INTRODUCTION TO APPLIED STATISTICS FOR PSYCHOLOGY STUDENTS



Gordon E. Sarty
University of Saskatchewan

LibreTexts™

Introduction to Applied Statistics for Psychology Students (Sarty)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 06/17/2026

TABLE OF CONTENTS

Licensing

About This Book

Acknowledgements

Statistical Software Used in this Book

Data Sets

1: Background and Motivation

- 1.1: Overview
- 1.2: Basic Definitions
- 1.3: Summation Convention

2: Descriptive Statistics - Frequency Data (Counting)

- 2.1: Frequency Tables
- 2.2: Plotting Frequency Data
- 2.3: SPSS Lesson 1- Getting Started with SPSS

3: Descriptive Statistics- Central Tendency and Dispersion

- 3.1: Central Tendency- Mean, Median, Mode
- 3.2: Dispersion- Variance and Standard Deviation
- 3.3: z-score / z-transformation
- 3.4: SPSS Lesson 2- Combining variables and recoding

4: Probability and the Binomial Distributions

- 4.1: Probability
- 4.2: Binomial Distribution
- 4.3: SPSS Lesson 3- Combining variables - advanced

5: The Normal Distributions

- 5.1: Discrete versus Continuous Distributions
- 5.2: **The Normal Distribution as a Limit of Binomial Distributions
- 5.3: Normal Distribution

6: Percentiles and Quartiles

- 6.1: Discrete Data Percentiles and Quartiles
- 6.2: Finding Outliers Using Quartiles
- 6.3: Box Plots
- 6.4: Robust Statistics
- 6.5: SPSS Lesson 4- Percentiles

7: The Central Limit Theorem

- 7.1: Using the Normal Distribution to Approximate the Binomial Distribution
- 7.2: The Central Limit Theorem

8: Confidence Intervals

- 8.1: Confidence Intervals Using the z-Distribution
- 8.2: **Bayesian Statistics
- 8.3: The t-Distributions
- 8.4: Proportions and Confidence Intervals for Proportions
- 8.5: Chi Squared Distribution

9: Hypothesis Testing

- 9.1: Hypothesis Testing Problem Solving Steps
- 9.2: z-Test for a Mean
- 9.3: t-Test for Means
- 9.4: z-Test for Proportions
- 9.5: Chi Squared Test for Variance or Standard Deviation
- 9.6: SPSS Lesson 5- Single Sample t-Test

10: Comparing Two Population Means

- 10.1: Unpaired z-Test
- 10.2: Confidence Interval for Difference of Means (Large Samples)
- 10.3: Difference between Two Variances - the F Distributions
- 10.4: Unpaired or Independent Sample t-Test
- 10.5: Confidence Intervals for the Difference of Two Means
- 10.6: SPSS Lesson 6- Independent Sample t-Test
- 10.7: Paired t-Test
- 10.8: Confidence Intervals for Paired t-Tests
- 10.9: SPSS Lesson 7- Paired Sample t-Test

11: Comparing Proportions

- 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 8- One-way ANOVA
- 12.4: Two-way ANOVA
- 12.5: SPSS Lesson 9- Two-way ANOVA
- 12.6: Higher Factorial ANOVA
- 12.7: Between and Within Factors
- 12.8: *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: Linear Regression
- 14.5: r^2 and the Standard Error of the Estimate of y'
- 14.6: Confidence Interval for y' at a Given x
- 14.7: SPSS Lesson 11- Linear Regression
- 14.8: Multiple Regression
- 14.9: SPSS Lesson 12- Multiple Regression

15: Chi Squared- Goodness of Fit and Contingency Tables

- 15.1: Goodness of Fit

16: Non-parametric Tests

- 16.1: New Page
- 16.2: New Page
- 16.3: New Page
- 16.4: New Page
- 16.5: New Page
- 16.6: New Page
- 16.7: New Page
- 16.8: New Page
- 16.9: New Page

17: Overview of the General Linear Model

- 17.1: New Page
- 17.2: New Page

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

About This Book

Acknowledgements

Statistical Software Used in this Book

Data Sets

CHAPTER OVERVIEW

1: Background and Motivation

1.1: Overview

1.2: Basic Definitions

1.3: Summation Convention

This page titled 1: Background and Motivation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

1.1: Overview

1.1: Overview is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

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.

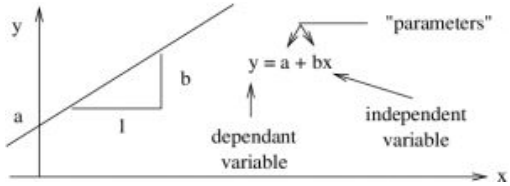


Figure 1.4: In the equation of a line, y is the dependent variable, x is the independent variable.

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 $image$" title="Rendered by QuickLaTeX.com" height="12" width="12" style="vertical-align: -2px;"> 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 you 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 years[1] 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” 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 nit 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. Bluman AG, Elementary Statistics: A Step-by-Step Approach, numerous editions, McGraw-Hill Ryerson, circa 2005. ↵

This page titled 1.2: Basic Definitions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

This page titled 1.3: Summation Convention is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

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

2.2: Plotting Frequency Data

2.3: SPSS Lesson 1- Getting Started with SPSS

This page titled 2: Descriptive Statistics - Frequency Data (Counting) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$W = \frac{R+1}{G} \tag{2.1.1}$$

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 numbers[1]. 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 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
				⋮	⋮

				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 count[2] 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 :

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. □

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.

---

1. Whole numbers are 0 and the positive integers. ↵
2. The frequency of A is the number of times A is in the dataset, etc. ← **the take-home concept here.** ↵

This page titled 2.1: Frequency Tables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

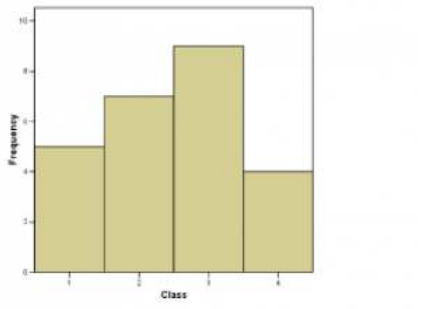


Figure 2.1 : Straight Forward histogram. A box or "step function" is used to show the frequency of each class. In this image, generated with SPSS, the classes are labelled with 1, 2, 3, and 4 which correspond to the classes A, B, O and AB. If we take these discrete quantitative class values literally, 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.

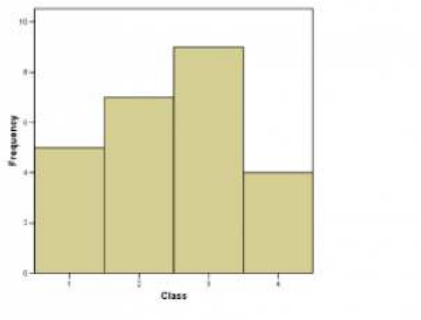


Figure 2.2 : Relative frequency histogram for the blood type data.

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 0 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.

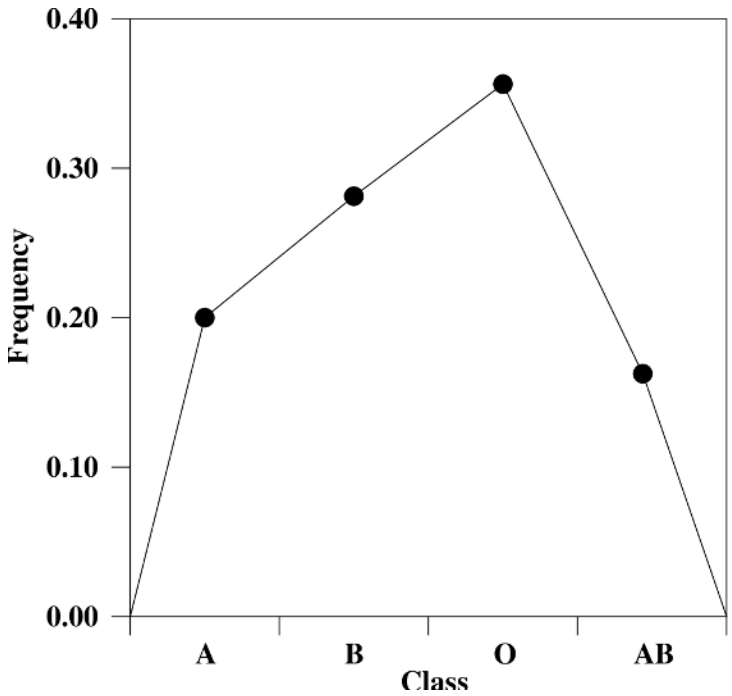


Figure 2.3 : Relative frequency polygon for the blood type data. Plot a dot at the center of each class at the y-value of the relative frequency then connect the dots as shown.

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 with 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.)

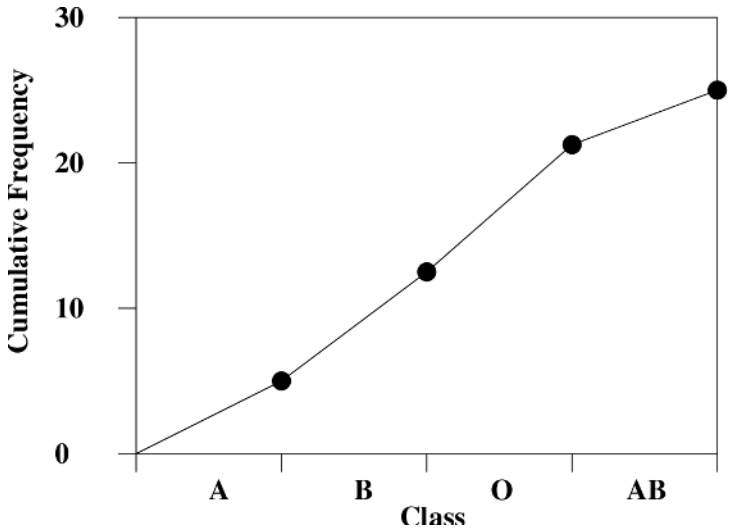


Figure 2.4 : Cumulative frequency graph for the blood sample data. Plot a dot at the end of the relevant class at a y-value equal to the cumulative frequency. Then connect the dots as shown.

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 $= 360^\circ$

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

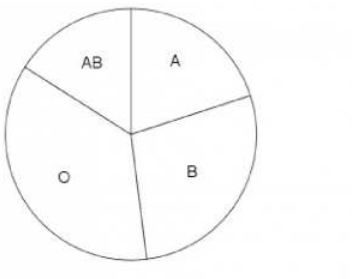


Figure 2.5 : Pie chart for the blood type data. It is a very good representation of the probability aspect of relative frequency. If you made the pie chart into a dart board and threw darts at it in a random fashion, then the probability of the dart landing in each class is equal to that class's relative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

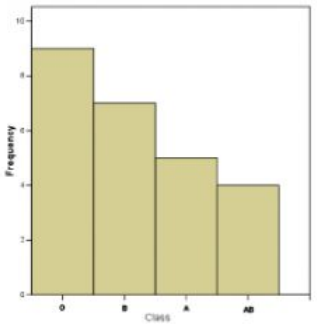


Figure 2.6: Pareto chart for the blood type data.

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,|65,65,66,66,67,68,69,69|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.

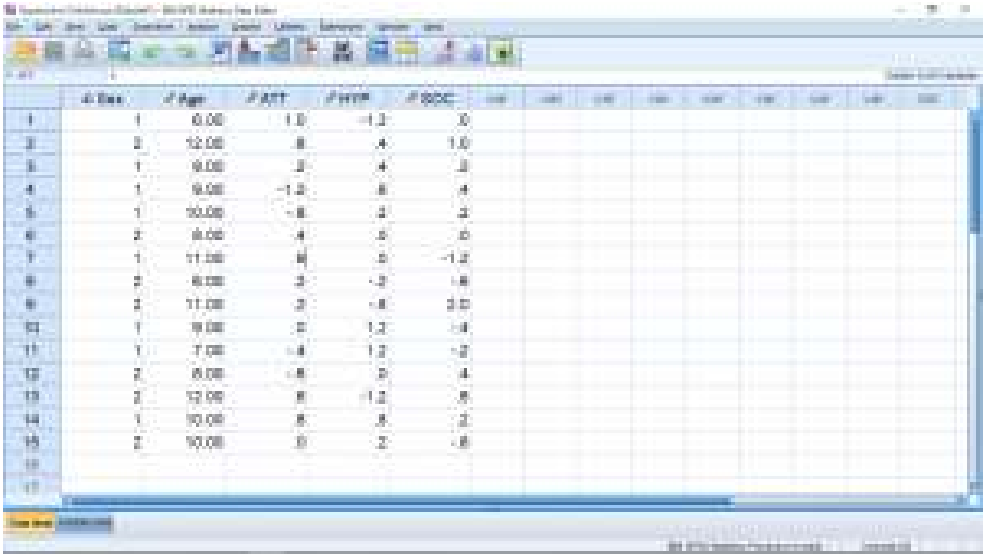
This page titled 2.2: Plotting Frequency Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

2.3: SPSS Lesson 1- 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:



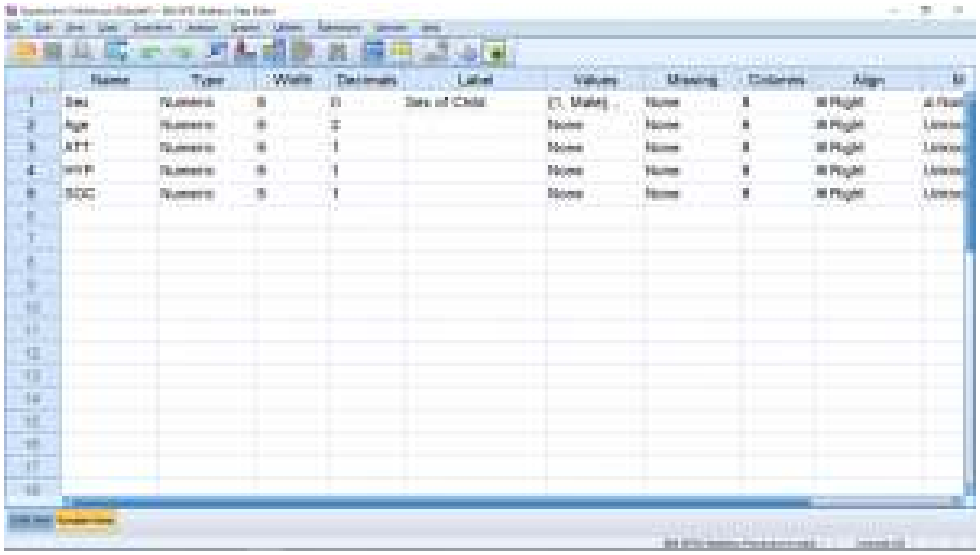
SPSS screenshot © International Business Machines Corporation.

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 :

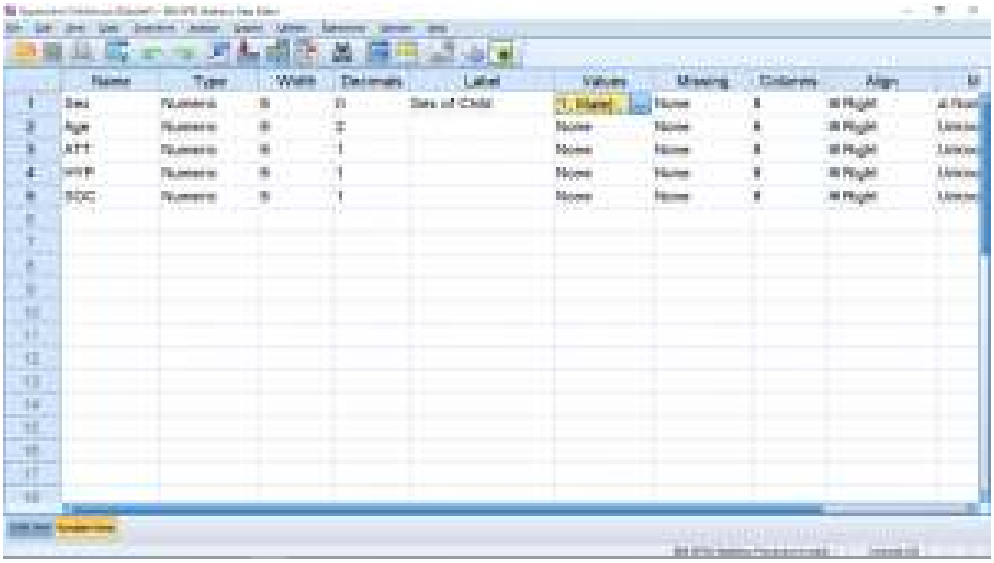


SPSS screenshot © International Business Machines Corporation.

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.

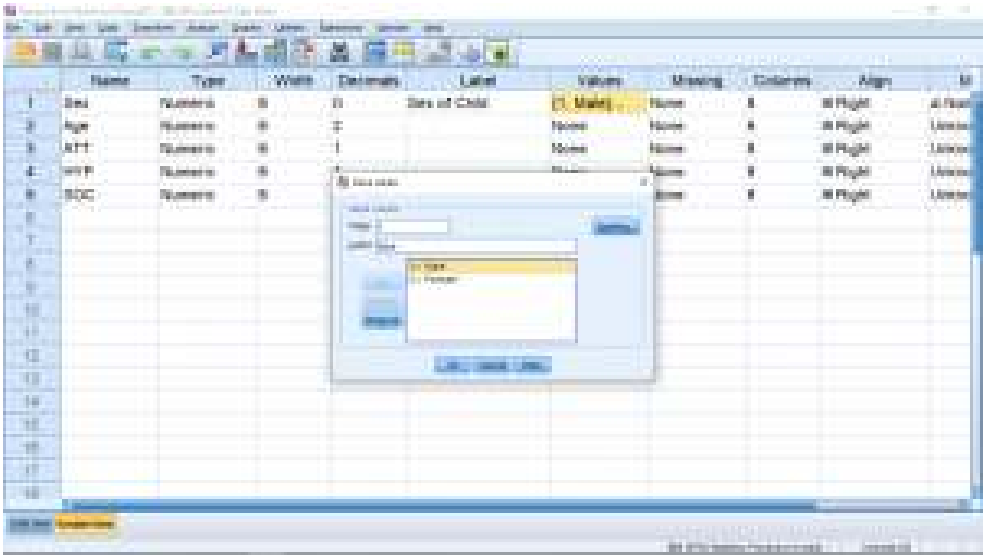
The Name attribute gives the name of the variable as it appears at the top of the columns in the Data View window. Type should be Numeric if you want to use the variable in any kind of statistical calculation. Having this set to String will cause errors if you are trying to use the variable as a qualitative 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 the 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.

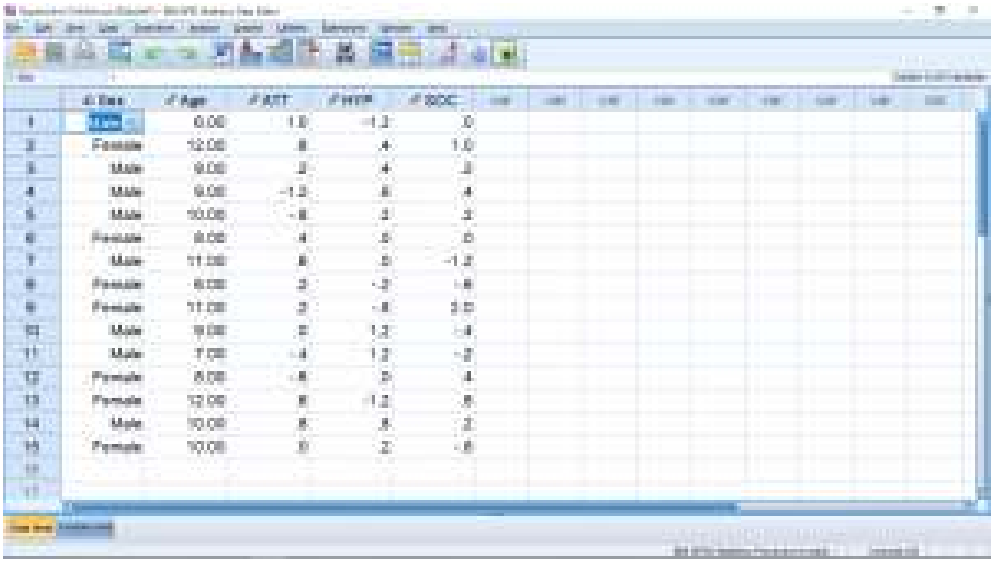
Clicking on the thing with three dots brings up a menu where you can define the connection between the qualitative description and your discrete number assignments :



SPSS screenshot © International Business Machines Corporation.

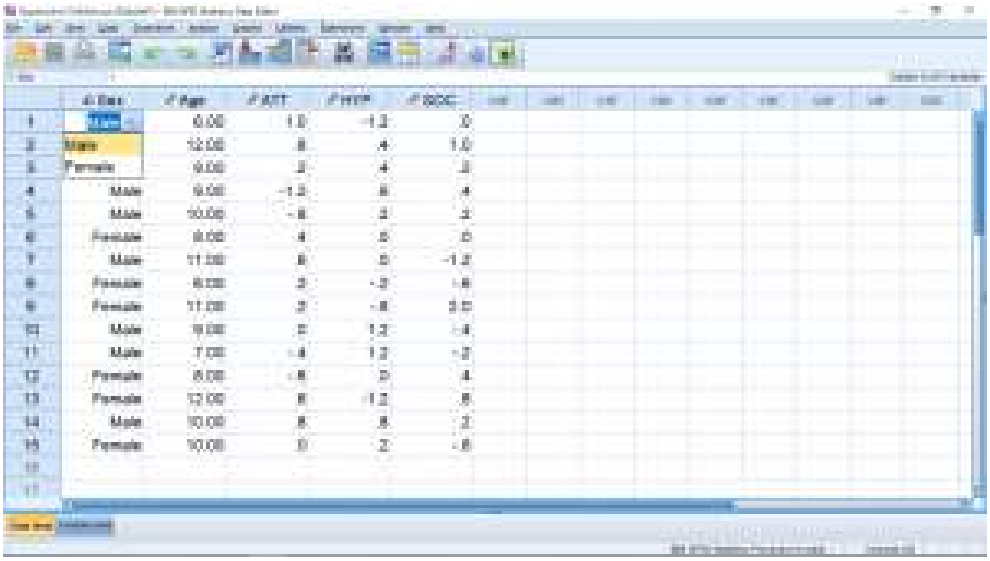
Here I have clicked on the 1.00 = “Male” line to show that the Value is 1 (arbitrary discrete quantitative) and the Label is Male (qualitative). To enter new values, type them in the Value and Label box and then click Add to add them to the list.

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.

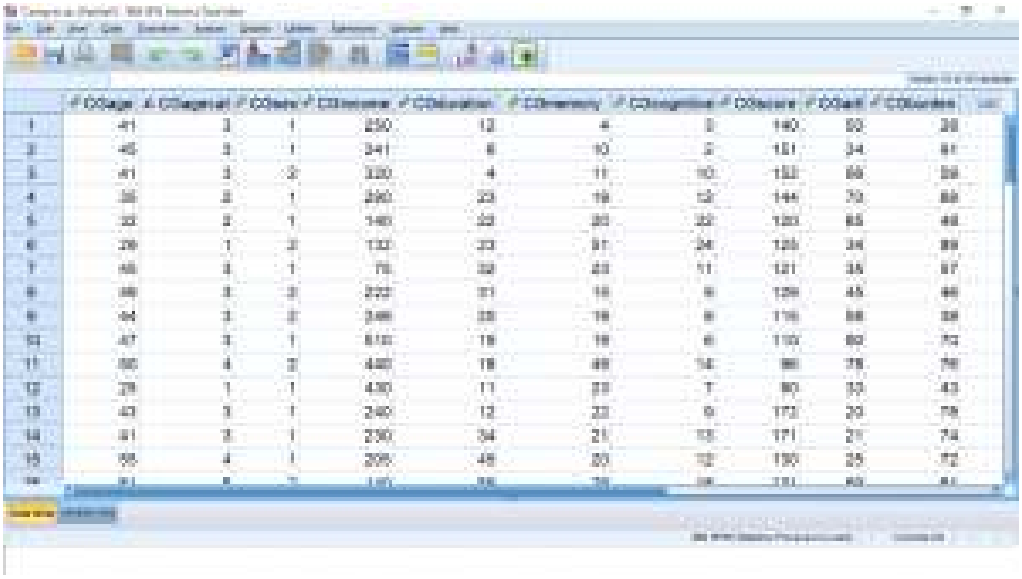
There's more. If you click on a cell in the gender variable, you will get a thing on the side of the cell and if you click on that thing, you will see:



SPSS screenshot © International Business Machines Corporation.

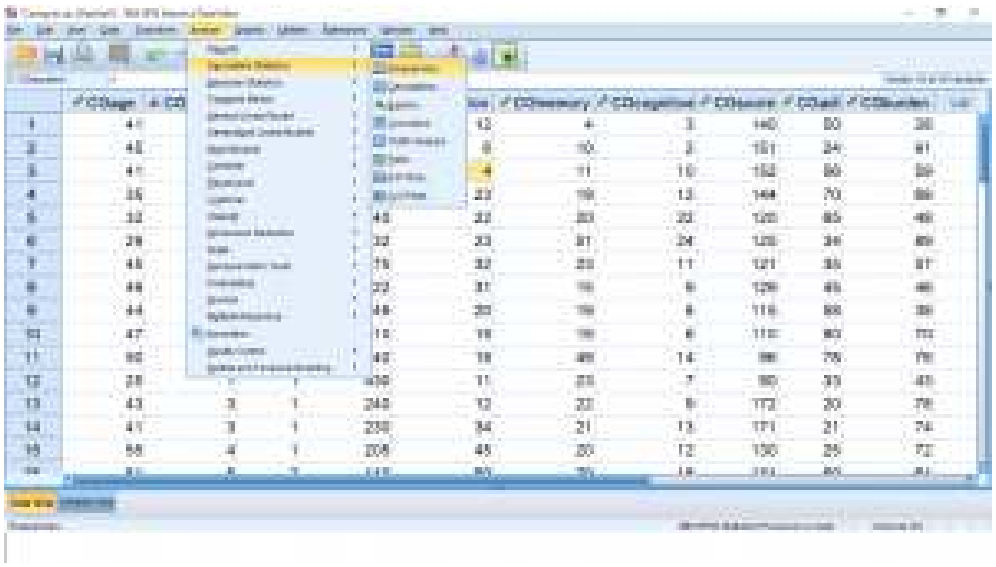
This pop-up allows you to change the value by clicking on the appropriate value. In one of your assignments you will get practice with entering qualitative data this way. In general, to enter data into SPSS from scratch, you can start by typing data into the Data View window and then fix up the attributes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



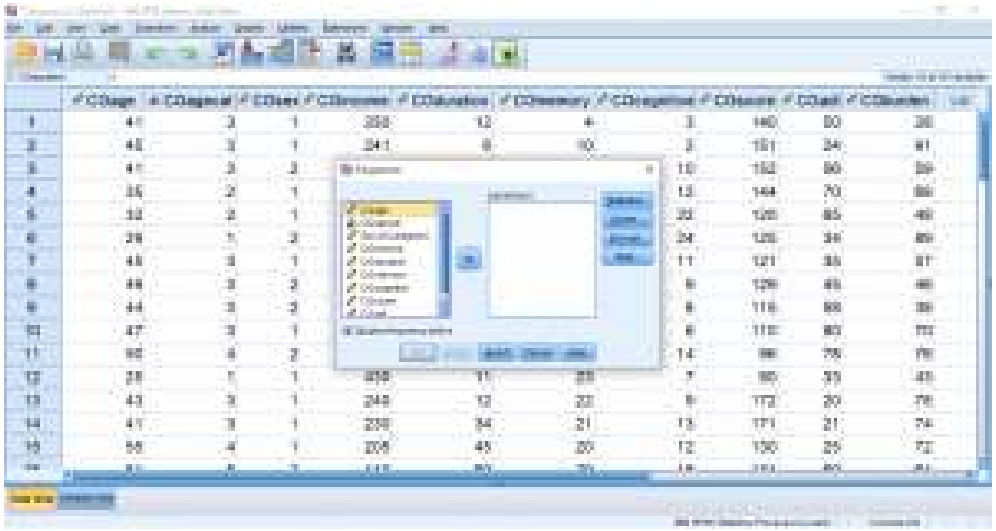
SPSS screenshot © International Business Machines Corporation.

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, lets 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):



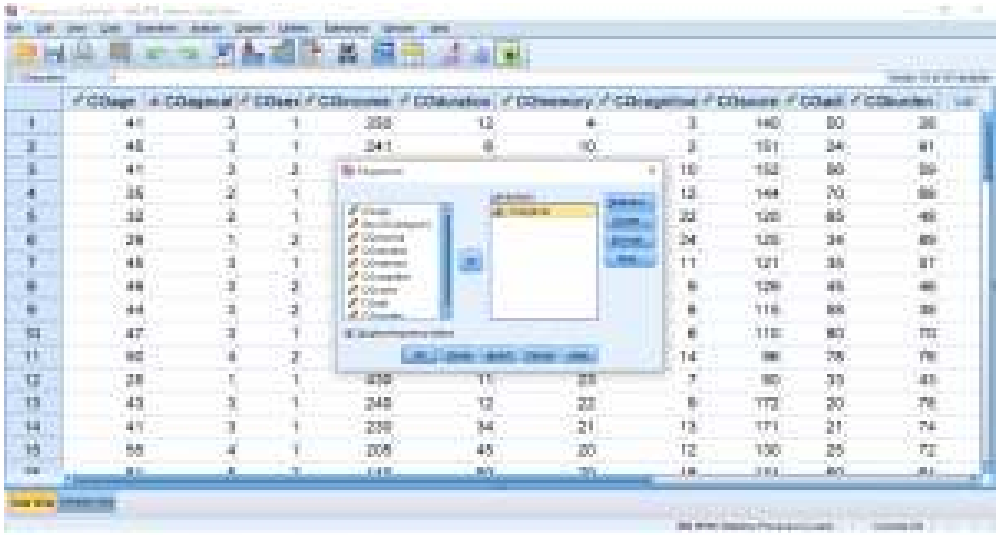
SPSS screenshot © International Business Machines Corporation.

Pick Frequencies… which brings up:



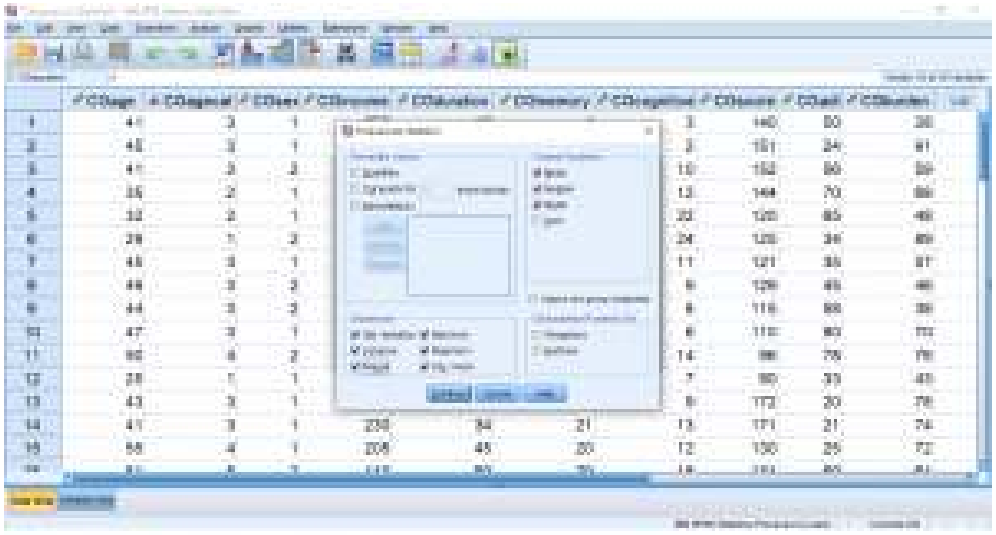
SPSS screenshot © International Business Machines Corporation.

Move the CGagecat variable over by clicking on the variable then the arrow button or just drag the variable over to get:



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 ($\overline{x}$), 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.



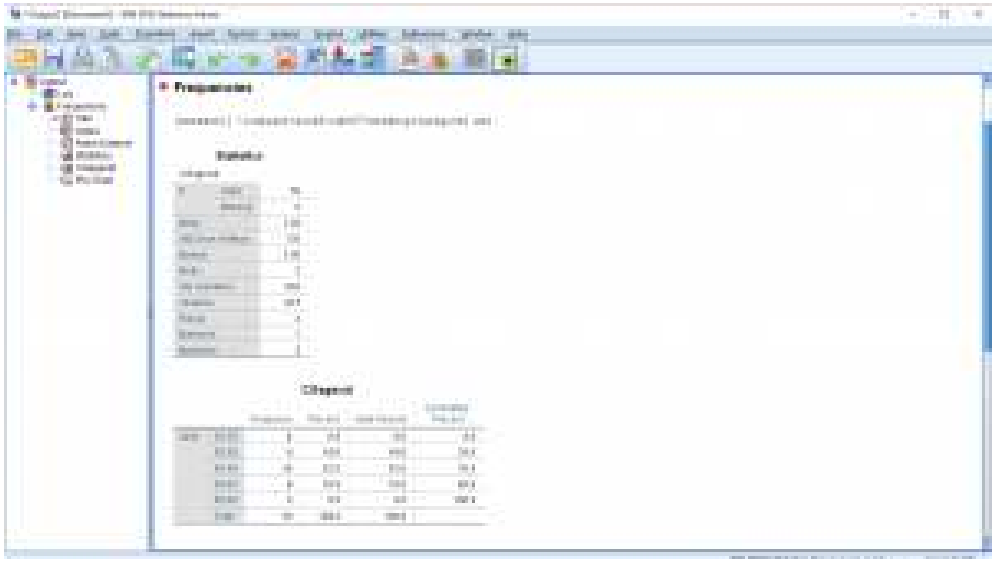
SPSS screenshot © International Business Machines Corporation.

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



SPSS screenshot © International Business Machines Corporation.

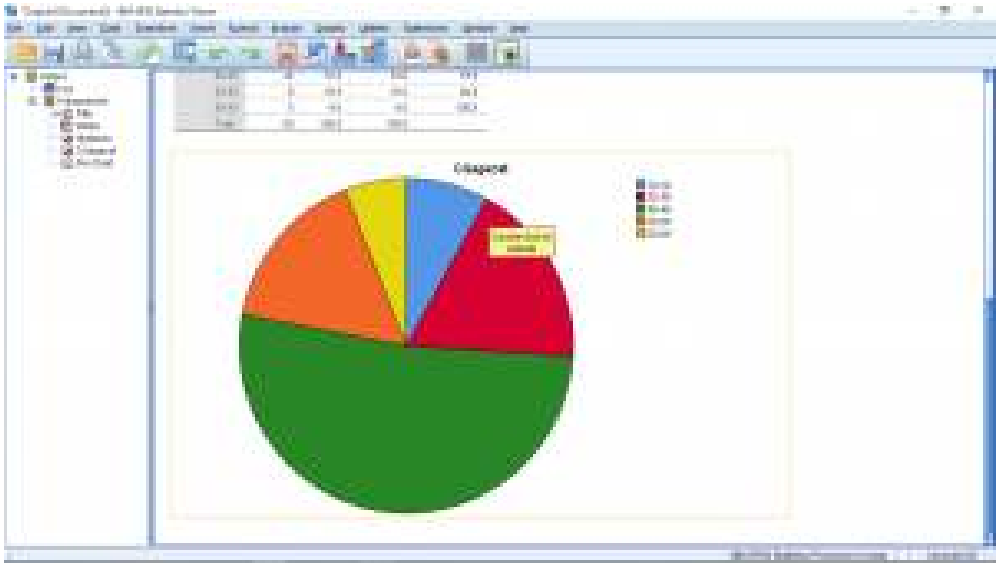
Hit Continue. You can look at the Format… and Style… menus if you want, they are not particularly interesting. Make sure "Display frequency tables" is checked (this will be important when you do the assignments), then hit OK. The Output window will pop up and in that window you will see:



SPSS screenshot © International Business Machines Corporation.

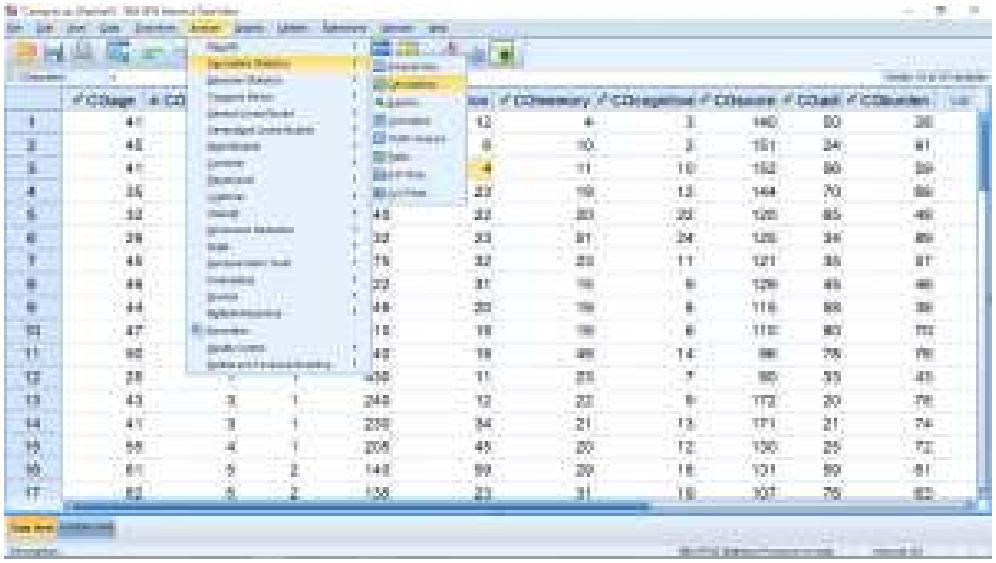
The first table, Statistics, shows the descriptive statistics you asked for. Note, especially, for future reference (when we hit skewness in Chapter 3), the value of the skewness. It is 0.411. More to the point it is > 0, or positive, meaning that the data set (CGagecatn) is right skewed or positively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

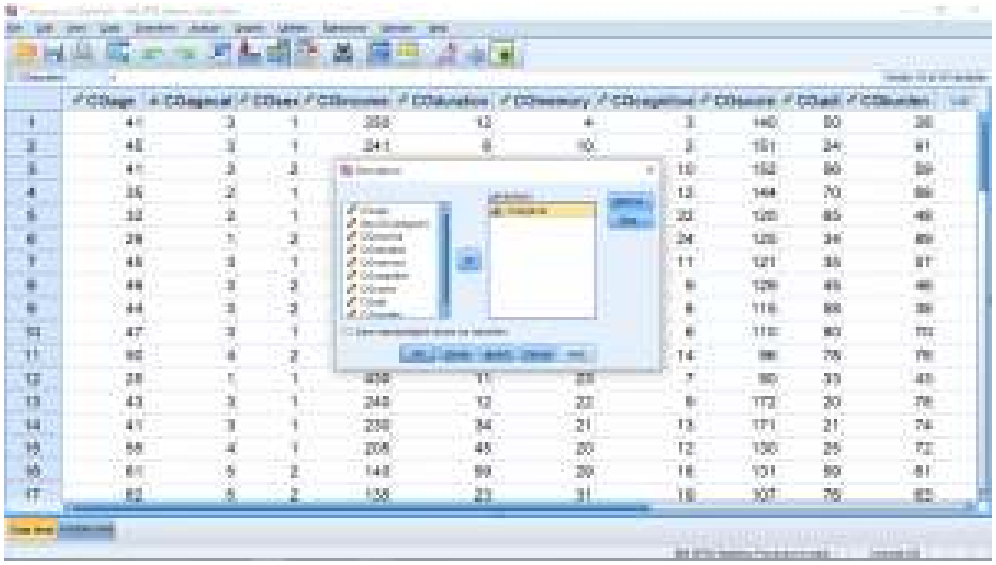


SPSS screenshot © International Business Machines Corporation.

Lets look at the Descriptives… menu next:



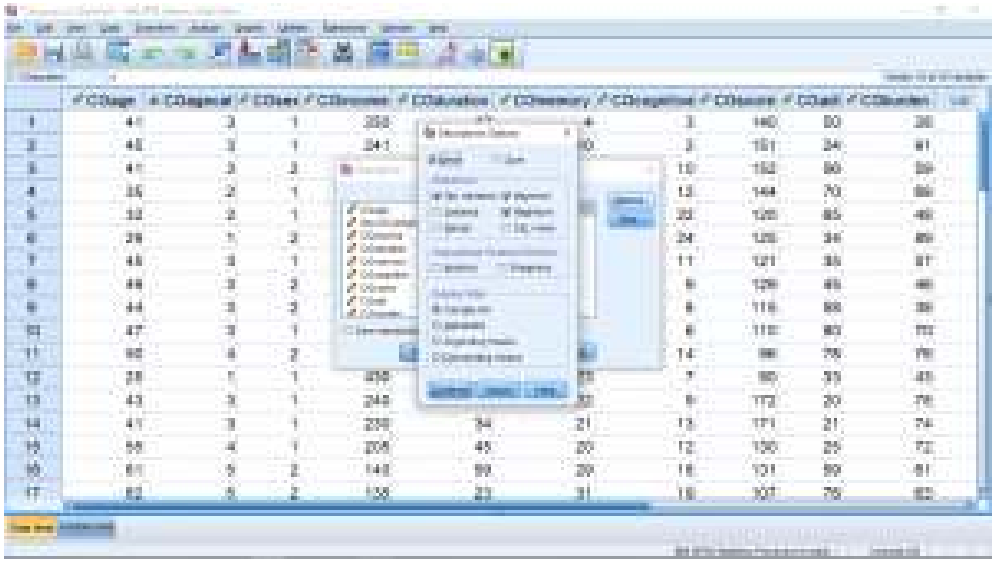
SPSS screenshot © International Business Machines Corporation.



SPSS screenshot © International Business Machines Corporation.

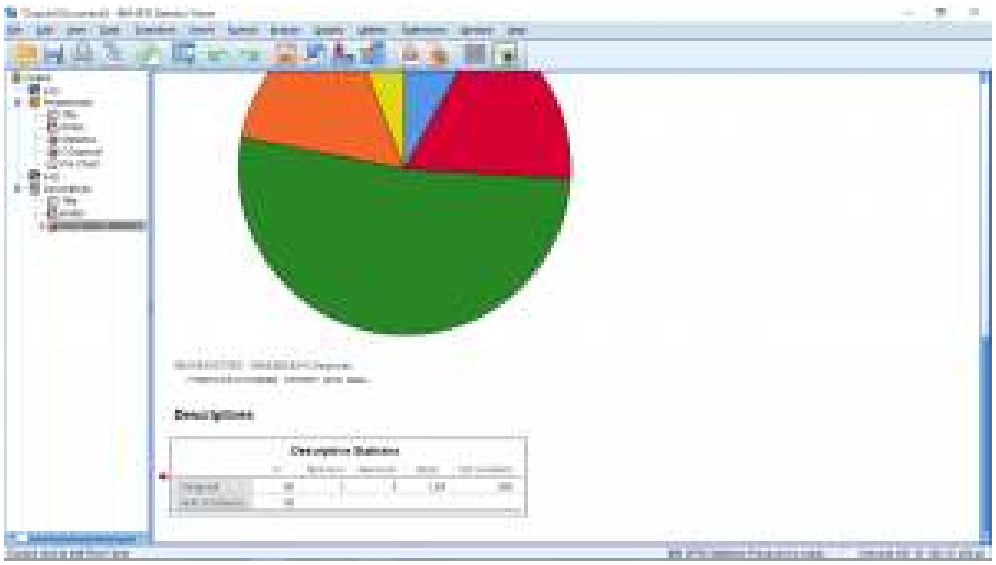
Move the CGagecat variable over as before and make sure to check off the “Save standardized values as variables”. We’ll learn about standardized values (z-values) in Chapter 3. Take note, this is the only way to get SPSS to compute z-values :

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.

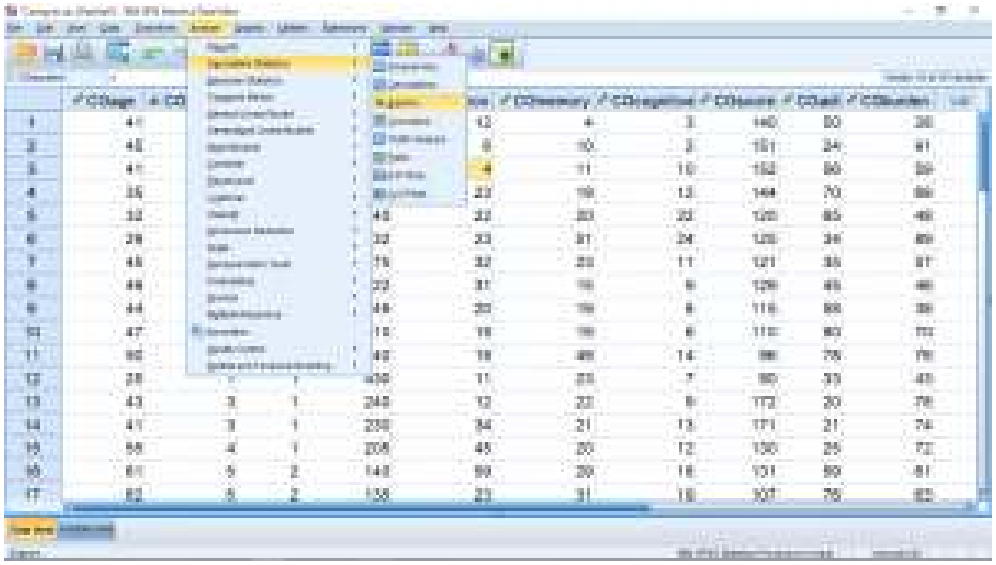
Hit Continue then OK and look at the results in the Output window. The output is straightforward:



SPSS screenshot © International Business Machines Corporation.

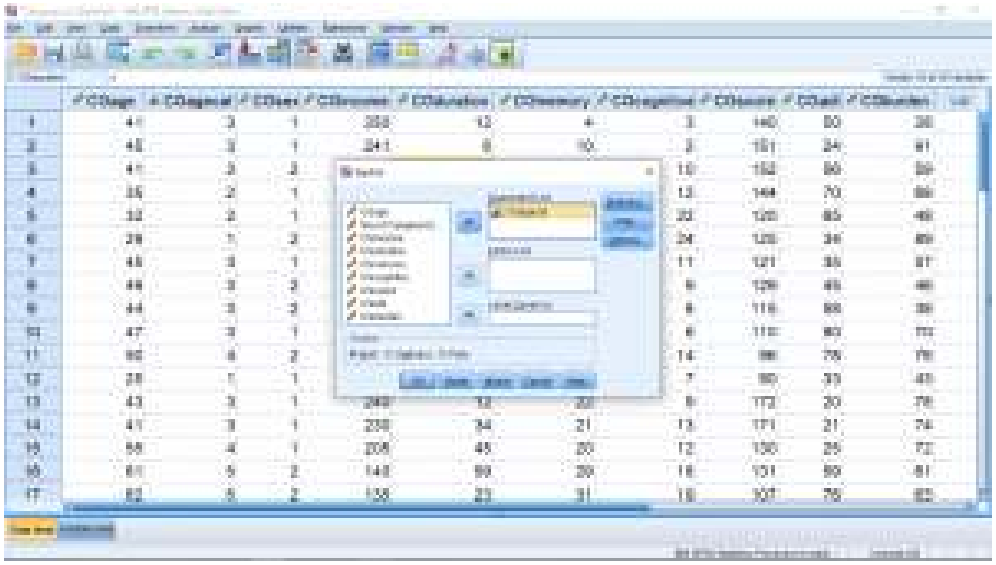
In Chapter 3 we will learn that the mean of a z-transformed variable is zero and the standard deviation is one. That is confirmed here. If you left the "Save standardized values as variables" box checked when you ran this, you'll get another variable added in the Data View window — the z-transform of 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:



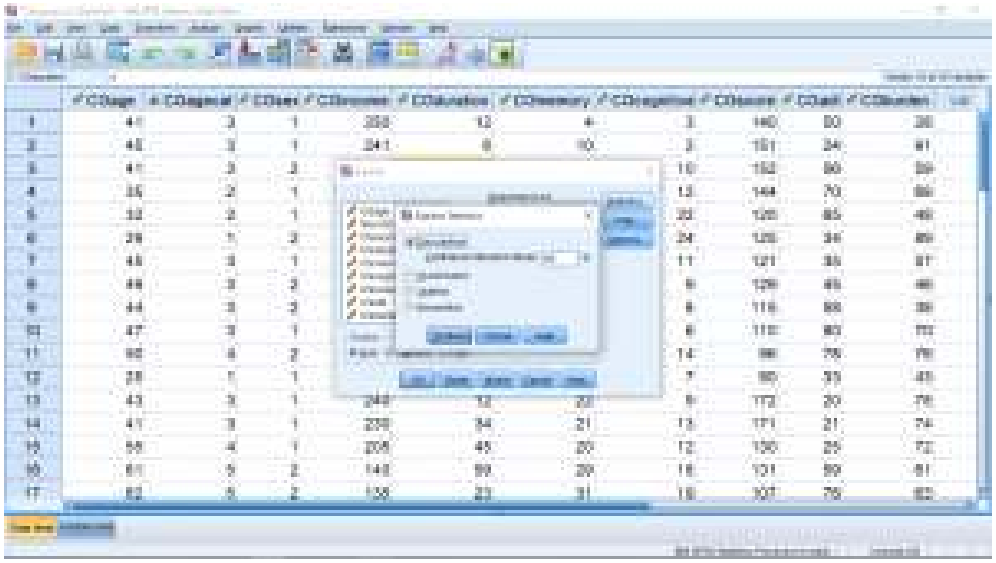
SPSS screenshot © International Business Machines Corporation.

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”):



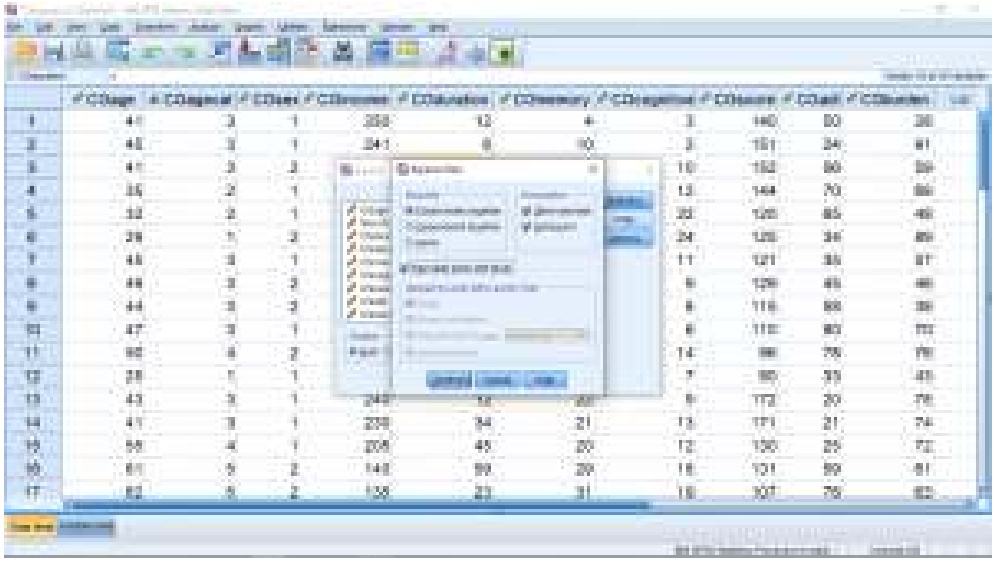
SPSS screenshot © International Business Machines Corporation.

