7, 88.0, 88.7, 88.7. The second column contains values: 88.0, 91.0, 88.0, 71.2, 88.0, 48.1, 88.0, 88.0, 88.0, 91.0, 88.0, 88.0.

Case #	Column 1	Column 2
1	28.0	88.0
2	71.2	91.0
3	98.7	88.0
4	27.1	71.2
5	88.0	88.0
6	77.1	71.2
7	91.0	88.0
8	88.7	88.0
9	42.7	48.1
10	88.0	88.0
11	88.7	91.0
12	88.7	88.0

SPSS
screenshot ©
International
Business
Machines
Corporation.

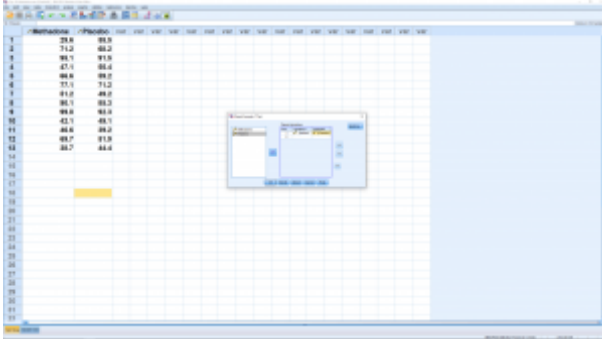
As set up, the file has two dependent variables. This “within subjects” dataset is fundamentally multivariate. When we did the paired t -test by hand we converted the multivariate data to univariate data by taking differences. SPSS will do the differences behind the scene and you won’t actually see them. Run the t -test by picking Analyze → Compare Means → Paired -Samples T-Test:



The screenshot shows the SPSS Data Editor window with the 'Analyze' menu open. The 'Compare Means' option is selected, and the 'Paired-Samples T-Test' option is highlighted. The data table from the previous screenshot is visible in the background.

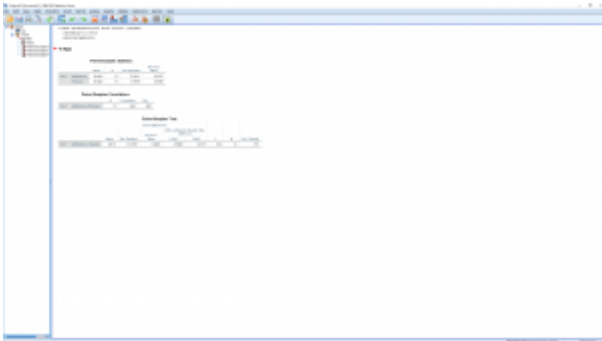
SPSS
screenshot ©
International
Business
Machines
Corporation.

Move the two variables into Pair 1 and hit OK (Options again allows you to specify a confidence intervals percentage):



SPSS screenshot © International Business Machines Corporation.

The output is:



SPSS screenshot © International Business Machines Corporation.

The first two tables are descriptive statistics. The last table gives the stuff we want: $\bar{D} = 0.9615$, $s_D = 10.7067$, the confidence interval

$$(10.11) \quad -5.5084 < \mu_D < 7.4315,$$

$t_{\text{test}} = 0.324$, $\nu = 12$ and $p = 0.002$ for the two-tailed hypotheses pair

$$H_0 : \mu_D = 0$$

$$(10.12) \quad H_1 : \mu_D \neq 0.$$

The very low p -value (0 in this case) and the absence of 0 in the confidence interval guide us to reject H_0 , the differences are significantly different from zero.

The standardized effect size and strength of association for the paired t -test are

$$(10.13) \quad d = \frac{t}{\sqrt{n}} = \frac{\bar{D}}{s_D}$$

and

$$(10.14) \quad \eta^2 = \frac{t^2}{t^2 + n - 1}$$

respectively.

10.11 RStudio Lesson 7: Paired Sample t-Test

[Coming soon]

II. COMPARING PROPORTIONS

In this Chapter we will use a χ^2 test to compare proportions and extend what we do here with the z -distribution.

11.1 z-Test for Comparing Proportions

11.2 Confidence Interval for the Difference between Two Proportions

12. ANOVA

12.1 One-way ANOVA

12.2 Post hoc Comparisons

12.3 SPSS Lesson 9: One-way ANOVA

12.4 R Lesson 9: One-way ANOVA

12.5 Two-way ANOVA

12.6 SPSS Lesson 9: Two-way ANOVA

12.7 R Lesson 9: Two-way ANOVA

12.8 Higher Factorial ANOVA

12.9 Between and Within Factors

12.10 *Contrasts

13. POWER

13.1 Power

14. CORRELATION AND REGRESSION

14.1 Scatter Plots

14.2 Correlation

14.3 SPSS Lesson 10: Scatterplots and Correlation

14.4 R Lesson 10: Scatterplots and Correlation

14.5 Linear Regression

14.6 r-squared and the Standard Error of the Estimate of y -prime

14.7 Confidence Interval for y-prime at a Given x

14.8 SPSS Lesson II: Linear Regression

14.9 R Lesson 11: Linear Regression

14.10 Multiple Regression

14.II SPSS Lesson 12: Multiple Regression

14.12 R Lesson 12: Multiple Regression

15. CHI SQUARED: GOODNESS OF FIT AND CONTINGENCY TABLES

15.1 Goodness of Fit

15.2 Contingency Tables

15.3 SPSS Lesson 13: Proportions, Goodness of Fit, and Contingency Tables

15.4 R Lesson 13: Proportions, Goodness of Fit, and Contingency Tables

16. NON-PARAMETRIC TESTS

16.1 How to Rank Data

16.2 Median Sign Test

16.3 Paired Sample Sign Test

16.4 Two Sample Wilcoxon Rank Sum Test (Mann-Whitney U Test)

16.5 Paired Wilcoxon Signed Rank Test

16.6 Kruskal-Wallis Test (H Test)

16.7 Spearman Rank Correlation Coefficient

16.8 SPSS Lesson 14: Non-parametric Tests

16.9 R Lesson 14: Non-parametric Tests

16.10 Runs Test

17. OVERVIEW OF THE GENERAL LINEAR MODEL

17.1 Linear Algebra Basics

17.2 The General Linear Model (GLM) for Univariate Statistics

Appendix: Tables

- Binomial Distribution Table ([PDF](#)) ([Word](#))
- Standard Normal Distribution Table ([PDF](#)) ([Word](#))
- t Distribution Table ([PDF](#)) ([Word](#))
- Chi Squared Distribution Table ([PDF](#)) ([Word](#))
- F Distribution Table ([PDF](#)) ([Word](#))