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:



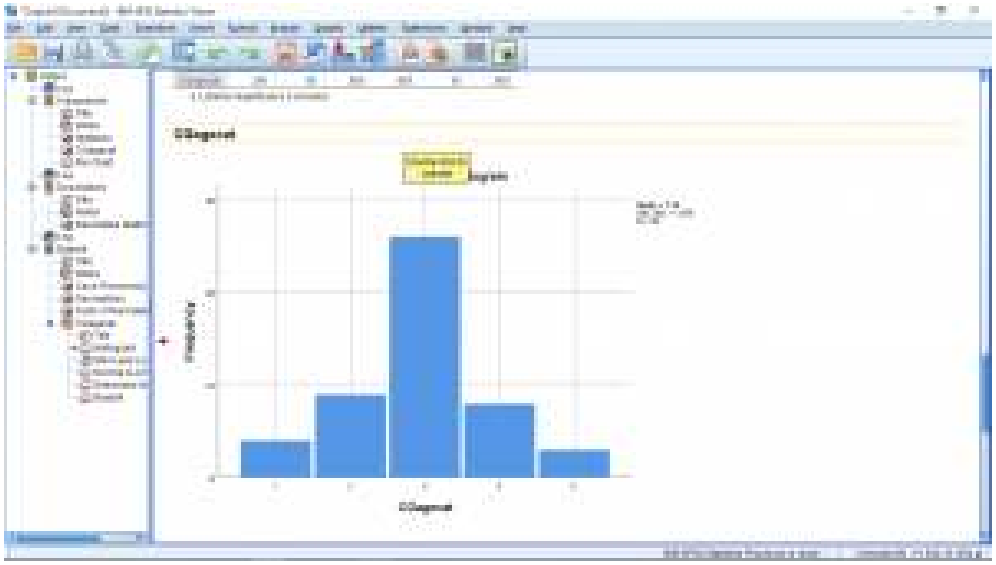
SPSS screenshot © International Business Machines Corporation.

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.

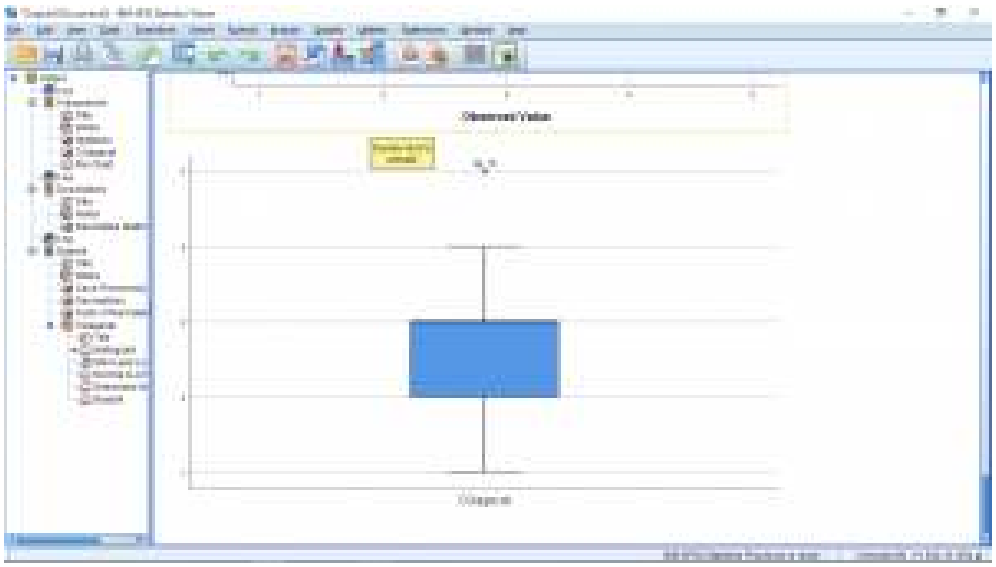
After we cover skewedness in Chapter 3, come back to this picture and note how the histogram is right skewed.

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 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.

This page titled 2.3: SPSS Lesson 1- Getting Started with SPSS is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

3: Descriptive Statistics- Central Tendency and Dispersion

This page titled 3: Descriptive Statistics- Central Tendency and Dispersion is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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, there 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 :

$$\bar{x} = sample\ mean \tag{3.1.1}$$

$$\mu = population\ mean \tag{3.1.2}$$

The formula for a sample mean is :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1.3}$$

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

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.1.4}$$

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

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

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

$$\bar{x} = \frac{\sum_{i=1}^{G} f_i x_{m_i}}{\sum_{i=1}^{G} f_i} \tag{3.1.5}$$

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 :

$$\bar{x} = \frac{\sum_i f_i x_{m_i}}{n} \tag{3.1.6}$$

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 :

$$\bar{x} = \frac{\sum_i f_i x_{m_i}}{n} = \frac{490}{20} = 24.5 \tag{3.1.7}$$

□

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{j=1}^{G} (\sum_{k=1}^{f_i} x_k)}{n}. \tag{3.1.8}$$

So the approximation is that

$$\sum_{k=1}^{f_i} x_k = f_i x_{m_i} \tag{3.1.9}$$

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

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

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

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3.1.10}$$

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

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{869}{12} = 72.4 \tag{3.1.11}$$

Note, that the unweighted mean for these data is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{297}{4} = 74.3 \tag{3.1.12}$$

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

$$\underline{8},\ 9,\ 9,\ 14,\ \underline{8},\ \underline{8},\ 10,\ 7,\ 6,\ 9,\ 7,\ \underline{8},\ 10,\ 14,\ 11,\ \underline{8},\ 14,\ 11$$

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

$$15,\ \underline{18, 18, 18,}\ 20,\ 22,\ \underline{24, 24, 24,}\ 26,\ 26$$

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

$$\text{MR} = \frac{H + L}{2} \tag{3.1.13}$$

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

$$\text{MR} = \frac{8 + 1}{2} = 4.5 \tag{3.1.14}$$

□

3.1.5 Mean, Median and Mode in Histograms: Skewness

If the shape of the histogram of a dataset is not too bizarre[1] (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.

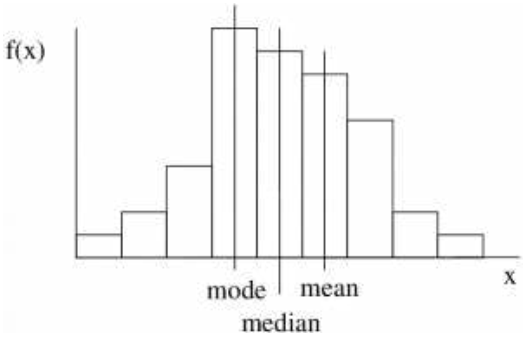


Figure 3.1: A right skewed histogram (or distribution) generally has the mean and median to the right, or positive side of the mode. The tail of the histogram stretches to the right or positive side.

Symmetric distribution

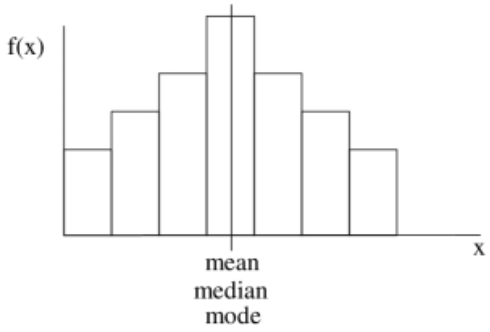


Figure 3.2: A symmetric distribution (histogram) has the mean, median and mode all in the same place. Its shape is symmetric.

Negatively skewed or left skewed histograms

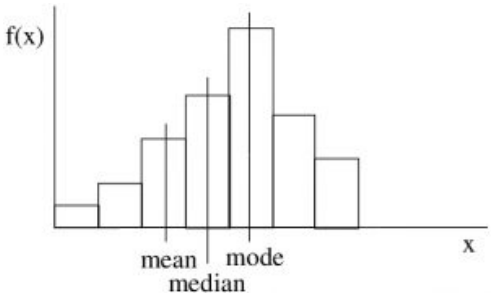


Figure 3.3: A left skewed histogram (or distribution) generally has the mean and median to the left, or negative side of the mode. The tail of the histogram stretches to the left or negative side.

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 histogram[2] (so if there are multiple modes, they don't all have to have the same maximum value).

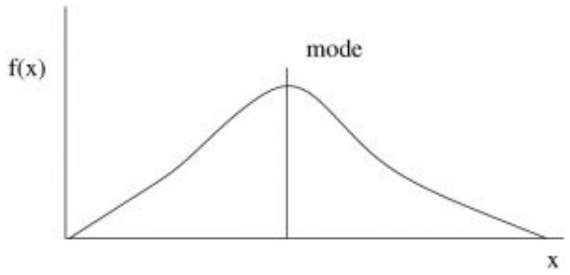


Figure 3.4: The mode is the maximum of the histogram (distribution).

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.

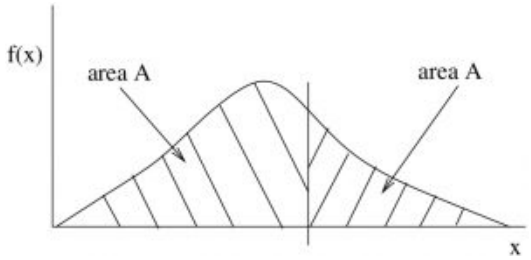


Figure 3.5: The median divides the area under the histogram into two equal areas A.

Mean

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

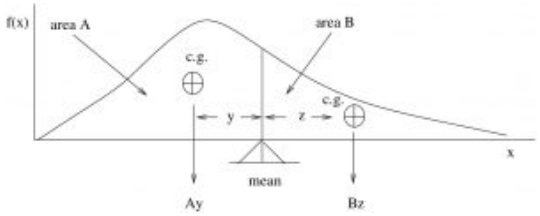


Figure 3.6: The mean is the balance point of the histogram. It is where the "first moments" of the area of the histogram balance. Here the moments are Ay and Bz balance. $Ay = Bz$.

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

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

$$M = Wx \tag{3.1.15}$$

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).

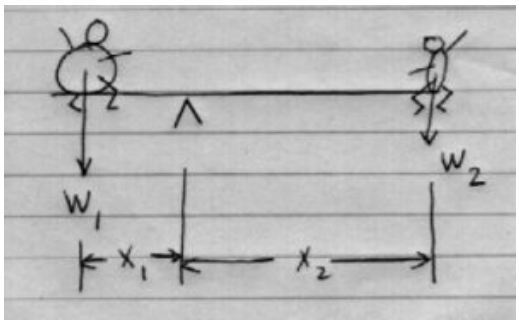


Figure 3.7

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

To balance the teeter-totter we must have

$$W_1 x_1 = W_2 x_2. \tag{3.1.16}$$

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 :

$$\sum W_i x_i = W_t x_g \tag{3.1.17}$$

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):

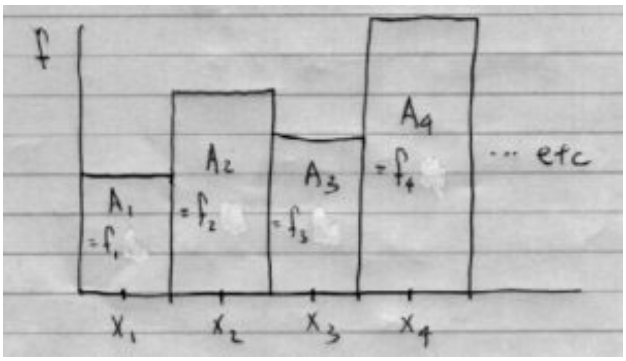


Figure 3.8

(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,

$$\sum W_i x_i = W_t x_g \tag{3.1.18}$$

translates to

$$\sum A_i x_i = A_t x_g$$
$$\sum f_i x_i = (\sum f_i) x_g$$
$$\sum f_i x_i = n x_g$$
$$x_g = \frac{\sum f_i x_i}{n}$$

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

$$x_g = \overline{x} = \frac{\sum f_i x_i}{n}. \tag{3.1.19}$$

Because our "weight" is area, $\overline{x}$ is technically called the "1st moment of area". (Variance, covered next, is the "2nd moment of area about the mean".)

□

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

This page titled 3.1: Central Tendency- Mean, Median, Mode is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

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

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 :

$$\sigma = \sqrt{\sigma^2} \tag{3.2.1}$$

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 :

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

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 :

$$s = \sqrt{s^2} \tag{3.2.2}$$

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

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

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 :

Misplaced &

Misplaced &

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 :

$$\mu = \frac{\sum x_i}{N} = \frac{210}{6} = 35. \tag{3.2.3}$$

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

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{1750}{6} = 291.7 \tag{3.2.4}$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{291.7} = 17.1. \tag{3.2.5}$$

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

$$\text{CVar} = \frac{\sigma}{\mu} \times 100\% = \frac{17.1}{35} \times 100\% = 48.9\%. \tag{3.2.6}$$

□

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	$(x_i - \overline{x})^2$	x_i^2
11.2		11.2^2 = 125.44
11.9		11.9^2 = 141.161
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

$$s^2 = \frac{\sum x_i^2 - (\frac{(\sum x_i)^2}{n})}{n-1} = \frac{958.94 - (\frac{5715.36}{6})}{6-1} = \frac{958.94 - 952.56}{5} = 1.28 \tag{3.2.7}$$

So

$$s = \sqrt{s^2} = \sqrt{1.28} = 1.13. \tag{3.2.8}$$

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:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{75.6}{6} = 12.60. \tag{3.2.9}$$

And with the mean we can then compute

$$\text{CVar } = \frac{s}{\bar{x}} = \frac{1.13}{12.6} \times 100\% = 9.0\% \tag{3.2.10}$$

□

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

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

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

$$n = \sum_{i=1}^{G} f_i \tag{3.2.11}$$

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

$$s^2 = \frac{\sum_{i=1}^{G} f_i (x_{m_i} - \mu)^2}{n-1} \tag{3.2.12}$$

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

$$\sigma^2 = \frac{\sum_{i=1}^{G} f_i (X_{m_i} - \mu)^2}{N} \tag{3.2.13}$$

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

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

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

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

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	$2 \cdot 38 = 76$	$38^2 = 1444$	$2 \cdot 1444 = 2888$
		$\sum f = 20$		$\sum f x_m = 490$		$\sum f x_m^2 = 13310$

The formula

$$s^2 = \frac{\sum (f x_m^2) - [\frac{(\sum f x_m)^2}{n}]}{n-1}$$

tells us that we need the sums of $f x_m^2$ and $f x_m$ 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

Misplaced &

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.

This page titled 3.2: Dispersion- Variance and Standard Deviation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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

$$z = \frac{x - \bar{x}}{s} \text{ (sample)} \tag{3.3.1}$$

$$z = \frac{x - \mu}{\sigma} \text{ (population)} \tag{3.3.2}$$

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

$$\bar{x} = \frac{\sum x_i}{n} = \frac{99}{10} = 9.9 \tag{3.3.3}$$

and

Misplaced &

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.)

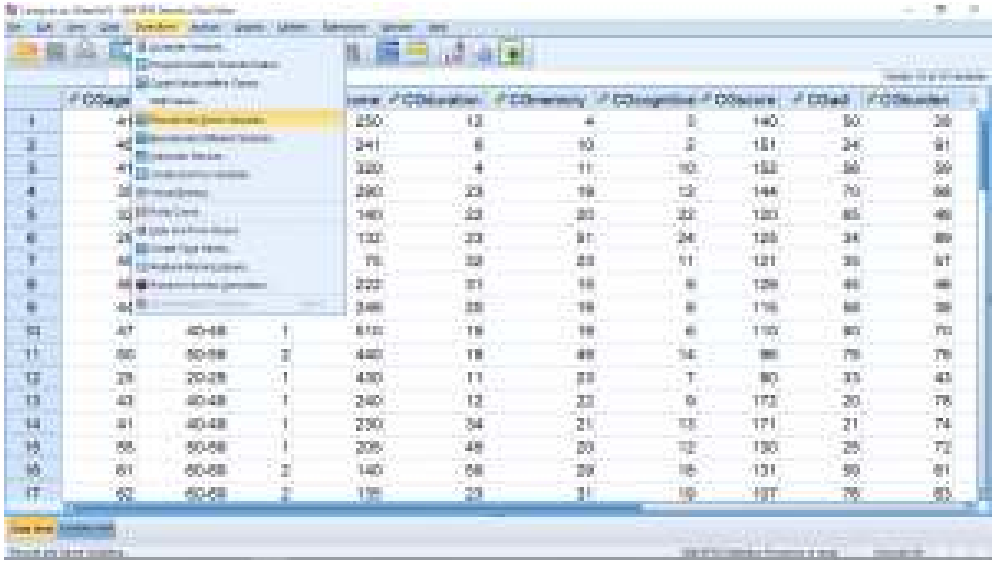
□

This page titled 3.3: z-score / z-transformation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

3.4: SPSS Lesson 2- Combining variables and recoding

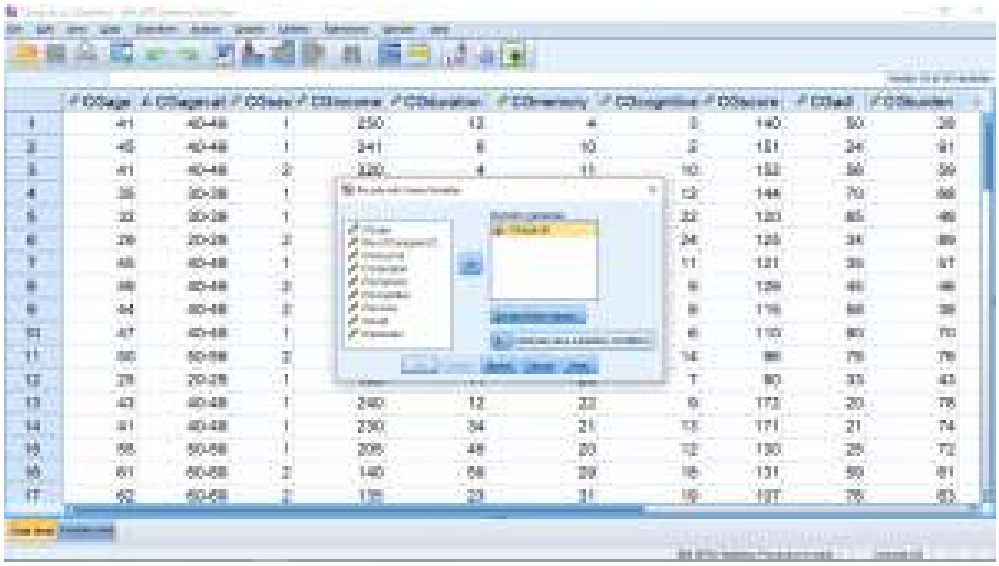
Frequently data collection results in a collection of many variables. This happens, for example, with tests or surveys where people answer questions on a 5 or 7 point *Lickert scale* where questions range from, say, “strongly agree” to “somewhat agree” to ... to “strongly disagree”. A bunch of those 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 :



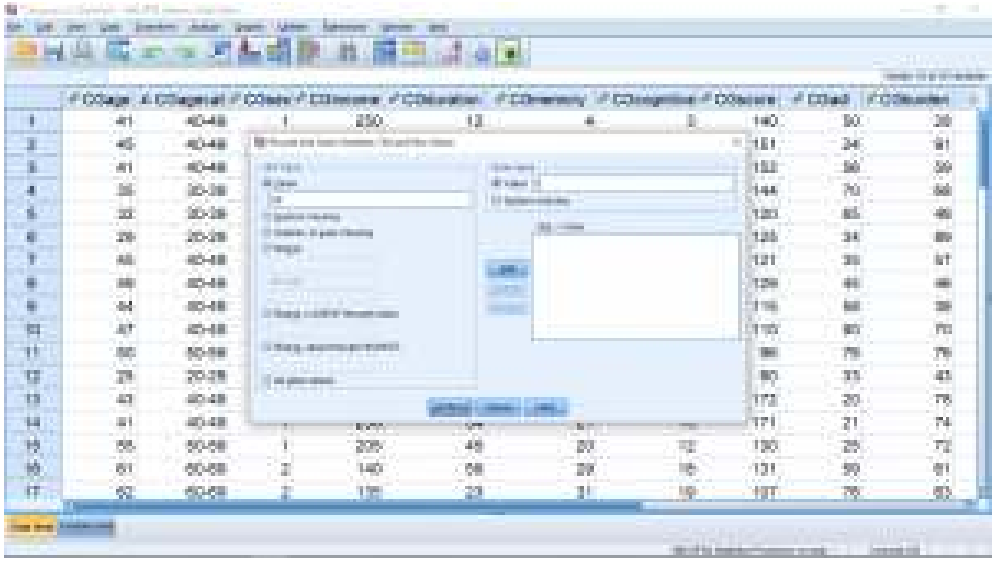
SPSS screenshot © International Business Machines Corporation.

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 :



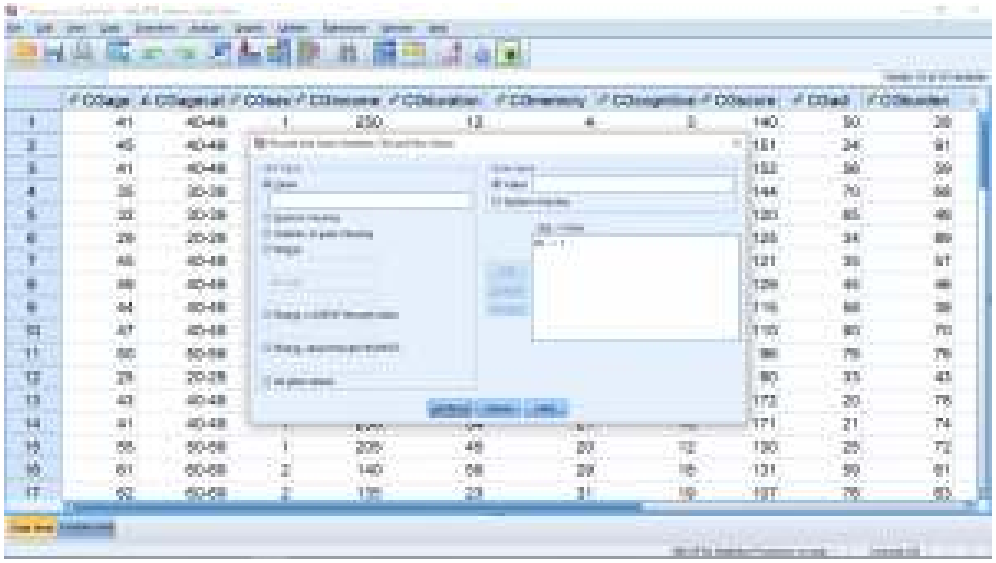
SPSS screenshot © International Business Machines Corporation.

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 :



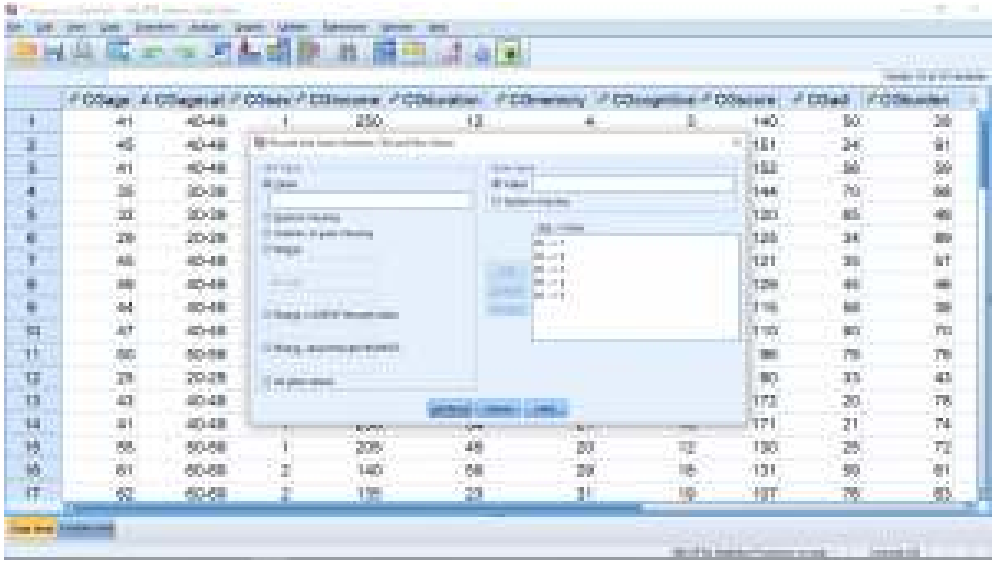
SPSS screenshot © International Business Machines Corporation.

then hit Add :



SPSS screenshot © International Business Machines Corporation.

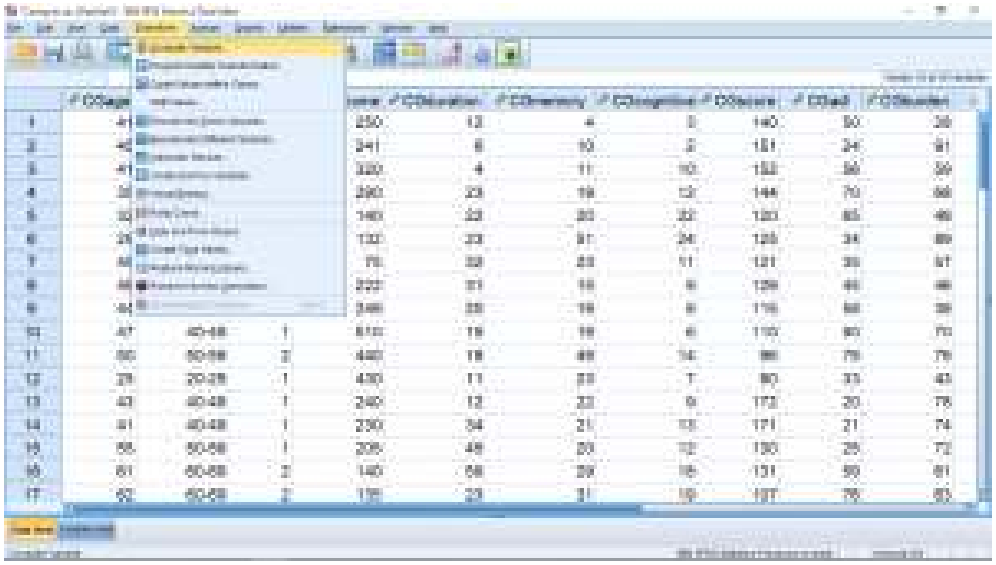
Continue this way to complete the recoding list :



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 :



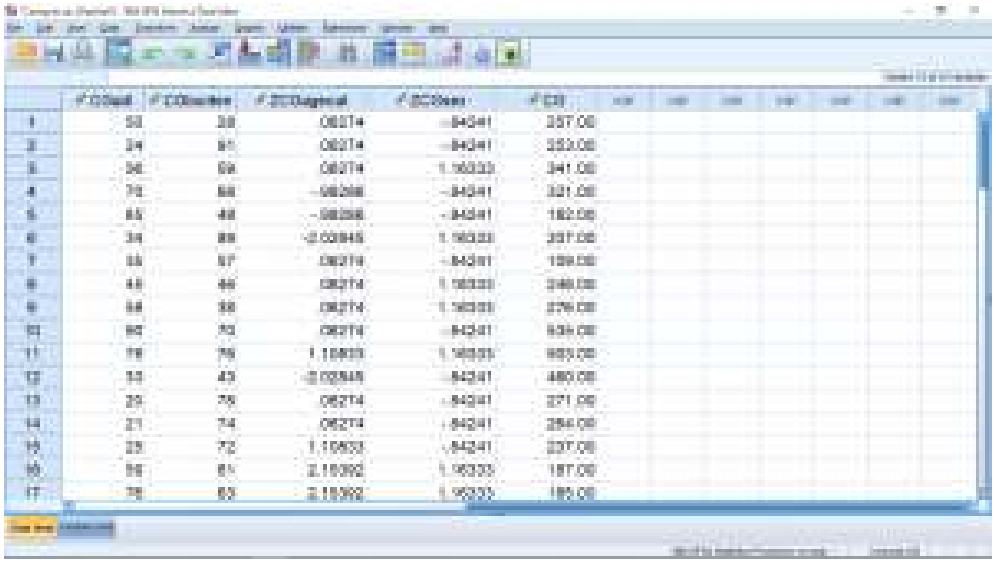
SPSS screenshot © International Business Machines Corporation.

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



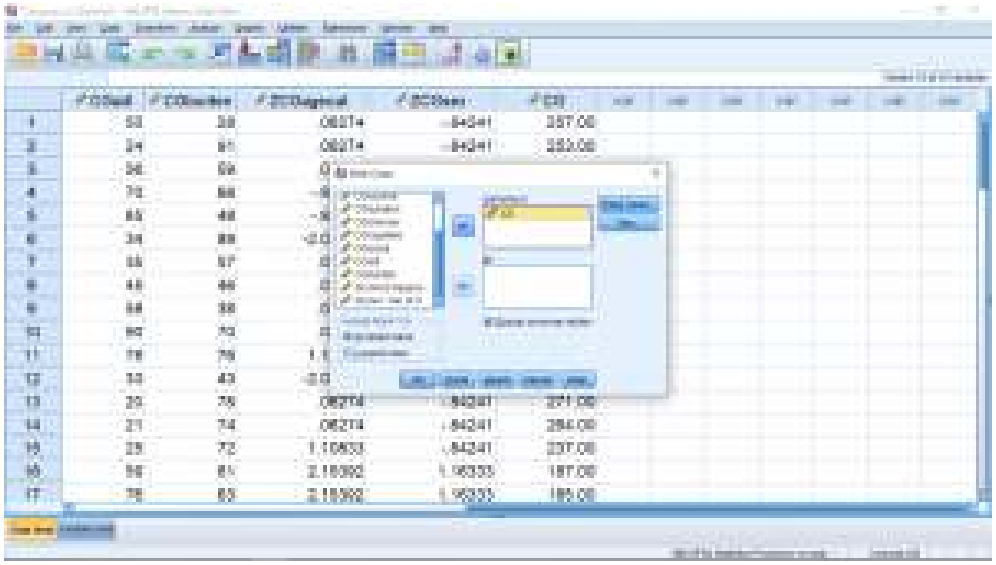
SPSS screenshot © International Business Machines Corporation.

Fill in the menu as shown above. You can move variable names into the Numeric Expression box by double clicking on the variable name, by clicking on the variable name and the arrow or by simply typing it. There are fancier ways to get a sum of variables experesion in Numeric Expression, but we will keep it simple for now. The target variable name is totalscore which, after you hit OK, shows up as a new variable, ready for statistical analysis, in the last column in the Data View window :



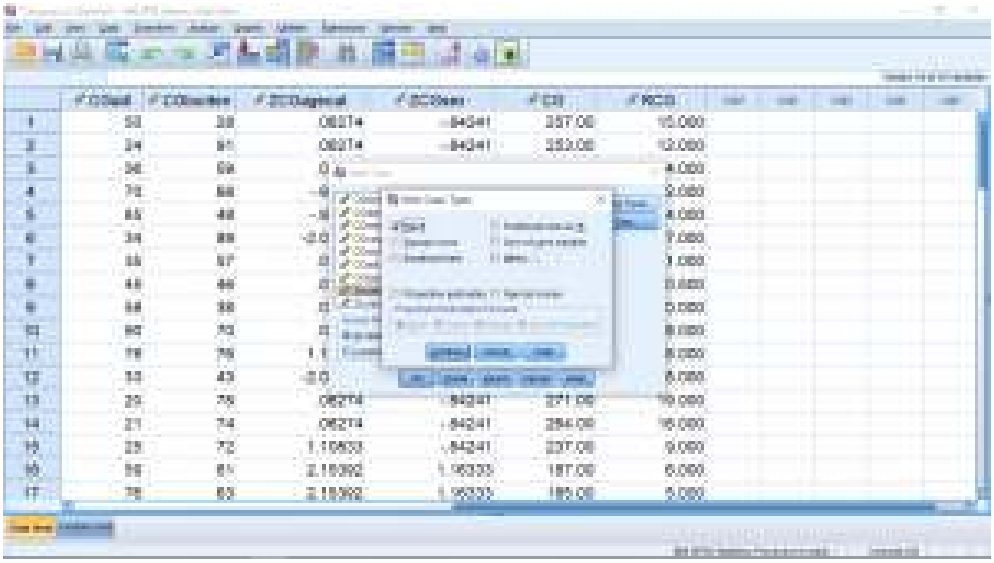
SPSS screenshot © International Business Machines Corporation.

Let's do a couple of (descriptive) analysis with this new variable. Let's take Caregiver as our dataset. Suppose we want to find the median of the totalscore values. To do this task by hand, we need to put the 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 $\rightarrow$ 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 :



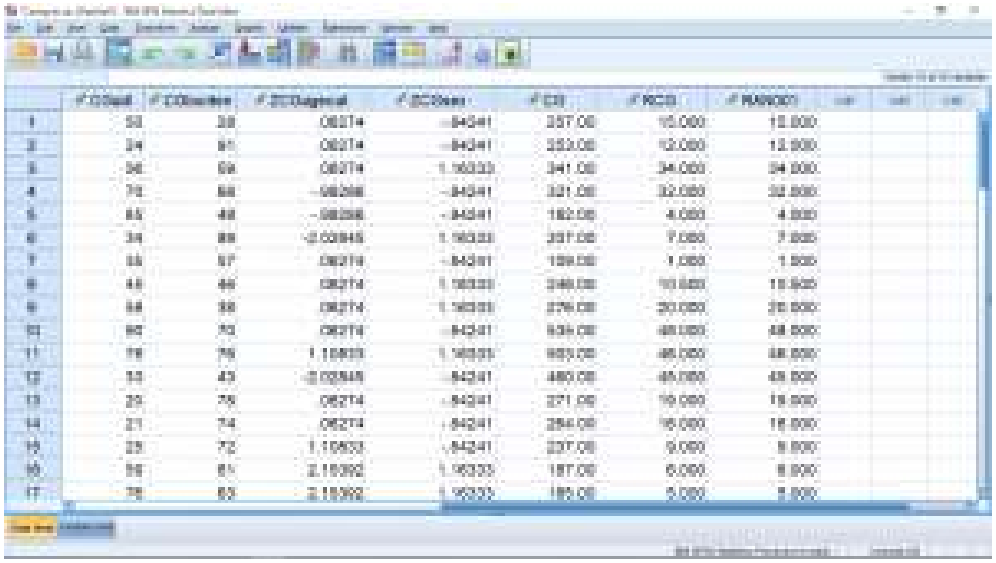
SPSS screenshot © International Business Machines Corporation.

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



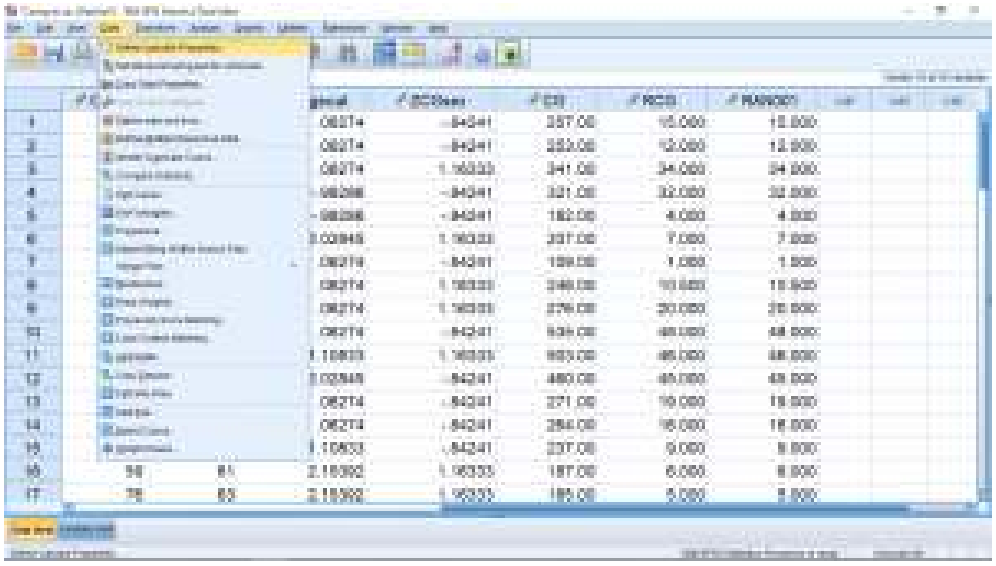
SPSS screenshot © International Business Machines Corporation.

We will assign the average (mean) rank to ties in out 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 Rtotalscore with be formed in the Data View menu :



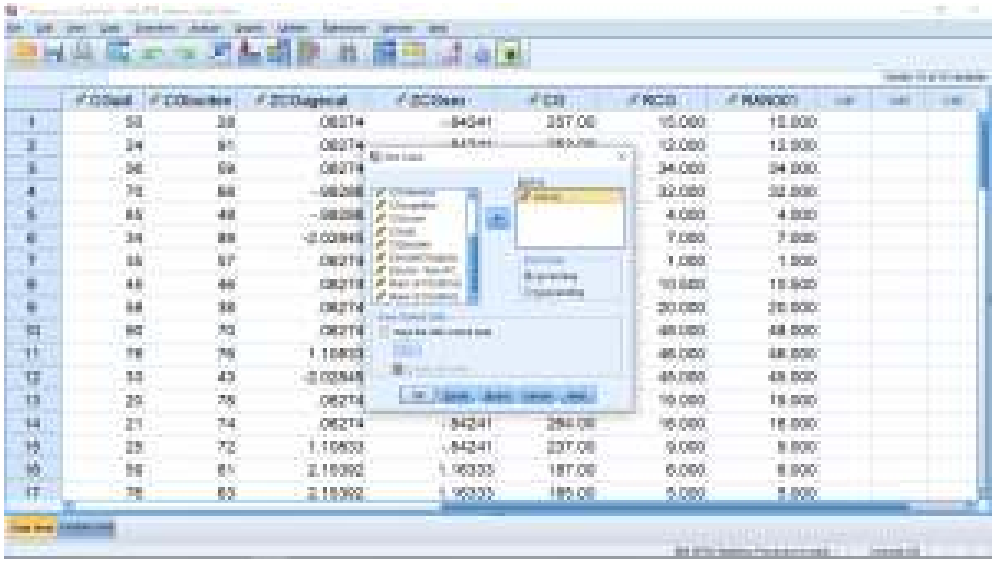
SPSS screenshot © International Business Machines Corporation.

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 :



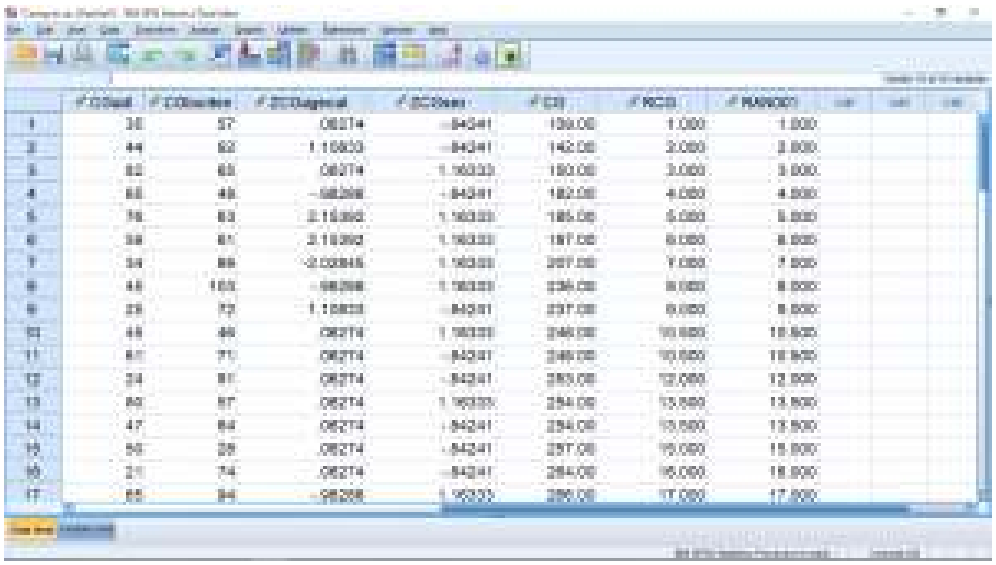
SPSS screenshot © International Business Machines Corporation.

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



SPSS screenshot © International Business Machines Corporation.

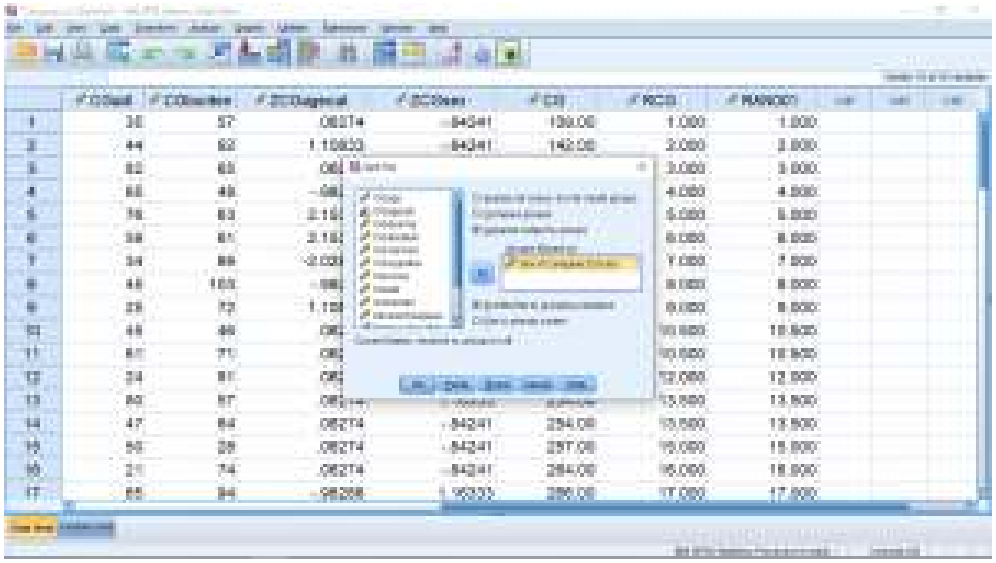
Keep the ascending button selected (sort from lowest to highest), then hit OK to sort the file :



SPSS screenshot © International Business Machines Corporation.

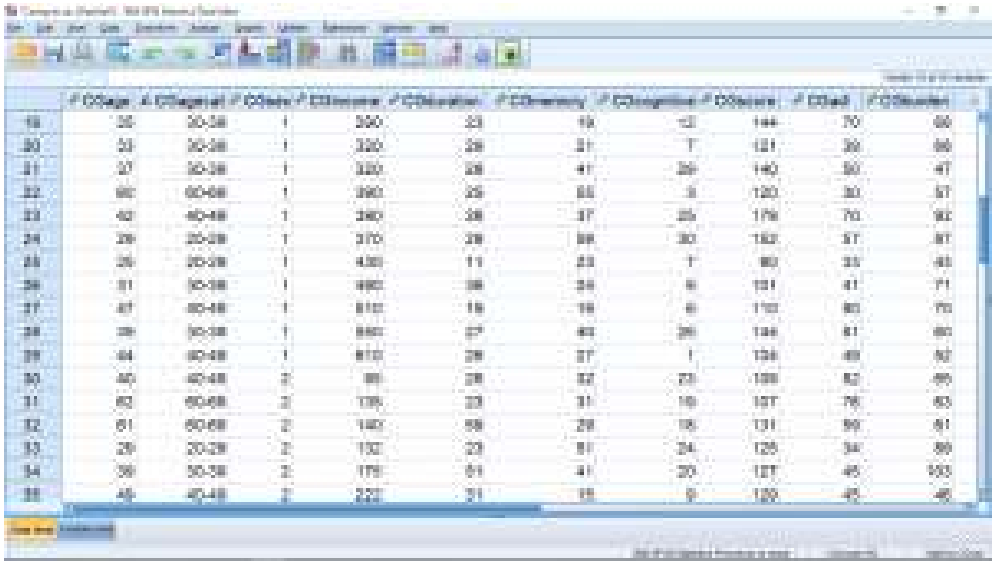
Everything is sorted now. (Note how useful the id variable is now. If that wasn't there, we'd lose track of who's data was what.) Now if we scroll down, we will find that the middle two total test scores are both 210. Thus the median of total score is 210.

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



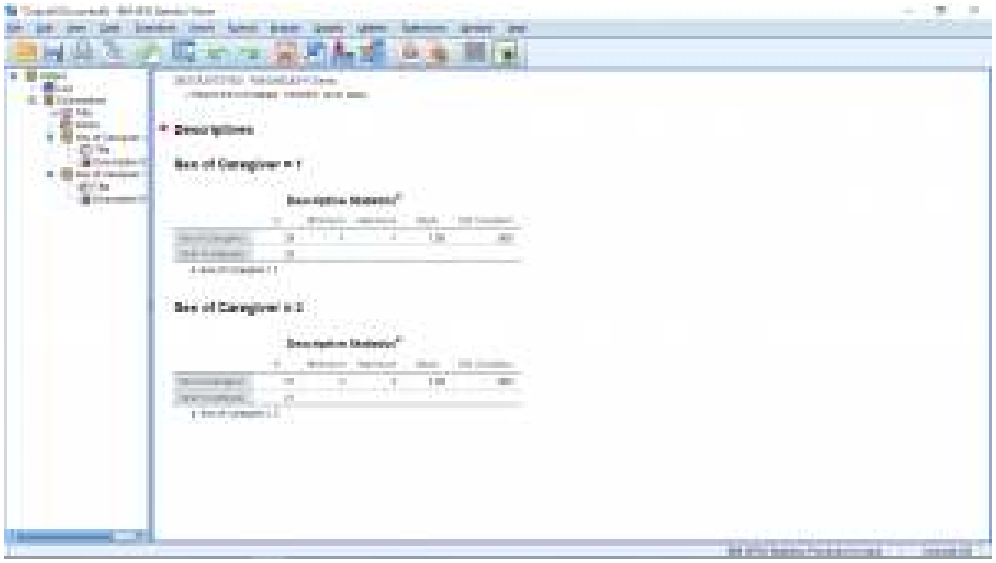
SPSS screenshot © International Business Machines Corporation.

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



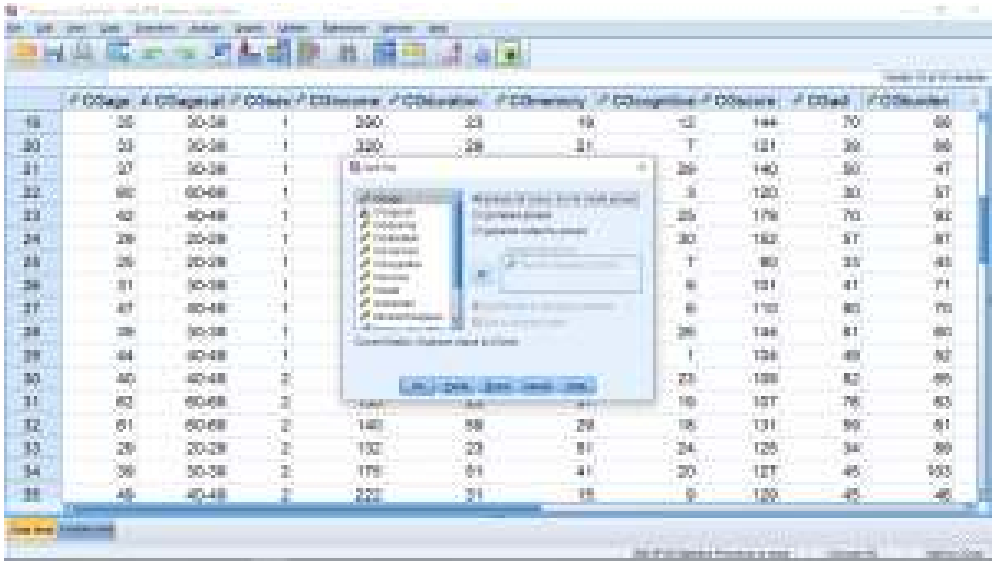
SPSS screenshot © International Business Machines Corporation.

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 :



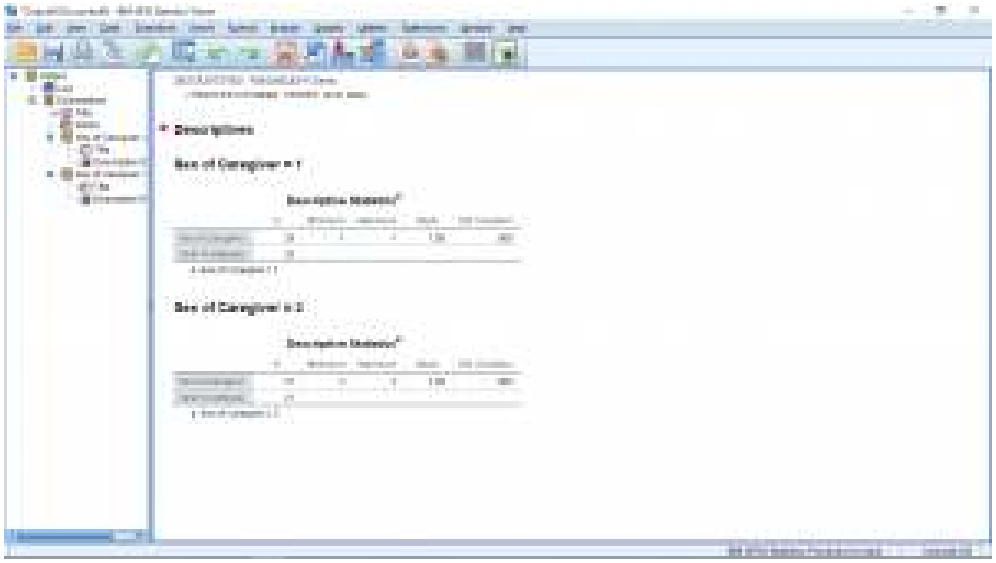
SPSS screenshot © International Business Machines Corporation.

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 :



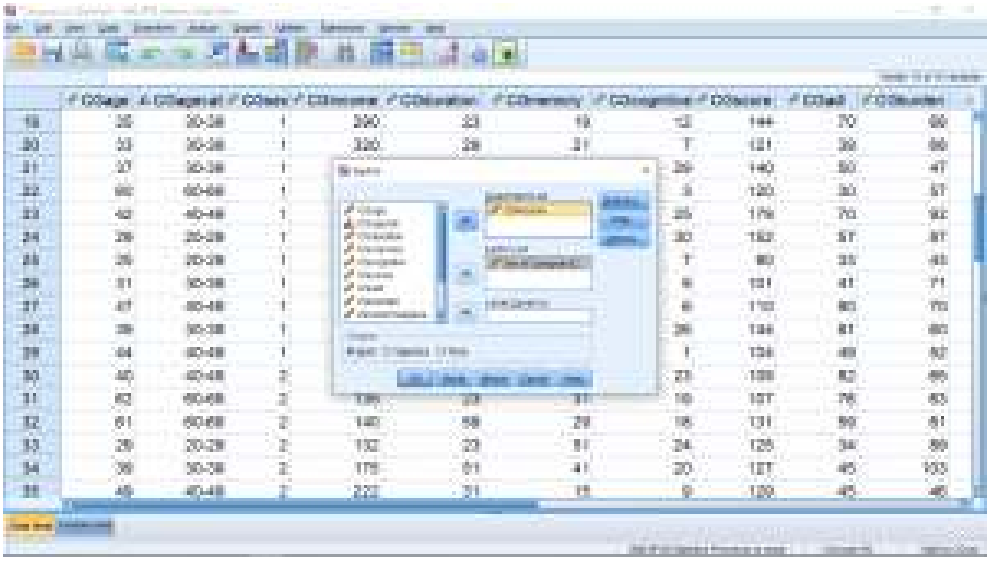
SPSS screenshot © International Business Machines Corporation.

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 :



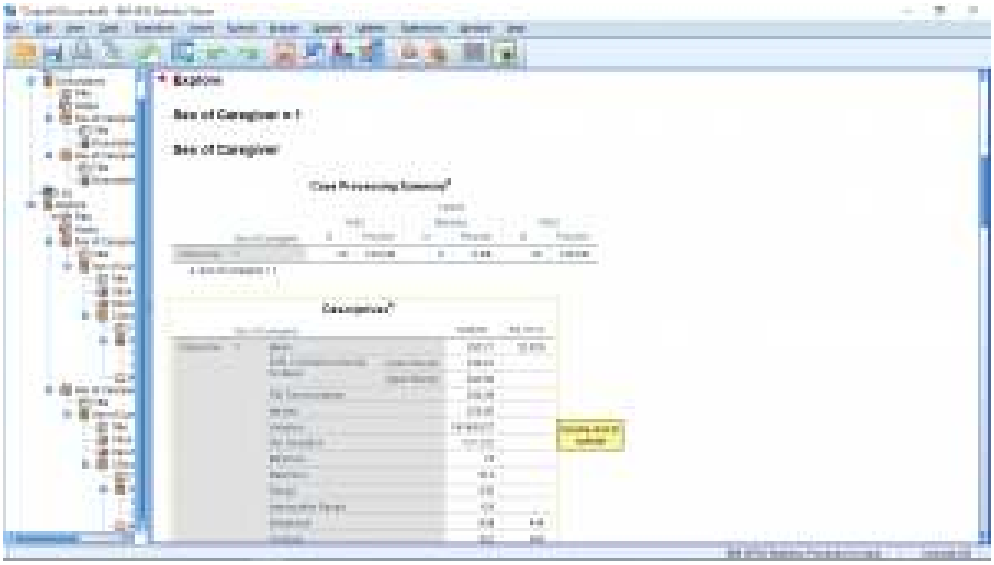
SPSS screenshot © International Business Machines Corporation.

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.

This page titled 3.4: SPSS Lesson 2- Combining variables and recoding is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

4: Probability and the Binomial Distributions

This page titled 4: Probability and the Binomial Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Probability

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

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

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 :

$$P(E) = \frac{4}{52} = 0.077 \text{ (7.7\% if we were to express the result in percentages)} \tag{4.1.1}$$

□

To use $P(E)$ mathematically we set

$$0 \leq P(E) \leq 1 \tag{4.1.2}$$

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 :

$$k_1 \cdot k_2 \cdot k_3 \backslash \text{hdots} k_n = \prod_{i=1}^{n} k_i \tag{4.1.3}$$

(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

$$k_1 k_2 k_3 = 10 \cdot 10 \cdot 10 = 10^3 = 1000 \tag{4.1.4}$$

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:

$$7 \cdot 2 \cdot 3 \cdot 2 = 84 \tag{4.1.5}$$

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 or n objects, *while keeping track of the order of selection* is [1]

$$ {}_nP_r = \frac{n!}{(n-r)!} \tag{4.1.6}$$

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 $k_r = n-r+1$:

$$ {}_nP_r = (n)(n-1)(n-2)\hdots(n-r+1) \tag{4.1.7}$$

$$ = \frac{(n)(n-1)\hdots(2)(1)}{(n-r)(n-r-1)\hdots(2)(1)} \tag{4.1.8}$$

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 :

\[{}_nC_x = \frac{n!}{(n-x)!x!} = \frac{3,628,800} = \underline{17.3 \times 10^{12}} \]

The symbol $\left(\begin{array}{c} n \\ x \end{array} \right)$ is also known as the *binomial coefficient* because it shows up in algebra when you expand expressions of the form $(x+y)^n$. For example[2]

$$ (x+y)^n = x^2 + 2xy + y^2 \tag{4.1.9}$$

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

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

$$ \begin{array}{ccccccccccccc} & & & & & & & & & & & & n = \\ & & & & & & 1 & & & & & & 0 \\ & & & & & 1 & & 1 & & & & & 1 \\ & & & & 1 & & 2 & & 1 & & & & 2 \\ & & & 1 & & 3 & & 3 & & 1 & & & 3 \\ & & 1 & & 4 & & 6 & & 4 & & 1 & & 4 \\ & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & 5 \\ 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \quad 6 \\ & & & & & & \text{etc.} & & & & & & \end{array} \tag{4.1.10}$$

Referring to Pascal's triangle we can quickly write

$$ (x+y)^6 = x^6 + 6x^5 y + 15x^4 y^2 + 20x^3 y^3 + 15x^2 y^4 + 6xy^5 + y^6 \tag{4.1.11}$$

for example.

1. Recall that the definition of factorial follows $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ etc. ↵
2. You don't need this algebra for this statistics course. It's just interesting. ↵

This page titled 4.1: Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 symbols[1] what is $P(x$ successes $\mid n$ trials)? Or with simpler notation, what is $P(x \mid n)$? The answer is :

$$P(x \mid n) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}.$$

***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

$$P(1 \mid 1) = p = \left(\begin{array}{c} 1 \\ 1 \end{array} \right) p^1 q^0. \tag{4.2.1}$$

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

$$P(0 \mid 2) = q^2 = \left(\begin{array}{c} 2 \\ 0 \end{array} \right) p^0 q^2. \tag{4.2.2}$$

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

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

$$P(1 \mid 2) = 2 \cdot p \cdot q = \left(\begin{array}{c} 2 \\ 1 \end{array} \right) p^1 q^1. \tag{4.2.3}$$

For $x = 2$ we have

$$P(2 \mid 2) = \left(\begin{array}{c} 2 \\ 2 \end{array} \right) p^2 q^0. \tag{4.2.4}$$

We can continue this way for $n = 3, 4, ...$ 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 $\left(\begin{array}{c} n \\ x \end{array} \right)$ *specific* sequences with x successes. Putting it all together :

$$P(x \mid n) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}. \tag{4.2.5}$$

□

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 draw[2]

Say we want to know P(3 successes |10 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

$$P(3 \mid 10) = \left(\begin{array}{c} 10 \\ 3 \end{array} \right) 0.2^3 0.8^7 = 0.201. \tag{4.2.6}$$

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.

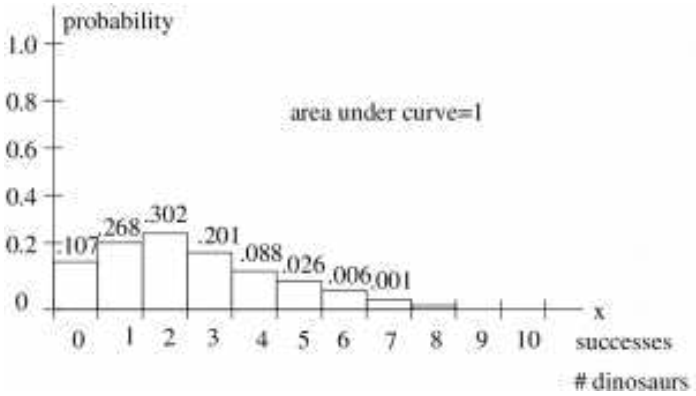


Figure 4.1 : The binomial distribution for the example of forming samples of $n = 10$ toys with x representing the number of dinosaurs in the sample and $p = 0.2$ being the probability of selecting a dinosaur in forming the sample. Note that the probability of x = 8, 9 or 10 is not zero, just less than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

$$\mu = \sum_x x \cdot P(x) \tag{4.2.7}$$

$$\sigma^2 = \sum_x (x-\mu)^2 \cdot P(x) = \left[\sum_x x^2 \cdot P(x)\right] - \mu^2 \tag{4.2.8}$$

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

$$\mu = np \tag{4.2.9}$$

$$\sigma^2 = npq \tag{4.2.10}$$

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:

$$\mu = n \cdot p = 10 \cdot 0.20 = 2 \tag{4.2.11}$$

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, \hdots, 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 :

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

□

---

1. Here the | is read as "given". ↵
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 hypergeometic 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 $\begin{pmatrix} 100 \\ 10 \end{pmatrix} = 17.3 \times 10^{12}$ 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 $\begin{pmatrix} 20 \\ 3 \end{pmatrix} = 1140$ ways to get 3 dinosaurs and $\begin{pmatrix} 80 \\ 7 \end{pmatrix} = 3.18 \times 10^9$ ways to get 7 bugs from the 80 in the bucket. So there are $\begin{pmatrix} 20 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 80 \\ 7 \end{pmatrix} = 3.62 \times 10^{12}$ ways to draw 3 dinosaurs and 7 bugs from the bucket. This number is $n(E)$ in Equation (4.1). And so

$$P(3 \text{ dinosaurs} \mid 10 \text{ toys}) = \frac{\begin{pmatrix} 20 \\ 3 \end{pmatrix} \begin{pmatrix} 80 \\ 7 \end{pmatrix}}{\begin{pmatrix} 100 \\ 10 \end{pmatrix}} = \frac{3.62 \times 10^{12}}{17.3 \times 10^1 2} = 0.209 \qquad (4.2.12)$$

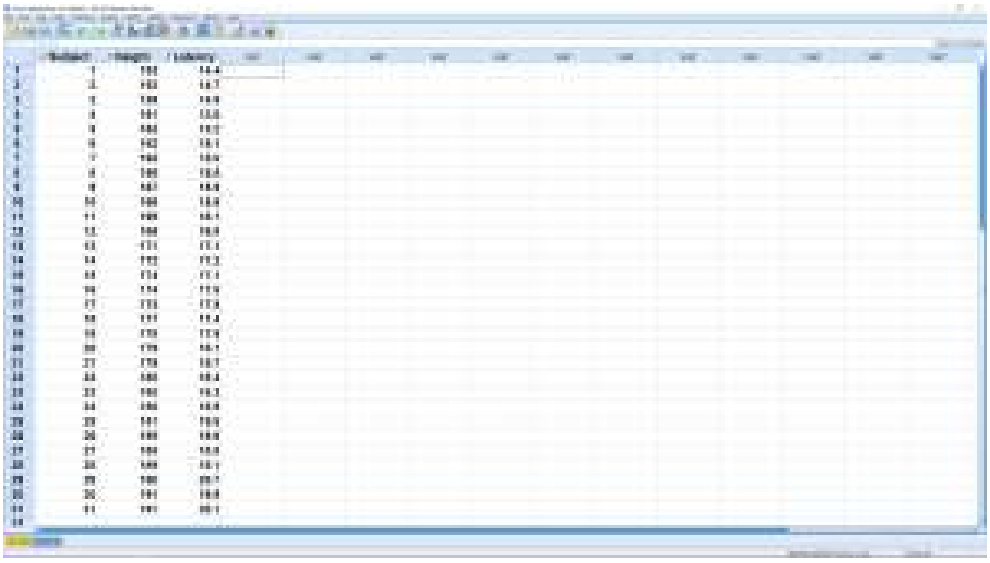
   Note how close this is to the answer from the binomial distribution of 0.201. ↵
3. Sample space is the set of all possible samples. ↵

This page titled 4.2: Binomial Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

4.3: SPSS Lesson 3- Combining variables - advanced

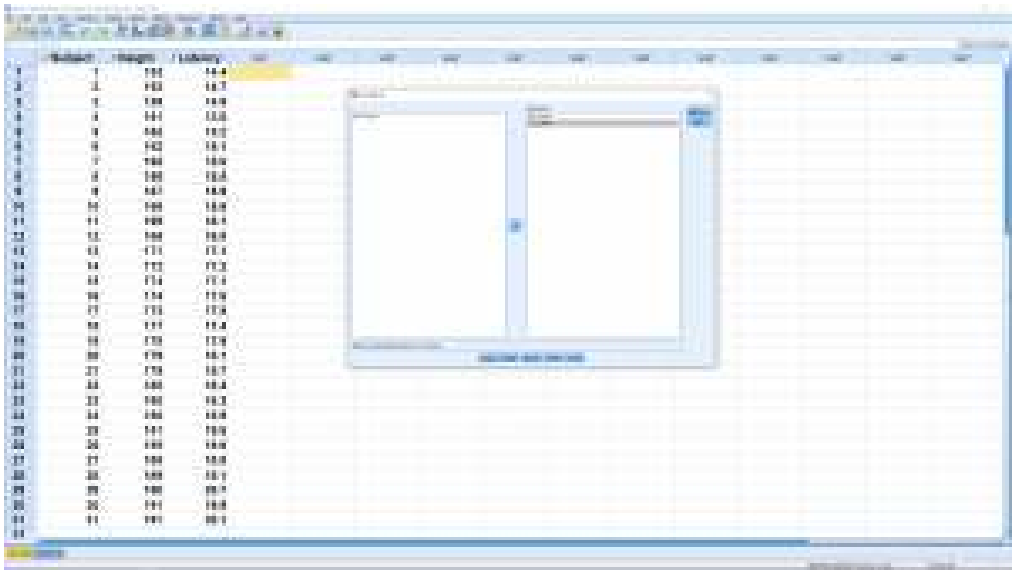
In SPSS Lesson 2 we saw how we can take variables defined on a Lickert scale and add them together, reverse scaling if necessary, to produce a single, better, variable for analysis. This works because the Lickert scale variables all have the same “units” (number of answer choices). You can combine any 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.



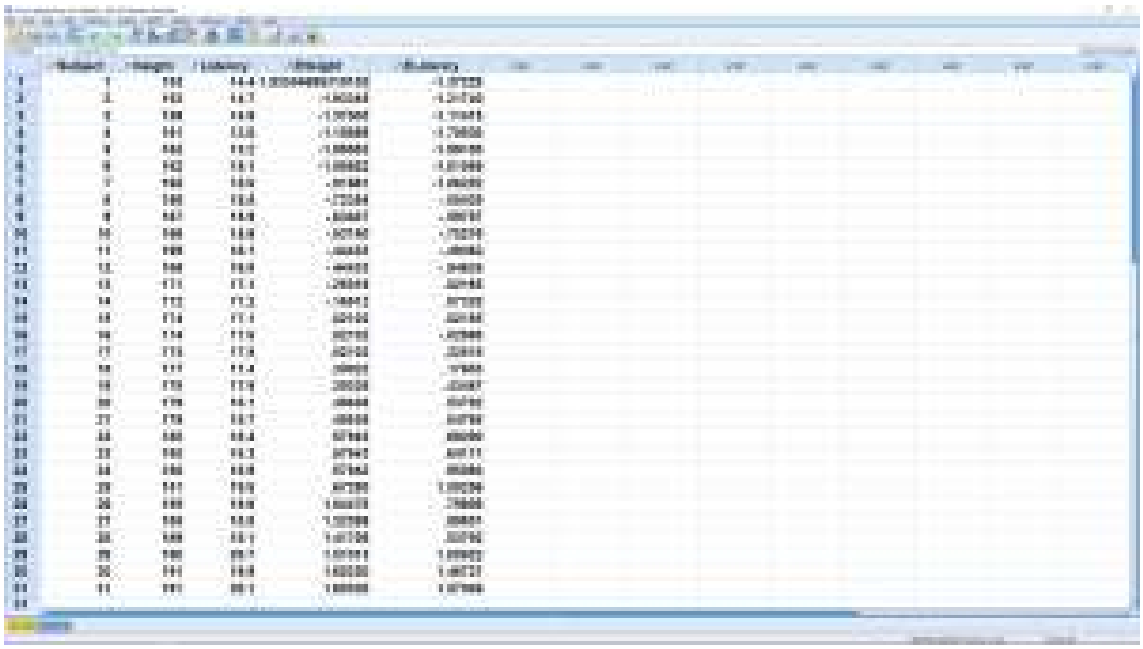
SPSS screenshot © International Business Machines Corporation.

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 :



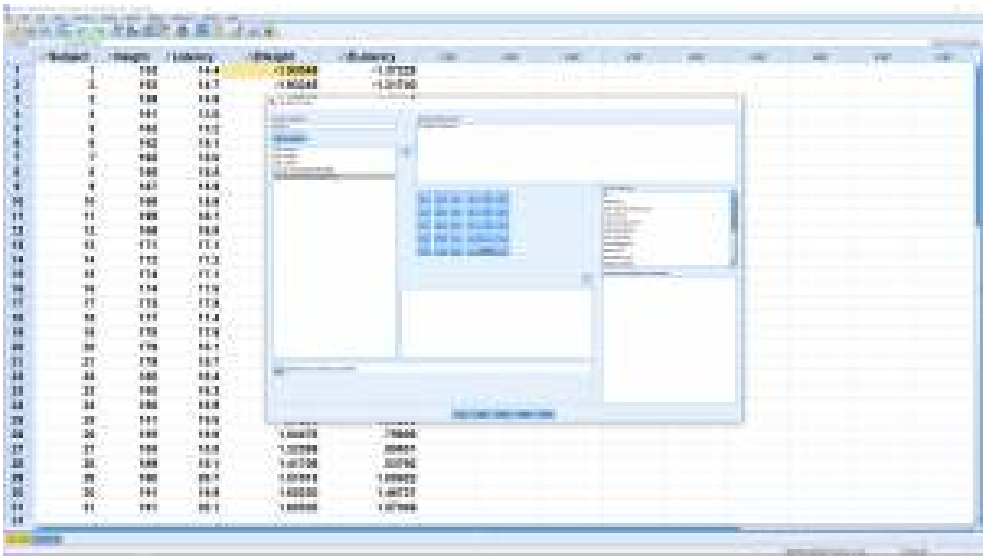
SPSS screenshot © International Business Machines Corporation.

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 :



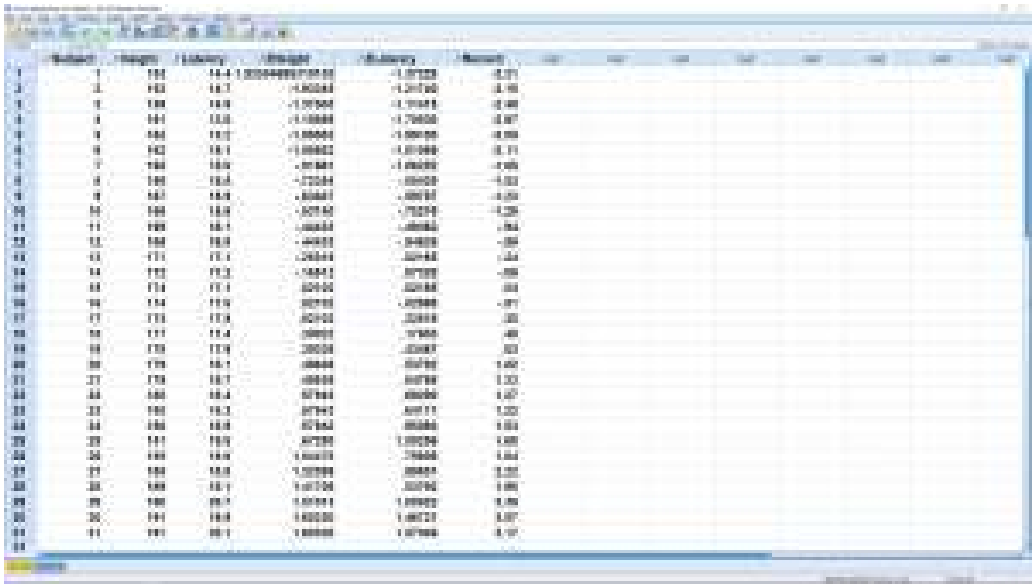
SPSS screenshot © International Business Machines Corporation.

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 :



SPSS screenshot © International Business Machines Corporation.

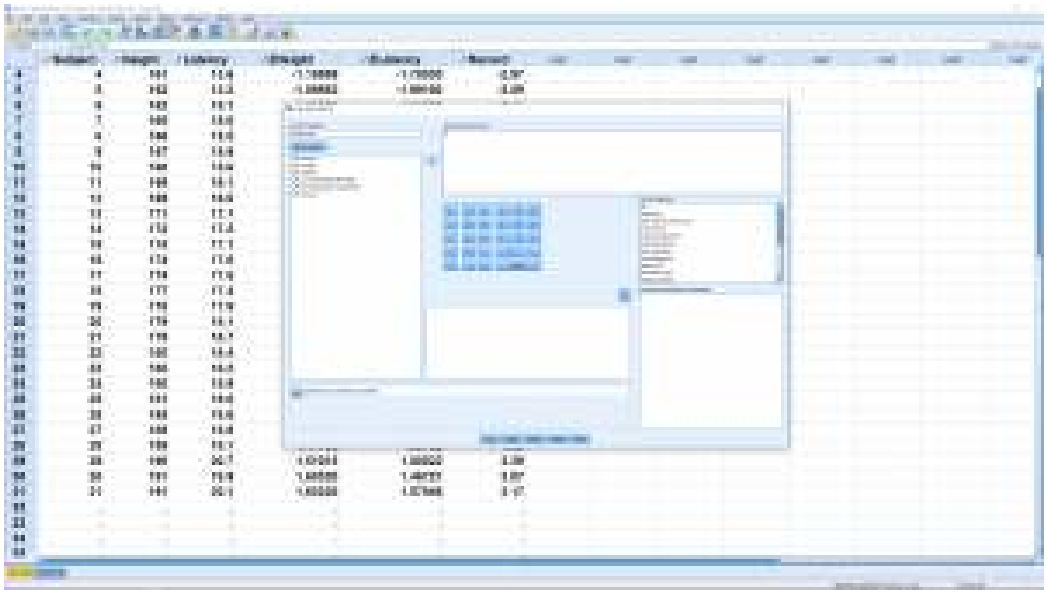
The new variable shows clearly on SPSS sheet :



SPSS screenshot © International Business Machines Corporation.

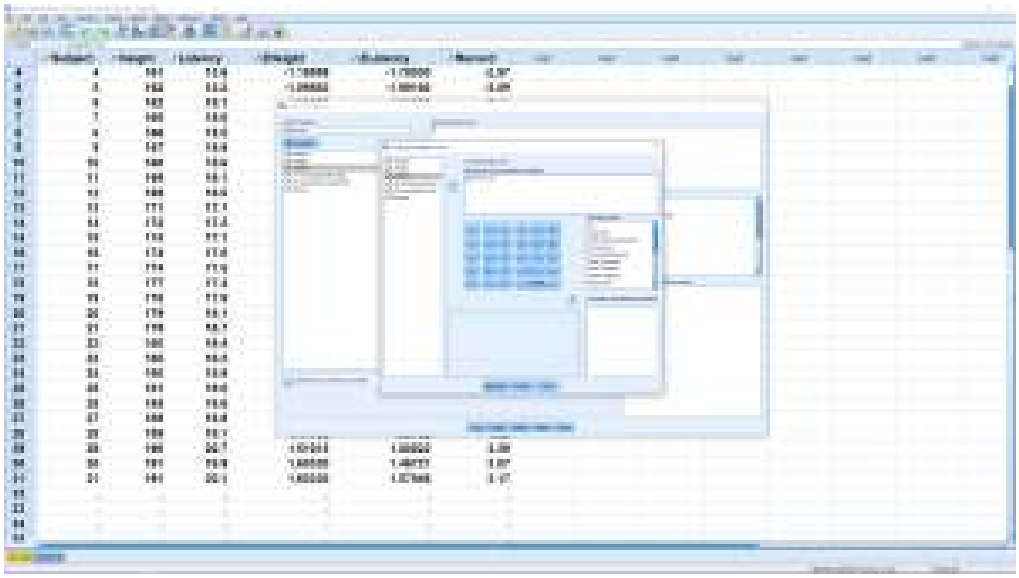
Next we will make a conversion from a quantitative variable to a qualitative variable essentially by dividing the data into classes. First a simple case. Create the new variable Life from the variable Latency as the following :

$$\text{Latency} = \begin{cases} 1 \text{ if } & \text{Latency} < 17.5 \\ 2 \text{ if } & \text{Latency} \geq 17.5 \end{cases} \tag{4.3.1}$$



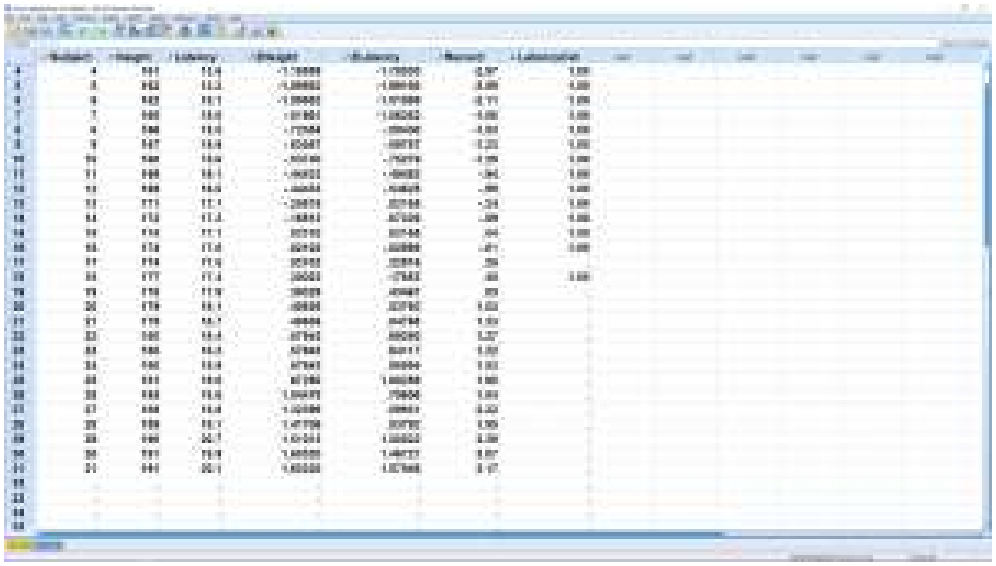
SPSS screenshot © International Business Machines Corporation.

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 :



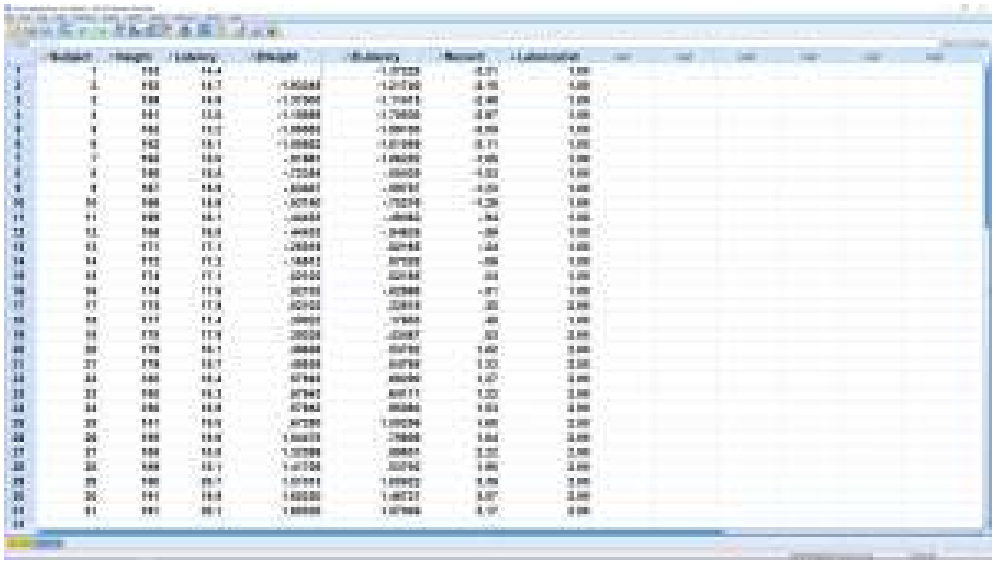
SPSS screenshot © International Business Machines Corporation.

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.



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 :



SPSS screenshot © International Business Machines Corporation.

This page titled 4.3: SPSS Lesson 3- Combining variables - advanced is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

5: The Normal Distributions

This page titled 5: The Normal Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 integrals[1] 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.

Discrete	Continuous
• 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**.	• We have an infinite number of values between the high and low values. • With continuous random variables we have a **probability density**.
By increasing the number of values in an appropriate limiting way you make $\longrightarrow$ the discrete probability distribution $\longrightarrow$ approach a probability density.	
• The units of $P(x)$ are probability.	• 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 :

$$\mu = \sum_x x \cdot P(x) \quad \sigma^2 = \sum_x [(x-\mu)^2 \cdot P(x)] \tag{5.1.1}$$

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

$$\mu = \int x \cdot P(x)\, dx \quad \sigma^2 = \int (x-\mu)^2 \cdot P(x)\, dx \tag{5.1.2}$$

  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) Rendered by QuickLaTeX.com

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 an on up) of: μ, σ^2, 0, 1, 0, 1, 0, ... The normal distribution is special that way among probability distributions.

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

This page titled 5.1: Discrete versus Continuous Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 Distributions[1]. A mathematical "trick" using logarithmic differentiation will be used.

First, recall the definition of the Binomial Distribution[2] as

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

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

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

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 Distribution[3] In other words as we let $n \to \infty$ we need to come up with a way to let Δx shrink[4] 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 $\overline{x} = np$ and variance $\sigma^2 = npq$. We will show how $lim_{n\to\infty} w_n(x) = w(x)$ where each Binomial Distribution $w_n(x)$ also has mean $\overline{x} = np$ and variance $\sigma^2 = npq$.

The heart of the trick is to notice[5] that

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

This is perfectly true for the density $w(x)$. The trick is to substitute the distribution $w_n(x)$ for the density $w(x)$ in the RHS of Equation (5.4) to get :

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

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 write[6]

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

$$= \lim_{n\to\infty} \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^x q^{n-x}} - 1 \tag{5.7}$$

$$= \lim_{n\to\infty} \frac{n-x}{x+1}\frac{p}{q} - 1. \tag{5.8}$$

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

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

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

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

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

$$= \frac{\left(nq - \frac{u}{h}\right)}{\left(\frac{uq}{\cdot} + nq + \frac{q}{\cdot}\right)} \tag{5.12}$$

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

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

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

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

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

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

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

$$= \lim_{n\to\infty} -\frac{u}{nhpq}. \tag{5.19}$$

where $O\left(\frac{1}{n}\right)$ means terms that will go to zero as $n \to \infty$, and we have used the relation $\frac{1}{1+x} = 1 - x + x^2 - x^3 + ...$ 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

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

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

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

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

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

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

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

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

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

so

$$C = \frac{1}{\sqrt{2\pi\sigma^2}}$$

and

$$w(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\overline{x})^2}{2\sigma^2}}$$

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

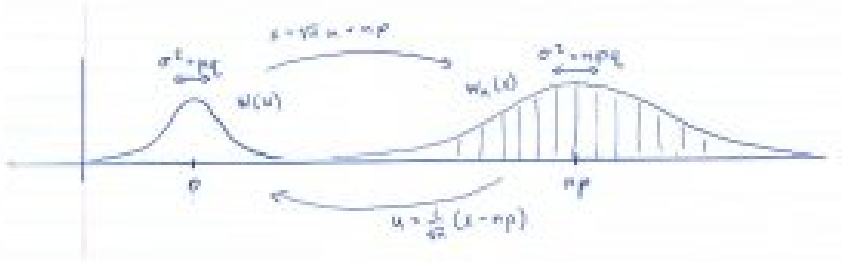


Figure 5.1 : The transformation $u = \frac{(x-np)}{\sqrt{n}}$ effectively shrinks the Δx of the Binomial Distribution with mean $\overline{x} = np$ and variance $\sigma^2 = npq$ by pulling a continuous version $w_n(x)$ back to the constant Normal Distribution $w(u)$. Another way of thinking about it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

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 $\overline{x} = np$) 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 Distribution[7] $w(x)$ with $x = \sqrt{n}u + np$ is an approximation for Binomial Distribution $w_n(x)$ that "asymptotically" approaches $w_n(x)$ as $n \to \infty$.

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.

---

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 $P(x \mid n)$. Here we use $w_n(x) = P(x \mid n)$ to avoid using too many P's and p's. ↵
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. ↵
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, $\lim_{n\to\infty} w_n(x)$ should exist at all $x \in \mathbb{R}$. ↵
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 $u_n(x) = \frac{(x-np)}{\sqrt{n}}$. ↵

This page titled 5.2: **The Normal Distribution as a Limit of Binomial Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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) Rendered by QuickLaTeX.com

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

$$\int P(x)\,dx = 1. \tag{5.3.1}$$

Without that factor we just have a bell-shaped curve [1] 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: $e = 2.718\ldots$ [2], $\pi = 3.1415\ldots$ [3], μ = mean and σ = standard deviation of the normal distribution. The normal distribution's shape is as shown in Figure 5.2.

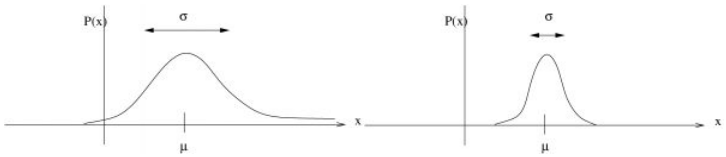


Figure 5.2: The normal distribution. It is a bell-shaped curve with its mode (= mean and median because it's symmetric, $\mu_3 = 0$) centred on its mean μ. On the left is a distribution with a large σ^2 and on the right one with a smaller σ^2.

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:

$$z = \frac{x - \mu}{\sigma} \tag{5.3.2}$$

applying this to $P(x)$ gives

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

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

(5.30) Rendered by QuickLaTeX.com

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.

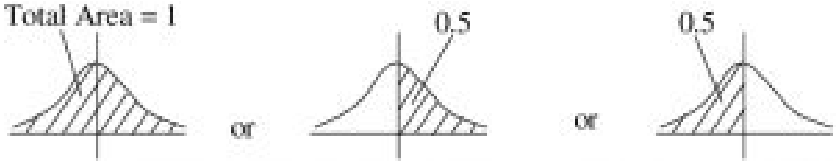


Figure 5.3 : The z-distribution is a probability distribution (total area = 1) and symmetric, so the area on either side of the mean (which is 0) is a half. You will need to remember this information as you calculate areas using the Standard Normal Distribution Table.

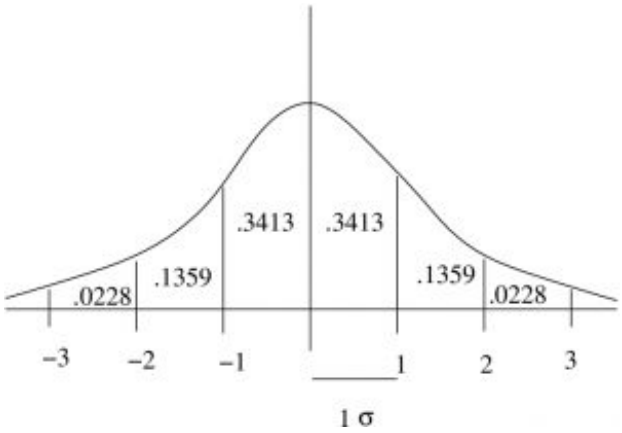


Figure 5.4 : The units of z in $P(z)$ are standard deviations. No matter what the measurement units of x were before the z-transformation, the units of z are "standardized" to be standard deviation units. With SPSS you will learn how to standardize (z-transform) variables so that you can sensibly combine 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 Bluman[4]. 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 tabulated values, $P(0 \le x \le z) = A(z)$, see Figure 5.5. This case also covers when z is a negative number: $P(-z \le x \le 0) = A(z)$.

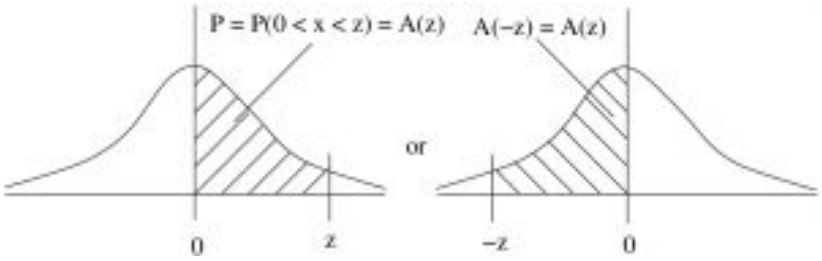


Figure 5.5 : Case 1: Areas on one side of the mean.

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 $\le$ because the probability of a single value is 0. That's why we need to use areas.)

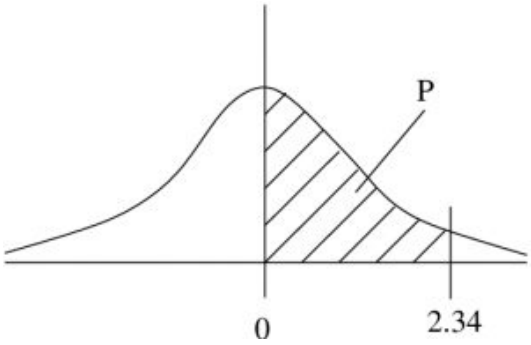


Figure 5.6 : The situation for Example 5.1.

□

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.

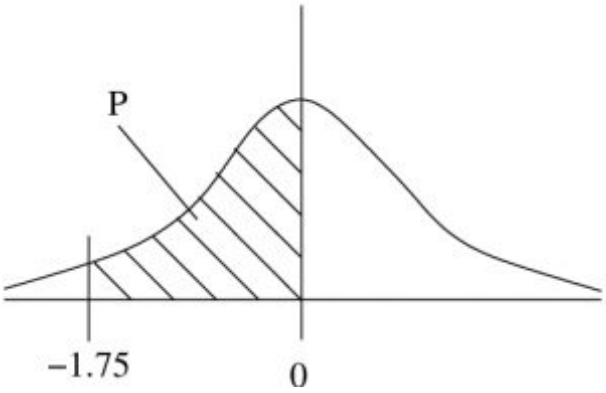


Figure 5.7 : The situation for Example 5.2.

□

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 $image$ z) = 0.5 - A(z)" title="Rendered by QuickLaTeX.com" height="19" width="211" style="vertical-align: -5px;"> or before a given negative value $-z$ is $P = P(x < -z) = 0.5 - A(z)$.

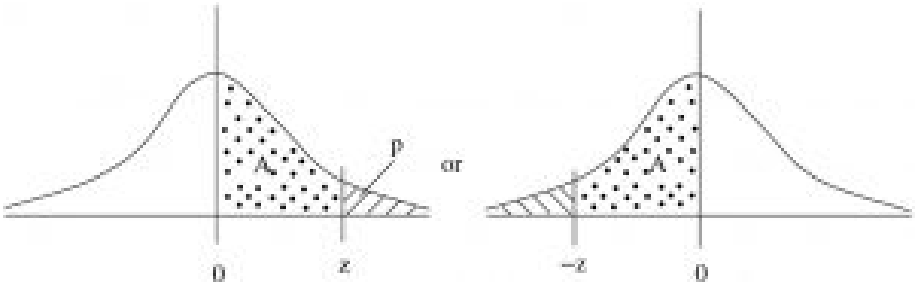


Figure 5.8 : Case 2 : Tail areas.

Example 5.3 : What is the probability that $image$ 1.11" title="Rendered by QuickLaTeX.com" height="14" width="64" style="vertical-align: -2px;">?

Solution : $image$ 1.11) = 0.5 - A(1.11) = 0.5 - 0.3665 = 0.1335" title="Rendered by QuickLaTeX.com" height="19" width="408" style="vertical-align: -5px;">, see Figure 5.9.

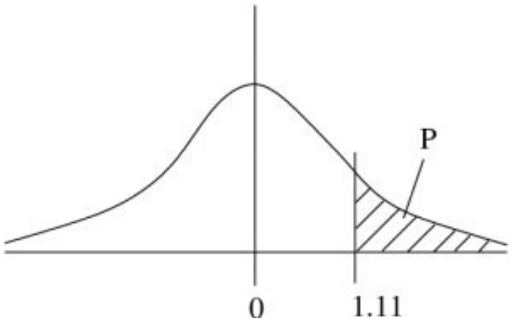


Figure 5.9 : The situation for Example 5.3.

□

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.

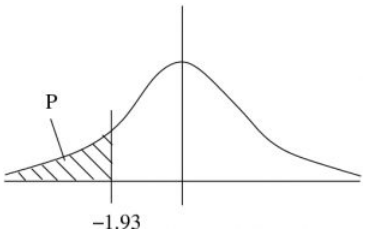


Figure 5.10 : The situation for Example 5.2.

□

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.

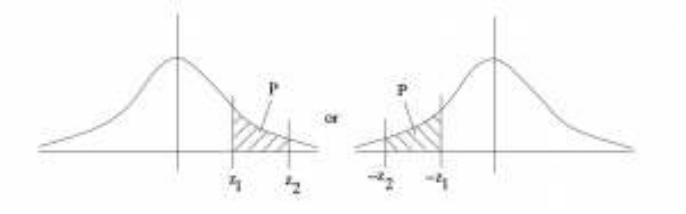


Figure 5.11: Case 3: An interval on one side of the mean.

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.

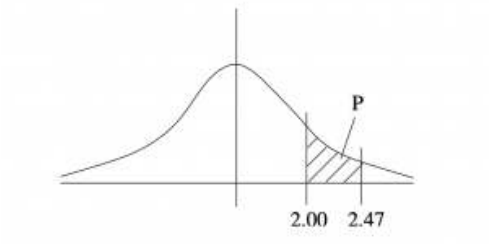


Figure 5.12: The situation for Example 5.5.

□

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.

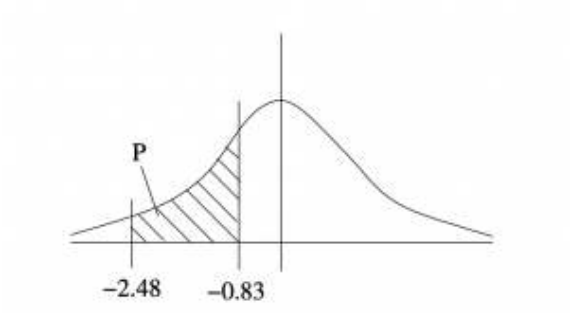


Figure 5.13: The situation of Example 5.6.

□

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 $P(-z_1 < x < z_2) = A(z_1) + A(z_2)$.

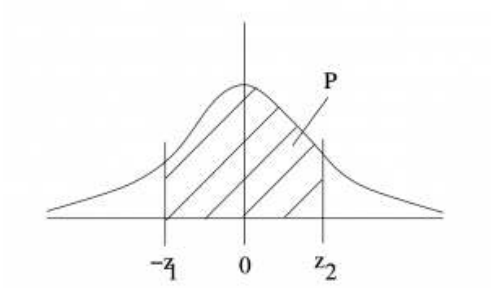


Figure 5.14: Case 4: An interval containing the mean.

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.

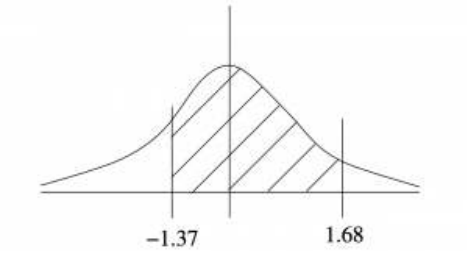


Figure 5.15: The situation for Example 5.7.

□

Cases 5 & 6 : Excluding tails. Case 5 is excluding the right tail, $P(x < z)$. Case 6 is excluding the left tail, $image$ -z)" title="Rendered by QuickLaTeX.com" height="19" width="83" style="vertical-align: -5px;">. See Figure 5.16. Case 5 is the situation which gives the percentile position of z if you multiply the are by 100. More about percentiles in Chapter 6. In either case, $P = 0.5 + A(z)$.

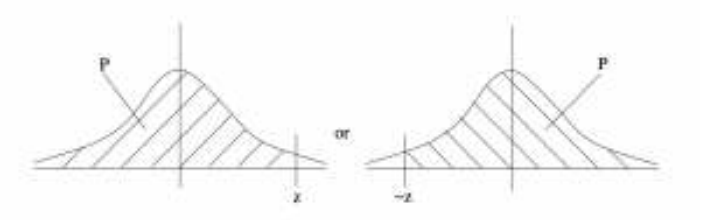


Figure 5.16: Left: Case 5. Right: Case 6.

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. $P = P(x < -z_1) + P(x < z_2) = (0.5 - A(z_1)) + (0.5 - A(z_2))$.

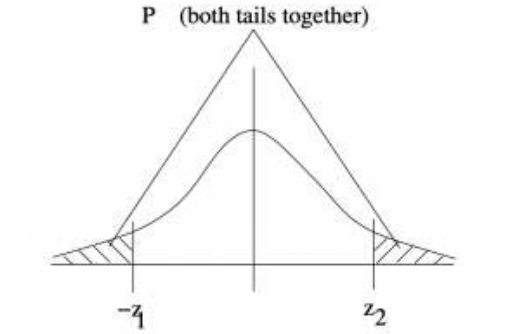


Figure 5.17: Case 7: Two unequal tails.

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

Solution :

$image$ 2.43)" title="Rendered by QuickLaTeX.com" height="19" width="197" style="vertical-align: -5px;">
$= (0.5 - A(3.01)) + (0.5 - A(2.43))$
$= (0.5 - 0.4987) + (0.5 - 0.4925)$
$= 0.0013 + 0.0075$
$= 0.0088.$

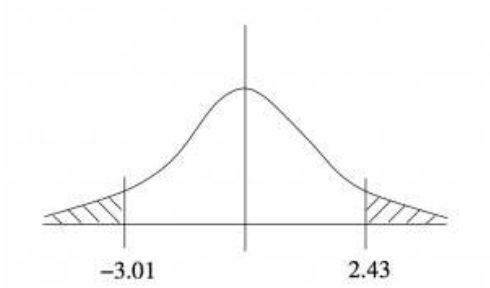


Figure 5.18: The situation for Example 5.8.

□

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 $z = 0.56$.

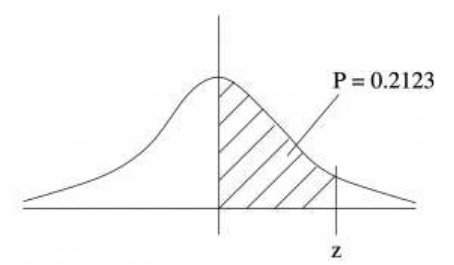


Figure 5.19: The situation for Example 5.9.

□

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:

$$z_1 = z(27) = \frac{27-28}{2} = \frac{-1}{2} = -0.5 \tag{5.3.3}$$

$$z_2 = z(31) = \frac{31-28}{2} = \frac{3}{2} = 1.5 \tag{5.3.4}$$

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

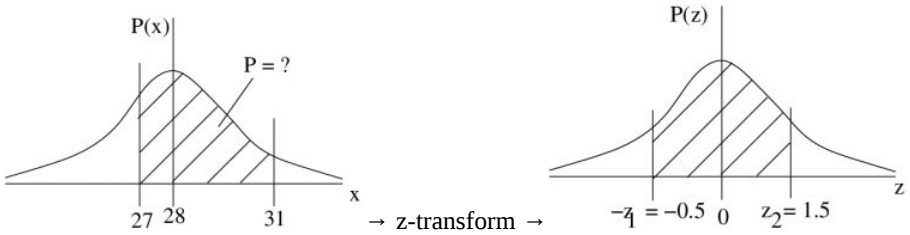


Figure 5.20 : The situation of Example 5.10. Left is the given population, $P(x)$. On the right is the z-transformed version of the population $P(z)$. The value 27 is z-transformed to -0.5 and 31 is z-transformed to 1.5.

□

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* :

$$x = z \cdot \sigma + \mu. \tag{5.3.5}$$

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 $image$ z) = 0.10" title="Rendered by QuickLaTeX.com" height="19" width="163" style="vertical-align: -5px;">. 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

$$x = 1.28 \times 20 + 200 = 25.60 + 200 = 225.60 \tag{5.3.6}$$

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.

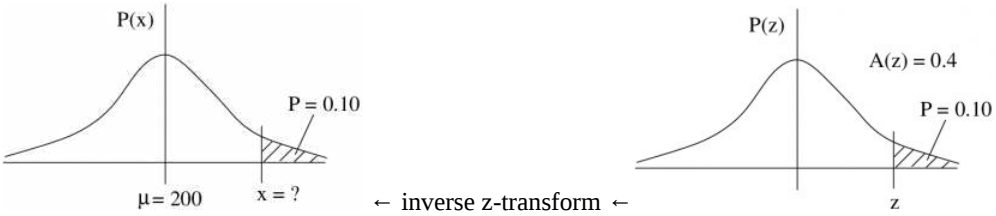


Figure 5.21 : The situation of Example 5.11

□

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: π= circumference/diameter. ↵
4. Bluman AG, *Elementary Statistics: A Step-by-Step Approach*, numerous editions, McGraw-Hill Ryerson, circa 2005. ↵

This page titled 5.3: Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

6: Percentiles and Quartiles

The concept of percentile[1] 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.

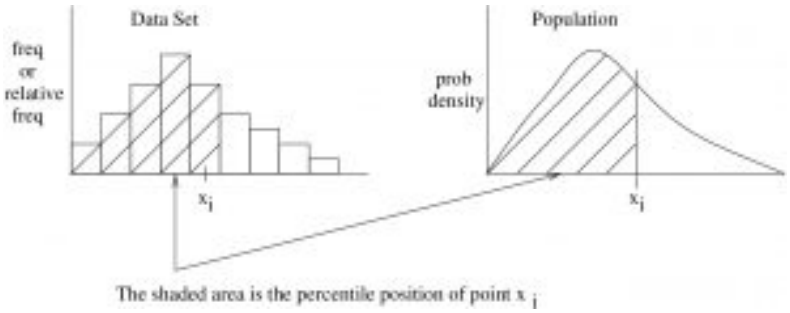


Figure 6.1: The concept of percentile applies to either a data set or to a continuous distribution.

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 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.

---

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. ↵

This page titled 6: Percentiles and Quartiles is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

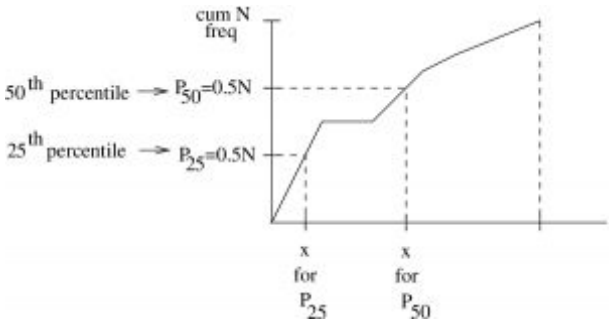


Figure 6.2 : With a cumulative frequency plot, we can read percentiles off the y axis. If you have a newborn baby and take it to the doctor for their first check up, they will measure the baby's head circumference and tell you the baby's head size percentile by looking at such a chart. The doctor's chart 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

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

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

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

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 to 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 25^{th} 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

$$D(x_i) = \frac{P(x_i)}{10} \qquad 0 \leq D(x_i) \leq 10 \tag{6.1.1}$$

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.

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

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

Q_0 &=& 0^{th} quartile
Q_1 & = & 1^{st} quartile
Q_2 & = & 2^{nd} quartile
Q_3 & = & 3^{rd} quartile
Q_4 & = & 4^{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:

$$\begin{aligned} Q_2 &= \text{ MD (median)} \\ Q_1 &= \text{ MD of values less than } Q_2 \\ Q_3 &= \text{ MD of values greater than } Q_2 \\ Q_0 &= L \\ Q_4 &= H \end{aligned}$$

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 :

$$5 \quad 6 \quad 12 \quad 13 \quad 15 \quad 18 \quad 22 \quad 50 \tag{6.1.2}$$

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

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

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

$Q_0 = L = 5$

$Q_4 = H = 50$

□

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 :

$$2 \quad 5 \quad 11 \quad 14 \quad 18 \quad 25 \quad 35 \tag{6.1.3}$$

$Q_1 = 5$

$MD = 14 = Q_2$

$Q_3 = 25$

$Q_0 = L = 2$

$Q_4 = H = 35$

□

This page titled 6.1: Discrete Data Percentiles and Quartiles is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$\text{Interquartile Range} = IQR = Q_3 - Q_1 \tag{6.2.1}$$

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

$$Q_1 = 9 \quad Q_2 = 14 \quad Q_3 = 20 \tag{6.2.2}$$

so,

$$IQR = Q_3 - Q_1 = 20 - 9 = 11. \tag{6.2.3}$$

Following our rules for finding outliers, we compute:

(a) lower acceptable value limit

$$\begin{aligned} &= Q_1 - (1.5 \times IQR) \\ &= 9 - (1.5 \times 11) \\ &= 9 - 16.5 = -7.5 \end{aligned}$$

(b) upper acceptable value limit

$$\begin{aligned} &= Q_3 + (1.5 \times IQR) \\ &= 20 + (1.5 \times 11) \\ &= 20 + 16.5 = 36.5 \end{aligned}$$

and 50 > 36.5 so 50 is considered an outlier.

□

This page titled 6.2: Finding Outliers Using Quartiles is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

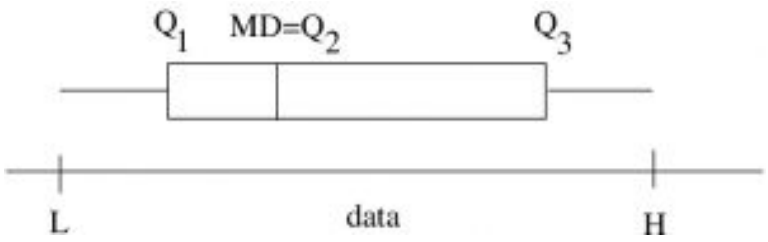


Figure 6.3: The features of a box plot, also known as a box-and-whiskers plot. When one of the whiskers is more than 1.5 times the length of the box (the IQR) then there are outliers by our definition in Section 6.2. The data line shown below the box plot is a construction line and not part of the box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

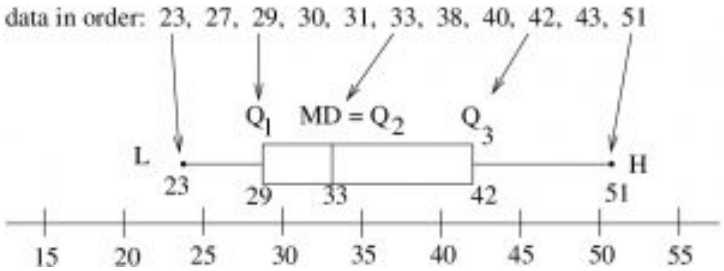


Figure 6.4: Construction of a box plot from the given data.

□

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

This page titled 6.3: Box Plots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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) 3 σ 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	$\bar{x}$
IQR	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:

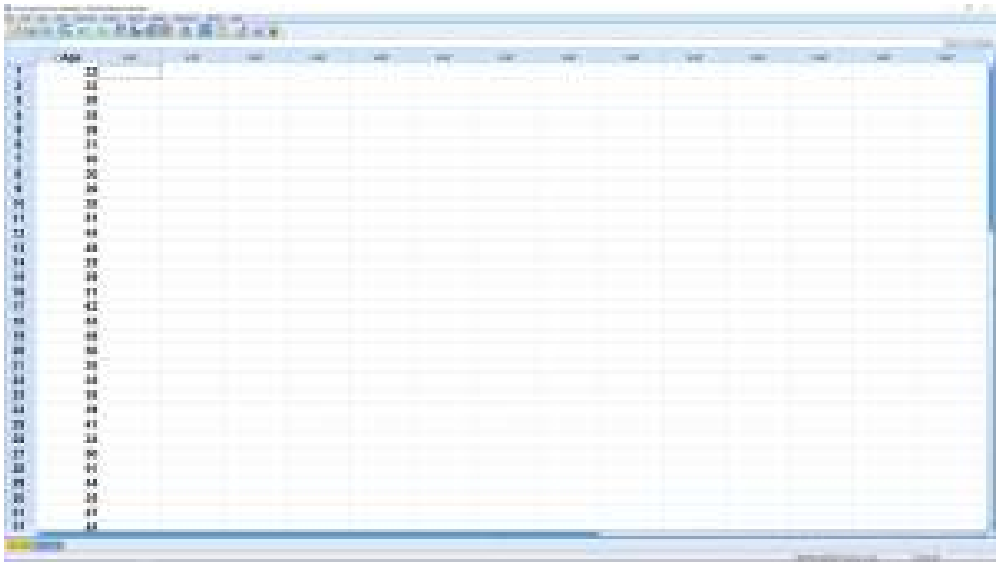
Traditional	Exploratory Data Analysis (EDA)
Frequency Tables	Stem and Leaf Plot
Histogram	Box Plot
Mean, $\bar{x}$	Median, MD
Standard Deviation, s	Interquartile Range, IQR

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

This page titled 6.4: Robust Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

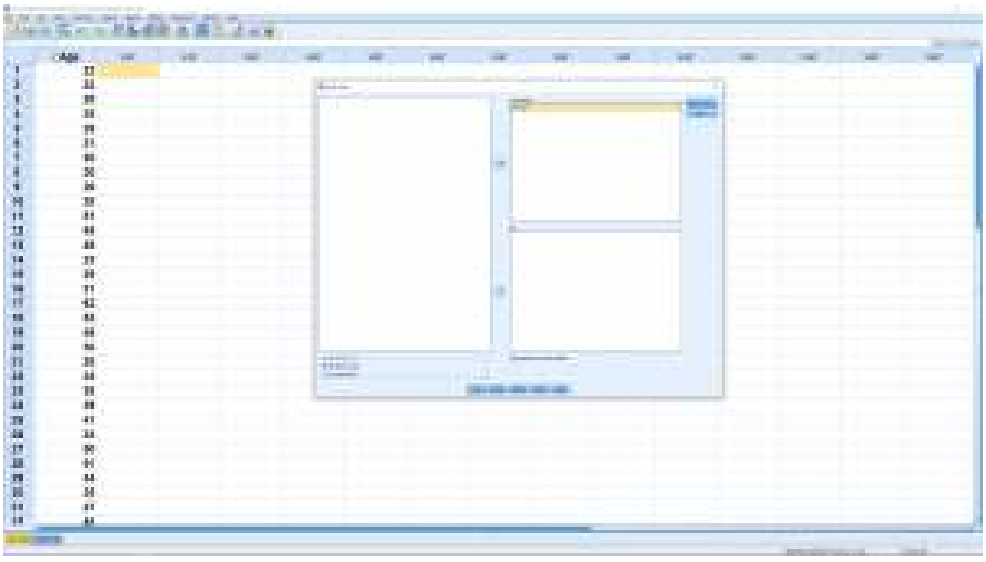
6.5: SPSS Lesson 4- Percentiles

To follow along, load in the file “AgeSmoker.sav” from the Data Sets. We will pick on the variable Age. We will compute the percentile rank of each value in the Age dataset two ways. One way, we will treat the data as a discrete data set and will compute the percentile position following Equation 6.1. The other way, we will treat the data as if they came from a normal population.



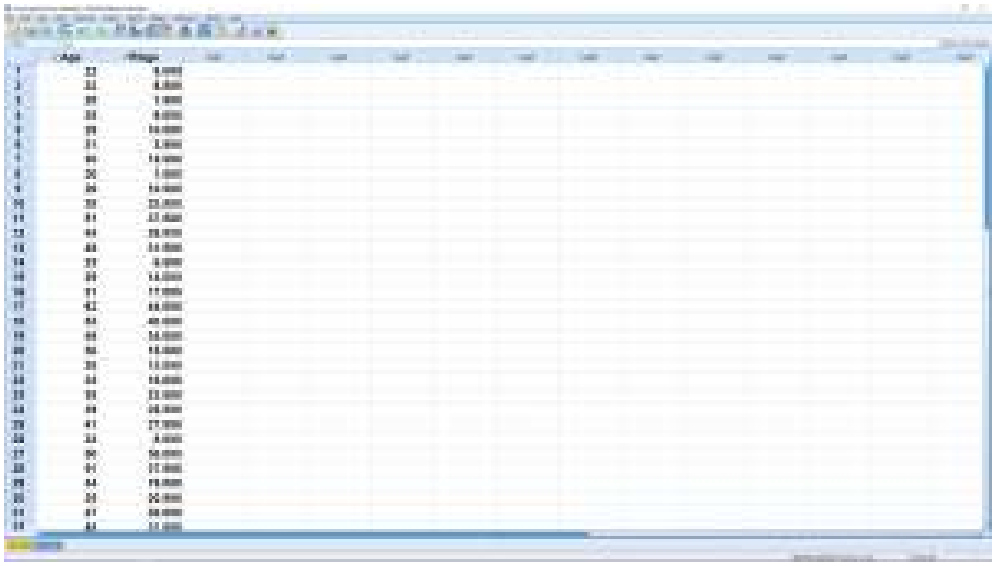
SPSS screenshot © International Business Machines Corporation.

First, treat the data as a stand alone discrete data set. First we need to rank the data; the ranks are the values i in Equation 6.1. Use Transform $\rightarrow$ Rank Cases :



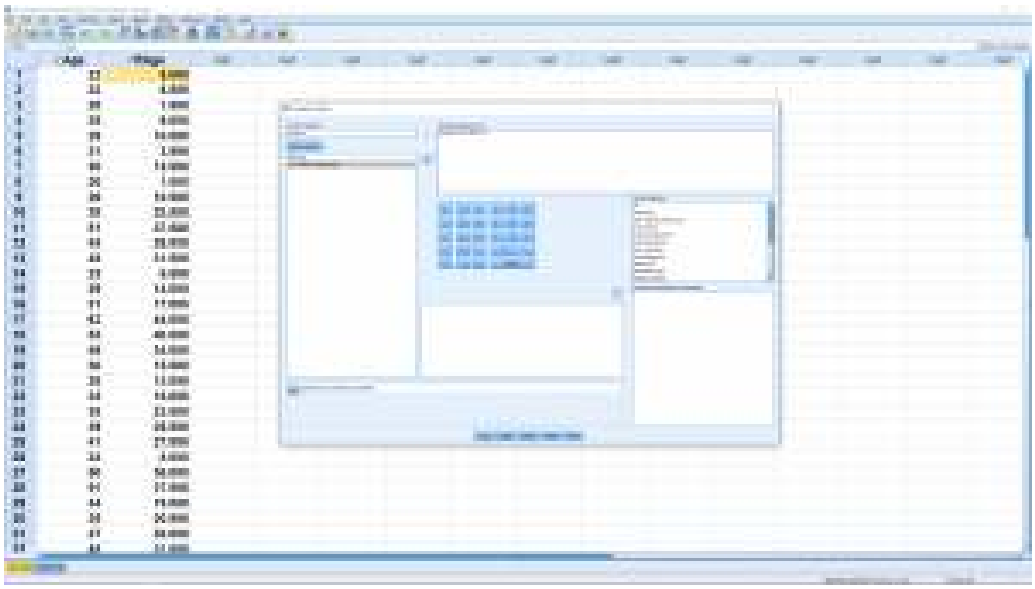
SPSS screenshot © International Business Machines Corporation.

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



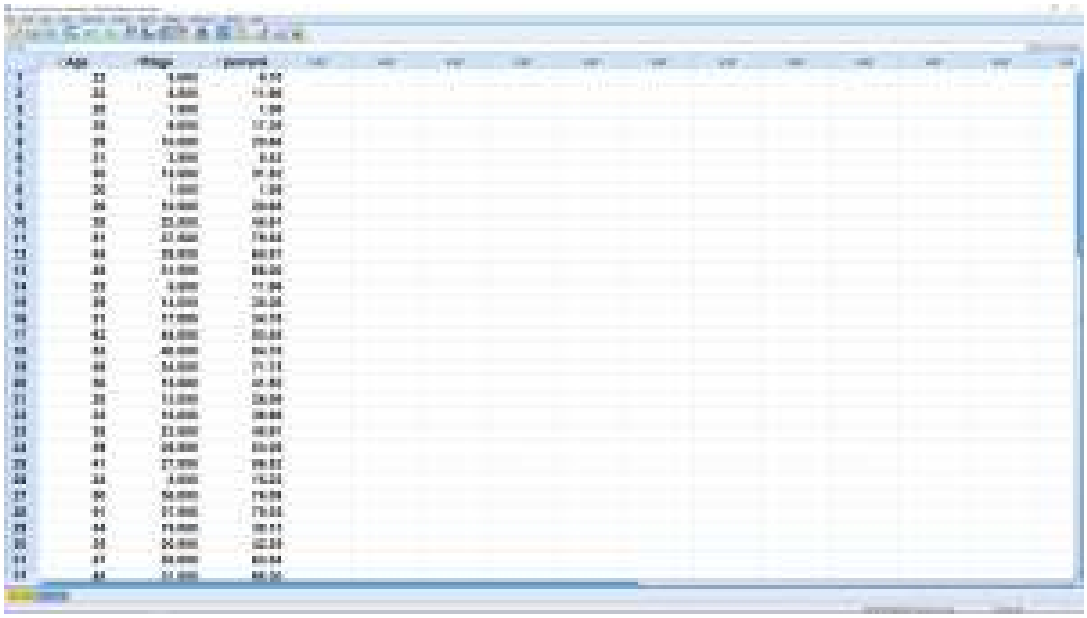
SPSS screenshot © International Business Machines Corporation.

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



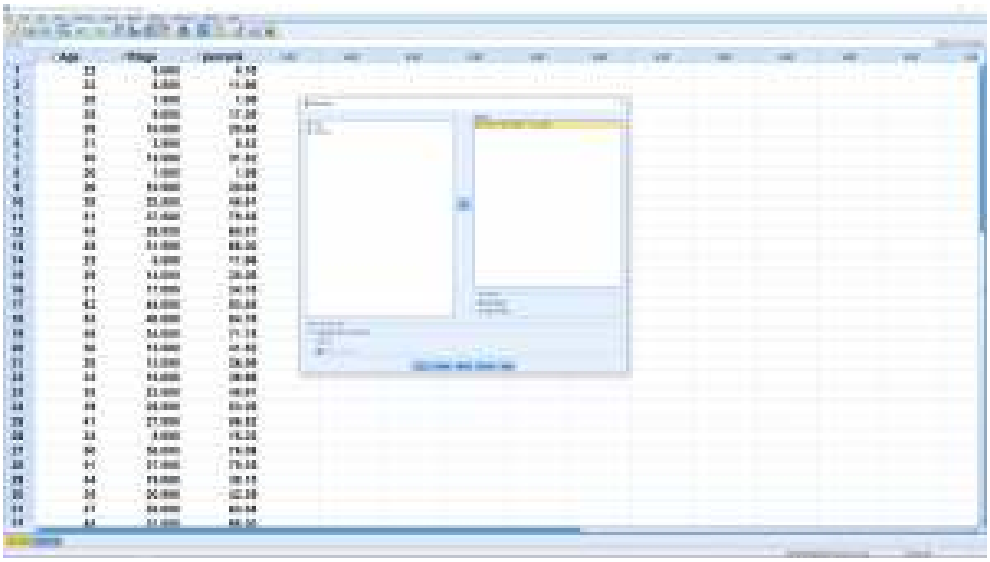
SPSS screenshot © International Business Machines Corporation.

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



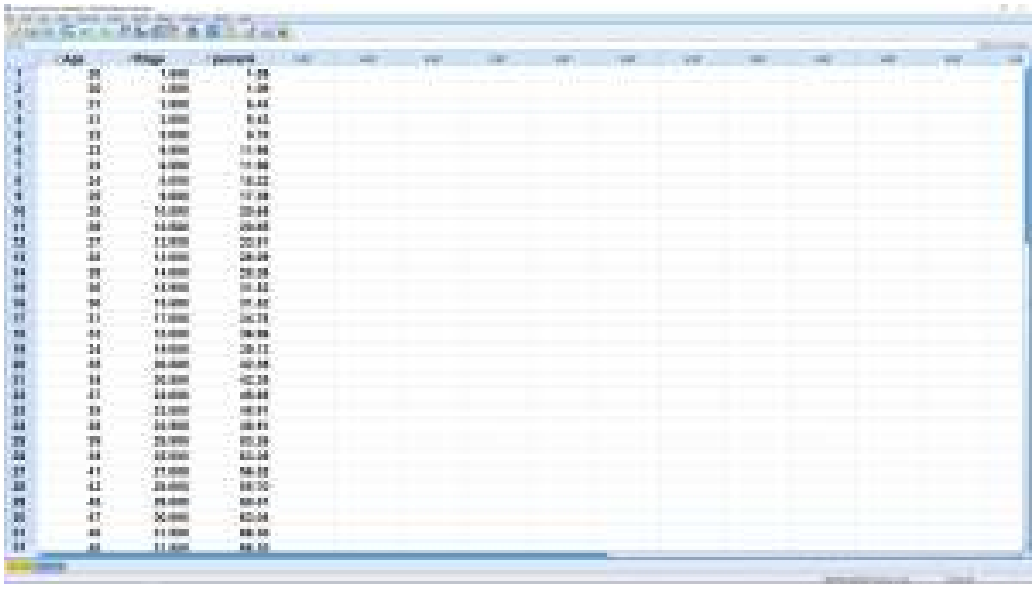
SPSS screenshot © International Business Machines Corporation.

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

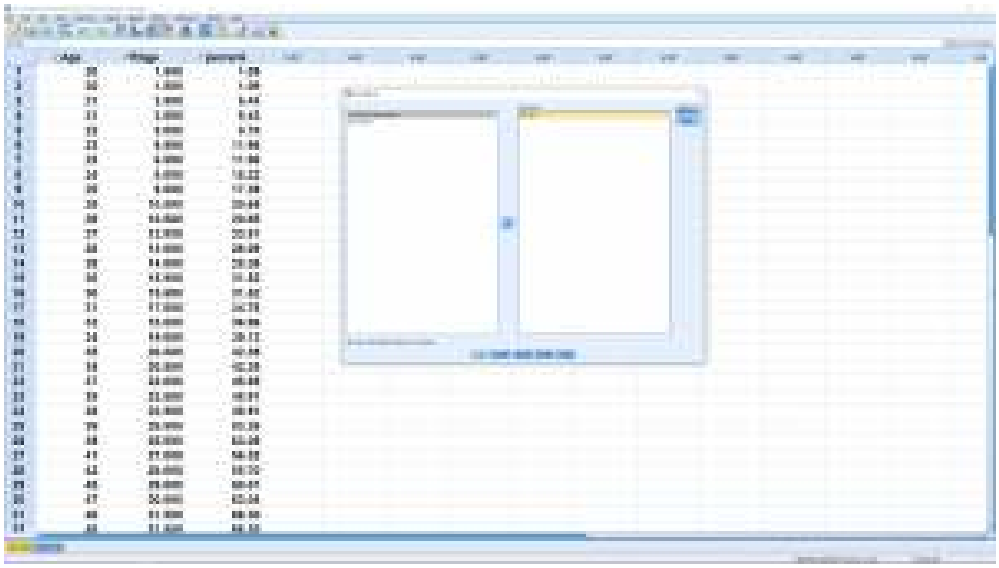


SPSS screenshot © International Business Machines Corporation.

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.

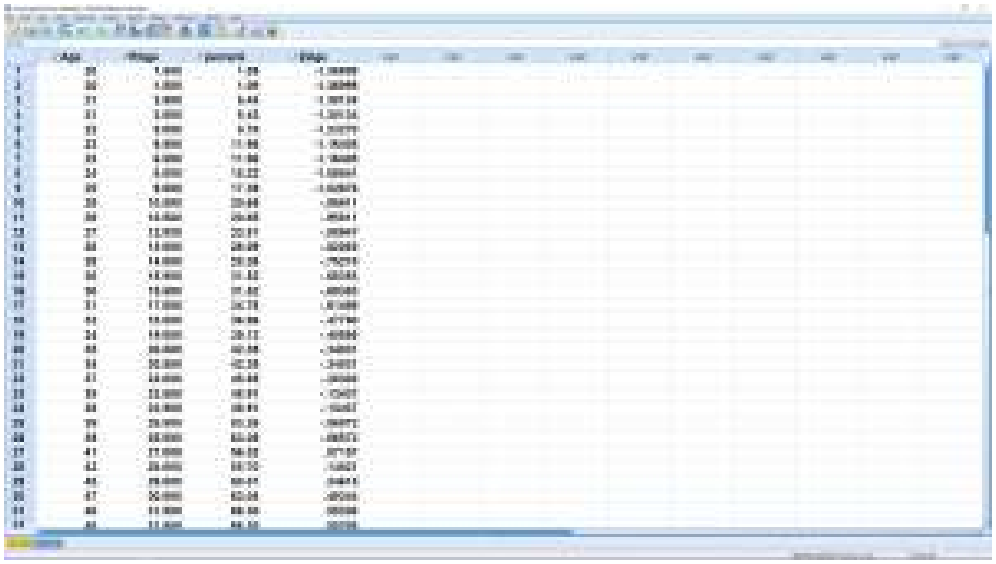


SPSS screenshot © International Business Machines Corporation.



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 :



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.

This page titled 6.5: SPSS Lesson 4- Percentiles is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

This page titled 7: The Central Limit Theorem is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

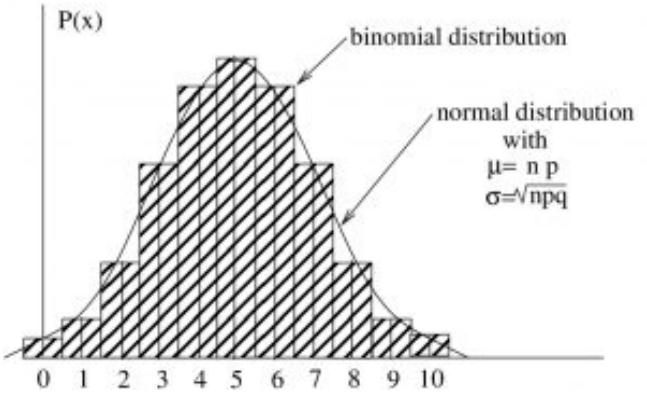


Figure 7.1: A normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$ is a good fit to the binomial distribution with the same mean and standard deviation as long as $np \geq 5$ and $nq \geq 5$.

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.

This page titled 7.1: Using the Normal Distribution to Approximate the Binomial Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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, $\overline{x}_1$, of that sample**. Put the sample back into the population[1]
3. Randomly pick another sample of size n. **Compute the mean of the new sample, $\overline{x}_2$**. Return the sample to the population..
4. Repeat step 3 an infinite number of times and **build up your collection of sample means $\overline{x}_i$**.
5. Then[2] 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

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.2.1}$$

where σ is the population's standard deviation. $\sigma_{\overline{x}} = \sigma/\sqrt{n}$ 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 $\overline{x}$ (see Figure 7.2a).

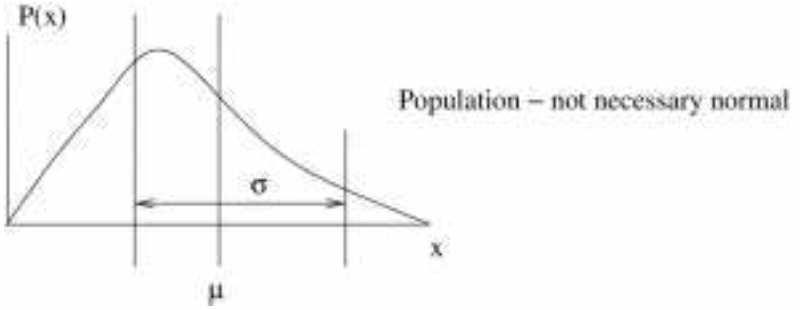


Figure 7.2a

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

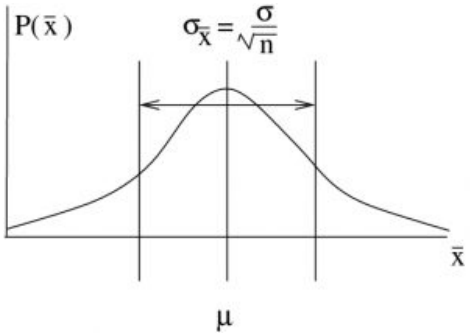


Figure 7.2b

The central limit theorem is our *fundamental sampling theory*. It tells us the *if* we know what the mean and standard deviation of a population[3] are *then* we can assign the probabilities of getting a certain mean $\overline{x}$ 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 $(\overline{x})$ to a standard normal z using the z-tranformation as usual: $z = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}}$. So, *for the distribution of sample means* the appropriate z-transformation is :

$$z = \frac{\overline{x} - \mu}{\sigma\sqrt{n}} \tag{7.2.2}$$

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.

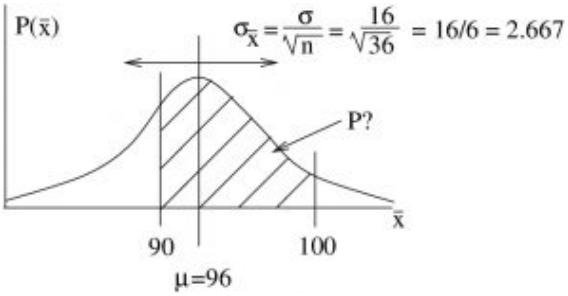
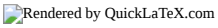


Figure 7.3 : Distribution of mean age from samples of 36 cars.

Convert 90 and 100 to z-scores as usual:



Then, the required probability using the **Standard Normal Distribution Table** is

$$\begin{aligned} P &= A(2.25) + A(1.50) \\ &= 0.4878 + 0.4332 \\ &= 0.921 \qquad (92.1\%) \end{aligned}$$

□

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 $n \to \infty$. 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. ↵
3. In hypothesis testing we know what the mean of the population in the null hypothesis is. ↵

This page titled 7.2: The Central Limit Theorem is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

8: Confidence Intervals

This page titled 8: Confidence Intervals is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 $\overline{x}$ 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 $\overline{x}$ 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 $\overline{x}$) then the central limit theorem gives the probability $P(D \mid H) = P(\overline{x} \mid \mu, \sigma)$. The confidence intervals that we'll look at first give $P(H \mid D) = P(\mu \mid \overline{x}, \sigma)$. 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 probability[1] of $\mathcal{C}$.

Typically $\mathcal{C}$ will be

$$\mathcal{C} = 0.90 \qquad (90\% \text{ confidence}) \tag{8.1.1}$$

$$\mathcal{C} = 0.95 \qquad (95\% \text{ confidence}) \tag{8.1.2}$$

$$\mathcal{C} = 0.99 \qquad (99\% \text{ confidence}) \tag{8.1.3}$$

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 $z_{\mathcal{C}} = z_{\alpha/2}$ where $\mathcal{C} = 1 - \alpha$ be the z-value, from the **Standard Normal Distribution Table** that corresponds to an area, between 0 and $z_{\mathcal{C}}$ of $\mathcal{C}/2$ as shown in Figure 8.1.

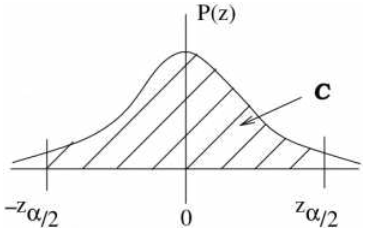


Figure 8.1 : The z-distribution areas of interest associated with $z_{\mathcal{C}} = z_{\alpha/2}$.

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 $\sigma/\sqrt{n}$ as shown in Figure 8.2.

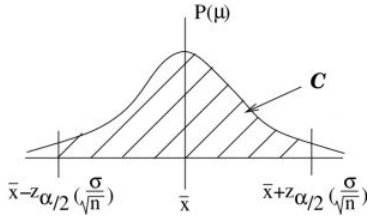
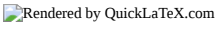


Figure 8.2 : The inverse z-transformation of Figure 8.1 gives the confidence interval for μ.

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 :



or using notation that we will use as a standard way of denoting symmetric confidence intervals

$$\bar{x} - E < \mu < \bar{x} + E$$

where

$$E = z_{\mathcal{C}} \left(\frac{\sigma}{\sqrt{n}} \right). \tag{8.1.4}$$

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

$$\mu = \bar{x} \pm E \tag{8.1.5}$$

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.

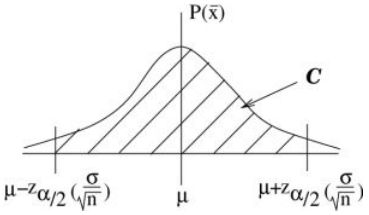


Figure 8.3 : The central limit theorem is about distributions of sample means.

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 $\mathcal{C} = 0.95$. First write down the formula prescription so you can see with numbers you need:

$$\bar{x} - E < \mu < \bar{x} + E \qquad \text{where} \qquad E = z_{95\%} \frac{\sigma}{\sqrt{n}}. \tag{8.1.6}$$

First determine $z_{\mathcal{C}} = z_{\alpha/2}$. 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 $z_{\mathcal{C}} = 1.96$. 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

$$E = 1.96(\frac{2}{\sqrt{50}}) = 0.6 \tag{8.1.7}$$

so

$$\begin{gathered} \bar{x} - E < \mu < \bar{x} + E \\ 23.2 - 0.6 < \mu < 23.2 + 0.6 \\ 22.6 < \mu < 23.8 \end{gathered}$$

with 95% confidence.

□

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". ↵

This page titled 8.1: Confidence Intervals Using the z-Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$P(A \mid B)P(B) = P(B \mid A)P(A). \tag{8.2.1}$$

Study Figure 8.4 to convince yourself that Bayes' rule is true. Notice that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{8.2.2}$$

and

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{8.2.3}$$

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

$$P(H \mid D) = P(D \mid H)\left[\frac{P(H)}{P(D)}\right].$$

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 $P(D \mid H)$ is what sampling theory, like the central limit theorem, gives and is known as the *likelihood*. Finally the quantity $P(H \mid D)$ 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).

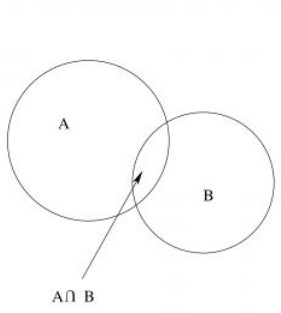


Figure 8.4 : Venn diagram illustration of Bayes rule.

If we assign $[P(H)/P(D)] = 1$ for the prior probability then $P(H \mid D) = P(D \mid H)$. 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.

This page titled 8.2: **Bayesian Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

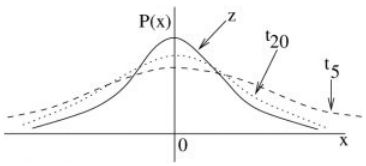


Figure 8.5 : The t-distributions are a family of distributions, labeled here by their degrees of freedom ν as in t_ν.

As the degrees of freedom, ν, increases, t_ν become closer to z, $\lim_{\nu\to\infty} t_\nu = z$. 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 $\mathcal{C}$ confidence interval for the mean μ for small n samples is given by

$$\bar{x} - E < \mu < \bar{x} + E \tag{8.3.1}$$

where, now

$$E = t_{\nu,\mathcal{C}} \left(\frac{s}{\sqrt{n}} \right). \tag{8.3.2}$$

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 $z_\mathcal{C}$ with $t_{\nu,\mathcal{C}}$. Some books use $t_{\nu,\mathcal{C}} = t_{\nu,\alpha/2}$ like the $z_\mathcal{C}$ of Section 8.1. We use $t_{\nu,\mathcal{C}}$ because we'll look up its value in the ***t* Distribution Table** in the column for $\mathcal{C}$ 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 :

$$\nu = n - 1. \tag{8.3.3}$$

The $t_{\nu,\mathcal{C}}$ specify a probability $\mathcal{C}$ as shown in Figure 8.6. As before, the inverse z-transform, in the form $x = t_{\nu,\mathcal{C}} s + \bar{x}$ 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.

\(" class="alignnone size-medium wp-image-1620" height="145" width="300" src="/@api/deki/files/38426/fig57png-300x145.jpg" /> → inverse

Figure 8.6 : Derivation of confidence intervals for means of small samples.

Example 8.2 : Given the following data:

$$5460 \quad 5900 \quad 6090 \quad 6310 \quad 7160 \quad 8440 \quad 9930 \tag{8.3.4}$$

find the 99% confidence interval for the mean.

Solution : First count $n = 7$ and then, with your stats calculator compute

$$\bar{x} = 7041.4 \quad \text{and} \quad s = 1610.3. \tag{8.3.5}$$

Using the ***t* Distribution Table** with $\nu = n - 1 = 6$ in the 99% confidence interval column, find

$$t_{n-1,\mathcal{C}} = t_{6,99\%} = 3.707. \tag{8.3.6}$$

With these numbers, compute

$$E = t_{n-1,\mathcal{C}} \left(\frac{s}{\sqrt{n}} \right) = 3.707 \left(\frac{1610.3}{\sqrt{7}} \right) = 2256.2 \tag{8.3.7}$$

so

$$\begin{aligned} \bar{x} - E &< \mu &< \bar{x} + E \\ 7041.4 - 2256.2 &< \mu &< 7041.4 + 2256.2 \\ 4785.2 &< \mu &< 9297.6 \end{aligned}$$

is the 99% confidence interval for μ.

□

This page titled 8.3: The t-Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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

$$p = \frac{X}{N} \tag{8.4.1}$$

Given a sample from the population of size n, the best estimate for p is:

$$\hat{p} = \frac{x}{n} \tag{8.4.2}$$

where x is the number of items in the sample having the given quality. To go along with $\hat{p}$ we also have

$$\hat{q} = 1 - \hat{p} \tag{8.4.3}$$

which is 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

$$E = z_{\mathcal{C}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{8.4.4}$$

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

$$\hat{p} - E < p < \hat{p} + E. \tag{8.4.5}$$

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

$$x_{\text{pop}} = \frac{n}{N} X = np \tag{8.4.6}$$

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 $\sigma = \sqrt{npq}$.

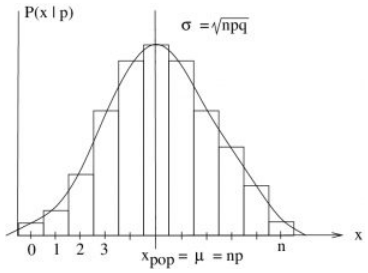


Figure 8.7 : The binomial distribution relevant to forming a sample of size n with x items of interest from a population with a proportion p of items of interest. The normal distribution with the same μ and σ is shown.

Moving to the normal approximation, we have the picture of Figure 8.8.

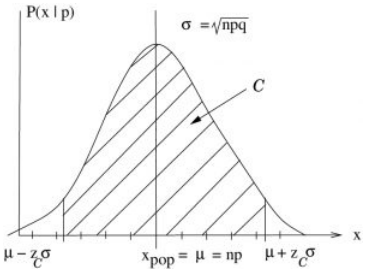


Figure 8.8 : The normal distribution relevant to forming a sample of size n with x items of interest from a population with a proportion p of items of interest. The boundaries of the area $\mathcal{C}$ follow from an inverse z-transform of the z-distribution to a normal distribution of mean μ and standard deviation σ, $x = z\sigma + \mu$.

Figure 8.8 says :

$$\begin{aligned} \mu - z_{\mathcal{C}}\sigma \ <x \ < \ \mu + z_{\mathcal{C}}\sigma \\ np - z_{\mathcal{C}}\sqrt{npq} \ <x \ < \ np + z_{\mathcal{C}}\sqrt{npq} \end{aligned}$$

with a (frequentist) probability of $\mathcal{C}$. This is our sampling theory. Divide by n:

$$\begin{aligned} p - z_{\mathcal{C}}\sqrt{\frac{pq}{n}} < \frac{x}{n} \ < \ p + z_{\mathcal{C}}\sqrt{\frac{pq}{n}} \\ p - z_{\mathcal{C}}\sqrt{\frac{pq}{n}} < \hat{p} \ < \ p + z_{\mathcal{C}}\sqrt{\frac{pq}{n}} \end{aligned}$$

Swapping the roles of the population and sample, we arrive at the confidence interval formula :

$$\hat{p} - z_{\mathcal{C}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \ < \ p \ < \ \hat{p} + z_{\mathcal{C}}\sqrt{\frac{\hat{p}\hat{q}}{n}}. \tag{8.4.7}$$

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

$$z_{\mathcal{C}} = 1.65 \tag{8.4.8}$$

and compute

$$\hat{p} = \frac{x}{n} = \frac{60}{500} = 0.12 \tag{8.4.9}$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.12 = 0.88 \tag{8.4.10}$$

Rendered by QuickLaTeX.com

Then

$$\hat{p} + E < p < \hat{p} - E \tag{8.4.11}$$

$$0.12 - 0.024 < p < 0.12 + 0.024 \tag{8.4.12}$$

$$0.096 < p < 0.144 \tag{8.4.13}$$

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 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 $\mathcal{C}$ is

$$n = \hat{p}\hat{q}\left(\frac{z_{\mathcal{C}}}{E}\right)^2. \tag{8.4.14}$$

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.

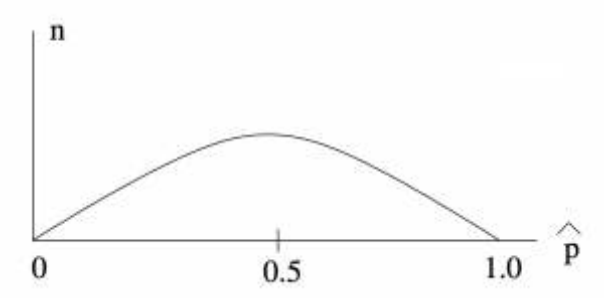


Figure 8.9 : The formula $n = \hat{p}\hat{q}\,(\frac{z_{\mathcal{C}}}{E})^2$ is a quadratic formula. Substitute $\hat{q} = 1 - \hat{p}$ to get $n = \hat{p}(1-\hat{p})\,(\frac{z_{\mathcal{C}}}{E})^2$ or $n = (\hat{p} - \hat{p}^2)\,(\frac{z_{\mathcal{C}}}{E})^2$. The maximum of $n_{\max} = \frac{1}{4}(\frac{z_{\mathcal{C}}}{E})^2$ 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 :

$$\hat{p} = 0.40, \quad \hat{q} = 0.60 \tag{8.4.15}$$

$$E = 0.02, \quad \alpha = 0.95 \tag{8.4.16}$$

From the ***t* Distribution Table** (or the **Standard Normal Distribution Table** if you think about the areas correctly) we find

$$z_{\mathcal{C}} = z_{95\%} = 1.960. \tag{8.4.17}$$

Therefore

Rendered by QuickLaTeX.com

Which we round up to a sample size of 2305 to ensure that $E < 0.02$.

□

1. We assume here that there are only two parties. For the real life situation of more than two parties we need the 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. ↵

This page titled 8.4: Proportions and Confidence Intervals for Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 $x_2 = \frac{(n-1)s_2^2}{\sigma}$ 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.

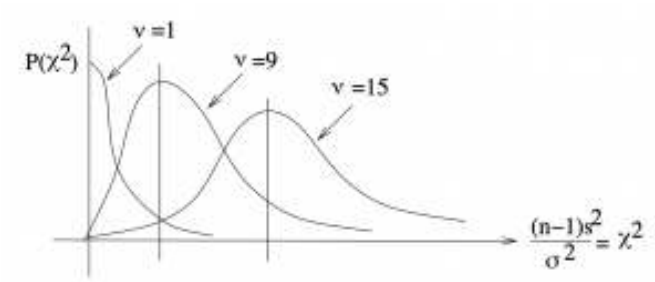


Figure 8.10 : The χ^2 distributions are enumerated by degrees of freedom.

The χ^2 distribution underlies why s is the best estimate for σ. It 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.

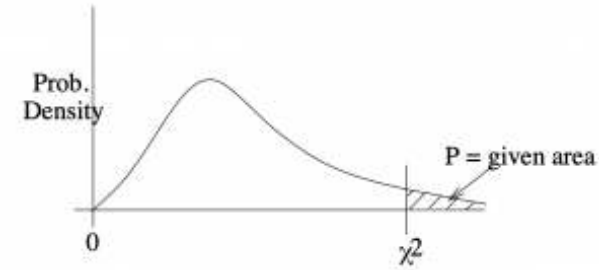


Figure 8.11 : The Chi-squared Distribution Table gives χ^2 associated with given right tail areas.

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 $\mathcal{C}$, see Figure 8.12.

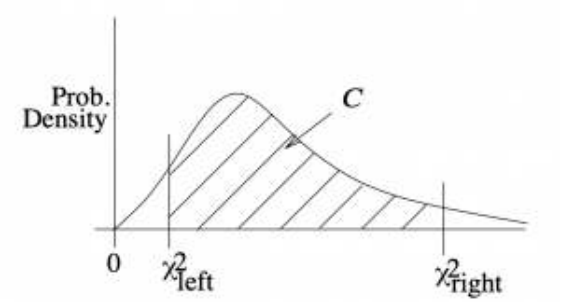


Figure 8.12 : The values χ^2_{left} and χ^2_{right} define the confidence region $\mathcal{C}$.

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**:

$$\chi^2_{\text{right}} = \chi^2\left(\frac{1-\mathcal{C}}{2}\right) \tag{8.5.1}$$

$$\chi^2_{\text{left}} = \chi^2\left(1-\left[\frac{1-\mathcal{C}}{2}\right]\right). \tag{8.5.2}$$

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

$$\frac{(n-1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^{-2}_{\text{left}}} \tag{8.5.3}$$

and for σ by:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{left}}}} \tag{8.5.4}$$

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

$$59, 54, 53, 52, 51, 39, 49, 46, 49, 48 \tag{8.5.5}$$

Solution : Compute, using your calculator :

$$s^2 = 28.2 \tag{8.5.6}$$

$$\nu = n-1 = 9. \tag{8.5.7}$$

From the **Chi-squared Distribution Table**, in the $\nu = 9$ line, find :

$$\chi^2_{\text{right}} = \chi^2\left(\frac{1-0.90}{2}\right) = \chi^2(0.05) = 16.919 \tag{8.5.8}$$

and

$$\chi^2_{\text{left}} = \chi^2(1-0.05) = \chi^2(0.95) = 3.325 \tag{8.5.9}$$

So

$$\begin{aligned} \frac{(n-1)s^2}{\chi^2_{\text{right}}} &< \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{left}}} \\ \frac{9\cdot 28.2}{16.919} &< \sigma^2 < \frac{9\cdot 28.2}{3.325} \\ 15.0 &< \sigma^2 < 76.3 \qquad \text{with } 90\% \text{ confidence.} \end{aligned}$$

Taking square roots:

$$3.87 < \sigma < 8.73 \qquad \text{with } 90\% \text{ confidence.} \tag{8.5.10}$$

□

This page titled 8.5: Chi Squared Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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, the 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 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.

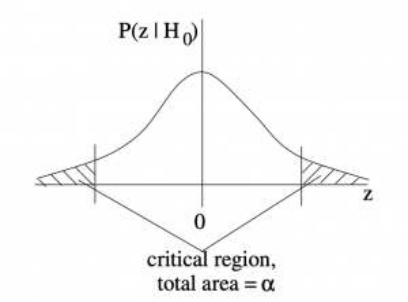
To fix ideas about hypothesis testing, we’ll first look at hypotheses on the means of populations (μ). Later we’ll consider hypotheses on σ and on p (proportions).

With means there are three combinations of H_0 and H_1 to consider :

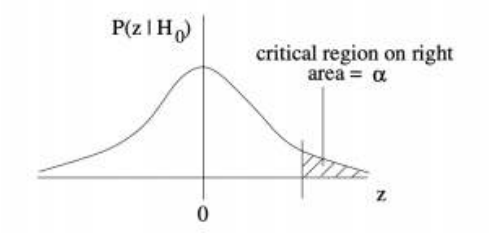
Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0 : \mu = k$	$H_0 : \mu \leq k$	$H_0 : \mu \geq k$
$H_1 : \mu \neq k$	$H_1 : \mu > k$	$H_1 : \mu < k$

Here k is a given number. Not 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 :

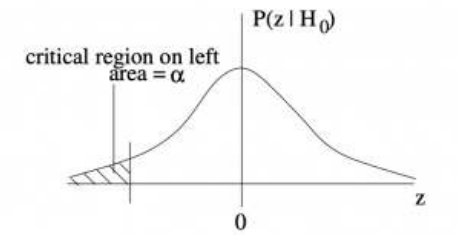
1. Two-tailed test:



2. Right-tailed test:



3. Left-tailed test:



The critical regions, or rejection regions, appear in the probability distributions $P(z \mid H_0)$, 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 $\mathcal{C} = 1 - \alpha$ 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-called[1] "confusion matrix" :

	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.

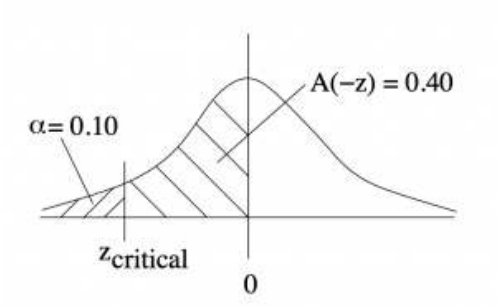
Back to understanding the meaning of hypothesis testing. As we said, a good experimental design will be set up so that H_1 is your favourite theory that there is an effect. In that case H_0 represents the case that there is *no effect* : the position of $\bar{x}$ away from k, or z away from 0 (in the 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* :



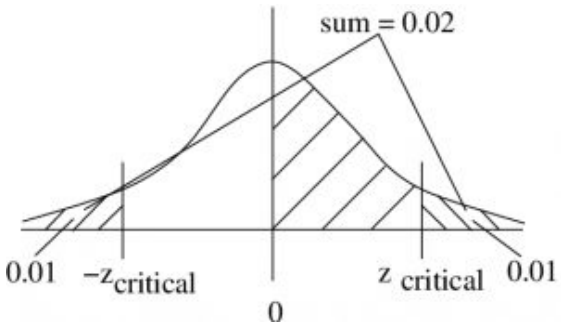
With the tables we have in the Appendix, there are two ways to find z_{critical} :

- 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.

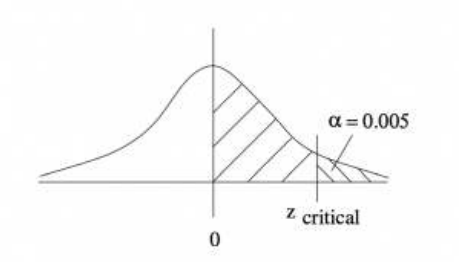
So the critical *areas* are those where

$$z > 2.326 \text{ and } z < -2.326 \tag{9.1}$$

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 $z > 2.576$ 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.

---

1. So called not because it is confusing but because you are never 100% sure which decision is correct. ↵

This page titled 9: Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 {\em 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 :

$$\text{test value} = \frac{(\text{observed value}) - (\text{expected } H_0 \text{ value})}{\text{standard error}}. \tag{9.1.1}$$

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.

This page titled 9.1: Hypothesis Testing Problem Solving Steps is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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
$H_0 : \mu = k$	$H_0 : \mu \leq k$	$H_0 : \mu \geq k$
$H_1 : \mu \neq k$	$H_1 : image$ k" title="Rendered by QuickLaTeX.com" height="16" width="43" style="vertical-align: -4px;">	$H_1 : \mu < k$

In all cases the test statistic is

(9.1) Rendered by QuickLaTeX.com

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 :

   $H_0 : \mu \leq 42,000$

   $image$ 42,000" title="Rendered by QuickLaTeX.com" height="16" width="124" style="vertical-align: -4px;"> (claim)

   (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: $z_{\text{critical}} = 1.645$.

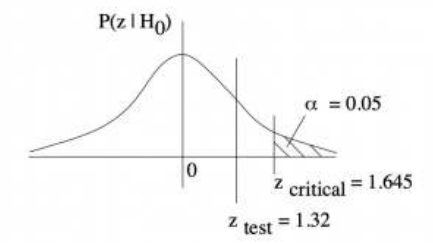
  Method (b) is the recommended method not only because it is faster but also because the procedure for the upcoming t-test will be the same for the z-test.

3. Test Statistic.

$$z_{\text{test}} = \frac{\bar{x} - k}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{43260 - 42000}{\left(\frac{5230}{\sqrt{30}}\right)} = 1.32 \qquad (9.2.1)$$

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.

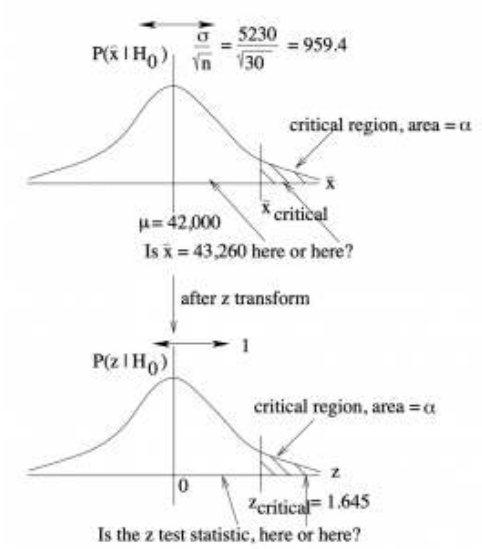


Figure 9.1: Derivation of the z test statistic.

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 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, if 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 tho think about this is to think of $\bar{x}$ as a fuzzy number. As a sharp number a statement like "$image$ k" title="Rendered by QuickLaTeX.com" height="14" width="43" style="vertical-align: -2px;">" 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 "$image$ k" title="Rendered by QuickLaTeX.com" height="14" width="43" style="vertical-align: -2px;">" may not be true when $\bar{x}$ is considered to be a fuzzy number[1]

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 H_0 under all circumstances but we use $\geq$ or $\leq$ is H_0 statements simply because they are the logical opposite of $<$ or $image$" title="Rendered by QuickLaTeX.com" height="12" width="12" style="vertical-align: -2px;"> 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:

$$\text{If } p \leq \alpha \text{ reject } H_0. \tag{9.2.2}$$

If $p > \alpha$ do not reject H_0. \alpha \mbox{ do not reject } H_{0}.\]" title="Rendered by QuickLaTeX.com">

The method of comparing test and critical statistics is the traditional approach, popular before computers because is 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 $z\text{test} = 1.32$ 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.

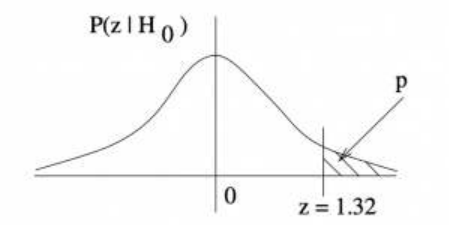


Figure 9.2 : The p-value associated with $z_{\text{test}} = 1.32$ in a one-tail test.

Using the **Standard Normal Distribution Table** to find the tail area associated with $z\text{test} = 1.32$, we compute :

$$\begin{aligned} p(z_{\text{test}}) &= 0.5 - A(z_{\text{test}}) \\ &= 0.5 - 0.4066 = 0.034 \end{aligned}$$

That is $p = 0.0934$. Since $image$ (\alpha = 0.05)" title="Rendered by QuickLaTeX.com" height="19" width="198" style="vertical-align: -5px;">, 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].

Don't try to remember these formula, draw a picture to see what the situation is.

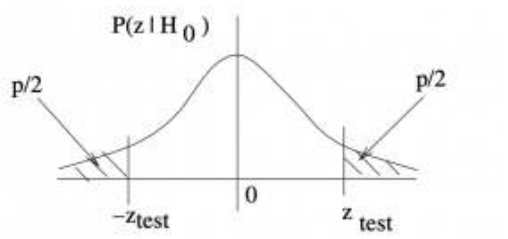


Figure 9.3 : The p-value associated with a two-tailed z_{test}. Since α is defined by, $\pm z_{\text{critical}}$, p is defined by $\pm z_{\text{test}}$.

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 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.

---

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. ↵
2. Of course substitute $-z$ in the formula for a left tail test. ↵

This page titled 9.2: z-Test for a Mean is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 statistic[1]. Specifically, the test statistic is

Rendered by QuickLaTeX.com

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***t* 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.

   $H_0 : \mu \leq 36.7$

   $H_1 : \mu > 36.7$ (claim)

2. Critical statistic.

   In the***t* 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

   $$t_{\text{critical}} = 1.761 \tag{9.3.1}$$

3. Test statistic.

Rendered by QuickLaTeX.com

   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

Rendered by QuickLaTeX.com

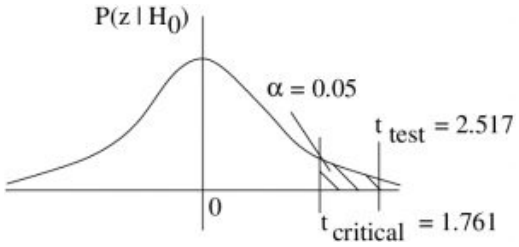
   At this point we can *estimate* the p-value using the***t* Distribution Table**, which doesn't have as much information about the t-distribution as the **Standard Normal Distribution Table** has about the z-distribution, so we can only estimate. The procedure is: In the $\nu = 14$ row, look for t values 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,

   $$0.010 < p < 0.025 \tag{9.3.2}$$

   is our estimate[2] for p.

4. Decision.



   Reject H_0. We can also base this decision on our p-value estimate since :

$$(0.010 < p < 0.025) < (\alpha = 0.05) \tag{9.3.3}$$

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 :

$$\begin{aligned} Z = &\text{ a random variable with a } z\text{-distribution} \\ X = &\text{ a random variable with a } \chi^2 \text{ distribution with } \nu \text{ degrees of freedom.} \end{aligned}$$

Then the random variable

$$T = \frac{Z}{X} \tag{9.3.4}$$

is a random variable that follows a t-distribution with ν degrees of freedom.

---

1. Again, SPSS applies the t-test, uses s directly, for any sample size. ↵
2. If you know how to interpolate then you can find a single value for p. ↵

This page titled 9.3: t-Test for Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$\text{test value } = \frac{\text{(observed value)-(expected } H_0 \text{ value)}}{\text{standard error}}. \tag{9.4.1}$$

but now the observed and expected values are proportions, $\hat{p}$ and p respectively. The standard error in this case is

$$\sqrt{\frac{pq}{n}} = \frac{\sigma_{binomial}}{n} = \frac{\sqrt{npq}}{n} \tag{9.4.2}$$

Using this information with the generic form, which mimics a t test statistic, the proportions test statistic is

$$z_{test} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \tag{9.4.3}$$

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 {\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.

$$H_0 : p \leq 0.25 \text{ , } H_1 : p > 0.25 \text{ (claim)}$$

2. Critical statistic.

Using the***t* Distribution Table** (last line) for a one tailed test at $\alpha = 0.05$ we find $z_{\text{critical}} = 1.645$

3. Test statistic.

$$z_{\text{test}} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \tag{9.4.4}$$

So using

$$\hat{p} = \frac{63}{200} = 0.315 \quad p = 0.25 \tag{9.4.5}$$

$$q = 1 - 0.25 = 0.75 \quad n = 200 \tag{9.4.6}$$

find

$$z_{\text{test}} = \frac{0.35 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{200}}} = 2.12. \tag{9.4.7}$$

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.

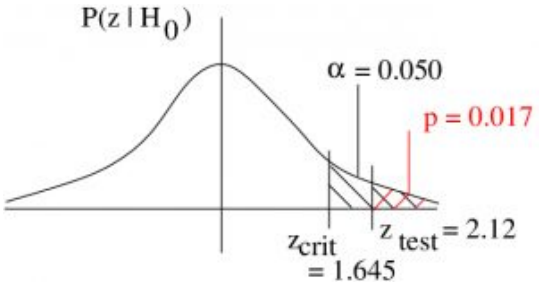


Figure 9.4 : The null hypothesis situation for Example 9.5

We come, of course, to the same decision by considering the p-value :

$$(p = 0.017) < (\alpha = 0.5) \tag{9.4.8}$$

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.

□

This page titled 9.4: z-Test for Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$\chi^2_{\text{test}} = \frac{(n-1)s^2}{\sigma^2} \tag{9.5.1}$$

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.

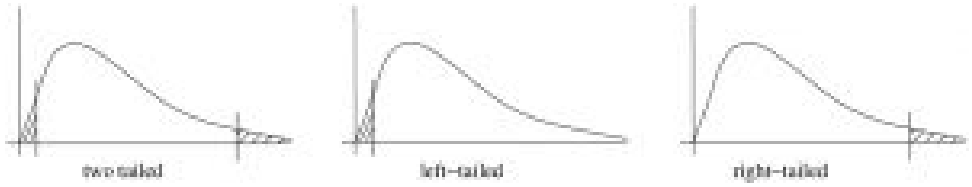


Figure 9.5 : Schematics of the critical regions for χ^2 tests of variance. In the two-tailed situation the tail areas are equal.

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.

$$H_0 : \sigma^2 \geq 225 \quad H_1 : \sigma^2 < 225 \tag{9.5.2}$$

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 $\chi^2_{\text{crit}} = 12.338$.

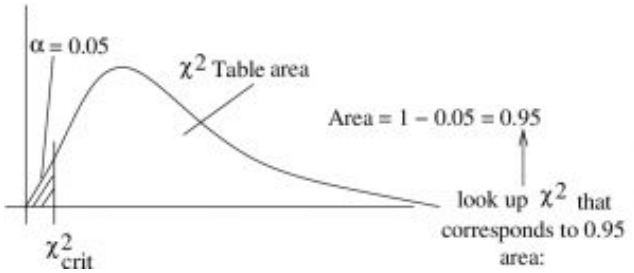


Figure 9.6 : Schematics of the critical regions for χ^2 tests of variance. In the two-tailed situation the tail areas are equal.

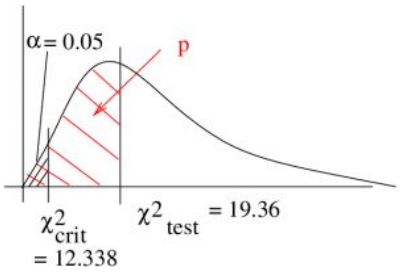
3. Test statistic.

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:

$$(0.10 < p < 0.90) > (\alpha = 0.05) \tag{9.5.3}$$

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.

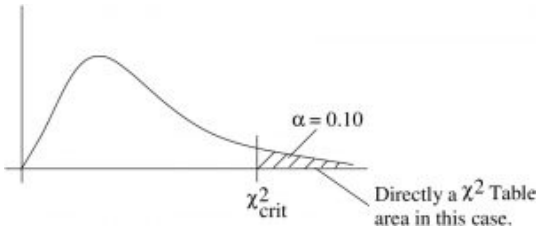
We'll introduce a step 0 when it looks like we should do some preliminary calculations with or data. In this case we should enter the dataset into our calculations and determine s. We find $s = 11.2$.

1. Hypotheses.

$$H_0 : \sigma^2 \leq 64 \quad H_1 : \sigma^2 > 64 \text{ (claim)} \tag{9.5.4}$$

Note conversion to σ^2 right away.

2. Critical statistic.



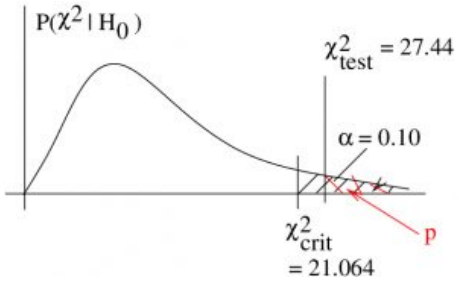
In the $\nu = 15 - 1 = 14$ line and $\alpha_T = 0.10$ column of the **Chi-squared Distribution Table**, look up $\chi^2_{\text{crit}} = 21.064$

3. Test statistic.

$$\chi^2_{\text{test}} = \frac{(n-1)s^2}{\sigma^2}$$
$$\chi^2_{\text{test}} = \frac{(14)(11.2)^2}{64} = 27.44$$

To estimate the p value, find the bracketing values of $\chi^2_{\text{test}} = 27.44$ 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 :

$$(0.010 < p < 0.025) < (\alpha = 0.10) \tag{9.5.5}$$

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.

$$H_0 : \sigma^2 = 0.644 \text{ (claim)} \quad H_1 : \sigma^2 \neq 0.64 \tag{9.5.6}$$

2. Critical statistic.

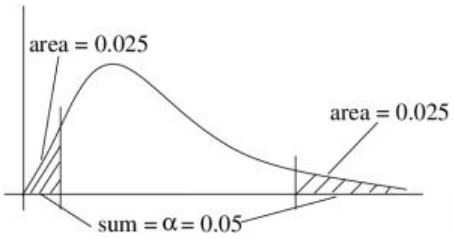


Figure 9.7 : Critical regions for a two tailed test.

Referring to Figure 9.7, we see that we need two χ^2_{crit} 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 $\chi^2_{\text{crit}} = 8.907$ from the $\alpha_T = 0.975$ column and $\chi^2_{\text{crit}} = 32.852$ from the $\alpha_T = 0.025$ column.

3. Test statistic.

$$\chi^2_{\text{test}} = \frac{(n-1)s^2}{\sigma^2}$$
$$\chi^2_{\text{test}} = \frac{(19)(1^2)}{(0.644)} = 29.50$$

To estimate the p value find the bracketing value of $\chi^2_{\text{test}} = 29.50$ 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 $0.10 < p < 0.20$.

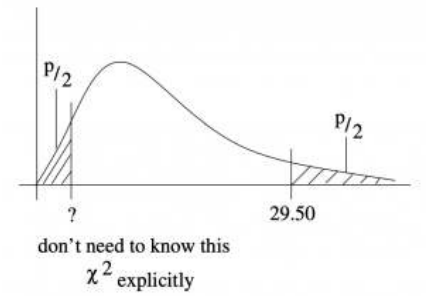
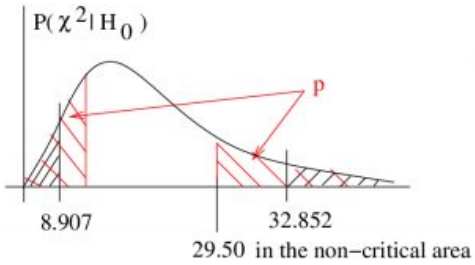


Figure 9.8 : Areas for p associated with the test statistic (29.50 here) in a two tail test.

4. Decision.



Do not reject H_0. The estimate p value leads to the same conclusion :

$$(0.10 < p < 0.20) > (\alpha = 0.05) \tag{9.5.7}$$

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.

This page titled 9.5: Chi Squared Test for Variance or Standard Deviation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

9.6: SPSS Lesson 5- Single Sample t-Test

Open “HeadCircum.sav” from the textbook Data Sets:



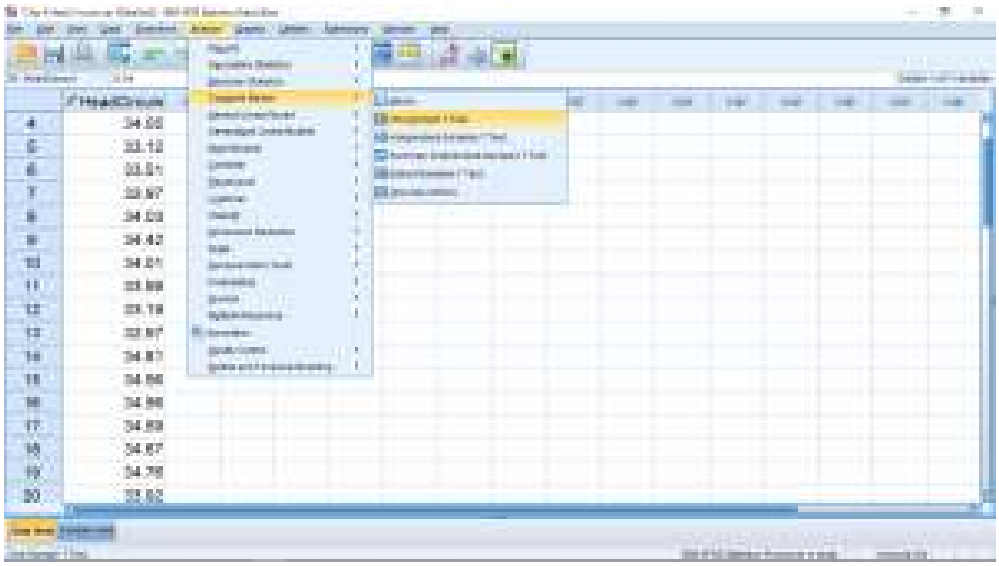
SPSS screenshot © International Business Machines Corporation.

Look at how simple it is! One variable. This is our single sample. Let’s do a t-test for the hypotheses:

$$H_0 : \mu = 33.8 \tag{9.6.1}$$

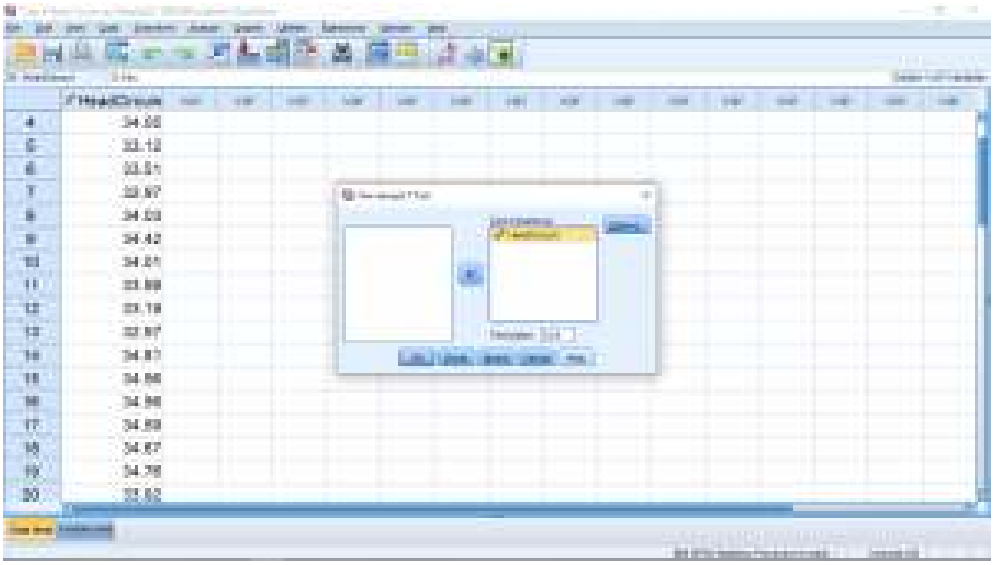
$$H_1 : \mu \neq 33.8$$

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 $\rightarrow$ Compare Means $\rightarrow$ One-Sample T Test:



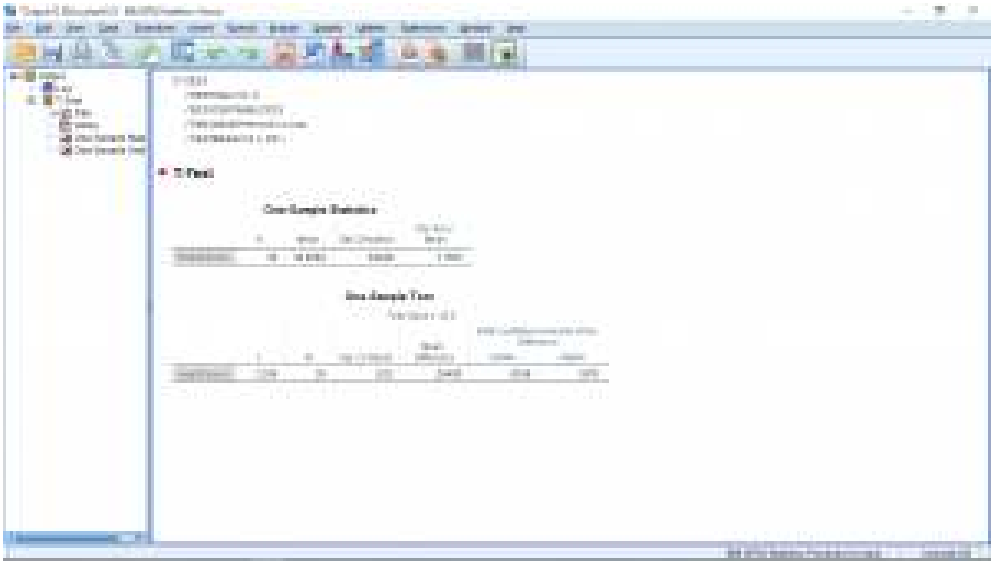
SPSS screenshot © International Business Machines Corporation.

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, $\overline{x}$ 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. Te columns give: t_{test}, ν, p and $\overline{x} - k$. Notice that the first column, t_{test} is the fourth column $\overline{x} - k$ divided by the last column of the first table, $s/\sqrt{n}$. The last two columns give the 95% confidence interval

$$0.0249 < \mu - 33.80 < 0.5031$$

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:

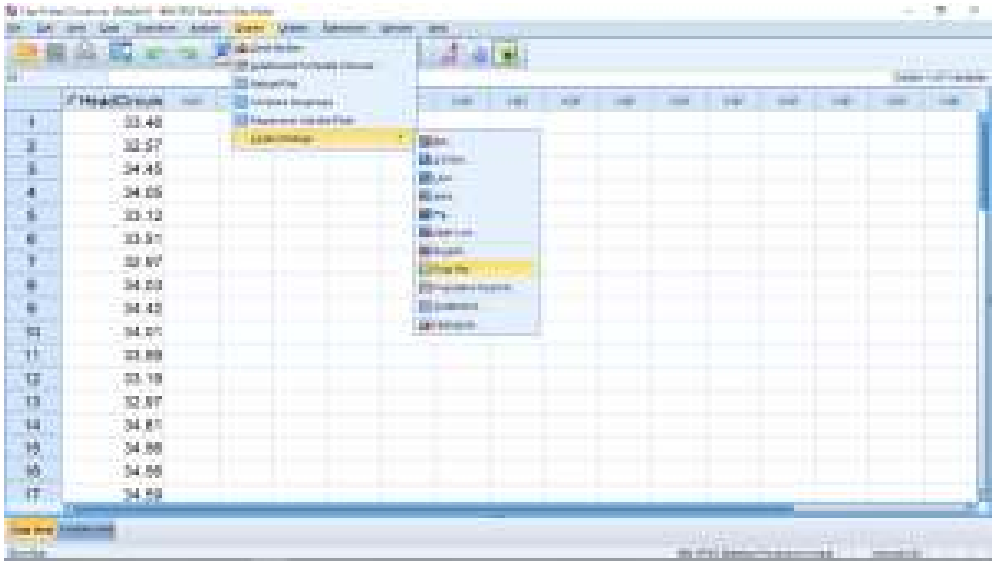
$$33.8249 < \mu < 34.3031$$

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

$$d = \frac{\overline{x} - k}{s} = \frac{t}{\sqrt{n}}$$

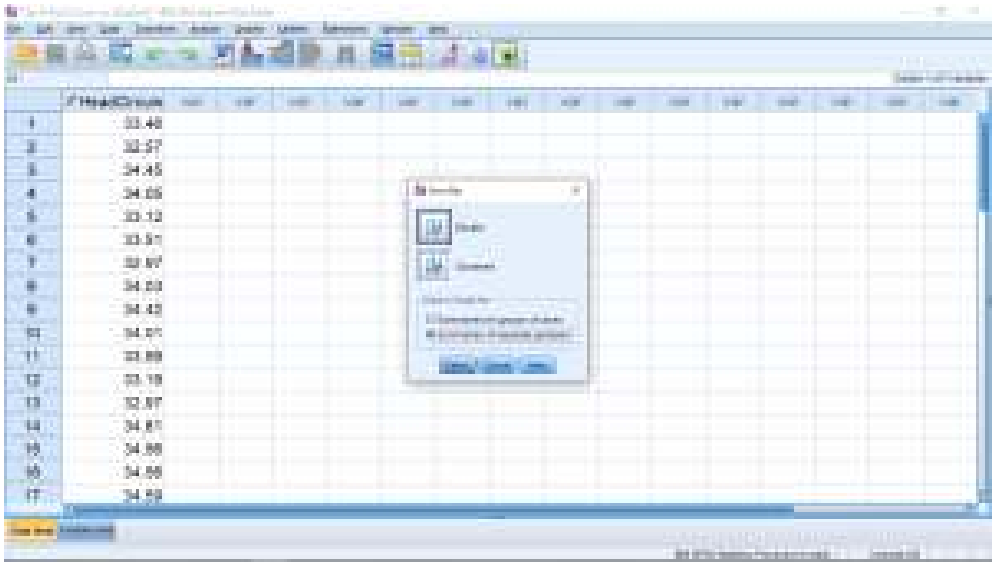
where, by t we mean t_{test}. Being a descriptive statistic, people use the following rule of thumb to describe d. If d is approximately 0.2 then d is considered “small”; if d is approximately 0.5 then d is considered “medium”; d is approximately 0.8 then d is considered “large”.

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 → Legacy Dialogs → Error Bar:



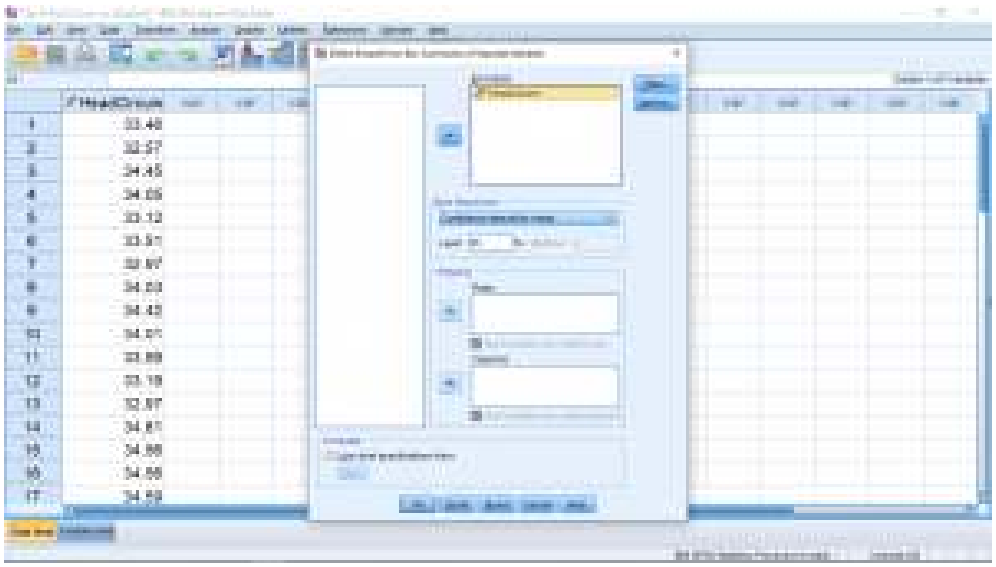
SPSS screenshot © International Business Machines Corporation.

Choose Simple and “Summaries of separate variables”:



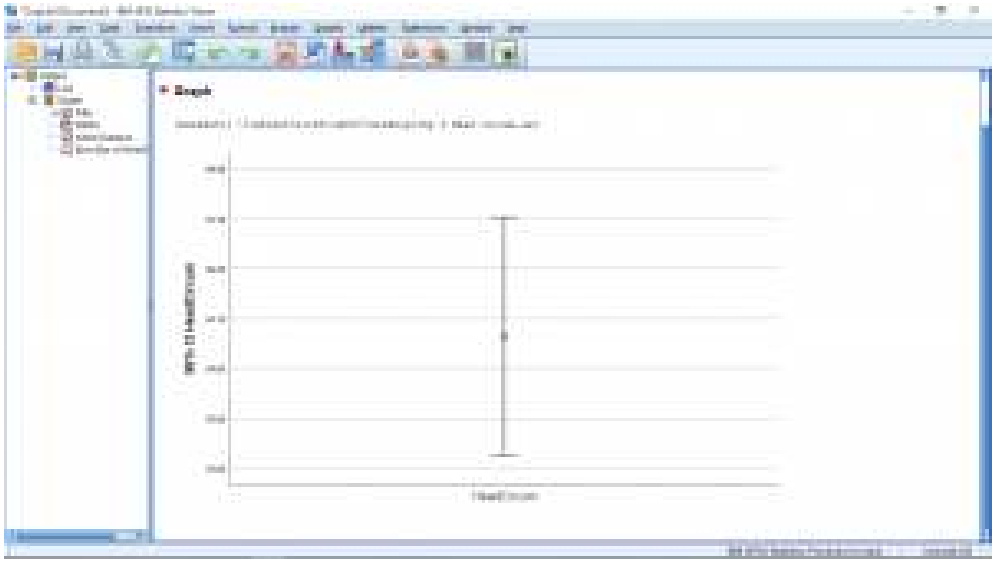
SPSS screenshot © International Business Machines Corporation.

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 $\overline{x} \pm \frac{s}{\sqrt{n}}$:

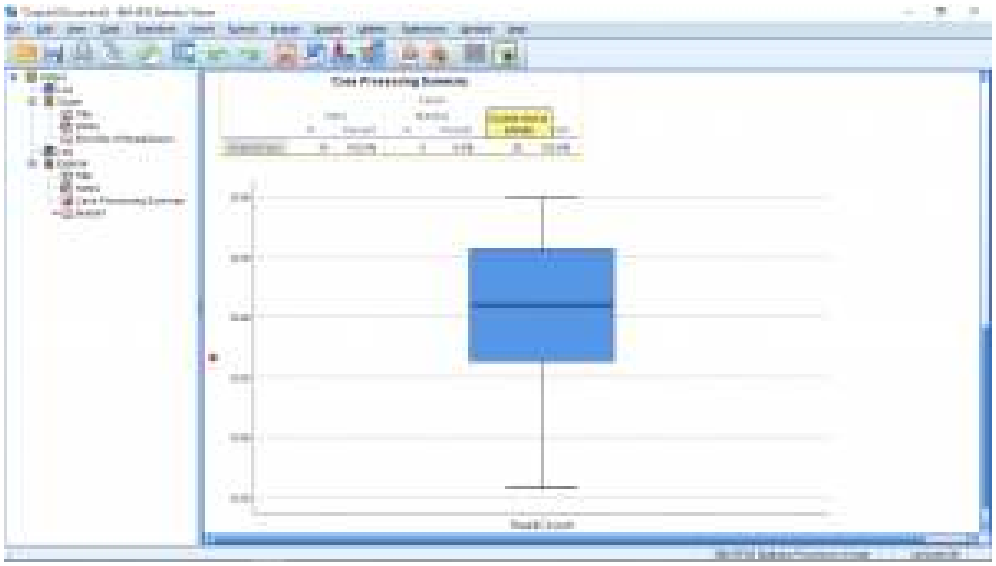


SPSS screenshot © International Business Machines Corporation.

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.

This page titled 9.6: SPSS Lesson 5- Single Sample t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

This page titled 10: Comparing Two Population Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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
$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 \leq$	$H_0 : \mu_1 - \mu_2 \geq 0$
$H_1 : \mu_1 - \mu_2 \neq 0$	$H_1 : \mu_1 - \mu_2 > 0$	$H_1 : \mu_1 - \mu_2 < 0$

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 cases[1] is

$$z_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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.

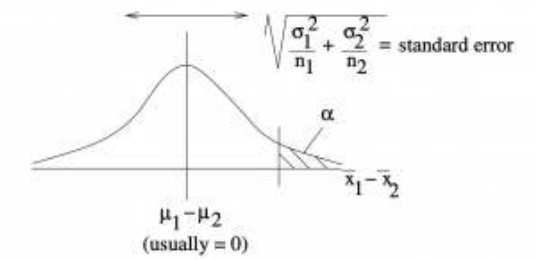


Figure 10.1 : The distribution of the difference of sample means $\bar{x}_1 - \bar{x}_2$ under the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$. A one-tail example is shown here. The test statistic of Equation 10.1 follows from a z-transformation of this picture.

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 :

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.

$$H_0 : \mu_1 \leq \mu_2 \qquad H_1 : \mu_1 > \mu_2 \text{ (claim)} \tag{10.1.1}$$

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

$$z_{\text{crit}} = 1.282 \tag{10.1.2}$$

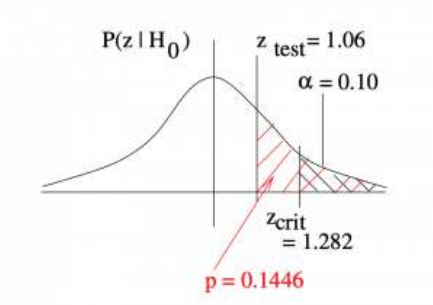
Note that $z_{critical}$ is positive because this is a right-tailed test. For left tailed tests make z_{crit} negative. For two-tailed tests you have $\pm z_{\text{crit}}$.

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 :

$$(p = 0.1446) > (\alpha = 0.10) \tag{10.1.3}$$

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.

□

1. You could specify a non-zero null hypothesis, e.g. $H_0 : \mu_1 - \mu_2 = k$, in which case you would have $z_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2) - k}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. We won't consider that case in this course. ↵

This page titled 10.1: Unpaired z-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \tag{10.2.1}$$

where

$$E = z_{\mathcal{C}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.2.2}$$

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:

$$\bar{x}_1 - \bar{x}_2 = 88.42 - 80.61 = 7.81 \tag{10.2.3}$$

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

$$7.81 - 2.05 < (\mu_1 - \mu_2) < 7.81 + 2.05 \tag{10.2.4}$$

or

$$5.76 < (\mu_1 - \mu_2) < 9.83 \tag{10.2.5}$$

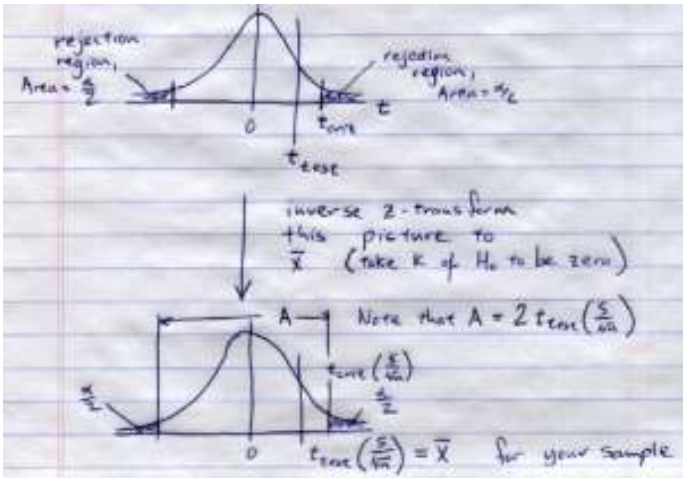
with 95% confidence. Notice that it is also correct to write $\mu_1 - \mu_2 = 7.81 \pm 2.05$ with 95% confidence.

□

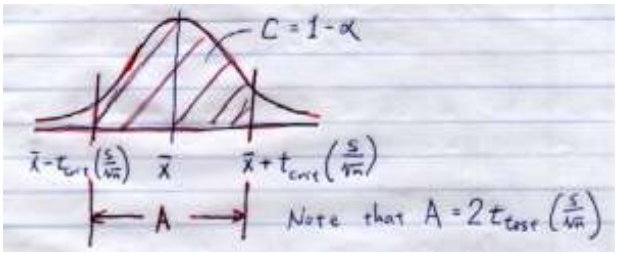
This is a good point to make an important observation. A *two-tailed* hypothesis test at a given α is complementary to a confidence interval of $\mathcal{C} = 1 - \alpha$ in the sense that if 0 is in the confidence interval then the complementary hypothesis test will not reject H_0.

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 :

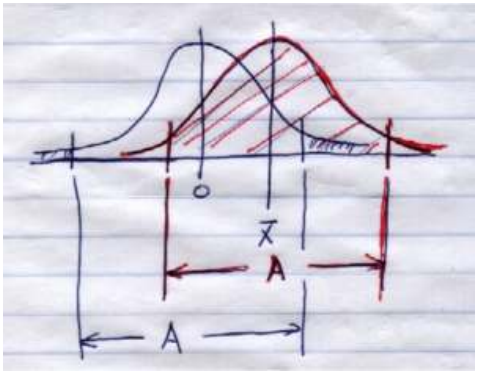
Case 1 : 0 in the confidence interval, fail to reject H_0. In the hypothesis test you would find :



In the confidence interval calculation you would find:

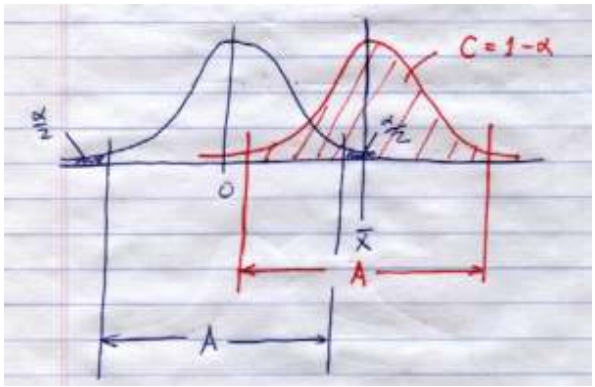


Putting the two pictures together gives:



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.

This page titled 10.2: Confidence Interval for Difference of Means (Large Samples) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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.

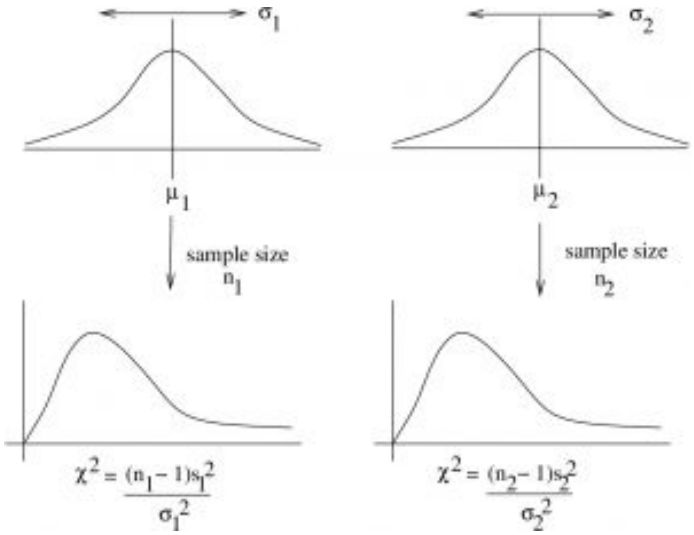


Figure 10.2: Two normal populations lead to two χ^2 distributions that represent distributions of sample variances. The F distribution results when you build up a distribution of the ratio of the two χ^2 sample values.

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.

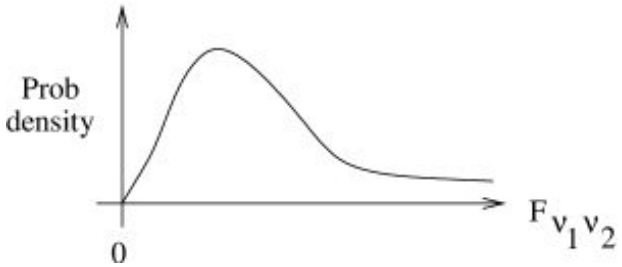


Figure 10.3: A generic F distribution.

F and t are related :

$$F_{1,\nu} = t_\nu^2 \tag{10.3.1}$$

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
$H_0 : \sigma_1^2 \leq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 > \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 \neq \sigma_2^2$

We'll always compare variances (σ^2) and not standard deviations (σ) to keep life simple.

The test statistic is

$$F_{\text{test}} = F_{\nu_1,\nu_2} = \frac{s_1^2}{s_2^2} \tag{10.3.2}$$

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 :

$$F_{\text{crit}} = F_{\alpha,\ \nu_1,\ \nu_2} \tag{10.3.3}$$

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 $\sigma_1^2 > \sigma_2^2$. 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 $s_1^2 > s_2^2$ 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:

$$\text{d.f.N.} = \nu_1 = n_1 - 1 = 26 - 1 = 25 \tag{10.3.4}$$

$$\text{d.f.D.} = \nu_2 = n_2 - 1 = 18 - 1 = 17 \tag{10.3.5}$$

So the critical statistic is

Rendered by QuickLaTeX.com

3. Test statistic.

$$F_{\nu_1,\nu_2} = \frac{s_1^2}{s_2^2} \tag{10.3.6}$$

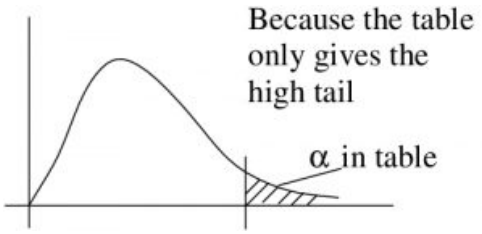
$$F_{\text{test}} = F_{25,17} = \frac{36}{10} = 3.6 \tag{10.3.7}$$

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 $p = \alpha_T$.)

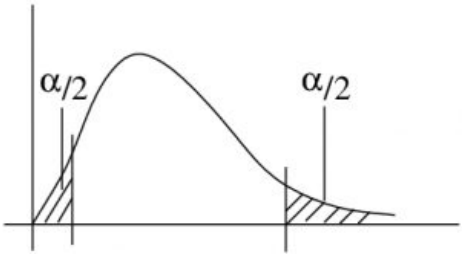
p α_T	0.20 0.10 0.05 0.02 0.01 0.10 0.05 0.025 0.01 0.005
F	1.84 2.19 2.56 3.08 3.51 3.6 is over here somewhere so $p < 0.01$

Notice how we put an upper limit on p because F_{test} was larger than all the F values in our little table.

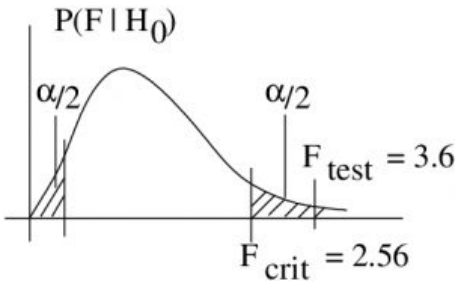
Let's take a graphical look at why we use $p = 2\alpha$ in the little table and $\alpha_T = \alpha/2$ for finding F_{crit} for two tailed tests :



But in a two-tailed test we want α split on both sides:



4. Decision.



Reject H_0. The p-value estimate supports this :

$$(p < 0.01) < (\alpha = 0.05) \tag{10.3.8}$$

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.

□

This page titled 10.3: Difference between Two Variances - the F Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

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 :

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} \tag{10.4.1}$$

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

$$\nu = \min(n_1 - 1, n_2 - 1).$$

This choice is a conservative approach (harder to reject H_0). SPSS uses a more accurate

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{s_1^2}{n_1}\right)}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)}{n_2 - 1}\right]}$$

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 know 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:

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10.4.2}$$

This test statistic formula can be made more intuitive by defining

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

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 $\nu_2 = n_1 - 1$; their sum is $\nu_1 + \nu_2 = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$. In other words

$$s_p^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} \tag{10.4.3}$$

and we can write the test statistic as

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

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 :

$$t_{\text{test}} = \frac{\text{Difference of means}}{\text{Standard error of the mean}} = \frac{\text{Signal}}{\text{Noise}}. \tag{10.4.4}$$

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 :

$s_1 = 38$	$\overline{x}_1 = 191$	$n_1 = 8$
$s_2 = 12$	$\overline{x}_2 = 199$	$n_2 = 10$

(Note that $image$ (s_{2}=12)" title="Rendered by QuickLaTeX.com" height="19" width="164" style="vertical-align: -5px;">. If that wasn't true, we could reverse the definitions of populations 1 and 2 so that $image$ 1" title="Rendered by QuickLaTeX.com" height="15" width="66" style="vertical-align: -3px;">.) Is $\overline{x}_1$ significantly different from $\overline{x}_2$? That is, is μ_1 different from μ_2? Test at $\alpha = 0.05$.

Solution :

So the question is to decide between

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2 \tag{10.4.5}$$

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.

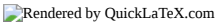
Test 1 : See if variances can be assumed equal or not.

1. Hypothesis.

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{10.4.6}$$

(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.



(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.

$$F_{\text{test}} = F_{7,9} = \frac{s_1^2}{s_2^2} = \frac{38^2}{12^2} = 10.03 \tag{10.4.7}$$

4. Decision.

$image$ 4.20" title="Rendered by QuickLaTeX.com" height="14" width="95" style="vertical-align: -2px;"> ($image$ F_{\rm crit}" title="Rendered by QuickLaTeX.com" height="15" width="89" style="vertical-align: -3px;"> — 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.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2 \tag{10.4.8}$$

2. Critical statistic.

From the ***t* Distribution Table**, with $\nu = \min(n_1 - 1, n_2 - 1) = \min(8 - 1, 10 - 1) \min(7, 9) = 7$, and a two-tailed test with $\alpha = 0.05$ we find

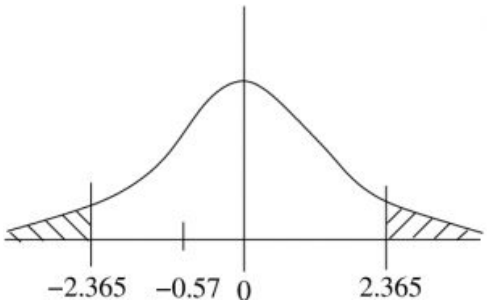
$$t_{\text{crit}} = \pm 2.365 \tag{10.4.9}$$

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 $image$ 0.50" title="Rendered by QuickLaTeX.com" height="17" width="66" style="vertical-align: -4px;">.

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 $image$ s_{2}^{2}" title="Rendered by QuickLaTeX.com" height="20" width="55" style="vertical-align: -5px;">. 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.

$$H_0 : \mu_1 \leq \mu_2 \tag{10.4.10}$$

$H_1 : \mu_1 > \mu_2$ \mu_{2}\]" title="Rendered by QuickLaTeX.com">

(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

$$t_{\text{crit}} = 2.583 \tag{10.4.11}$$

3. Test statistic.

$$t_{\text{test}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$t_{\text{test}} = \frac{(26,800 - 25,400)}{\sqrt{\frac{(10-1)600^2+(8-1)450^2}{10+8-2}}\sqrt{\frac{1}{10}+\frac{1}{8}}}$$

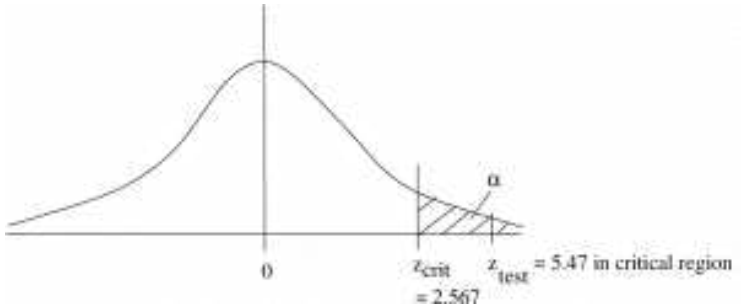
$$t_{\text{test}} = \frac{1400}{\sqrt{\frac{(9)(360000)+(7)(202500)}{16}}\sqrt{0.1+0.125}}$$

$$t_{\text{test}} = \frac{1400}{\sqrt{\frac{3240000+1417500}{16}}\sqrt{0.225}}$$

$$t_{\text{test}} = \frac{1400}{(\sqrt{291093.75})(\sqrt{0.225})} = 5.47$$

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 $(p < 0.005) < (\alpha = 0.01)$.

$t_{test} > t_{crit}$ $(5.47 > 2.583)$ t_{crit} \hspace{.25in} (5.47 > 2.583)\]" 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.

□

This page titled 10.4: Unpaired or Independent Sample t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

10.5: Confidence Intervals for the Difference of Two Means

The form of the confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \tag{10.5.1}$$

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}

$$E = t_{\mathcal{C}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.5.2}$$

where the degrees of freedom to use when looking up $t_{\mathcal{C}}$ in the ***t* Distribution Table** is

$$\nu = \min[(n_1 - 1), (n_2 - 1)] \tag{10.5.3}$$

Case 2 : Variances of the 2 populations equal

$$E = t_{\mathcal{C}} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{10.5.4}$$

where we use

$$\nu = n_1 + n_2 - 2 \tag{10.5.5}$$

when looking up $t_{\mathcal{C}}$.

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 $t_{\mathcal{C}}$ in the ***t* Distribution Table** for 95% confidence interval for $\nu = 7$:

$$t_{95\%} = 2.365 \tag{10.5.6}$$

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

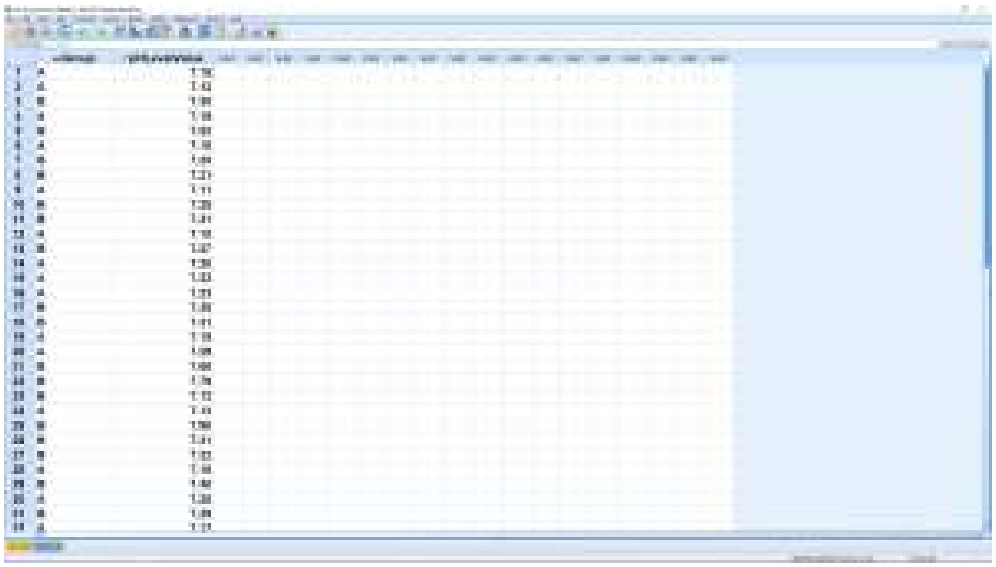
$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 &< (\bar{x}_1 - \bar{x}_2) + E \\ (191 - 199) - 33.02 < \mu_1 - \mu_2 &< (191 - 199) + 33.02 \\ -8 - 33.02 < \mu_1 - \mu_2 &< -8 + 33.02 \\ -41.02 < \mu_1 - \mu_2 &< 25.02 \end{aligned}$$

be careful of the order!

This page titled 10.5: Confidence Intervals for the Difference of Two Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

10.6: SPSS Lesson 6- Independent Sample t-Test

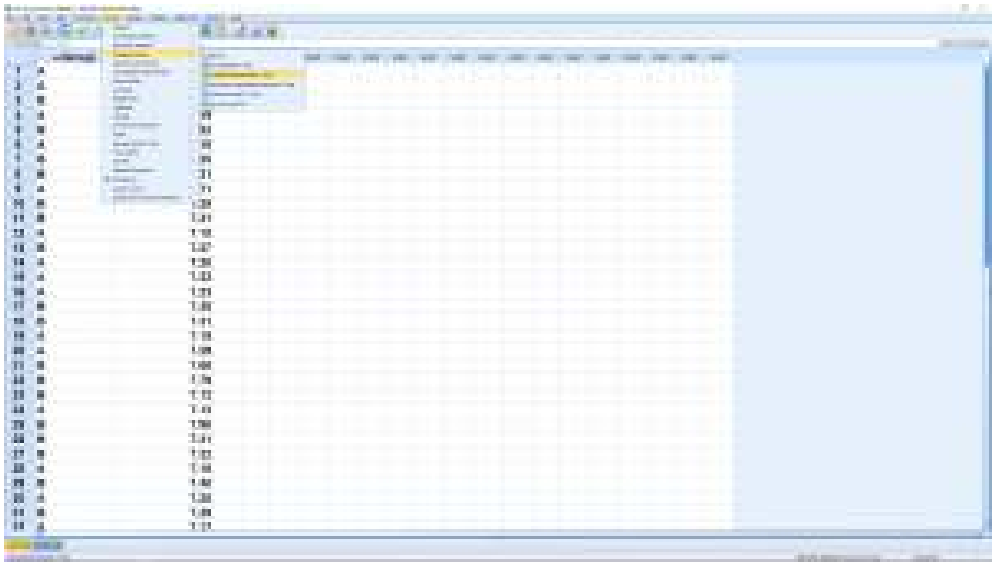
To follow along, load in the Data Set titled “pHLevel.sav”:



SPSS screenshot © International Business Machines Corporation.

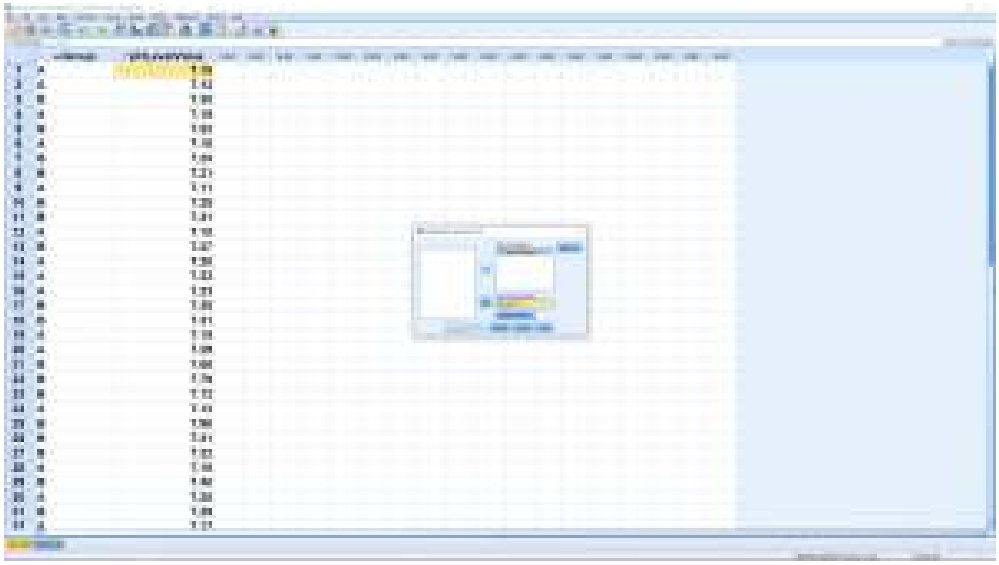
This is the first time we have an independent variable, Species in this case, and it has two values, setosa and versicolor, that label the two populations. Notice, especially, that we do not have separate columns for each sample. There is only one dependent variable, Sepal.Length in this case. As we cover the more advanced statistical tests in later chapters (part of PSY 234) the nature and complexity of the independent variable will evolve but we will always have just one independent variable.

Running the *t*-test is easy, pick Analyze → Compare Means → Independent Samples T Test :



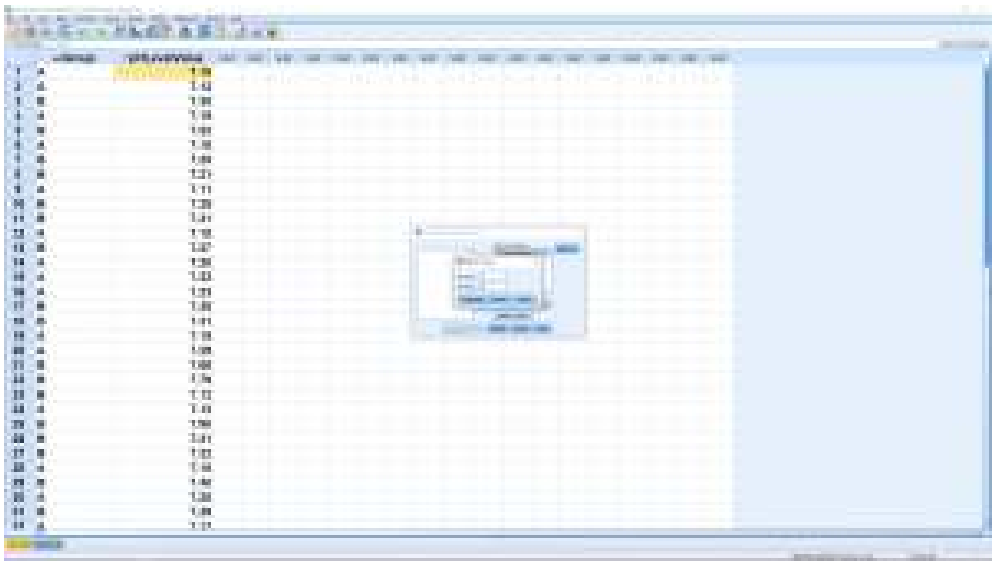
SPSS screenshot © International Business Machines Corporation.

Select Sepal.Length as the Test Variable (dependent variable) and Species as the group variable (independent variable) :



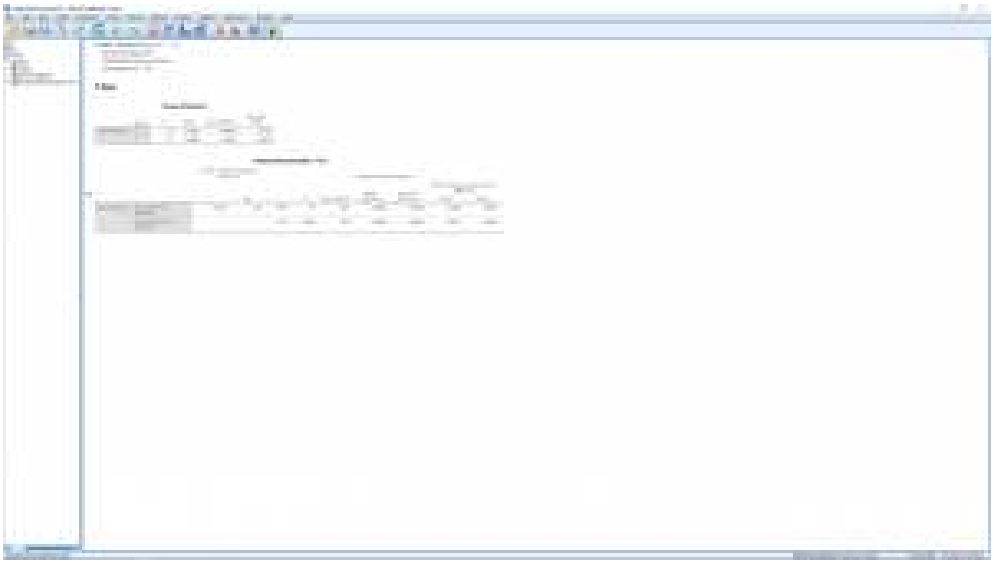
SPSS screenshot © International Business Machines Corporation.

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:

$$H_0 : \mu_1 - \mu_2 = 0 \tag{10.6.1}$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

This, as you recall, is our test of interest. When we did this test by hand, we had to do a preliminary F test the see if we could assume homoscedasticity or not :

Misplaced &

Misplaced &

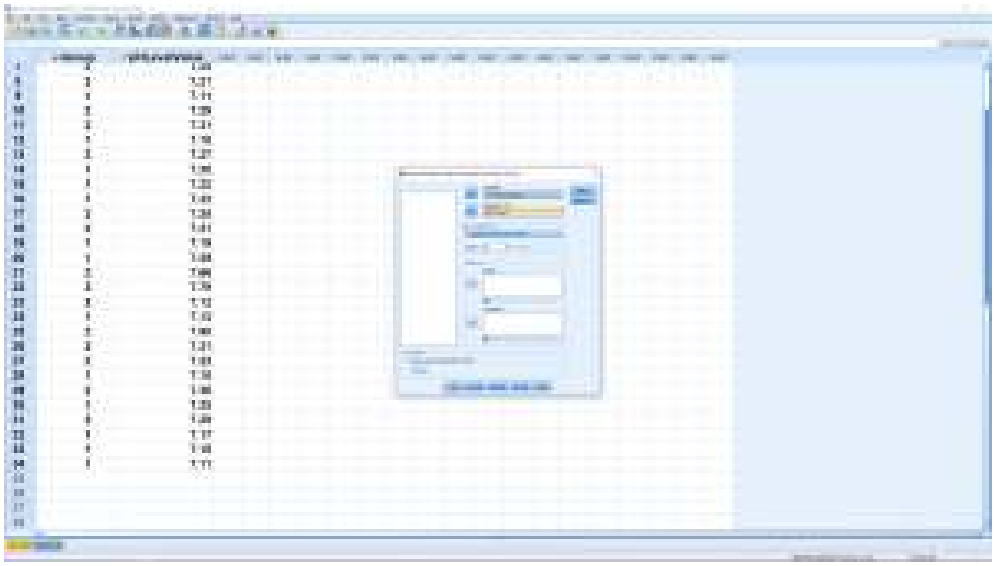
That preliminary test is given to us as Levine's test in the first two columns of the second table. Levine's test is similar to but not exactly the same as the F test we used but it also uses F as a test statistic. Here we see $F_{\text{test}} = 12.061$ with $p = 0.001$, so we reject H_0 and assume that 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), $\overline{x}_1 - \overline{x}_2 = -0.208$, 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

$$0.75429 < \mu_1 - \mu_2 < 1.10571$$

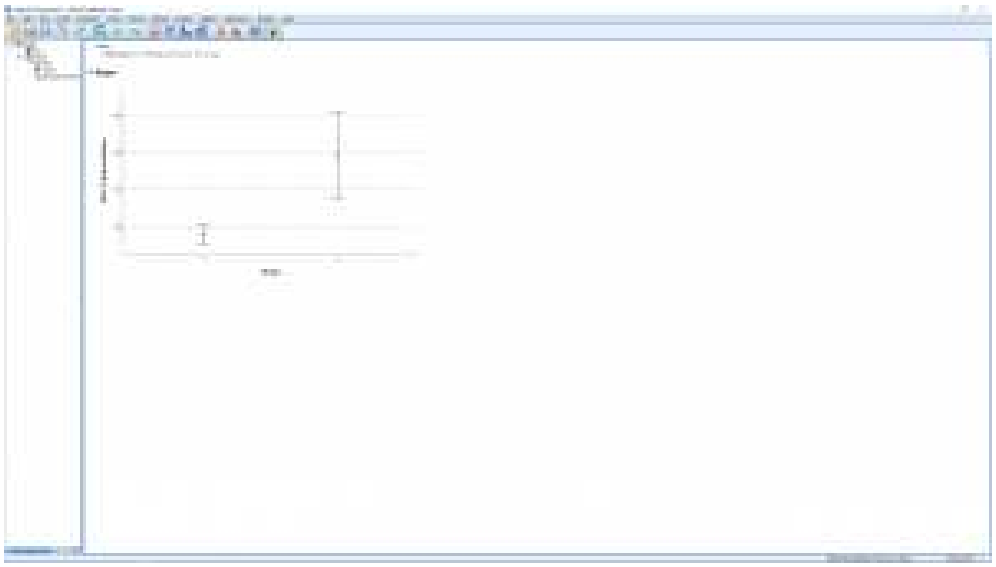
Notice that zero is not in the confidence interval, consistent with rejecting H_0.

We can also make an error bar plot. Go through Graphs → Legacy Dialogs → Errorbar and pick Simple and "Summaries for groups of cases" in the next menu and:



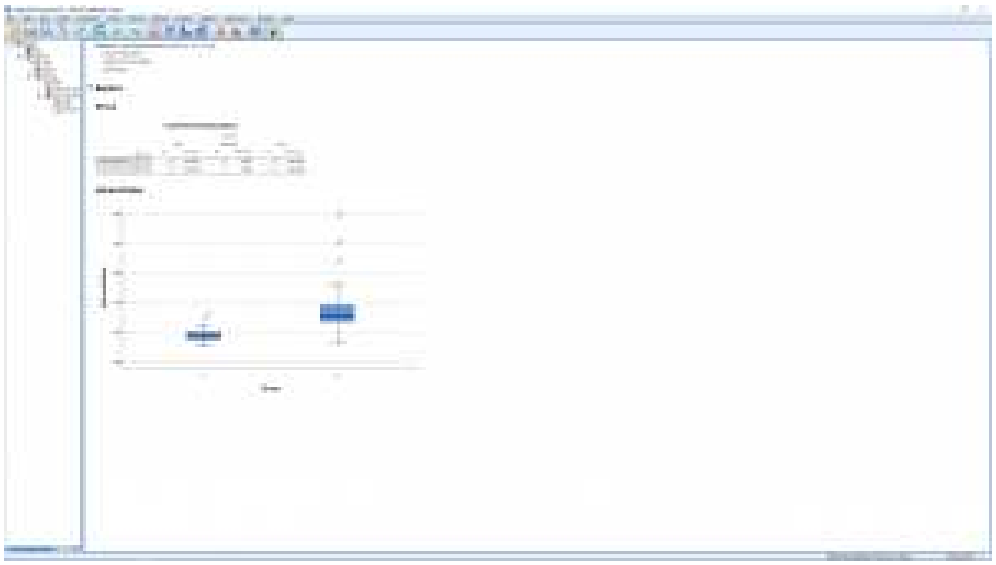
SPSS screenshot © International Business Machines Corporation.

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:

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = \frac{\overline{x}_1 - \overline{x}_2}{s_p}$$

where s_p is the pooled variance as given by Equation (10.4). Another measure is the strength of association

$$\eta^2 = \frac{t^2}{t^2 + (n_1 + n_2 - 2)}$$

which measures a kind of "correlation'" between x_1 and x_2. The larger t, the closer η^2 is to 1.

This page titled 10.6: SPSS Lesson 6- Independent Sample t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

10.7: 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 variable[1]. 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 :

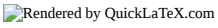
x_1	x_2	$D = x_1 - x_2$
1	2	-1
2	3	-1
3	5	-2
1	-2	3

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. 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



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 $\overline{D} = -2.375$, $s_D = 4.84$ by entering the difference data into your calculator.

1. Hypothesis.

$$H_0 : \mu_D \geq 0 \tag{10.7.1}$$

$$H_1 : \mu_D < 0 \tag{10.7.2}$$

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

$$t\text{crit} = -1.895 \tag{10.7.3}$$

(We added the negative sign because this is a left-tailed test.)

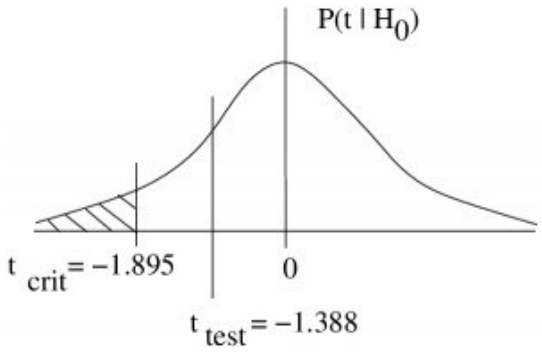
3. Test statistic.

$$t_{test} = \frac{\bar{D} - \mu_D}{\left(\frac{s_D}{\sqrt{n}}\right)} = \frac{-2.375 - 0}{\left(\frac{4.84}{\sqrt{8}}\right)} \tag{10.7.4}$$

$$t_{test} = -1.388 \tag{10.7.5}$$

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 $\alpha = 0.05$, the is not enough evidence to conclude that the vitamin increases strength.

□

1. An introduction to vectors will be given in Chapter 17. ↵

This page titled 10.7: Paired t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

10.8: Confidence Intervals for Paired t-Tests

The usual form applies :

$$\overline{D} - E < \mu_D < \overline{D} + E \tag{10.8.1}$$

where now

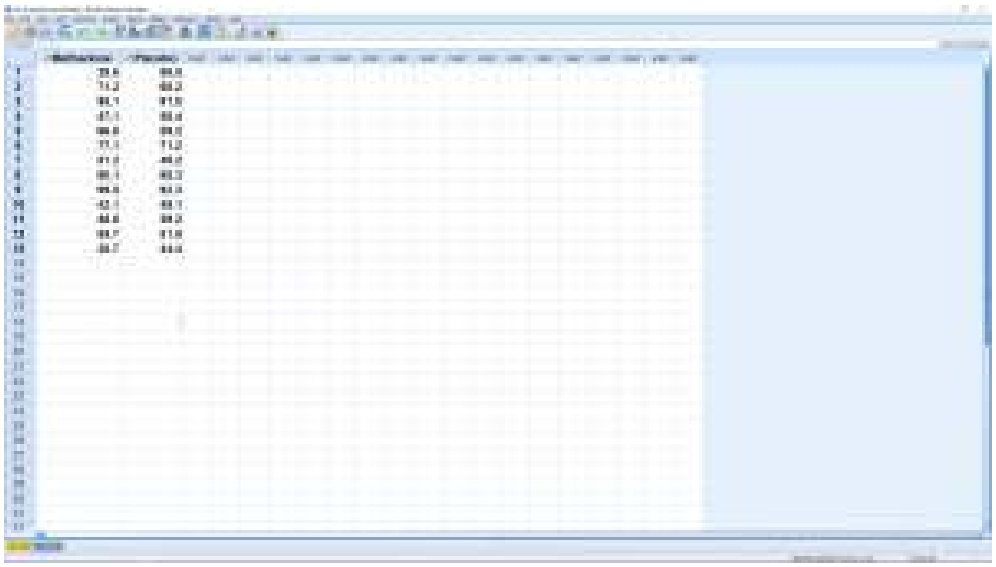
$$E = t_{\mathcal{C}} \left(\frac{s_D}{\sqrt{n}} \right) \tag{10.8.2}$$

and $t_{\mathcal{C}}$ can found from the ***t* Distribution Table** in the $\nu = n - 1$ line using the “confidence intervals” heading.

This page titled 10.8: Confidence Intervals for Paired t-Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

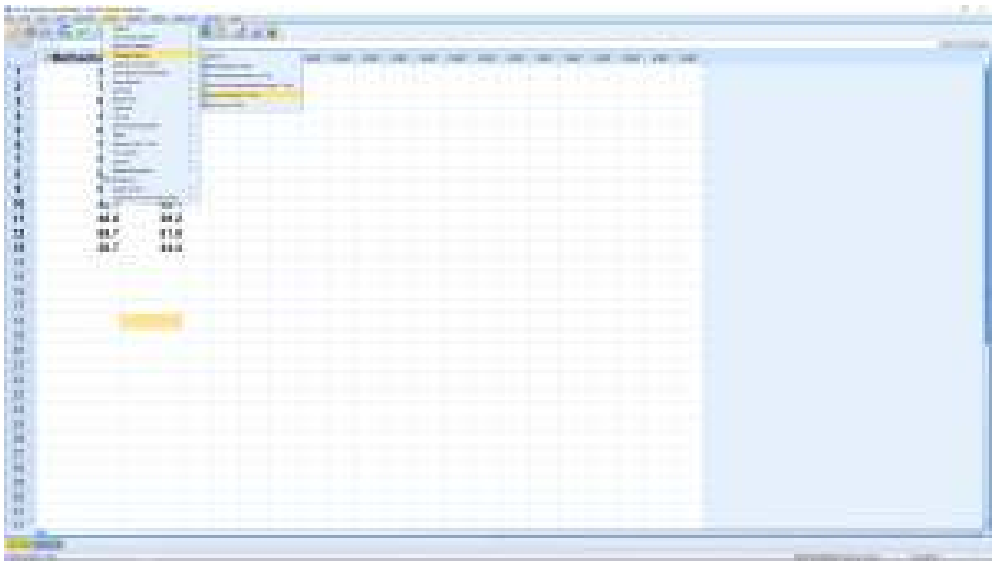
10.9: SPSS Lesson 7- Paired Sample t-Test

To follow along, load in the Data Set "Methadone.sav":



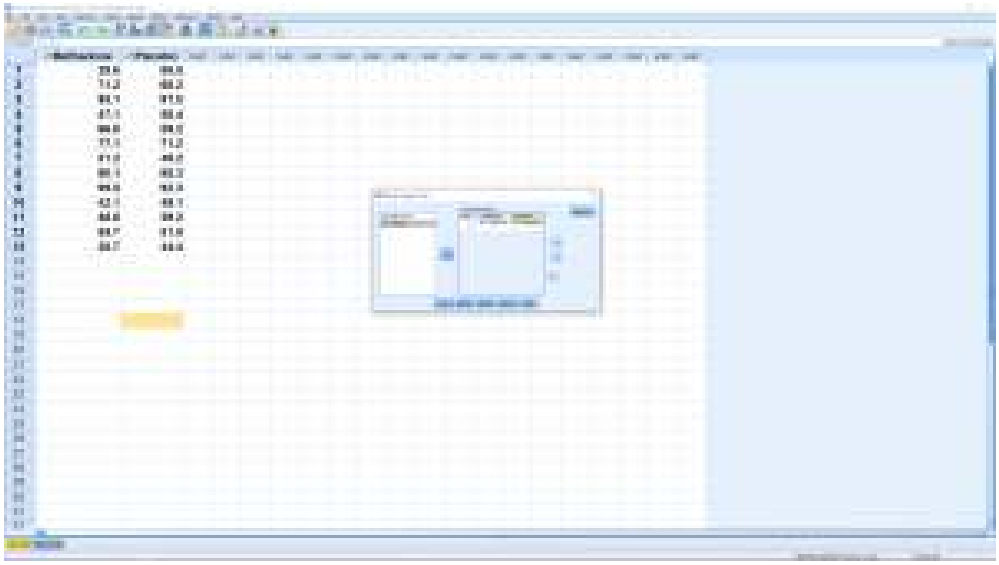
SPSS screenshot © International Business Machines Corporation.

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:



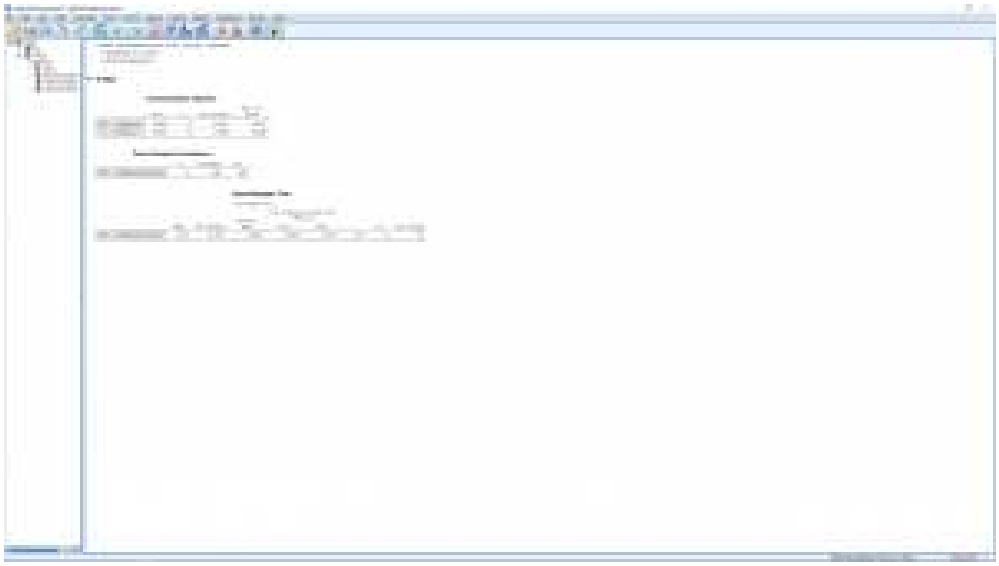
SPSS screenshot © International Business Machines Corporation.

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: $\overline{D} = 0.9615$, $s_D = 10.7067$, the confidence interval

$$-5.5084 < \mu_D < 7.4315,$$

$t_{\text{test}} = 0.324$, $\nu = 12$ and $p = 0.002$ for the two-tailed hypotheses pair

Misplaced &

Misplaced &

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

$$d = \frac{t}{\sqrt{n}} = \frac{\overline{D}}{s_D}$$

and

$$\eta^2 = \frac{t^2}{t^2 + n - 1}$$

respectively.

This page titled 10.9: SPSS Lesson 7- Paired Sample t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

11: 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.

This page titled 11: Comparing Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

11.1: z-Test for Comparing Proportions

In Section 9.4 we covered a one-sample test for proportions using the z approximation to the binomial distribution. Here we want to compare a proportion p_1 in one population with p_2 in another population, a two-sample test for proportions, also using the z approximation to the binomial distribution. Define

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2} \tag{11.1.1}$$

where x_1 and x_2 are the number of items of interest in the samples from the two populations and n_1 and n_2 are their sample sizes. Also define the corresponding $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$. The hypotheses we want to test is

Missing dimension or its units for \hspace

which is equivalent to

Missing dimension or its units for \hspace

If n_1p_1, n_1q_1, n_2p_2, and n_2q_2 are all > 5 then the appropriate normal distribution will provide a good approximation to the relevant binomial distribution and we can use the following test statistic to test the hypotheses

$$z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11.1.2}$$

where

Missing dimension or its units for \hspace

are the proportions of items of interest and not of interest in the two samples combined.

Example 11.1 : In a nursing home study we are interested in the proportions of nursing homes that have vaccination rates of less than 80%. The two populations we want to compare are small nursing homes and large nursing homes. In a sample of 34 small nursing homes, 12 were found to have a vaccination rate of less than 80%. In a sample of 24 large nursing homes, 17 were found to have a vaccination rate of less than 80%. At $\alpha = 0.05$ is there a difference in the proportions of small and large nursing homes with vaccination rates of less than 80%?

Solution :

0. Data reduction.

First define: population 1 = small nursing homes and population 2 = large nursing homes. Then compute the proportions:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{12}{34} = 0.35 \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{17}{24} = 0.71 \tag{11.1.3}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{12 + 17}{34 + 24} = \frac{29}{58} = 0.5 \qquad \bar{q} = 1 - \bar{p} = 1 - 0.5 = 0.5 \tag{11.1.4}$$

1. Hypotheses.

Missing dimension or its units for \hspace

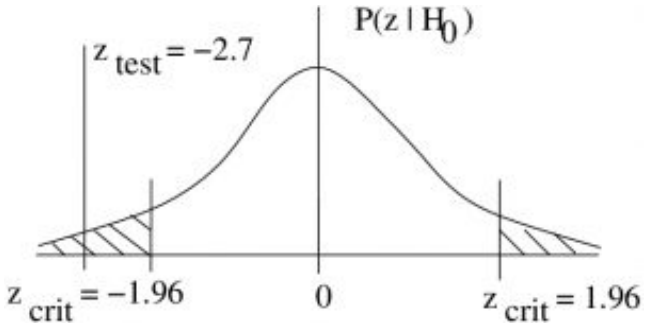
2. Critical statistic.

Use Table F, the last (z) line in the column for a two-tailed test at $\alpha = 0.05$: $z_{\text{crit}} = \pm 1.96$

3. Test statistic.

$$z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$z_{\text{test}} = \frac{0.35 - 0.71}{\sqrt{(0.5)(0.5)\left(\frac{1}{34} + \frac{1}{24}\right)}}$$
$$z_{\text{test}} = \frac{-0.36}{0.1333}$$
$$z_{\text{test}} = -2.7$$

4. Decision.



Reject H_0.

5. Interpretation.

There is enough evidence, from a z proportions test at $\alpha = 0.05$ to support the observation that large nursing homes have worse vaccination rates than small nursing homes. Make sure your parents end up in a small nursing home. (Note that rejection of H_0 in a one-tail test allows us to believe the direction of difference given by the sample data.)

This page titled 11.1: z-Test for Comparing Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

11.2: Confidence Interval for the Difference between Two Proportions

The form of the confidence interval is

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E \tag{11.2.1}$$

with

$$E = z_{\mathcal{C}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \tag{11.2.2}$$

where, as usual you can get $z_{\mathcal{C}}$ from the last line of the **t Distribution Table**.

Example 11.2 : Using the data from Example 11.1, find the 95% confidence interval for $p_1 - p_2$.

Solution : The relevant numbers from Example 11.1 are: $n_1 = 34$, $\hat{p}_1 = 0.35$, $\hat{q}_1 = 1 - 0.35 = 0.65$ and $n_2 = 24$, $\hat{p}_2 = 0.71$, $\hat{q}_1 = 1 - 0.71 = 0.29$.

Compute (after finding $z_{95\%} = 1.96$ from the **t Distribution Table**)

$$\begin{aligned} E &= z_{95\%} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\ E &= 1.96 \sqrt{\frac{(0.35)(0.65)}{34} + \frac{(0.71)(0.29)}{24}} \\ E &= 0.242 \end{aligned}$$

and

$$\hat{p}_1 - \hat{p}_2 = 0.35 - 0.71 = -0.36 \tag{11.2.3}$$

So

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) &< (\hat{p}_1 - \hat{p}_2) + E \\ -0.36 - 0.242 < (p_1 - p_2) &< -0.36 + 0.242 \\ -0.602 < (p_1 - p_2) &< -0.118 \end{aligned}$$

with 95% confidence. (Note that this corresponds with the rejection of H_0 in Example 11.1 since 0 is not in the confidence interval.)

□

This page titled 11.2: Confidence Interval for the Difference between Two Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12: ANOVA

This page titled 12: ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.1: One-way ANOVA

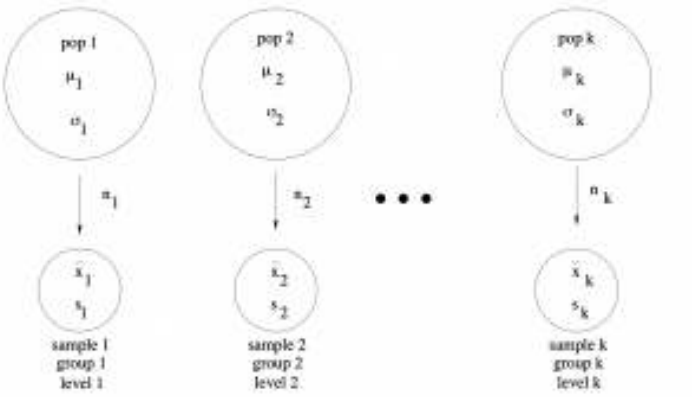
A one-way ANOVA (ANalysis Of VAriance) is a generalization of the independent samples t-test to compare more than 2 groups. (Actually an independent samples t-test and an ANOVA with two groups are the same thing). The hypotheses to be tested, in comparing the mean of k groups, with a one-way ANOVA are :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \text{ At least one of the means is different from the others.}$$

The following assumptions must be met for ANOVA (the version we have here) to be valid :

1. Normally distributed populations (although ANOVA is robust to violations of this condition).
2. Independent samples (between subjects).
3. Homoscedasticity : $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$. (ANOVA is robust to violations of this too, especially for larger sample sizes.)

The concept of ANOVA is simple but we need to learn some terminology so we can understand how other people talk about ANOVA. Each sample set from each population is referred to as a *group* or each population is called a group.



There will be k groups with sample sizes n_1, n_2, ..., n_k with the total number of data points being $N = \sum_{i=1}^{k} n_i$. For an ANOVA, the concept of independent variable (IV) and dependent variable (DV) become important (the IV in a single sample or a paired t-test is trivially a number like k or 0). The groups comprise different values of one IV. The IV is discrete with k values or *levels*.

In raw form, the test statistic for a one-way ANOVA is

$$F_{\text{test}} = F_{\nu_1,\nu_2} = \frac{s_{\text{B}}^2}{s_{\text{W}}^2} \tag{12.1.1}$$

where

$$\nu_1 = k - 1 \text{ (d.f.N.) } \nu_2 = N - k \text{ (d.f.D.)} \tag{12.1.2}$$

are the degrees of freedom you use when looking up F_{crit} in the **F Distribution Table** and where

$$s_{\text{B}}^2 = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{\text{GM}})^2}{k-1} \tag{12.1.3}$$

is the variance between groups, and

$$s_{\text{W}}^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} \tag{12.1.4}$$

is the variance within groups. Here n_i, $\bar{x}_i$ and s_i are the sample size, mean and standard deviation for sample i and $\bar{x}_{\text{GM}}$ is the grand mean:

$$\bar{x}_{\text{GM}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{N} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{N} \tag{12.1.5}$$

where x_{ij} is data point j in group i.

So you can see that ANOVA, the analysis of variance, is about comparing two variances. The within variance $s_{\rm W}^2$ is the variance of all the data lumped together, just as the grand mean $\bar{x}_{\rm GM}$ is the mean of all the data lumped together. It is the noise. You can see that the within variance is the weighted mean (weighted by $n_i - 1$) of the group sample variances — a little algebra shows that this is the variance of all the data lumped together. The between variance $s_{\rm B}^2$ a variance of the sample means $\bar{x}_i$. It is the signal. If the sample means were all exactly the same then the between variance $s_{\rm B}^2$ would be zero. So the higher $F_{\rm test}$ the more likely the means are different. $F_{\rm test}$ is a signal-to-noise ratio. If the means were all the same in the population then $s_{\rm B}^2$ would follow a χ_{k-1}^2 distribution and $s_{\rm W}^2$ (whether the population means were the same or not) would follow a χ_{N-k}^2 distribution. Thus if the population means were all the same (H_0) then the F test statistic follows a F_{ν_1,ν_2} distribution which has an expected value[1] (mean) of about 1. $F_{\rm test}$ must be sufficiently bigger than 1 to reject H_0.

The analysis of the variances can be broken down further, to sums of squares, with the following definitions[2] :

$$s_{\rm B}^2 = {\rm MS_B} \text{ between groups mean square}$$

and

$$s_{\rm W}^2 = {\rm MS_W} \text{ within groups mean square.}$$

Next we note that $\nu_1 = k - 1$ and $\nu_2 = N - k = \sum_{i=1}^{k}(n_i - 1)$ so

\[\begin{eqnarray*}{\rm MS}_{\rm B} = \frac{\nu_{1}}\end{eqnarray*}\]

and

\[\begin{eqnarray*}{\rm MS}_{\rm W} = \frac{\nu_{2}}\end{eqnarray*}\]

where

$$\text{SS}_{\rm B} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{\rm GM})^2 \text{ sum of squares between groups}$$

and

$$\text{SS}_{\rm W} = \sum_{i=1}^{k}(n_i - 1)s_i^2 \text{ sum of squares within groups}$$

so that

\[\begin{eqnarray*}F_{\rm test} = \frac\end{eqnarray*}\]

Why are sums of squares so prominent in statistics? (They will show up in linear regression too.) Because squares are the essence of variance. Look at the formula for the normal distribution, Equation 5.1. The exponent is a square. Mean and variance are all you need to completely characterize a normal distribution. Means are easy to understand, so sums of square focus our attention to the variance of normal distributions. If we make an assumption that all random noise has a normal distribution (which can be justified on general principles) then the sums of squares will tell the whole statistical story. Sums of squares also tightly links statistics to linear algebra (see Chapter 17) because the Pythagorus Theorem, which gives distances in ordinary geometrical spaces, is about sums of squares.

Computer programs, like SPSS, will output an ANOVA table that breaks down all the sums of squares and other pieces of the F test statistic :

Source	SS	ν	MS	F	p (sig)
Between (signal)	$\text{SS}_{\rm B}$	$\nu_1 = k - 1$	${\rm MS}_B = {\rm SS_B}/\nu_1$	Rendered by QuickLaTeX.com	p
Within (error)	$\text{SS}_{\rm W}$	$\nu_2 = N - k$	${\rm MS}_W = {\rm SS_W}/\nu_2$		
Totals	$\text{SS}_{\rm T}$	$\nu_T = N - 1$			

Here p is the p-value of $F_{\rm test}$, reported by SPSS as “sig” for significance. $F_{\rm test}$ is significant (you can reject H_0) if $p < \alpha$. You should be able to reconstruct an ANOVA table given only the SS values. Notice that the total degrees of freedom of the ANOVA is $\nu_T = \nu_1 + \nu_2 = N - 1$. One degree of freedom is used up in computing the grand mean, the rest in computing the variances, very similar to how $n - 1$ is the degrees of freedom for sample standard deviation s. If you think of degrees of freedom as the amount of

information in the data then the one-way ANOVA uses up all the information in the data. This point will come up again when we consider post hoc comparisons.

Example 12.1 : A state employee wishes to see if there is a significant difference in the number of employees at the interchanges of three state toll roads. At $\alpha = 0.05$ is there a difference in the average number of employees at each interchange between the toll roads?

The data are :

Road 1 (group 1)	Road 2 (group 2)	Road 3 (group 3)
7	10	1
14	1	12
32	1	1
19	0	9
10	11	1
11	1	11

Solution :

0. Data reduction.

Using your calculators, find

Missing dimension or its units for \hspace

Missing dimension or its units for \hspace

Missing dimension or its units for \hspace

$N = 18$.

1. Hypothesis.

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 :$ At least one of the means is different from the others.

2. Critical statistic.

Use the F Distribution Table with $\alpha = 0.05$; do not divide the table α (right tail area) by 2 in this case, there are no left and right tail tests in ANOVA. The degrees of freedom needed are $\nu_1 = k - 1 = 3 - 1 = 2$ (d.f.N.) and $\nu_2 = N - k = 18 - 3 = 15$ (d.f.D.). With that information

$$F_{\text{crit}} = 3.68 \tag{12.1.6}$$

3. Test statistic.

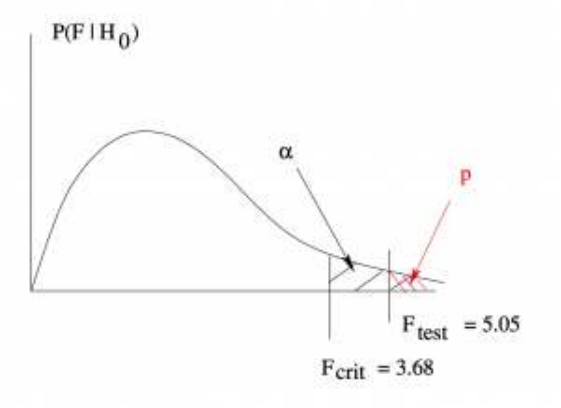
Compute, in turn :

\[\begin{eqnarray*} \bar{x}_{\rm GM} & = & \frac{n_{1} \bar{x}_{1} + n_{2} \bar{x}_{2} + n_{3} \bar{x}_{3}}{N} \ & = & \frac{(6) (15.5) + (6) (4.0) + (6) (5.83)}{18} \ & = & \frac{152}{18} = 8.4 \ s_{\rm B}^{2} & = & \frac{\sum_{i=1}^{k} n_{i} (\bar{x}_{i}- \bar{x}_{\rm GM})^{2}}{k-1} \ & = & \frac{(6)(15.5 - 8.4)^{2} + (6)(4.0-8.4)^{2} + (6)(5.8 - 8.4)^{2}}{3-1} \ & = & \frac{\nu_{1}} = \frac{459.18}{2} = 229.59 \ s_{\rm W}^{2} & = & \frac{\sum_{i=1}^{k}(n_{i}-1) s_{i}^{2}} {\sum_{i=1}^{k}(n_{i}-1)} \ & = & \frac{(6-1)(81.9) + (6-1)(25.6) + (6-1)(29.0)}{(18-3)} \ & = & \frac{\nu_{2}} = \frac{682.5} {15} = 45.5 \ \end{eqnarray*}\]

Note how we saved SS_B and SS_W for the ANOVA table. And finally

$$F_{\text{test}} = \frac{s_{\text{B}}^2}{s_{\text{W}}^2} = \frac{229.59}{45.5} = 5.05 \tag{12.1.7}$$

4. Decision.



Reject H_0.

5. Interpretation.

Using one-way ANOVA at $\alpha = 0.05$ we found that at least one of the toll roads has a different average number of employees at their interchanges. The ANOVA table is :

Source	SS	ν	MS	F	p (sig)
Between (signal)	459.18	2	229.59	5.05	$p < 0.05$
Within (error)	682.5	15	45.5		
Totals	1141.68	17			

We did not compute p but a computer program like SPSS will.

□

1. The mean of the F_{ν_1,ν_2} distribution is $\mu_F = \frac{\nu_2}{\nu_2 - 2}$ if $\nu_2 > 2$. ↵
2. You might have heard of RMS for "root mean square". RMS = \(\sqrt{\mbox{MS}} = \sqrt = s\). RMS is standard deviation. ↵

This page titled 12.1: One-way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.2: Post hoc Comparisons

If H_0 is rejected in a one-way ANOVA, you will frequently want to know where the differences in the means are. For example if we tested $H_0 : \mu_1 = \mu_2 = \mu_3$ and rejected H_0 in a one-way ANOVA then we will want to know if $\mu_1 \neq \mu_2$ or $\mu_2 \neq \mu_3$, etc.

To see which means are different after doing an ANOVA we could just compare all possible combinations of pairs using t-tests. But such an approach is no good because the assumed type I error rates, α, associated with the t-tests would be wrong. The α rate would be higher because in making such *multiple comparisons* you incur a greater chance of making an error.

So we need to correct our test statistic and/or the corresponding α value when we do such multiple comparisons. We will cover two such multiple comparison approaches in detail :

1. Scheffé test
2. Tukey test

and we will look at the Bonferroni approach.

Doing multiple comparisons after an ANOVA is known as post hoc testing. It is the traditional approach for comparing several means. The opening "omnibus" ANOVA lets you know if there are any differences at all. If you fail to reject the ANOVA H_0 then you are done. Only when you reject H_0 do you put in the effort of comparing means pairwise. This traditional approach, designed to minimize the necessary calculations, is not the only way to compare multiple means. The other approach is to forget about the ANOVA and then use t-tests to compare means pairwise on in combinations[1] of means until you use up the N degrees of freedom in the dataset. Here we will stick with the traditional approach.

12.2.1 Scheffé test

The test statistic for the Scheffé test is

$$F_{s,\text{test}} = F_s = \frac{(\bar{x}_i - \bar{x}_j)^2}{s_{\text{W}}^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{12.2.1}$$

Note that F_s is basically a t^2 quantity (recall that $F_{1,\nu} = t^2$) but with a pooled estimate s_p^2 of the common population variance σ given by the value of s_{W}^2 from the ANOVA. In other words F_s uses information from all of the data to estimate σ instead of from just groups i and j as a t-test would (see Equation 10.5). Note that the Scheffé test does not require equal group sizes n_i.

The critical statistic is a modification of the critical statistic from the ANOVA is

$$F'_{\text{crit}} = F' = (k-1)F_{\alpha,\nu_1,\nu_2} = (k-1)F_{\text{crit,ANOVA}} \tag{12.2.2}$$

where ν_1 and ν_2 are the ANOVA degrees of freedom. The critical statistic is the same for all pairwise comparisons regardless of the sample sizes, n_i and n_j, of the pair of groups being compared.

Example 12.2 : The ANOVA of Example 12.1 found that at least one of the three means was different from the others. Use the Scheffé test to find the significant differences between the means. There has to be at least one.

Solution :

0. Data reduction.

Collect the necessary information from the omnibus ANOVA. We'll need:

Missing dimension or its units for \hspace

Missing dimension or its units for \hspace

1. Hypotheses.

There are 3 hypotheses pairs to test :

$$H$$

2. Critical statistic.

One value for all three hypothesis tests:

$$F_{\text{crit}} = (k-1)F_{0.05,2,15} = (3-1)(3.68) = (2)(3.68) = 7.367 \tag{12.2.5}$$

3. Test statistic.

There are three of them:

μ_1 vs. μ_2 :

$$F_s = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_{\text{W}}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(15.5-4.0)^2}{45.5\left(\frac{1}{6} + \frac{1}{6}\right)} = 8.72 \tag{12.2.6}$$

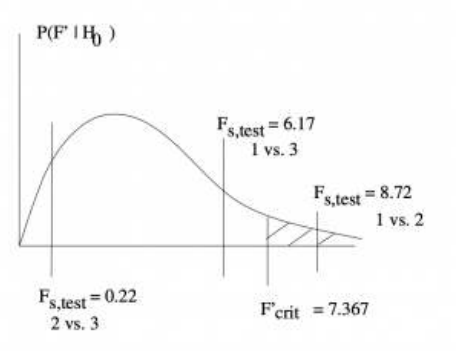
μ_1 vs. μ_3 :

$$F_s = \frac{(\bar{x}_1 - \bar{x}_3)^2}{s_{\text{W}}^2\left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = \frac{(15.5-5.83)^2}{45.5\left(\frac{1}{6} + \frac{1}{6}\right)} = 6.17 \tag{12.2.7}$$

μ_2 vs. μ_3 :

$$F_s = \frac{(\bar{x}_2 - \bar{x}_3)^2}{s_{\text{W}}^2\left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = \frac{(4.0-5.83)^2}{45.5\left(\frac{1}{6} + \frac{1}{6}\right)} = 0.22 \tag{12.2.8}$$

4. Decision.



For μ_1 vs. μ_2, reject H_0. For μ_1 vs. μ_3, do not reject H_0. For μ_2 vs. μ_3, do not reject H_0.

5. Interpretation.

The results of the Scheffé test at $\alpha = 0.05$ conclude that only the mean numbers of interchange employees between toll roads 1 and 2 are significantly different.

□

12.2.2 Tukey Test

The test statistic for the Tukey test is

Rendered by QuickLaTeX.com

where, again, s_{W}^2 is from the omnibus ANOVA, $\bar{x}_i$ is the mean of group i and we must have equal sample sizes for all groups: $n = n_i$ for all i. There is a Tukey test statistic for unequal n, and it is used by SPSS, but we won't cover that here.

The critical statistic, q_{crit}, comes from a table of critical values from a new distribution called the q distribution. The critical values are tabulated in the **Tukey Test Critical Values Table**. To use this table, you need two numbers going in :

1. k = number of groups
2. $\nu = \nu_2$ = degrees of freedom for s_W^2

Reject H_0 when $image q_{\mbox{crit}}$" title="Rendered by QuickLaTeX.com" height="16" width="66" style="vertical-align: -6px;">. In this case we don't have a picture of the q distribution handy (although it is basically the absolute value of t), so we just use the $image q_{\mbox{crit}}$" title="Rendered by QuickLaTeX.com" height="16" width="66" style="vertical-align: -6px;"> rule similar to how we use the p-value.

Example 12.3 : Repeat Example 12.2 using the Tukey test instead of the Scheffé test.

Solution : 0. Data Reduction.

We use the same data from the omnibus ANOVA :

Missing dimension or its units for \hspace

Missing dimension or its units for \hspace

1. Hypotheses.

The 3 hypotheses pairs to test are the same :

$$H_0 : \mu_1 = \mu_2, \qquad H_0 : \mu_1 = \mu_3, \qquad H_0 : \mu_2 = \mu_3 \tag{12.2.9}$$

$$H_1 : \mu_1 \neq \mu_2, \qquad H_1 : \mu_1 \neq \mu_3, \qquad \[H_1 : \mu_2 \neq \mu_3 \tag{12.2.10}$$

2. Critical statistic.

Use the **Tukey Test Critical Values Table**. Go into the table with

- Number of groups = $k = 3$.
- $\nu = \nu_2 = N - k = nk - k = (6)(3) - 3 = 18 - 3 = 15$.

and $\alpha = 0.05$ to find

$$q_{\text{crit}} = 3.67 \tag{12.2.11}$$

3. Test statistic.

Again, there are three of them :

μ_1 vs. μ_2 :

Rendered by QuickLaTeX.com

μ_1 vs. μ_3:

Rendered by QuickLaTeX.com

μ_2 vs. μ_3:

Rendered by QuickLaTeX.com

4. Decision.

Reject H_0 when $image$ q_{\mbox{crit}}" title="Rendered by QuickLaTeX.com" height="16" width="66" style="vertical-align: -6px;">. This only happens for one hypothesis pair : For μ_1 vs. μ_2, reject H_0. For μ_1 vs. μ_3, do not reject H_0. For μ_2 vs. μ_3, do not reject H_0.

5. Interpretation.

The results of the Tukey test at $\alpha = 0.05$ conclude that only the mean numbers of interchange employees between toll roads 1 and 2 are significantly different. (Same result as the Scheff\'e test. Usually this happens but when it doesn't, you need to use some kind of non-mathematical judgement.)

12.2.3 Bonferroni correction

A more conservative (less power) approach to multiple comparisons (post hoc testing) is to use Bonferroni's method. The fundamental idea of the Bonferroni correction is to add the probabilities of making individual type I errors to get an overall type I error rate.

Implementing the idea is simple. Do a bunch of t-tests and multiply the p-value by a correction factor C. There are a number of ways to choose C (you will have to dig to find out which method SPSS uses). The easiest (and most conservative) is to set C equal to the number of pairwise comparisons done. So if you have k groups then C is given by the binomial coefficient:

$$C = \left(\begin{array}{c} k \\ 2 \end{array} \right). \tag{12.2.12}$$

Another way is to look at the total degrees of freedom, ν_{pairs}, associated with the pairwise t-tests and compare it to the total degrees of freedom in the data, $\nu = N$ (or one could argue $\nu = N - 1$), to come up with

$$C = \frac{\nu_{\text{pairs}} + \nu}{\nu}. \tag{12.2.13}$$

Since there is some ambiguity as to what we should use for C, we will not do Bonferroni post hoc testing by hand. However, be able to recognize Bonferroni results in SPSS, treating the value of C as an SPSS blackbox parameter.

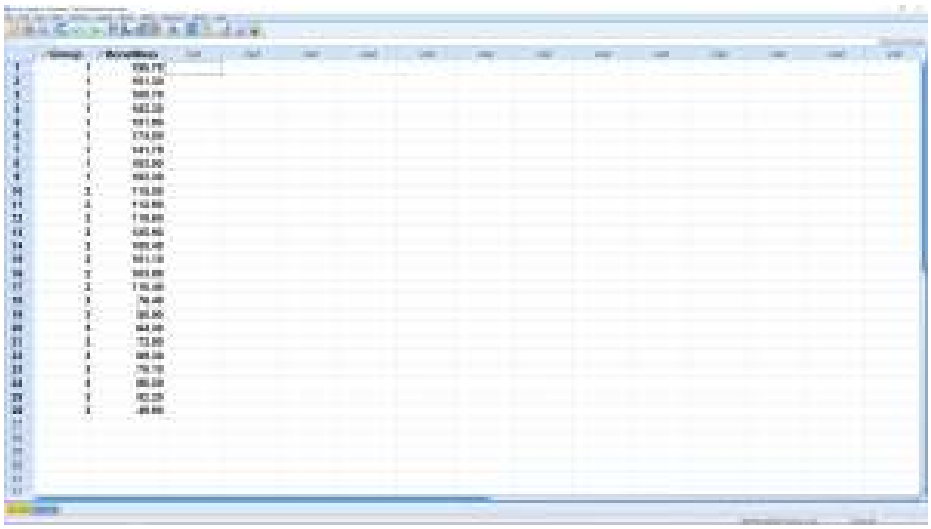
---

1. Combinations of means may be compared using "contrasts". For example $\mu_1 + \mu_2$ might be compared with $2\mu_3$. Contrasts are not covered in Psy 234. ↵

This page titled 12.2: Post hoc Comparisons is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.3: SPSS Lesson 8- One-way ANOVA

To follow along, load in the Data Set "BoneStrength.sav":

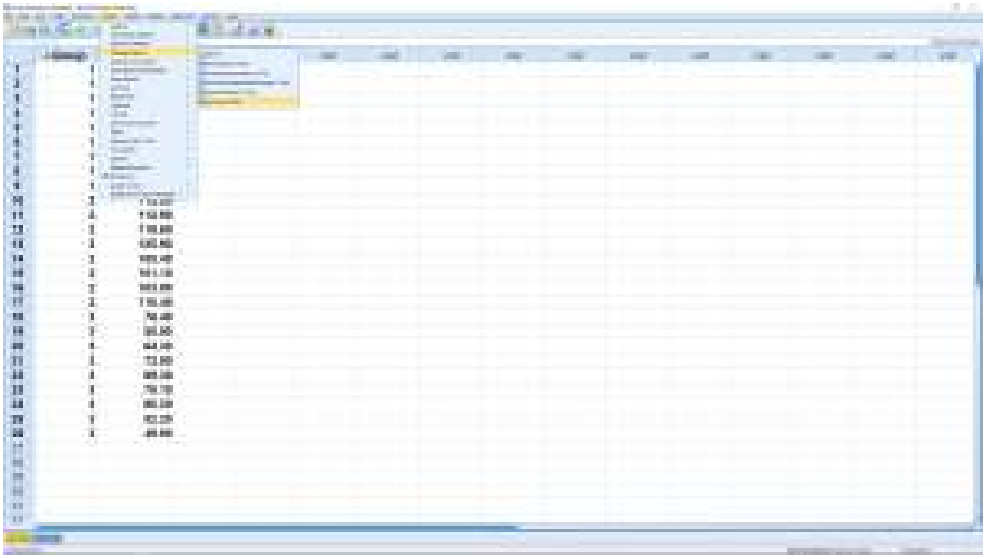


SPSS screenshot © International Business Machines Corporation.

The data format is exactly the same as used for an independent samples t-test, except now there are more than two groups in the independent variable, named group in this case. The dependent variable here is diff and we want to test the hypothesis

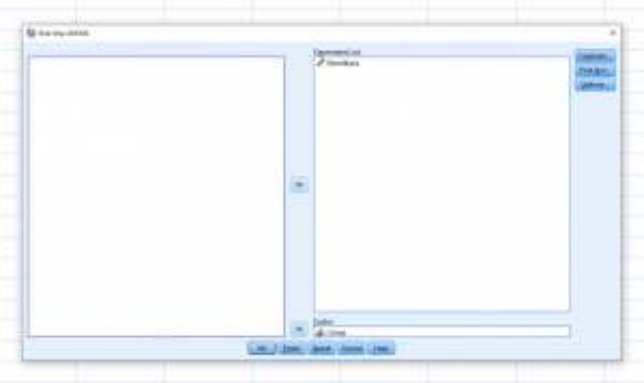
$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \mu_3 \\ H_1: &\quad \text{At least one of the means is different.} \end{aligned}$$

There are two ways to do this in SPSS. We'll cover each one. The first method is to go through Analyze → Compare means → One-Way ANOVA :



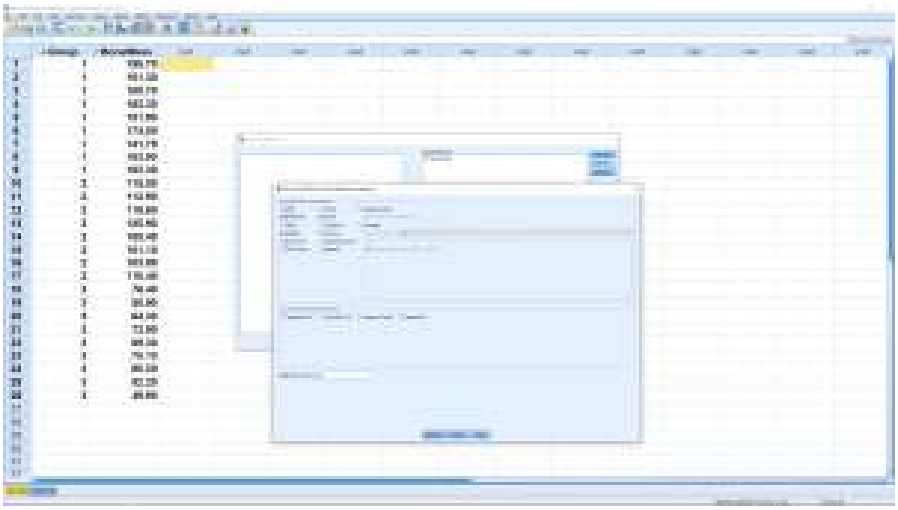
SPSS screenshot © International Business Machines Corporation.

Move the independent variable into the Factor box and the dependent variable into the Dependent List box :



SPSS screenshot © International Business Machines Corporation.

We will ignore the Contrasts menu but the Post Hoc menu is where you set things up for Post Hoc analyses :



SPSS screenshot © International Business Machines Corporation.

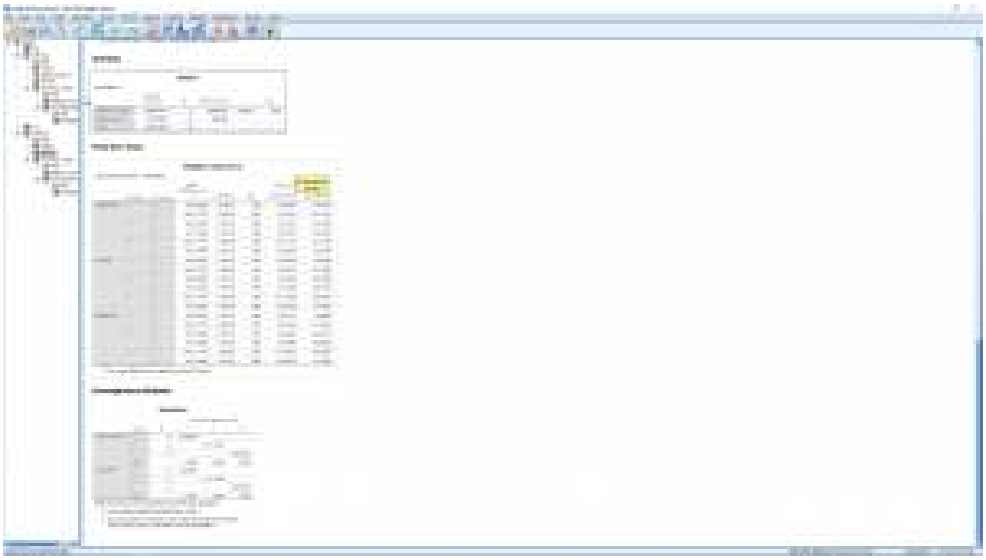
We have checked off Bonferroni, Scheffe and Tukey in the “Equal Variances Assumed” box — we will be assuming homoscedasticity for all our ANOVA work. Hit Continue then OK to get the output :

ANOVA

BoneMass

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	50909.763	2	25454.881	158.841	.000
Within Groups	3685.837	23	160.254		
Total	54595.600	25			

SPSS screenshot © International Business Machines Corporation.



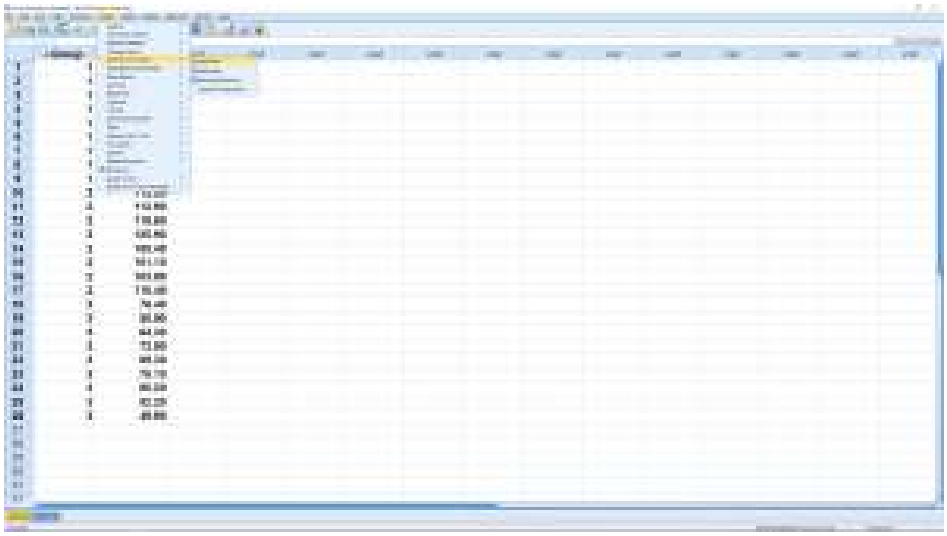
SPSS screenshot © International Business Machines Corporation.

The first table is the ANOVA table. The third column can be obtained from the first two since $\text{MS} = \text{SS}/\nu$. The fourth column is, of course, Rendered by QuickLaTeX.com. The Sig.~column gives $p < 0.001$ so we reject H_0.

The next table, which is not nonsense since the ANOVA tells us that at least one of the means is different, gives all the pairwise comparisons in a very redundant way. For each test we checked, all three pairs of means are compared — twice (hugely sloppy programming in my opinion). Looking through the table we see that $\overline{x}_1 - \overline{x}_2 = 58.75$ and $\overline{x}_1 - \overline{x}_3 = 106.17$ are significantly different from zero as indicated by the * or as can be seen by looking at the p-values. The difference $\overline{x}_2 - \overline{x}_3 = 47.41$ is significantly different from zero. Here all three post hoc methods disagreed. If there is ever a disagreement then you should choose the most conservative result, the one with the lease amount of significant differences.

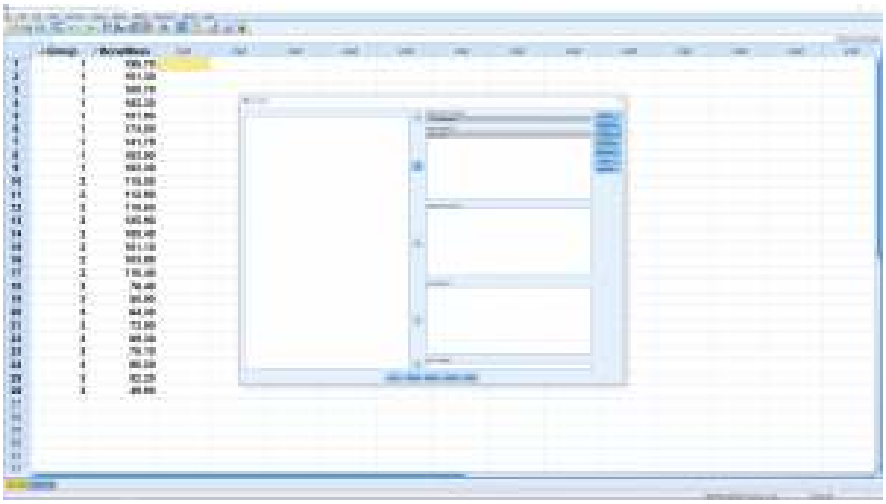
We won't worry too much about the last table. It merges groups that are not significantly different from each other into "homogeneous subsets". Here groups 2 and 3 are considered the same and merged into homogeneous subset 1 while group 1 stands on its own as homogeneous subset 2.

The other way of doing a one-way ANOVA is to pick Analyze → General Linear Model → Univariate :



SPSS screenshot © International Business Machines Corporation.

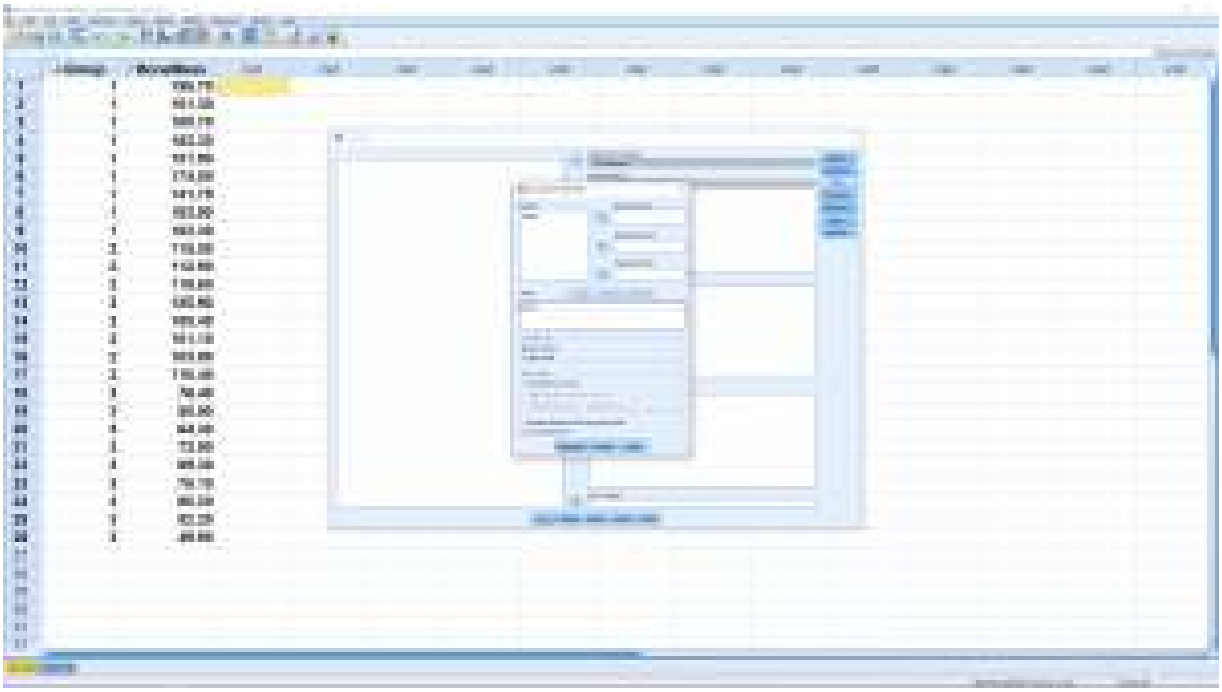
This brings up :



SPSS screenshot © International Business Machines Corporation.

Move the dependent variable into the Dependent Variable box and the independent variable, known as a factor in ANOVA language, into the Fixed Factor box. You will be entering two fixed factors in here when we get to 2-way ANOVA. The Random Factor is for the case where there are multiple populations (factors) and you do not get data from all of them, but only from a random sample of those populations. We will not cover this approach here but you can use it no problem in the future if you have to, it works the same way but the SS formulae are different. The Covariate box is for a method known as ANCOVA (analysis of covariance). We will not cover ANCOVA here, but note that it is a combination of ANOVA and linear regression.

Look at the Model menu and leave the button selected to “Full Factorial” (for one-way ANOVA this is the only choice anyway) and leave the “Include intercept in model” button as selected too. We’ll leave Contrasts as it is too. Open plots and set it up as, by clicking group into “Horizontal Axis” and then clicking Add, so we can get a profile plot :



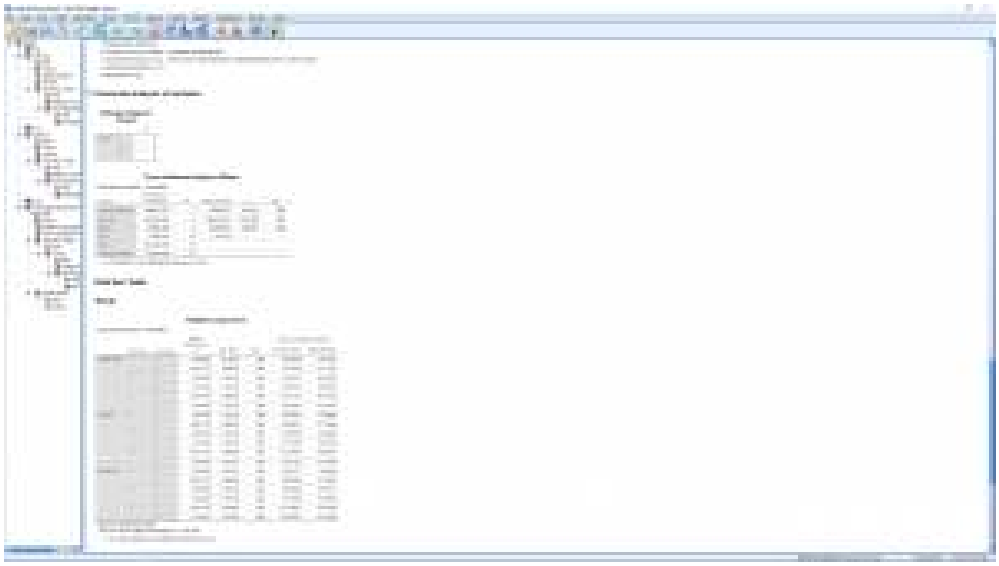
SPSS screenshot © International Business Machines Corporation.

Finally, open the Post Hoc menu and set it up the same way as we set up the Post Hoc menu above :



SPSS screenshot © International Business Machines Corporation.

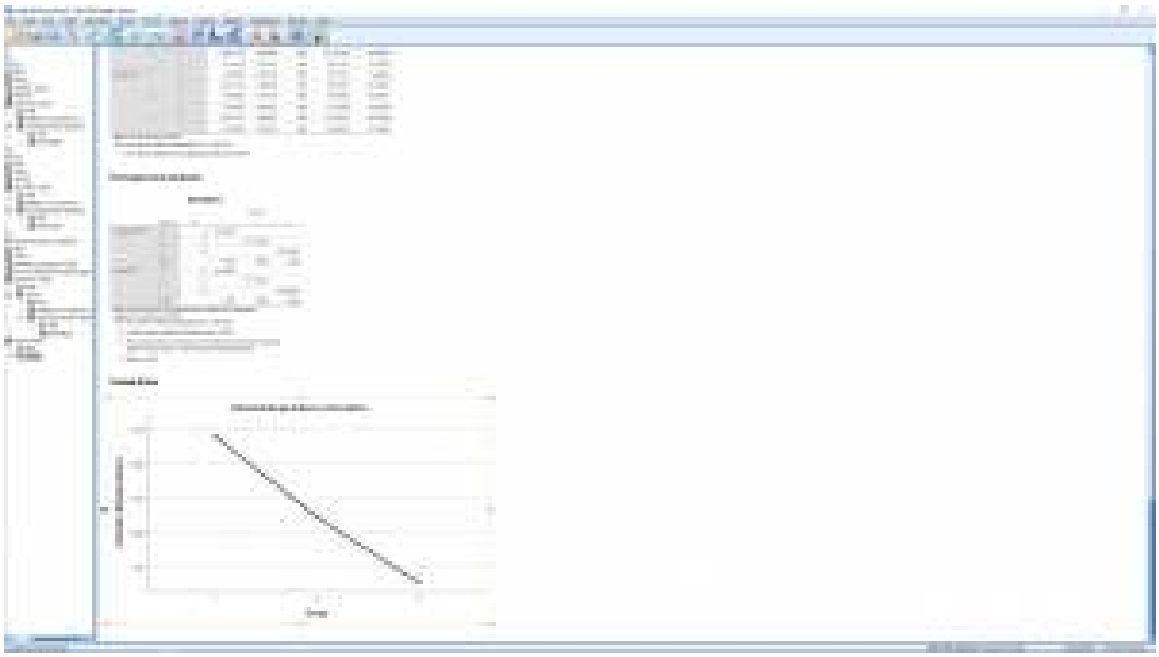
Hit Continue the OK to get the output :



SPSS screenshot © International Business Machines Corporation.

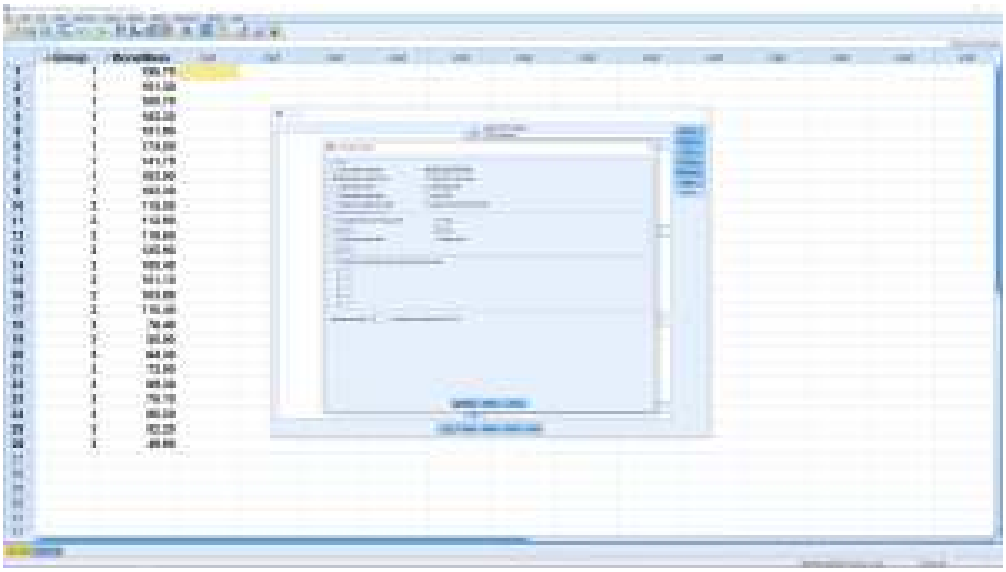
The output is pretty much the same as before (the homogeneous subsets is output also but it is not shown here) but the ANOVA table is a little different. In particular there are extra lines in the ANOVA table that you need to learn to ignore. The relevant lines are group (between), Error (within) and “Corrected Total”. When we run this analysis we had the “Include intercept in model” box checked, if we unchecked that box then the ANOVA table will not contain an Intercept line.

The output also contains a profile plot where we can clearly see that the mean for group 1 is different from the other two groups :



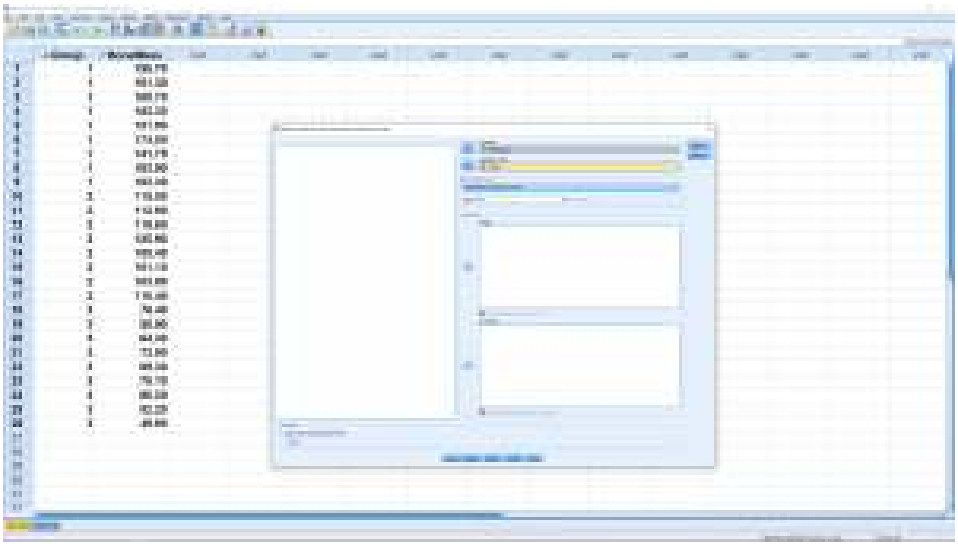
SPSS screenshot © International Business Machines Corporation.

We skipped the Save and the Options menu when we set up the test. Take a look at the Save menu. You will see a bunch of essentially descriptive statistics that we won't worry about here. In the Options menus, though, check off a couple of items, as shown here, and re-run :



SPSS screenshot © International Business Machines Corporation.

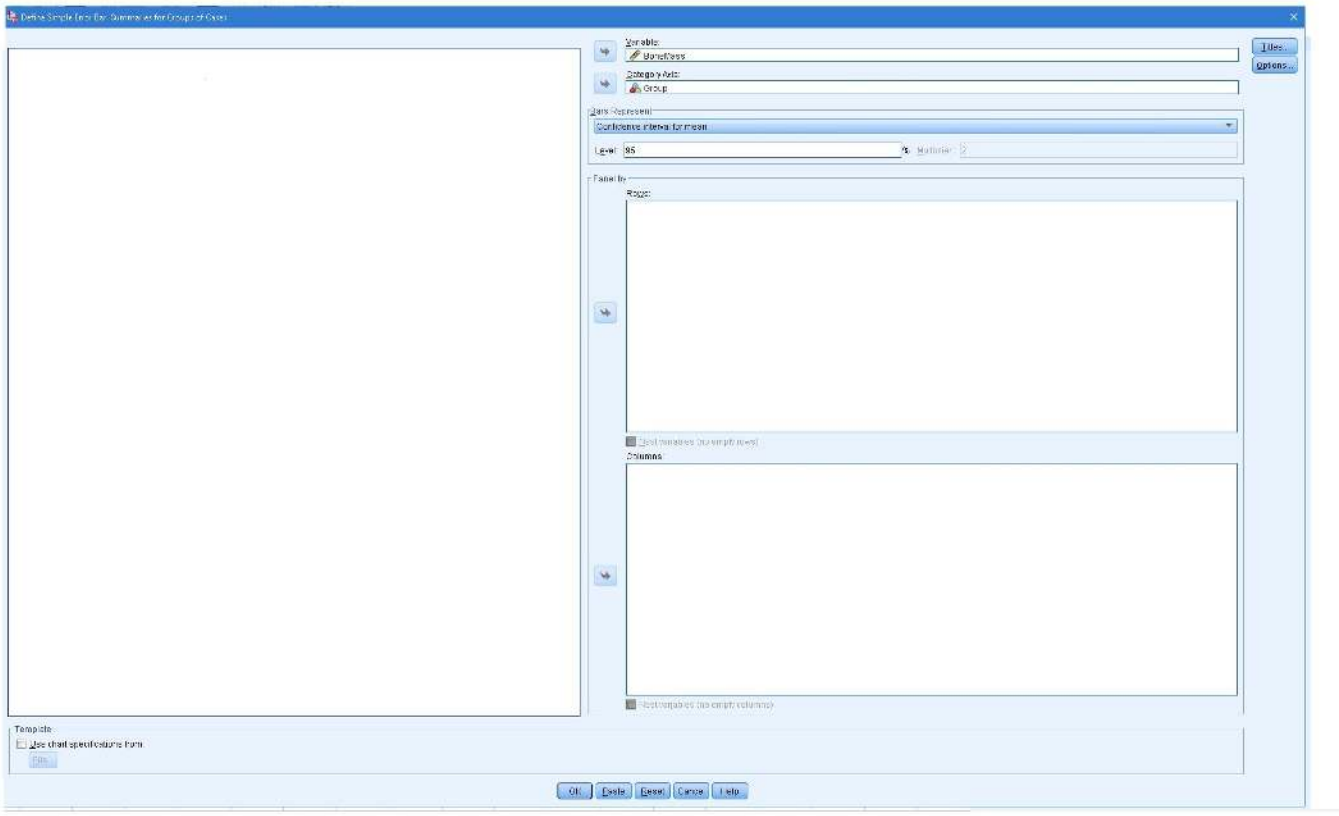
This time you get an additional table for Levine's test and more information in the ANOVA table :



SPSS screenshot © International Business Machines Corporation.

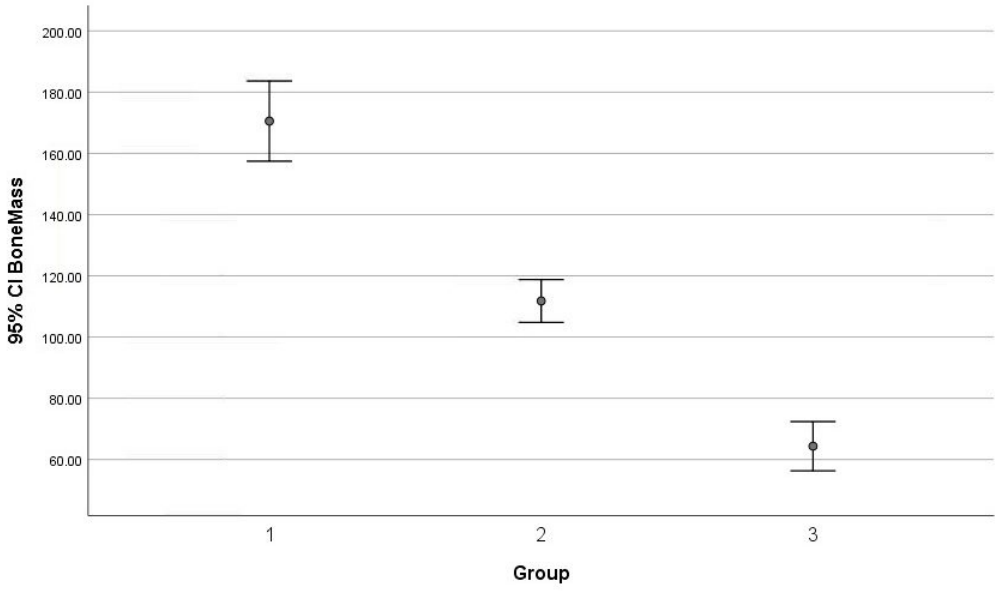
Levine's test has $p = 0$ so we do not reject the null hypothesis of homosecdasticity between the three groups. This is just a test of assumptions about the sums of squares formulae using in the analysis. The ANOVA table contains a column for the descriptive η^2 strength of association. Note that it is the same as R squared.

The profile plot produced by the ANOVA analysis can be miss leading if s_w is large. To see the scatter in the data more clearly, we can make profile plots with error bars. Go to Graphs → Legacy Dialogs → Error Bar, leave the default settings at Simple and "Summaries for groups of cases". Then set up the menu as follows.



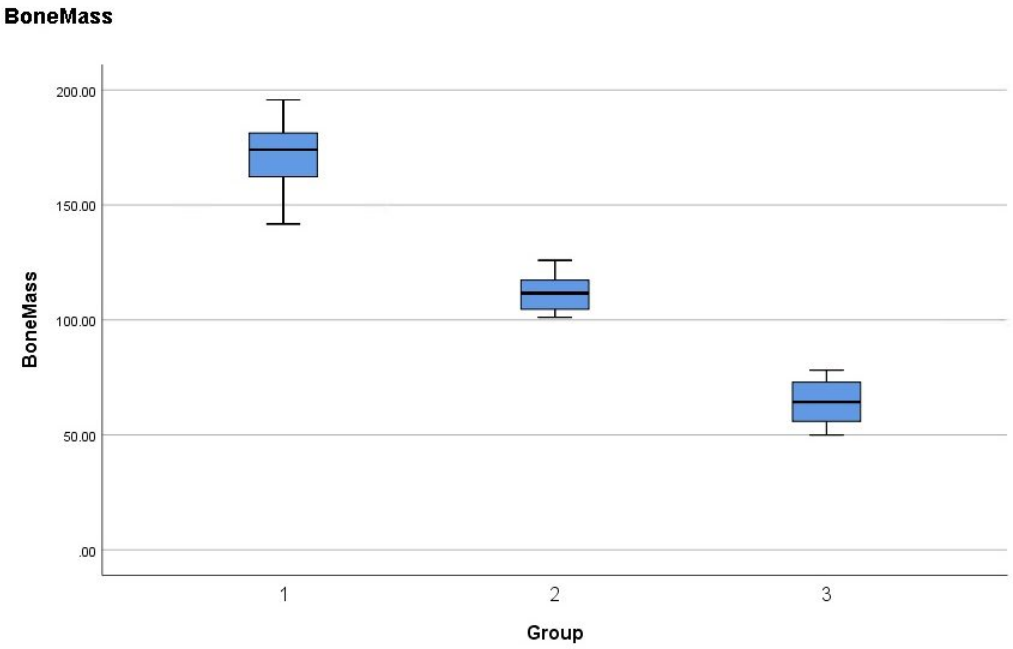
SPSS screenshot © International Business Machines Corporation.

This produces :



SPSS screenshot © International Business Machines Corporation.

Similarly we can produce a boxplot version :



SPSS screenshot © International Business Machines Corporation.

The ANOVA correctly identified the means of all groups as being different. This kind of information can drastically change your interpretation of the results (in this case that group C may not be as effective as the ANOVA indicates).

This page titled 12.3: SPSS Lesson 8- One-way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.4: Two-way ANOVA

In all the statistical testing we've done so far, and will do in Psy 233/234, there is only one *dependent variable* (DV) — we have been/are doing univariate statistics.

And so far, in all the tests we've seen there has only been one *independent variable* (IV). For the t-tests the IV is group or population with only two values[1] 1 and 2. In one-way ANOVA the single IV has k (number of groups) values. Also, so far, the IV has been a discrete variable (that will change when we get to regression). The graph to keep in mind for the one-way ANOVA is a profile graph as shown in Figure 12.1.

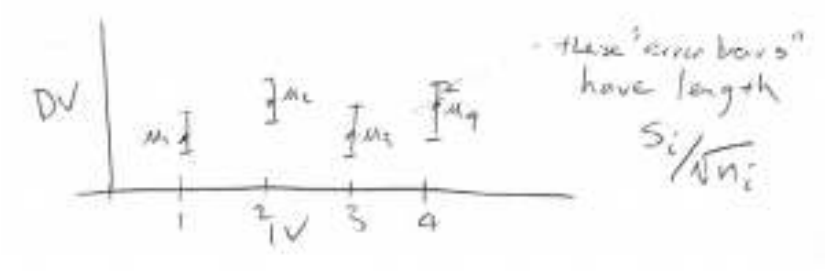
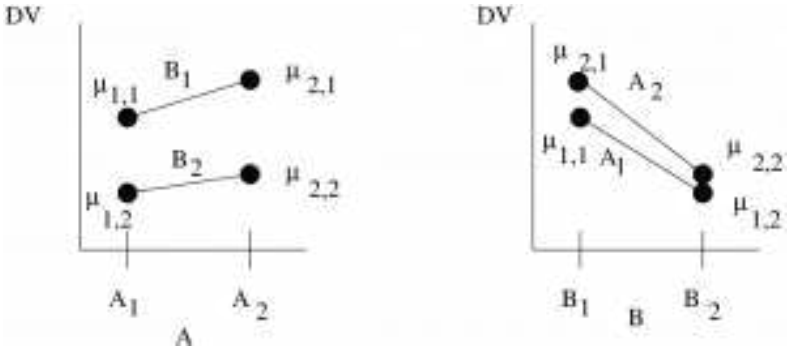


Figure 12.1: The profile plot is a good way to think of one-way ANOVA data with the IV on the x-axis and the DV on the y-axis. A one-way ANOVA tests the hypothesis: are all the means μ_i equal to each other? (An actual data profile graph can only have sample values $\overline{x}_i$, we show a kind of confidence interval plot here.)

With two-way ANOVA you have **two IVs**. Let's call the two IVs A and B. Each IV in two-way ANOVA is called a *factor*. A and B can each have several values (or "levels"). To introduce concepts, let's stick with the case were each of A and B hove only 2 values: A_1 and A_2 for A, and B_1 and B_2 for B. This is the 2×2 ANOVA case, where the 2 tells you haw many levels are in each factor. If, for example, A had 4 levels (values) and B had 3 levels then you'd have a 4×3 ANOVA. Let's stick with the 2×2 case for now.

There are several ways to think of a two-way ANOVA. Let's start with two-dimensional profile plots for a 2×2 ANOVA :

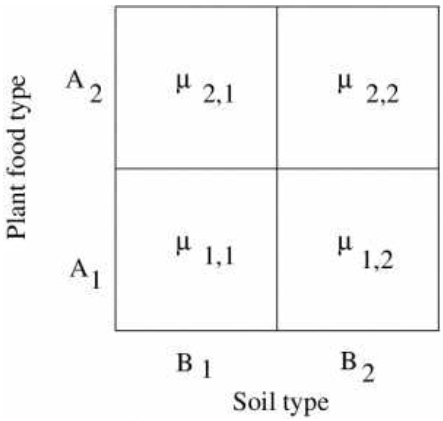


The profile plot can be done in one of two ways. The y axis represents the DV in both cases. On the left, the x axis represents the IV A and the two values of the other IV, B, are represented as lines. On the right, the x axis represents the IV B and the two values of the other IV, A, are represented as lines. Look closely at the plots. The dots represent the population values[2] with $\mu_{i,j}$ being the value of the population labelled by $A = i, B = j$. The means in the two plots are exactly the same. Each combination of IVs, i, j, defines a *treatment group*. For a 2×2 ANOVA there are four treatment groups.

Two way ANOVA supposedly had one of its first applications to agriculture. So, to fix ideas, let's take our two IV's, also known as two ***factors*** as :

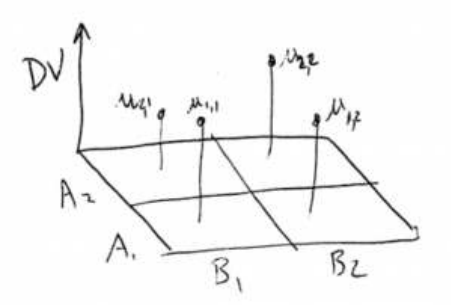
$$\begin{aligned} A = &\text{ Plant food type} \\ B = &\text{ Soil type} \end{aligned}$$

Then, with two levels for each factor, we can visualize the setup as fields where you would grow plants :



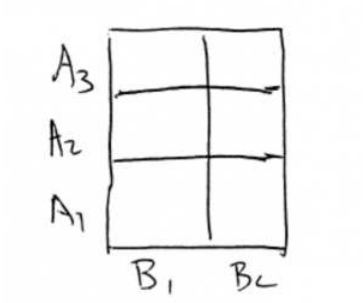
Each field, or treatment group, is also known as a ***cell***.

Now, lets use the x and y axes to represent the IVs (B and A in this case). Then we can use the z axis to represent the DV in a 3D plot :

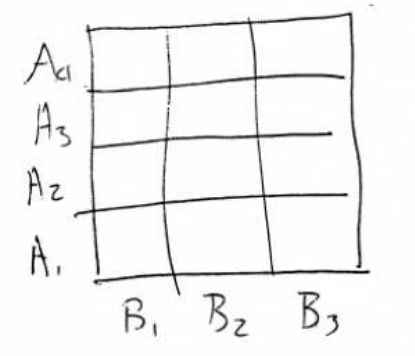


This makes sense. A two-way ANOVA has three variables, two IVs and on DV, so the data are 3D data and the plot above shows how those data appear in 3D space. If you look at the 3D plot from the front you see the profile plot with B on the x axis. If you look at the 3D plot from the (right) side, you see the profile plot with A on the x axis.

We've focused on 2×2 designs. But the two IVs can have any number of discrete values or levels. For example, a 3×2 cell diagram would look like :



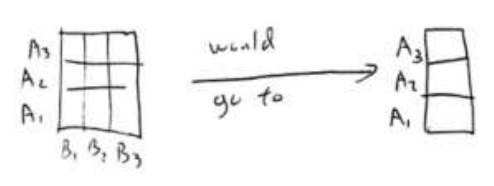
And a 4×3 design would look like :



Now that we understand what kind of data we have, it's time to move onto hypothesis testing. In two-way ANOVA there are three hypotheses to test :

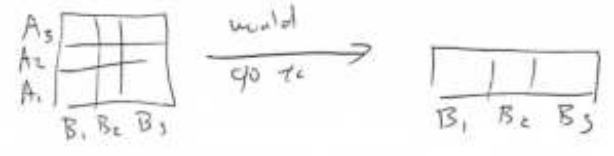
1. Is there a "main effect" of A?
2. Is there a "main effect" of B?
3. Is there an *interaction* of $A \times B$?

In all cases, H_0 is that there is no effect or interaction. As we will see, each hypothesis is a one-way ANOVA of the two-way data suitably collapsed into a one-way design. Let's begin with the main effect of A. The hypothesis is equivalent to collapsing the design across B :



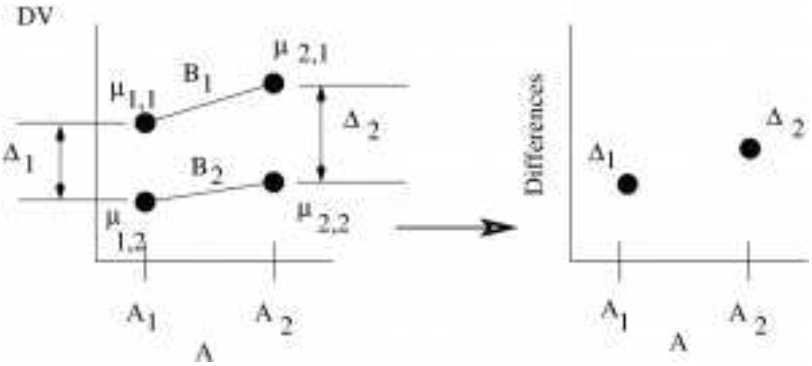
and then doing a one-way ANOVA with the one IV equal to A. The collapse is done by averaging over B which is the same as removing the cell boundaries between the B cells and only categorizing the data by the A levels.

The hypothesis for the main effect of B is similarly equivalent to collapsing across A :



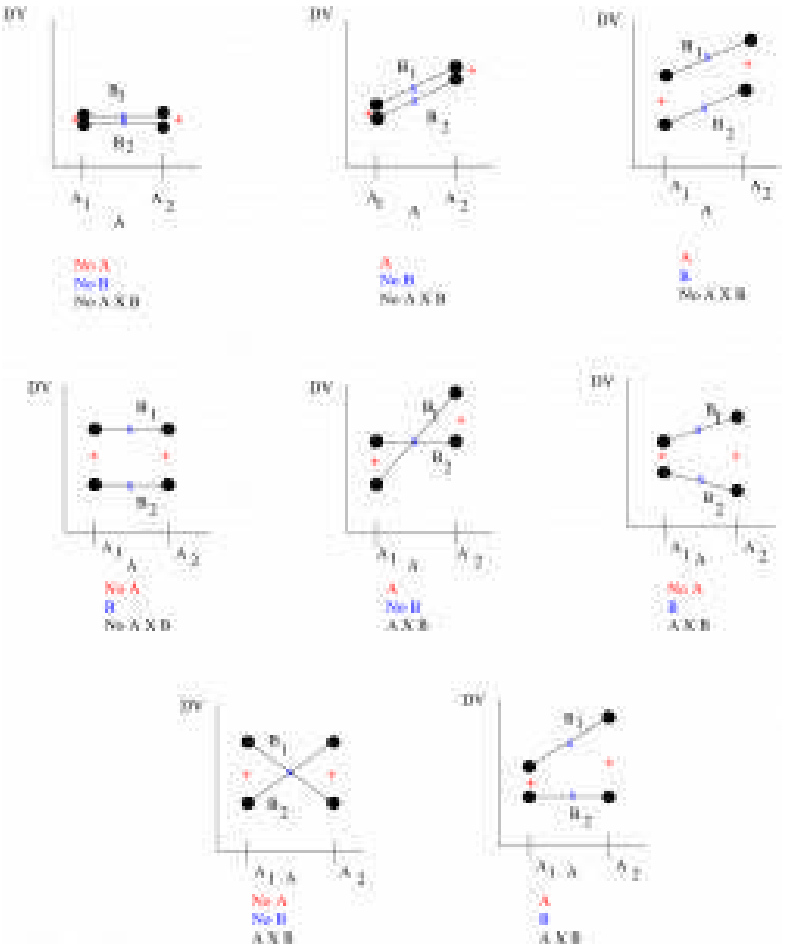
and then doing a one-way ANOVA with the one IV equal to B.

The hypothesis test for the interaction is a one-way ANOVA on the "difference of differences". The idea in interpreting a significant[3] interaction is that the effect of changing IV A *depends on* the effect of changing B. Let's see how the differences arise in a 2×2 ANOVA :



The interaction tests if there is a significant difference between the two differences $\Delta_1 = (\mu_{1,1} - \mu_{1,2})$ and $\Delta_2 = (\mu_{2,1} - \mu_{2,2})$. Note that I could have set up the differences on the profile plot with B on the x axis. It does not matter, the resulting one-way ANOVA turns out to be the same. For $A \times 2$ or $2 \times B$ ANOVAs you get more than two differences to compare with a one-way ANOVA. With a generic $A \times B$ ANOVA you need to take a mean of differences to compare to each other. The interpretation of a generic $A \times B$ can be tricky.

The interpretation of 2×2 interactions is, however, pretty straightforward. You need to consider all the possible outcome of the 2×2 ANOVA with its three hypotheses. Since any hypothesis can be significant or not we have $2^3 = 8$ possible outcomes[4]. Let's look at *generic* cases of all the combinations of the outcomes of a 2×2 ANOVA using A, B and $A \times B$ to denote that H_0 has been rejected and significant effects have been found :



The first thing to remember about these diagrams is that they are for *interpretation* — for step 5 of our hypothesis testing procedure. You have to do the actual hypothesis test with three F_{test} statistics to decide which case you have. Secondly, note that the graphs are *generic*. In statistics numbers are fuzzy. That is, every mean is fuzzy by a standard deviation. So think of the dots on the graphs as fuzzy balls and that the lines do not have to go to the centers of the fuzzy balls. Now look at the + and x symbols. The + symbols show what happens when you collapse the design over B to see the main effect of A; what is left are the two averages[5] for A_1 and A_2. The two + means are then compared with a one-way ANOVA (essentially a t-test since $t_\nu^2 = F_{1,\nu}$) to see if there is a main effect of A. Similarly the x symbols show what happens when you collapse the design across A to see the main effect of B. The means for B_1 and B_2 will be halfway[6] along the B_1 and B_2 lines. The two x means are then compared with a one-way ANOVA to see if there is a main effect of B. Finally lets look at the interactions. There are four cases in the diagrams that show interactions. In two cases the diagrams have crossed lines where the differences at either end are the negative of each other (and so are different) and in two cases the magnitudes of the differences are different. Looking at all the cases we see that there will be an interaction if the lines are not *statistically* parallel. The concept of statistically parallel is important here. Your actual data profile plot may not look like it has parallel lines but there will be no significant interaction if the lines are not statistically distinguishable from being parallel — this is the information that the hypothesis test gives you.

Before we move on, let's consider post-hoc testing for two-way ANOVAs. This usually means comparing means pairwise cell by cell. As with one-way ANOVA that would mean also finding a suitable correction for the p-value if t-tests are used. We won't cover post-hoc testing for two-way ANOVAs in any detail here except to point out that post hoc testing for a 2×2 ANOVA is redundant. A 2×2 ANOVA is essentially three t tests. If there is any interesting cell by cell difference, there will be an interaction. With a 2×2 ANOVA comparing cells is an interpretation problem, not one of statistical testing. The post-hoc test for a 2×2 ANOVA is really to figure out what generic profile plot matches your data.

Next, let's look at the ANOVA table for a two-way ANOVA. It looks like :

Source	Sum of Squares	Degrees of Freedom	Mean Square	F_{test}
A	SS_A	$v_A = a - 1$	MS_A	F_A
B	SS_B	$v_B = b - 1$	MS_B	F_B
$A \times B$	$SS_{A\times B}$	$v_{A\times B} = (a-1)(b-1)$	$MS_{A\times B}$	$F_{A\times B}$
Within (error)	SS_W	$v_W = ab(n-1)$	MS_W	
Totals	SS_T	N-1		

The two-way ANOVA table is very similar to the one-way ANOVA table except that there is now one line for each of the three hypotheses (three signals) plus a line that essentially quantifies the noise. We could also add another column for the p-value of the three effects. In the degrees of freedom formula, a is the number of levels for the A factor and b is the number of levels for the B factor. The formula for ν_W in the table is for a *balanced design* that has the same number, $n_{i,j} = n$, of data points in each cell (i, j). The total number of data points in a balanced design is $N = abn$. For a generic design, $N = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{i,j}$ and $\nu_W = \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{i,j} - 1)$.

The formulae in the other columns are the same for any ANOVA table: MS = SS/ν for each line, or effect, and F_{effect} =$\text{MS}_{\text{effect}}$/ MS_W. Explicitly:

$$\text{MS}_A = \frac{\text{SS}_A}{\nu_A} \qquad \text{MS}_B = \frac{\text{SS}_B}{\nu_B} \qquad \text{MS}_{A\times B} = \frac{\text{SS}_{A\times B}}{\nu_{A\times B}} \qquad \text{MS}_W = \frac{\text{SS}_W}{\nu_W} \tag{12.4.1}$$

and the F test statistics are

$$F_A = \frac{\text{MS}_A}{\text{MS}_W} \qquad F_B = \frac{\text{MS}_B}{\text{MS}_W} \qquad F_{A\times B} = \frac{\text{MS}_{A\times B}}{\text{MS}_W}. \tag{12.4.2}$$

For the critical statistics, which you look up in the ***F* Distribution Table**, the degrees of freedom to use are $\nu_A = a - 1$, $\nu_B = b - 1$, $\nu_{A\times B} = (a-1)(b-1)$ and $\nu_W = ab(n-1)$.

Now all we need are the formulae for the sums of squares. These sums of squares formulae, and the two-way ANOVA that you are responsible for in this class are for a *between subjects* design. That is, the samples for each cell are *independent*, every data point is from a different individual. We also assume homoscedasticity, $\sigma_{i,j}^2 = \sigma^2$ for all cells (i, j). Now to the SS formulae, we'll just give them for a balanced design[7]. To do this we need to label the data points this way: use $x_{i,j,k}$ where i and j label the cell ($(1 \le i \le a)$ and $(1 \le j \le b)$) and k labels the data point within the cell ($(1 \le k \le n)$)). First we define a "correction term", C, to keep the formulae simple:

$$C = \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{i,j,k}\right)^2 \tag{12.4.3}$$

With this, the formulae for the sums of squares in a balanced design two-way ANOVA are:

$$\mathrm{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{i,j,k}^2 - C$$

$$\mathrm{SS}_A = \frac{1}{bn}\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\left(\sum_{k=1}^{n} x_{i,j,k}\right)\right)^2 - C$$

$$\mathrm{SS}_B = \frac{1}{an}\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\left(\sum_{k=1}^{n} x_{i,j,k}\right)\right)^2 - C$$

$$\mathrm{SS}_{A\times B} = \frac{1}{n}\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{k=1}^{n} x_{i,j,k}\right)^2\right) - C - \mathrm{SS}_A - \mathrm{SS}_B$$

$$\mathrm{SS}_W = \mathrm{SS}_T - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_{A\times B}$$

Relax, you won't have to chug your way through these sums of squares formulae in an exam. That would be way too tedious even if you are comfortable with all those summation signs. But we will take a look at using them in an example where we set up cell diagrams and use marginal sums to help us along. On an exam, you will be able to simply read the values for the sums of squares from an SPSS ANOVA table output.

Example 12.4 : A researcher wishes to see whether the type of gasoline used and the type of automobile driven have any effect on gasoline consumption. Two types of gasoline, regular and high octane, will be used and two types of automobiles, two-wheel drive and four-wheel drive, will be used in each group. There will be two automobiles in each group for a total of eight automobiles used. The data, in cell form are (the DV is miles per gallon) :

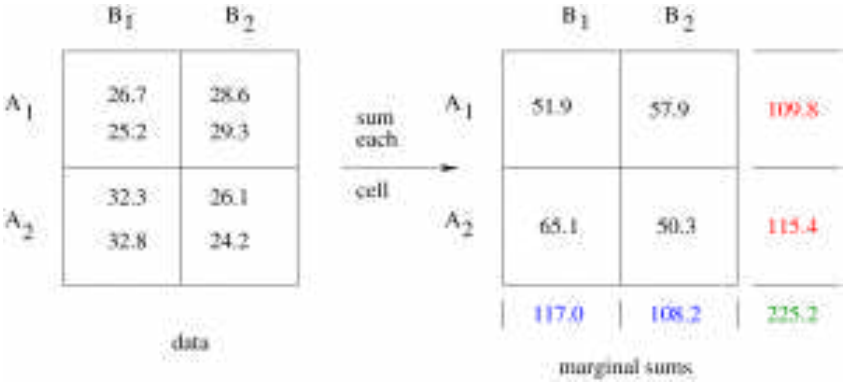
		Type of Automobile (B)	
		2-Wheel	4-Wheel
Gas (A)	Regular	26.7 25.2	28.6 29.3
	High Octane	32.3 32.8	26.1 24.2

Using a two-way ANOVA at $\alpha = 0.05$ test the effects of gasoline and automobile types on gas millage.

Solution :

0.Data Reduction.

Here we will calculate the sums of squares, SS_A, SS_B, $\mathrm{SS}_{A\times B}$ and SS_W (and SS_T) by hand using marginal sums. Again, in an exam you will be given the sums of squares. But we will see marginal sums again when we do χ^2 contingency tables in Chapter 15.



Beginning with the data on the left, sum each cell to give the numbers on the right. In summing each cell you are computing the terms $\sum_{k=1}^{n} x_{i,j,k} = \sum_{k=1}^{2} x_{i,j,k}$, $1 \le i, j \le 2$ in the sum of squares equations on page 234. (Note that these sums are n times the means, $\overline{x}_{i,j}$, of the cells, which is what the two-way ANOVA compares: $\sum_{k=1}^{n} x_{i,j,k} = n\overline{x}_{i,j}$.) Next, compute the marginal sums, the **sums of the rows**, on the far right, and the **sums of the columns**, on the bottom. Then compute the **grand sum**, the sum of everything, which is the the sum of the marginal sums on the right which equals the sums on the bottoms (which should be equal — a check).

The marginal sums show up in the second inner brackets in the sums of squares formula. Notice that the **sums of the rows** collapse the design across B to give a one-way ANOVA for A (main effect of A) and the **sums of the columns** collapse the design across A to give a one-way ANOVA for B (main effect of B). With the marginal sums we compute:

\[C = \frac{1}{N} \left(\sum \sum \sum x_{i,j,k} \right)^{2} = \frac{8} = 6339.38 \]

$$\begin{aligned} \text{SS}_T &= \sum\sum\sum x_{i,j,k}^2 - C \text{ (This one's not from the marginal sums.)} \\ &= (26.7^2 + 25.2^2 + 28.6^2 + 29.3^2 + 32.3^2 + 32.8^2 + 26.1^2 + 24.2^2) - 6339.38 \\ &= 6410.36 - 6339.38 \\ &= 70.98 \end{aligned}$$

$$\begin{aligned} \text{SS}_A &= \frac{1}{bn} \sum_{i=1}^{a} \left(\sum_{j=1}^{b} \left(\sum_{k=1}^{n} x_{i,j,k} \right) \right)^2 - C \\ &= \frac{1}{(2)(2)} \left[(109.8)^2 + (115.4)^2 \right] - 6339.38 \\ &= 6343.30 - 6339.38 \\ &= 3.92 \end{aligned}$$

$$\begin{aligned} \text{SS}_B &= \frac{1}{an} \sum_{j=1}^{b} \left(\sum_{i=1}^{a} \left(\sum_{k=1}^{n} x_{i,j,k} \right) \right)^2 - C \\ &= \frac{1}{(2)(2)} \left[(117.0)^2 + (108.2)^2 \right] - 6339.38 \\ &= 6349.06 - 6339.38 \\ &= 9.68 \end{aligned}$$

$$\begin{aligned} \text{SS}_{A\times B} &= \frac{1}{n} \left(\sum\sum \left(\sum x_{i,j,k} \right)^2 \right) - C - \text{SS}_A - \text{SS}_B \text{ (No marginal sums.)} \\ &= \frac{1}{2} \left(51.9^2 + 57.9^2 + 65.1^2 + 50.3^2 \right) - 6339.38 - 3.92 - 9.68 \\ &= 6407.06 - 6339.38 - 3.93 - 9.68 \\ &= 54.08 \end{aligned}$$

$$\begin{aligned} \text{SS}_W &= \text{SS}_T - \text{SS}_A - \text{SS}_B - \text{SS}_{A\times B} \\ &= 70.89 - 3.92 - 9.68 - 54.08 \\ &= 3.30 \end{aligned}$$

There. Now the sums of squares are ready for computing the test statistics. At this point you can start making your ANOVA table to keep track of your calculations. Here we'll see the ANOVA table at the last step.

1. Hypotheses.

H_0 : No main effect of A. (Changing gas type doesn't change mileage.)
H_1 : Main effect of A.

H_0 : No main effect of B. (Changing auto type doesn't change mileage.)
H_1 : Main effect of B.

H_0 : No interaction $A \times B$.
H_1 : Interaction $A \times B$. (The effect of gas type on mileage depends on auto type.)

2. Critical statistics.

There are three of them, one for each hypothesis pair. Use the *F* Distribution Table with the α labelling the table equal to the test $\alpha = 0.05$ since there are no such things as one and two tailed tests for ANOVA. From the ***F* Distribution Table** find:

For A:

$$\begin{aligned} \nu_1 &= a - 1 = 2 - 1 = 1 \text{ (d.f.N)} \\ \nu_2 &= ab(n-1) = (2)(2)(2-1) = 4 \text{ (d.f.D.)} \\ F_{\text{crit}} &= 7.71 \end{aligned}$$

For B:

$$\begin{aligned} \nu_1 &= b-1 = 2-1 = 1 \text{ (d.f.N)} \\ \nu_2 &= ab(n-1) = (2)(2)(2-1) = 4 \text{ (d.f.D.)} \\ F_{\text{crit}} &= 7.71 \end{aligned}$$

For $A \times B$:

$$\begin{aligned} \nu_1 &= (a-1)(b-1) = (2-1)(2-1) = 1 \text{ (d.f.N)} \\ \nu_2 &= ab(n-1) = (2)(2)(2-1) = 4 \text{ (d.f.D.)} \\ F_{\text{crit}} &= 7.71 \end{aligned}$$

The critical statistics are all the same for a 2×2 ANOVA ($\nu_1 = 1$ for all the hypotheses pairs — essentially three t-tests because $t_\nu^2 = F_{1,\nu}$). For bigger designs, the critical statistics will, in general, be different for each hypothesis pair.

3. Test statistics.

Use the sums of squares to compute:

$$\text{MS}_A = \frac{\text{SS}_A}{a-1} = \frac{3.920}{2-1} = 3.920 \tag{12.4.4}$$

$$\text{MS}_B = \frac{\text{SS}_B}{b-1} = \frac{9.680}{2-1} = 9.680 \tag{12.4.5}$$

$$\text{MS}_{A\times B} = \frac{\text{SS}_{A\times B}}{(a-1)(b-1)} = \frac{54.080}{(2-1)(2-1)} = 54.080 \tag{12.4.6}$$

$$\text{MS}_W = \frac{\text{SS}_W}{ab(n-1)} = \frac{3.300}{4} = 0.825 \tag{12.4.7}$$

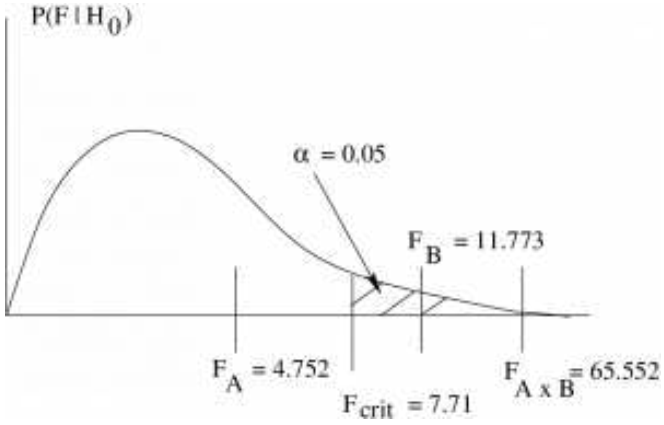
$$F_A = \frac{\text{MS}_A}{\text{MS}_W} = \frac{3.920}{0.825} = 4.752 \tag{12.4.8}$$

$$F_B = \frac{\text{MS}_B}{\text{MS}_W} = \frac{9.680}{0.825} = 11.773 \tag{12.4.9}$$

$$F_{A\times B} = \frac{\text{MS}_{A\times B}}{\text{MS}_W} = \frac{54.080}{0.825} = 65.552 \tag{12.4.10}$$

4. Decision.

In general, we need three diagrams, but in this case all the critical statistics are the same so we can draw :



So :

- For A, do not reject H_0, there is no main effect of A.
- For B, reject H_0, there is a main effect of B.
- For $A \times B$, reject H_0, there is an interaction.

5. Interpretation.

Simply put, at $\alpha = 0.05$ there is no effect of gas type (factor A) on mileage; there is an effect of auto type (factor B) on gas mileage and; there is an interaction between gas type and mileage, the change in mileage with auto type depends on the gas type used. We'll look at the profile plot to see if we really understand what this means but first we should complete the ANOVA table :

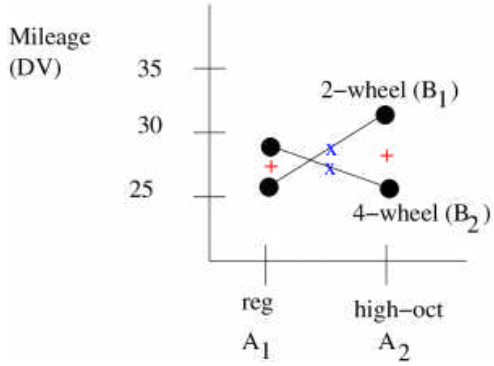
Source	Sum of Squares	Degrees of Freedom	Mean Square	F_{test}	p
A (gas)	3.92	1	3.92	F_A	$p > 0.05$
B (auto)	9.68	1	9.68	F_B	$p < 0.05$
$A \times B$	54.08	1	54.08	$F_{A \times B}$	$p < 0.05$
Within (error)	3.30	4	0.825		
Totals	70.98	7			

To draw the profile plot, we need to do one more data reduction :

		2–wheel B_1	4–wheel B_2
regular	A_1	25.95	28.95
high–octane	A_2	32.55	25.15

cell means

So the profile plot (without error bars — but remember the numbers are fuzzy) is :



To interpret this fully, remember the rules from previously about collapsing by looking at the midpoints between the two A values and the midpoint of the lines. Here we see that averaged over auto types, it looks like there is no difference in gas mileage between gas type. That conclusion is statistically confirmed by the fact that we found no main effect of factor A, gas — the two + values are not significantly different. The centres of the lines marked by the **x** are, however, significantly different because we found a main effect of B, auto type. And the nature of the statistically significant interaction is obvious, the gas mileage can go up or down when you change gas types depending on what kind of car you drive. Switching from regular gas to high octane gas will improve your mileage if you drive a 4-wheel drive car but the mileage will get worse if you drive a 2-wheel drive car.

□

You will see bigger designs than a 2×2. Collapsing across A or B in those larger designs to get to one-way ANOVAs is conceptually straightforward. The interaction is trickier, but the idea of an interaction existing when there are statistically non-parallel lines still holds. The 2×2 ANOVA is essentially three t-tests. This makes the 2×2 ANOVA powerful and easy to interpret.

As always, in statistics simpler is more powerful. We will take a brief quantitative look at statistical power in Chapter 13 but qualitatively, simpler is more powerful.

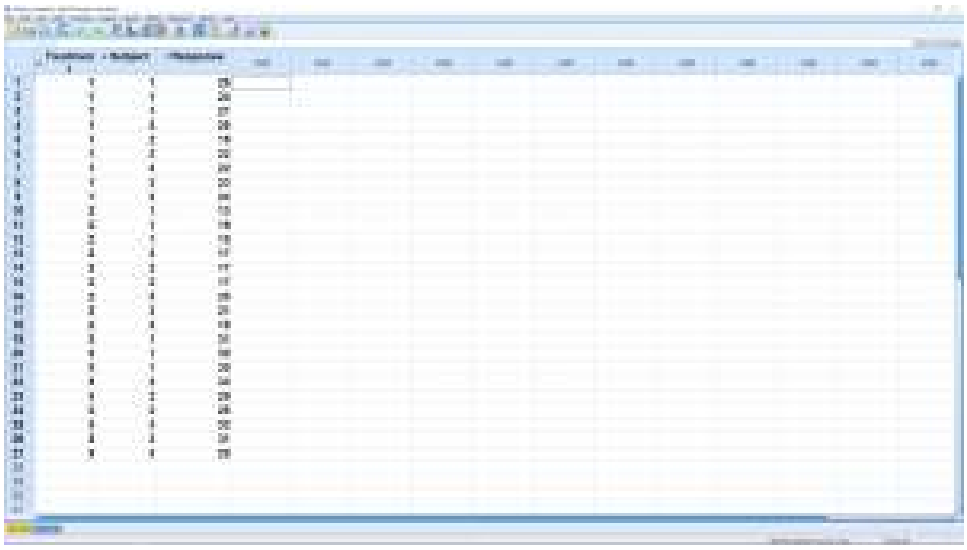
---

1. For the single sample t-test the two values of the IV were the population of interest and a hypothetical population representing H_0 having the mean k. ↵
2. Population values are used for this illustration. When you plot profile plots like this you will use sample means. ↵
3. Remember that "significant" means "reject H_0". ↵
4. Remember the counting rule! ↵
5. If the cell sizes, $n_{i,j}$, are all the same then the average is exactly halfway between the dots. ↵
6. For equal cell sizes. For unequal cell sizes the x will still be somewhere along the line. ↵
7. Of course, if you're using SPSS you don't need to restrict yourself to a balanced design. SPSS knows the generic SS formulae. ↵

This page titled 12.4: Two-way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

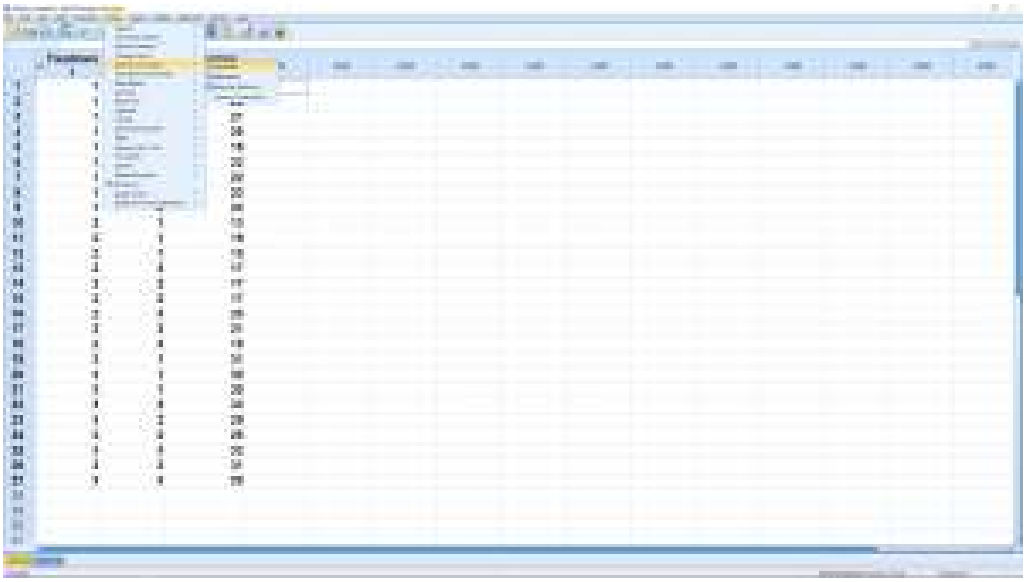
12.5: SPSS Lesson 9- Two-way ANOVA

From the Data Sets, open the file “Relief.sav” :



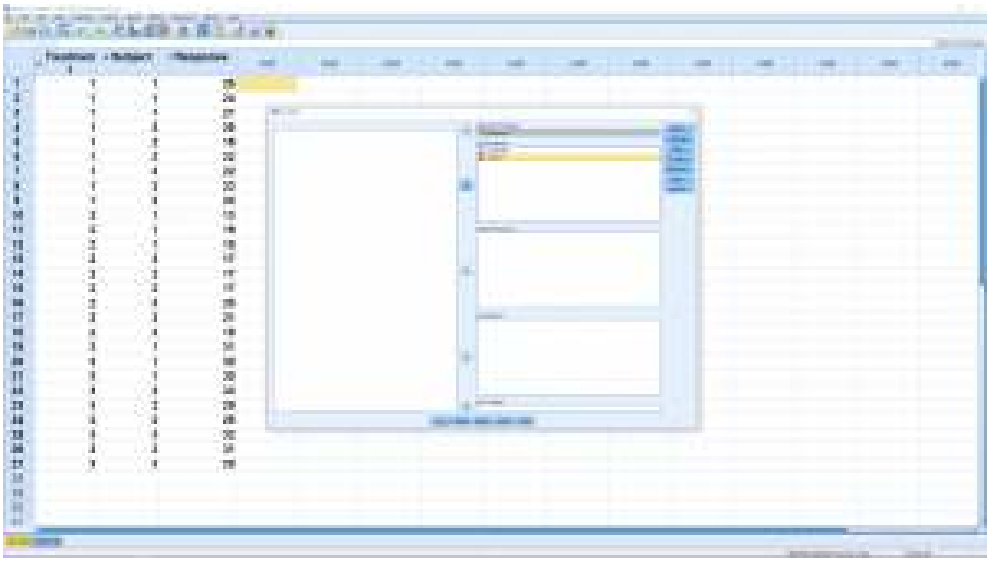
SPSS screenshot © International Business Machines Corporation.

Note that there are now two independent variables, Treatment and Subject; Treatment has 3 levels, Subject has 3 levels. The cell structure is somewhat hard to see so you will have to be organized when you enter data on your own. The dependent variable is Response. To run the ANOVA, select Analyze → General Linear Model → Univariate :



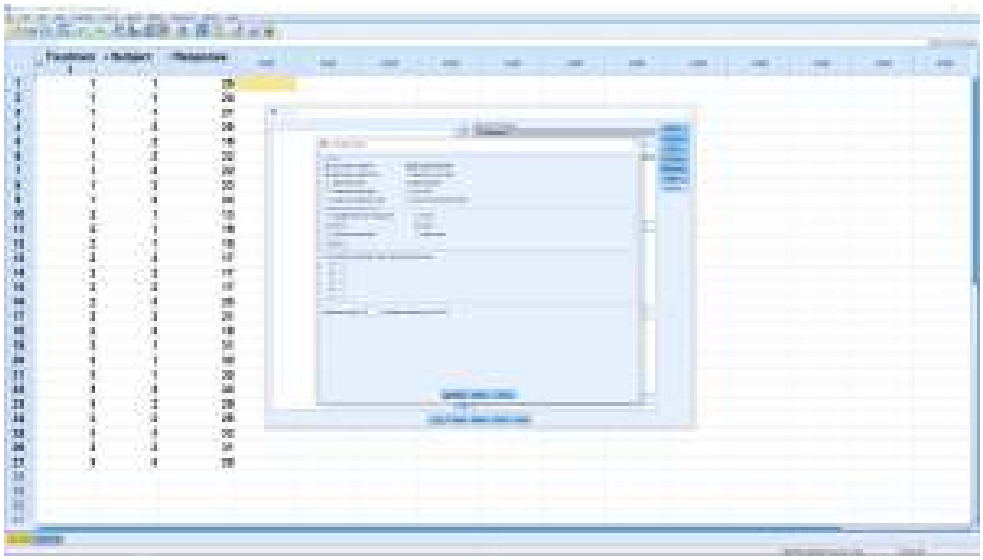
SPSS screenshot © International Business Machines Corporation.

which will give you this menu :



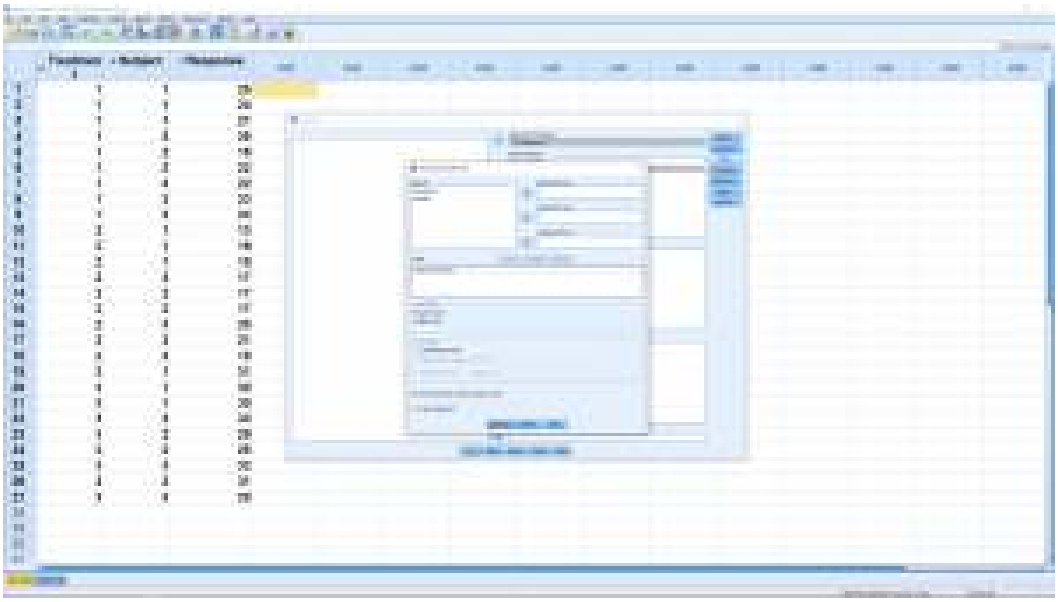
SPSS screenshot © International Business Machines Corporation.

where we have entered the independent variable, the factors, into the Fixed Factor box and the dependent variable into the dependent box. The submenus setups will be left pretty much alone, as with the one-way ANOVA. There are post-hoc tests available but we will not worry about that for this course. In the Model menu, there is a check box for “intercept” which, if checked, will result in an extra line in the ANOVA table output that we will need to ignore. We will look at the output, below, that is generated if that box is checked. In the Options menu, check off Descriptive statistics (this will give cell means), Estimates of effect size (this will give η^2) and Homogeneity tests (Levine’s test for homoscedasticity) :



SPSS screenshot © International Business Machines Corporation.

The Plots menu is where you set up the profile plot output. Recall that you can view the 3D profile plot from two directions: along the A factor axis with the B factor plotted as separate lines or along the B factor axis with the A factor as horizontal lines. Here is what the menu looks like just before you hit the Add button :



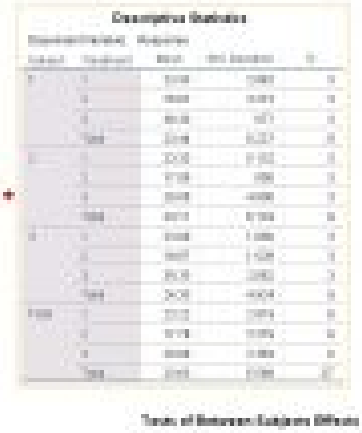
SPSS screenshot © International Business Machines Corporation.

Finally, hit the OK button to get the output. First comes the descriptive statistics where you can see the cell means and sample standard deviations :

Between-Subjects Factors

		N
Treatment	1	9
	2	9
	3	9
Subject	1	9
	2	9
	3	9

SPSS screenshot © International Business Machines Corporation.



The important ANOVA table output looks like :

Tests of Between-Subjects Effects

Dependent Variable: Response

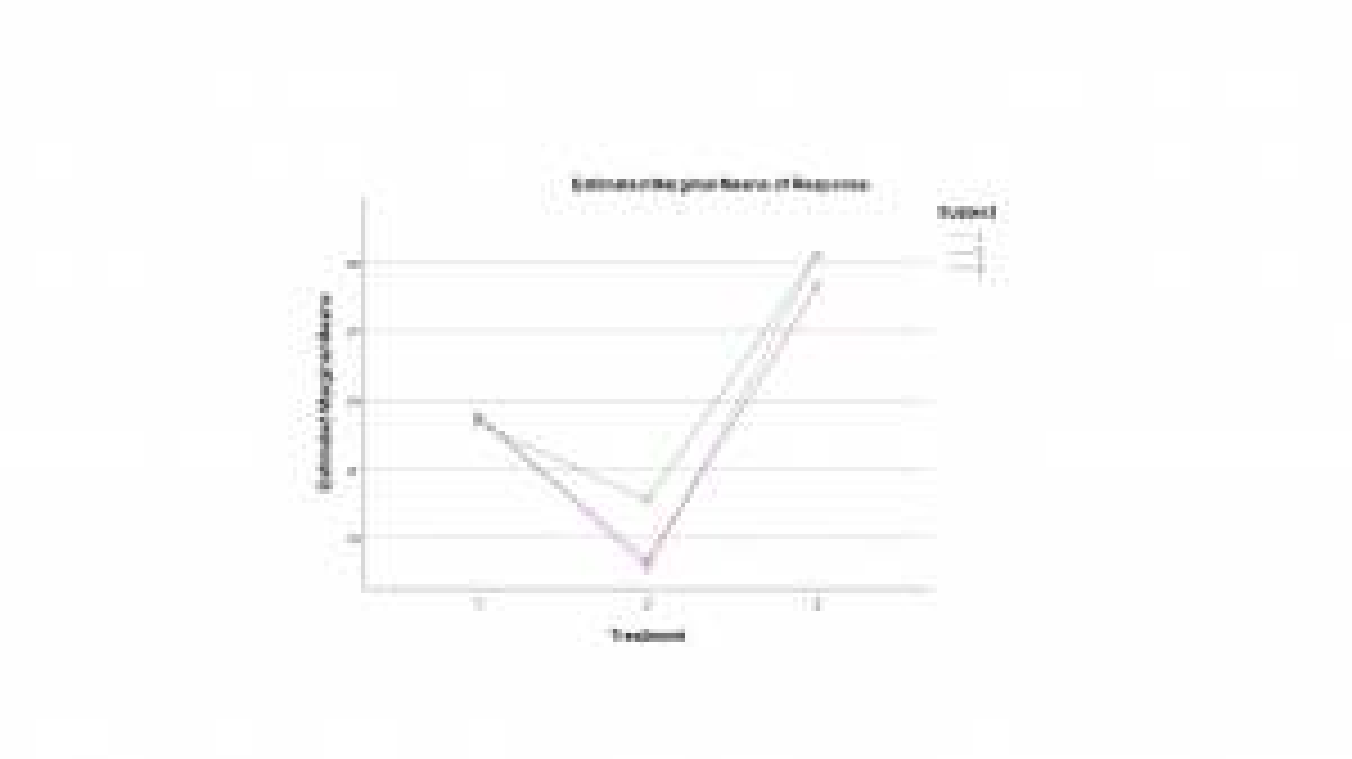
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	682.296[a]	8	85.287	11.808	.000
Intercept	15075.704	1	15075.704	2087.405	.000
Subject	7.185	2	3.593	.497	.616
Treatment	662.296	2	331.148	45.851	.000
Subject * Treatment	12.815	4	3.204	.444	.776
Error	130.000	18	7.222		
Total	15888.000	27			
Corrected Total	812.296	26			

a. R Squared = .840 (Adjusted R Squared = .769)

SPSS screenshot © International Business Machines Corporation.

As usual, ignore the Corrected Model, the Intercept and the Total source lines. Factor A is the gender source line, factor B is the method source line, treatment*group is the $A \times B$ interaction source and Error is the within variance source. The Corrected Total is the correct total of the A, B, $A \times B$ and error SS and degrees of freedom. Interpretation is the thing you want out of this so looking at the p values we see that there is no main effect of A, group, there is a main effect of B, method and there is an interaction. No A, B and $A \times B$. The η^2 of the ANOVA as a whole shows up on the corrected model line and is the same as the r^2 reported at the bottom of the table, $\eta^2 = r^2 = 0.840$; it is a measure of how well the data fit the group means – a measure of how well the data fit the ANOVA model. We'll explore the general linear ANOVA model in Chapter 17.

The profile plots come out as :

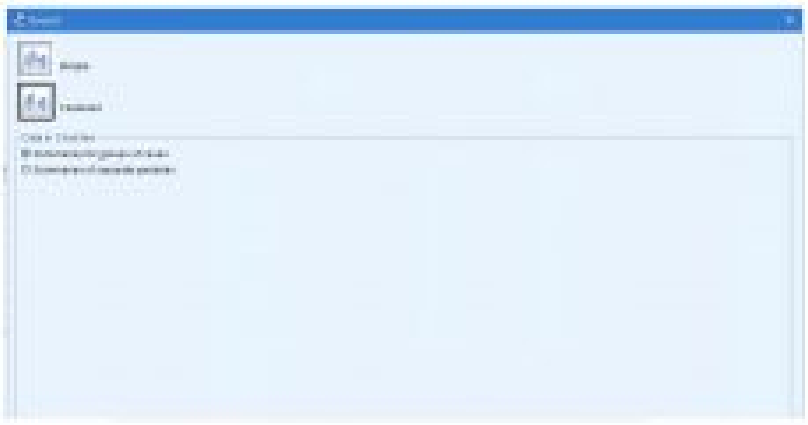


SPSS screenshot © International Business Machines Corporation.

Look at the two line plot on the left and we clearly see the interaction in the “non-parallel” lines. Look at that interaction in another way: look at the difference, the separation, between the group values for each of the 3 methods and image a profile plot of those values. A one-way ANOVA on those values (remember, this is what the interaction hypothesis test does) finds a significant difference; in particular the difference between groups is greater for group 2 than for the other groups.

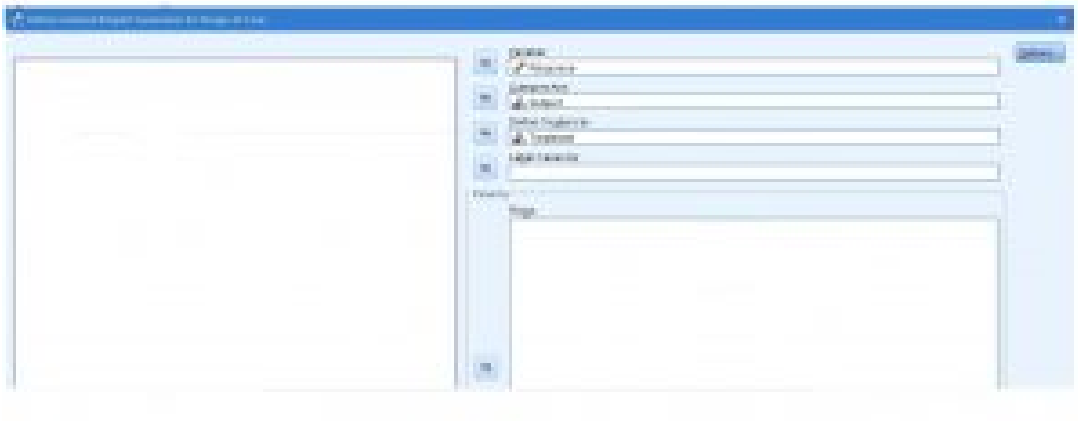
Plotting this the other way, as on the right above, we see the interaction manifest as non-parallel lines, but the difference of differences angle is harder to see. What you need to do to see it is, for each of the groups, look at the average difference of methods with the mean of the methods. There is a significant difference between the average difference value for the groups. The main effect of method shows up as a significant difference between the center of the three lines.

If you want to present a profile plot in a paper, you should show some error bars. Here’s how to get such a plot out of SPSS: first pick Graphs → Legacy Dialogs → Error Bars (pick boxplots to do boxplots). Then pick clustered with “Summaries are for groups of cases” :



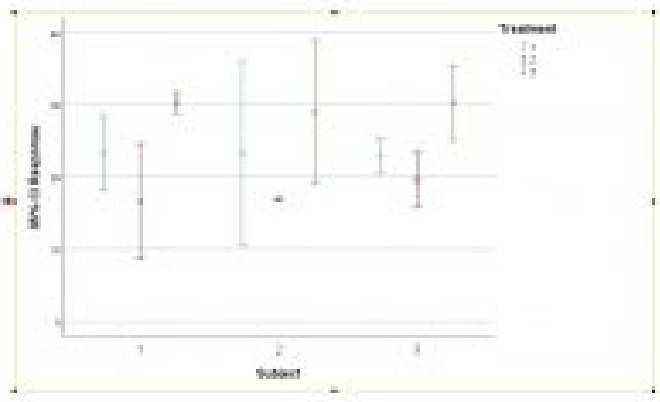
SPSS screenshot © International Business Machines Corporation.

Set up the plot as follows :



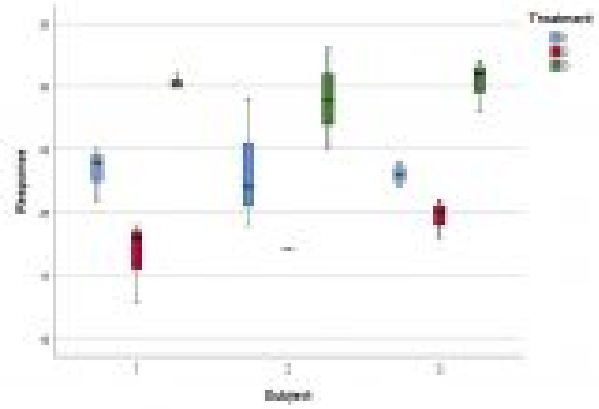
SPSS screenshot © International Business Machines Corporation.

The resulting plot is not that great (there's no way here to create the lines):



SPSS screenshot © International Business Machines Corporation.

The boxplot version looks a little better, at least there are clearer colors there to show the line factor:



SPSS screenshot © International Business Machines Corporation.

Compare this boxplot profile plot to the profile plot that came from running the two-way ANOVA.

This page titled 12.5: SPSS Lesson 9- Two-way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.6: Higher Factorial ANOVA

We've seen 1-way ANOVA and 2-way ANOVA but it doesn't have to stop there. We can have any number of factors, or independent variables. We can have 3-way ANOVA, 4-way ANOVA, etc. In general we can have an m-way ANOVA. An m-way ANOVA will have m IVs (m factors) but still only one DV.

12.8.1 3-way ANOVA

A 3-way ANOVA will have 3 factors (IVs): A, B, and C with a, b and c levels respectively. A 3-way ANOVA will test **7** hypotheses (all of which are one-way ANOVAs) :

1. Main effect of A (collapse across B and C).
2. Main effect of B (collapse across A and C).
3. Main effect of C (collapse across A and B).
4. 2-way interaction $A \times B$ (collapse across C).
5. 2-way interaction $A \times C$ (collapse across B).
6. 2-way interaction $B \times C$ (collapse across A).
7. 3-way interaction $A \times B \times C$.

So there will be 7 test statistics to consider:

$$F_A, \quad F_B, \quad F_C, \quad F_{A\times B}, \quad F_{A\times C}, \quad F_{A\times C}, \quad F_{B\times C}, \quad F_{A\times B\times C} \tag{12.6.1}$$

The profile plots for a 3-way ANOVA are intrinsically 4-dimensional and so can be difficult to draw. One approach is to make c 2-way style ANOVA plots :



The interpretation of a 3-way interaction can be tough and there will be many post-hoc pairwise comparisons of cells that may be meaningful. For these reasons it is best to be more reductionist in your experiment designs so that you never have to use a 3-way ANOVA. A design that uses preplanned *contrasts* is usually better than one that requires a 3 (or higher) way ANOVA.

For an m-way ANOVA, there will be

$$\binom{m}{1}\binom{m}{2}\cdots+\binom{m}{m}=\sum_{i=1}^{m}\binom{m}{i} \tag{12.6.2}$$

hypotheses to test, each with an associated F test statistic. The number of profile plots to consider will be large and will necessarily involve collapsing factors because the data exist in an $m+1$ dimensional space (number of IVs plus DV). Interpretation will be a nightmare. An m-dimensional ANOVA for $m \geq 3$ is more of a mathematical curiosity than a useful scientific tool.

This page titled 12.6: Higher Factorial ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.7: Between and Within Factors

So far, all of our factors have been *between subject*, or independent, factors. But it is possible to have any or all of the factors as *within subject*, or dependent, factors in a so-called *repeated measures* design. In a repeated measures design you will have more than one DV, more than one measurement from each subject. When you have more than one DV per subject, you have measured a *vector*[1] from each subject not just a number. When you measure a vector instead of a number you need multivariate statistics. The repeated measures approach is an approach that is between univariate and multivariate statistics. With repeated measures you set up your ANOVA as if it were a univariate design and use modified sums of squares and corresponding F test statistics. Certain assumptions need to be satisfied before you can do repeated measures ANOVA with the most important criteria being one known as "sphericity"[2]. If sphericity fails then you need to use the full-blown multivariate approach known as MANOVA (Multivariate ANOVA). If you use SPSS to do a within subjects ANOVA then you can use the sphericity hypothesis test output in the same way that you used the Levine's test output when deciding to use the homoscedastic or heteroscedastic t-test result from SPSS. Sphericity is H_0 so if SPSS fails to reject H_0 then you can use the repeated measures results. If $p < \alpha$ for the sphericity test then you reject H_0 and you will need to set up a MANOVA.

When you have a two-way (or higher factorial) ANOVA then *mixed designs* are possible where one factor is a between subjects factor and the other is a within subjects factor.

12.9.1 *One-way ANOVA with between factors

To be completed in a later edition of this text.

1. A vector is a collection of numbers. We will have more to say about vectors in Chapter 17. ↵
2. Sphericity will be covered in a later edition of this text in a MANOVA chapter. ↵

This page titled 12.7: Between and Within Factors is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

12.8: *Contrasts

To be completed in a later edition of this text.

This page titled 12.8: *Contrasts is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

13: Power

This page titled 13: Power is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

13.1: Power

Power is a concept that applies to all statistical testing. Here we will look at power quantitatively for the z-test for means (t-test with large n). We will see explicitly in that case some principles that apply to other tests. These principles are: the bigger your sample size (n), the higher the power; the larger α is, the more power there is[1]; the larger the "effect size" is the more power there is. A final principle, that we can't show by restricting ourselves to a z-test, is that the simpler the statistical test, the more power it has — being clever doesn't get you anywhere in statistics.

Let's being by recalling the "confusion matrix" (here labelled a little differently than the one shown in Chapter 9 to emphasize the decision making). Note: The α, β, etc. quantities are the probabilities that each conclusion will happen.

		Reality	
		H_0	H_1
Conclusion of Test	H_1	Type I error α	Correct decision $1-\beta$
	H_0	Correct decision $1-\alpha$	Type II error β

Recall that $1-\beta$ is the power, the probability of correctly rejecting H_0. With the definition of H_1 as not H_0, we cannot actually compute a power because this definition is too vague. The confusion matrix with H_0 and H_1 as given here is purely a conceptual device. To actually compute a power number we need to nail down a specific *alternate hypothesis* H_a and compute β for the more specific confusion matrix:

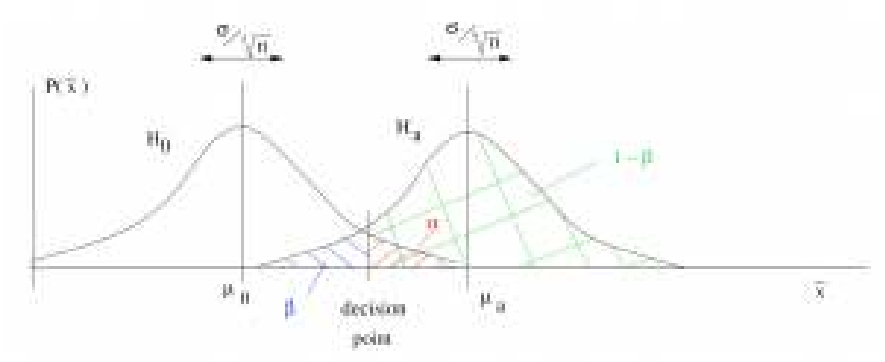
		Reality	
		H_0	H_a
Conclusion of Test	H_a	Type I error α	Correct decision $1-\beta$ **(power)**
	H_0	Correct decision $1-\alpha$	Type II error β

We will define H_0 and H_a be three parameters. The first is that we assume that the populations associated with H_0 and H_a both have the same standard deviation σ. Then, assuming that both populations are normal, H_0 is defined by its population mean μ_0 (we used k in Chapter 9) and H_a is defined by its population mean μ_a.

We can define two flavors of power :

1. **Predicted power.** Based on a *pre-defined* alternate mean μ_a of interest and an estimate of $\sigma/\sqrt{n}$. The population standard deviation σ is frequently estimated from the sample standard deviation s of a small pilot study.
2. **Observed power.** Based on the *observed* sample mean $\overline{x}$ which is then used as the alternate mean μ_a and sample standard deviation s which is used for σ.

The type II error rate β (and power $1-\beta$) is calculated by considering the populations associated with H_0 and H_1 :



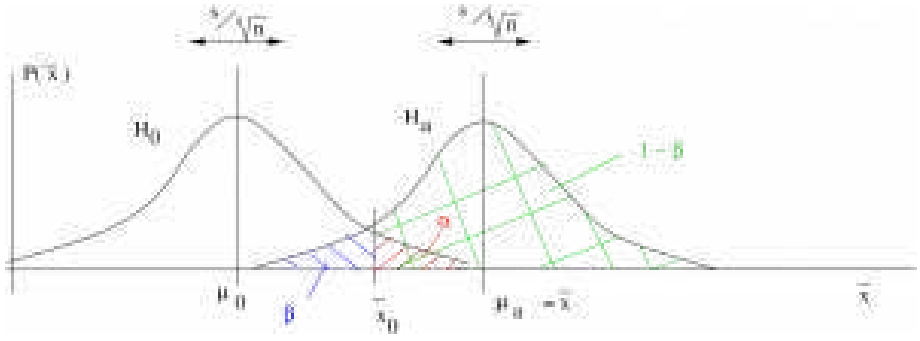
This picture follows directly from the Central Limit Theorem. Hypothesis testing is a decision process. In the picture above, which shows a one-tailed z-test for means, you reject H_0 if $\overline{x}$ falls to the right of the decision point. The decision point is set by the value of α. Note that the alternate mean μ_a needs to be in the rejection region of H_0 for the picture to make sense. The value of β (and hence the power $1-\beta$) depends on the magnitude of the effect size[2] $\mu_a - \mu_0$. We can see that power will increase if the effect size that we are looking for in our experiment increases. This makes sense because larger differences should be easier to measure. Also note that if n increases, as it would by replicating an experiment with a larger sample size, then the two distributions of sample means will get skinner and, for a given effect size, the power will increase. Again, this makes intuitive sense because more data is always better. We will illustrate these features in the numerical examples that follow.

For the purpose of learning the mechanics of statistical power we focus on **observed power**. With observed power we use the sample data for the power calculations; set $\mu_a = \overline{x}$ and $\sigma = s$. Since μ_a needs to be in the rejection region of H_0, observed power can only be computed when the conclusion of the hypothesis test is to reject H_0. In real life if you reject H_0 you don't care about what power the experiment had to reject H_0. It's a bit like calculating if you have enough gas to drive to Regina after you've arrived at Regina. In real life you will care about power only if you fail to reject H_0 because you will want to know the problem was that you tried to measure too small of an effect size or if a larger sample might lead to a decision to reject H_0. In that case you will need to decide what effect size, or sample size, to use in computing a predicted power. You will use predicted power in your experiment design. If your experiment design has a predicted power of about 0.80 then you have a reasonable chance of rejecting the null hypothesis. If your research involves invasive intervention with people (needles, surgery, etc.) then you may need to present a power calculation to prove to an ethics committee that your experiment has a reasonable chance of finding what you think it will find.

In addition to $\mu_a = \overline{x}$ and $\sigma = s$ we need the value of the decision point $\overline{x}_0$ which is the inverse z-transform of $z_\alpha = z_{\text{crit}}$. We'll consider three cases :

***Case 1*. Right tailed test:**

$H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0$ \mu_{0} \]" title="Rendered by QuickLaTeX.com">

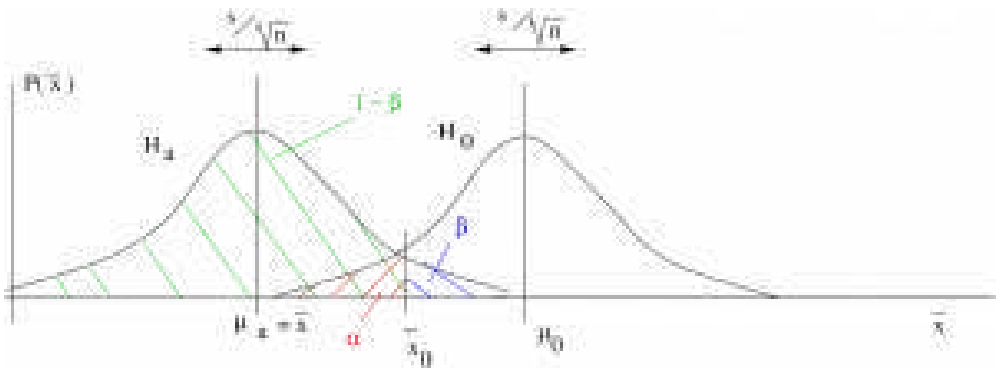


where, In this case

$$\overline{x}_0 = \mu_0 + z_\alpha \left(\frac{s}{\sqrt{n}} \right) \tag{13.1.1}$$

***Case* 2. Left tailed test:**

$$H_0 : \mu \geq \mu_0 \qquad H_1 : \mu < \mu_0 \tag{13.1.2}$$



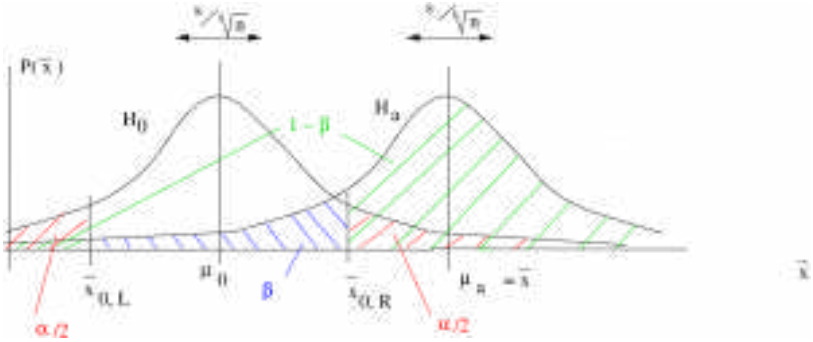
where, In this case

$$\overline{x}_0 = \mu_0 - z_\alpha \left(\frac{s}{\sqrt{n}} \right) \tag{13.1.3}$$

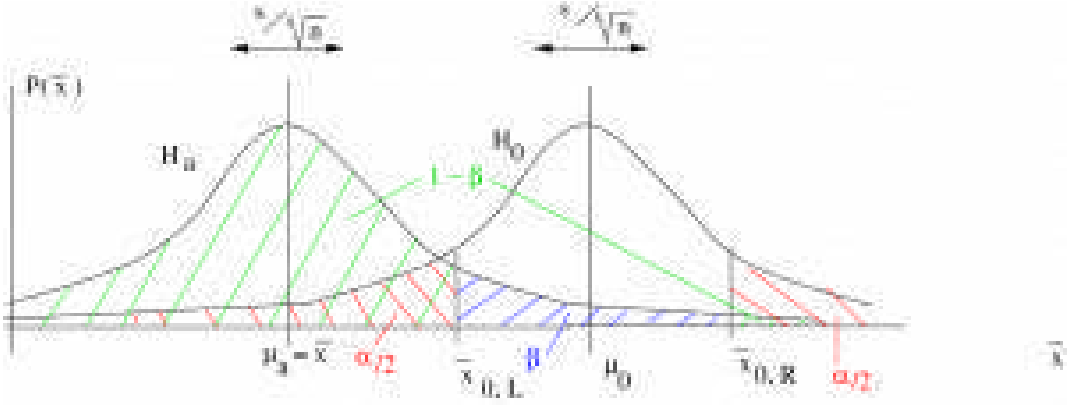
***Case* 3. Two-tailed test:**

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0 \tag{13.1.4}$$

(a) $\overline{x}$ in the right tail :



(b) $\overline{x}$ in the left tail:



where, in both cases:

Rendered by QuickLaTeX.com

In both two-tailed cases, notice the small piece of $1 - \beta$ area on the side of the H_a distribution on the opposite side from $\overline{x}$. It turns out that the area of that small part is so incredibly small that we can take it to be zero. This will be obvious she we work through the examples. So the upshot is that going from a one-tailed test to a two-tailed test effectively decreases α to $\alpha/2$ which increases β and decreases the power $1 - \beta$. One-tailed tests have more power than two-tailed tests for the same α.

Example 13.1 Right tailed test.

Given :

$image$ 150" title="Rendered by QuickLaTeX.com" height="17" width="179" style="vertical-align: -4px;">

$n = 50,\ \alpha = 0.05,\ s = 15,\ \overline{x} = \mu_a = 155$

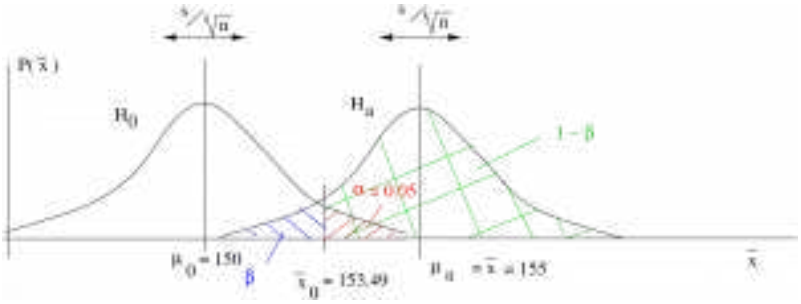
Find the observed power.

Step 1 : Look up $z_\alpha = z_{0.05}$ in the ***t* Distribution Table** for a one-tailed test: $z_\alpha = 1.645$.

Step 2 : Compute :

$$\begin{aligned}\overline{x}_0 &= \mu_0 + z_\alpha \left(\frac{s}{\sqrt{n}}\right) \\ &= 150 + (1.645)\left(\frac{15}{\sqrt{50}}\right) \\ &= 153.49\end{aligned}$$

Step 3 : Draw picture :



Step 4 : Compute the z-transform of $\overline{x}_0$ relative to H_a :

Rendered by QuickLaTeX.com

Step 5 : Look up the area $A(-z_a)$ in the **Standard Normal Distribution Table**. That area will be $0.5 - \beta$: $0.5 - \beta = 0.2611$, so $\beta = 0.5 - 0.2611 = 0.2389$ and **power** = $1 - \beta = 1 - 0.2389 =$ **0.7611**.

□

Example 13.2 : Another right tailed test with the data the same as in Example 13.1 but with a smaller α. This example shows how reducing α will reduce the power. With reduced power, it is harder to reject H_0.

Given :

$image$ 150" title="Rendered by QuickLaTeX.com" height="17" width="179" style="vertical-align: -4px;">

$n = 50,\ \alpha = 0.01,\ s = 15,\ \overline{x} = \mu_a = 155$

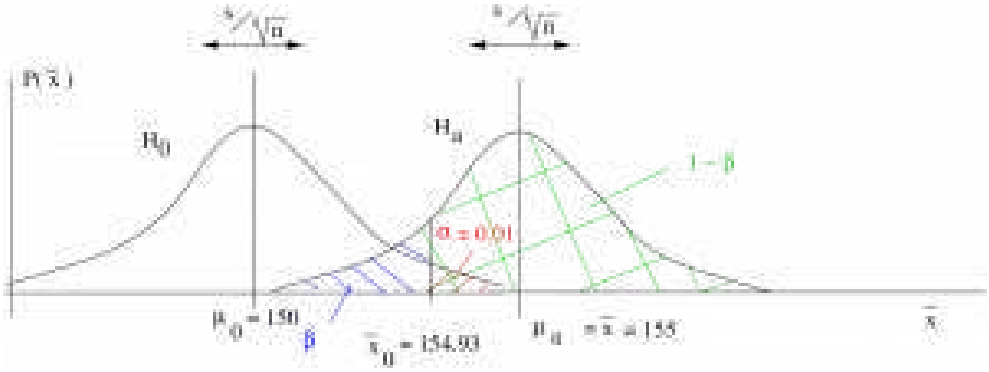
Find the observed power.

Step 1 : Look up $z_\alpha = z_{0.01}$ in ***t* Distribution Table** for a one-tailed test: $z_\alpha = 2.326$.

Step 2 : Compute :

$$\begin{aligned}\overline{x}_0 &= \mu_0 + z_\alpha \left(\frac{s}{\sqrt{n}}\right) \\ &= 150 + (2.326)\left(\frac{15}{\sqrt{50}}\right) \\ &= 154.93\end{aligned}$$

Step 3 : Draw picture :



Step 4 : Compute the z-transform of $\overline{x}_0$ relative to H_a :

Rendered by QuickLaTeX.com

Step 5 : Look up the area $A(0.03)$ in the **Standard Normal Distribution Table**. That area will be $A(0.03) = 0.0120$. So $\beta = 0.5 - 0.0120 = 0.4880$ and **power** = $1 - \beta =$ **0.5120** which is smaller than the power found in Example 13.1.

□

Example 13.3 : Another right tailed test with the data the same as in Example 13.2 but with larger n. This example shows how increasing the sample size increases the power. This makes sense because more data is always better.

Given :

$image$ 150" title="Rendered by QuickLaTeX.com" height="17" width="179" style="vertical-align: -4px;">

$n = 150$, $\alpha = 0.01$, $s = 15$, $\overline{x} = \mu_a = 155$

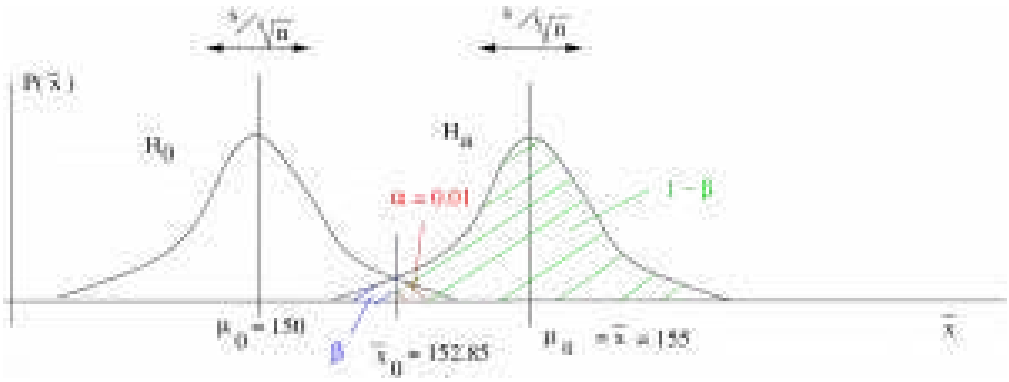
Find the observed power.

Step 1 : Look up $z_\alpha = z_{0.01}$ in ***t* Distribution Table** for a one-tailed test: $z_\alpha = 2.326$.

Step 2 : Compute :

$$\begin{aligned}\overline{x}_0 &= \mu_0 + z_\alpha \left(\frac{s}{\sqrt{n}}\right) \\ &= 150 + (2.326)\left(\frac{15}{\sqrt{150}}\right) \\ &= 152.85\end{aligned}$$

Step 3 : Draw picture :



Step 4 : Compute the z-transform of $\overline{x}_0$ relative to H_a :

Rendered by QuickLaTeX.com

Step 5 : Look up the area $A(0.03)$ in the **Standard Normal Distribution Table**. That area will be $A(0.03) = 0.4808$. So $\beta = 0.5 - 0.4808 = 0.0392$ and **power** = $1 - \beta =$ **0.9608** which is larger than the power found in Example 13.2.

□

Example 13.4 : Another right tailed test with the data the same as in Example 13.3 but with a smaller value for $\overline{x} = \mu_a$ which leads to a smaller effect size. This example shows how decreasing the effect size decreases the power. This makes sense because it is harder to detect a smaller signal.

Given :

$image$ 150" title="Rendered by QuickLaTeX.com" height="17" width="179" style="vertical-align: -4px;">

$n = 150$, $\alpha = 0.01$, $s = 15$, $\overline{x} = \mu_a = 153$

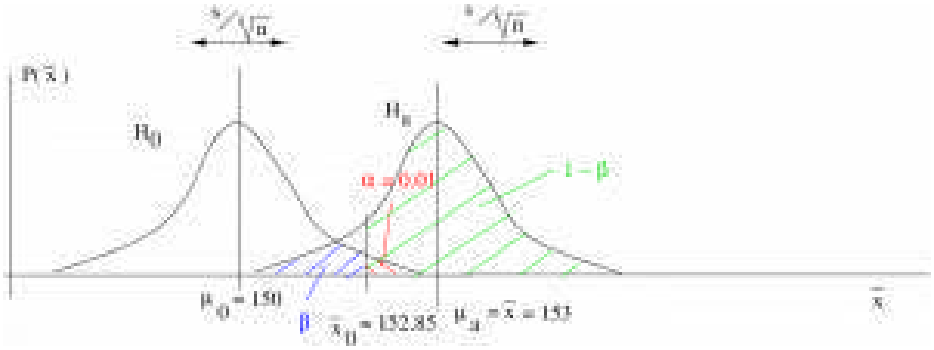
Find the observed power.

Step 1 : Look up $z_\alpha = z_{0.01}$ in the ***t* Distribution Table** for a one-tailed test: $z_\alpha = 2.326$.

Step 2 : Compute :

$$\begin{aligned}\overline{x}_0 &= \mu_0 + z_\alpha \left(\frac{s}{\sqrt{n}}\right) \\ &= 150 + (2.326)\left(\frac{15}{\sqrt{150}}\right) \\ &= 152.85\end{aligned}$$

Step 3 : Draw picture :



Step 4 : Compute the z-transform of $\overline{x}_0$ relative to H_a :

Rendered by QuickLaTeX.com

Step 5 : Look up the area $A(0.12)$ in the **Standard Normal Distribution Table**. That area will be $A(0.12) = 0.0478$. So $\beta = 0.5 - 0.0478 = 0.4522$ and **power** = $1 - \beta =$ **0.5478** which is smaller than the power found in Example 13.3.

□

Example 13.5 : Left tailed test.

Given :

$H_0 : \mu \geq 150$, $H_1 < 150$

$n = 50$, $\alpha = 0.05$, $s = 15$, $\overline{x} = \mu_a = 144$

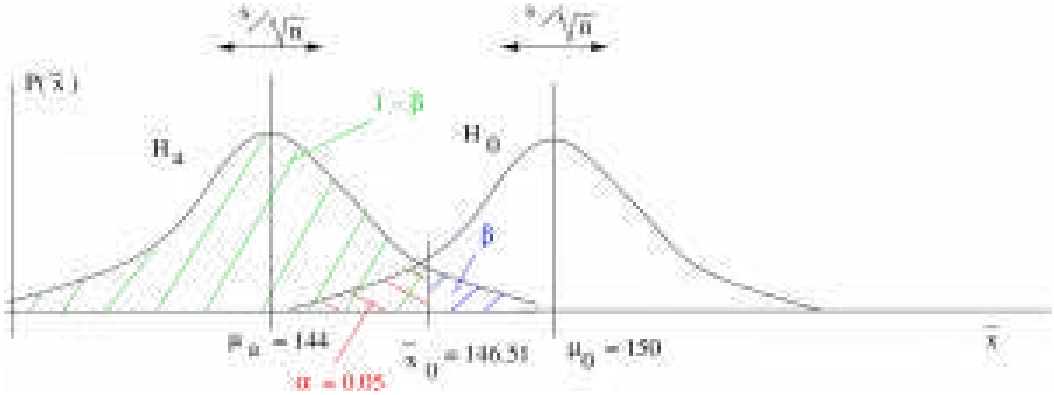
Find the observed power.

Step 1 : Look up $z_\alpha = z_{0.05}$ in the ***t* Distribution Table** for a one-tailed test: $z_{0.05} = 1.645$.

Step 2 : Compute :

$$\begin{aligned}\overline{x}_0 &= \mu_0 - z_\alpha \left(\frac{s}{\sqrt{n}}\right) \\ &= 150 - (1.645)\left(\frac{15}{\sqrt{50}}\right) \\ &= 146.51\end{aligned}$$

Step 3 : Draw picture :



Step 4 : Compute the z-transform of $\overline{x}_0$ relative to H_a :

Rendered by QuickLaTeX.com

Step 5 : Look up the area $A(1.18)$ in the **Standard Normal Distribution Table**. That area will be $A(1.18) = 0.3810$. So $\beta = 0.5 - 0.3810 = 0.1190$ and **power** = $1 - \beta =$ **0.8810**.

□

Example 13.6 : Two tailed z-test with data the same as Example 13.5.

Given :

$H_0 : \mu = 150$, $H_1 \neq 150$

$n = 50$, $\alpha = 0.05$, $s = 15$, $\overline{x} = \mu_a = 144$

Find the observed power.

Step 1 : Look up $z_{\alpha/2} = z_{0.025}$ in the ***t* Distribution Table** for a one-tailed test: $z_{0.025} = 1.960$.

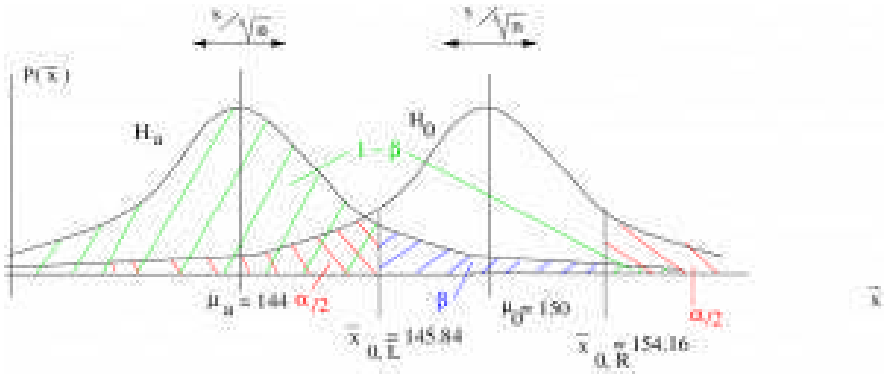
Step 2 : Compute:

Rendered by QuickLaTeX.com

and

Rendered by QuickLaTeX.com

Step 3 : Draw picture :



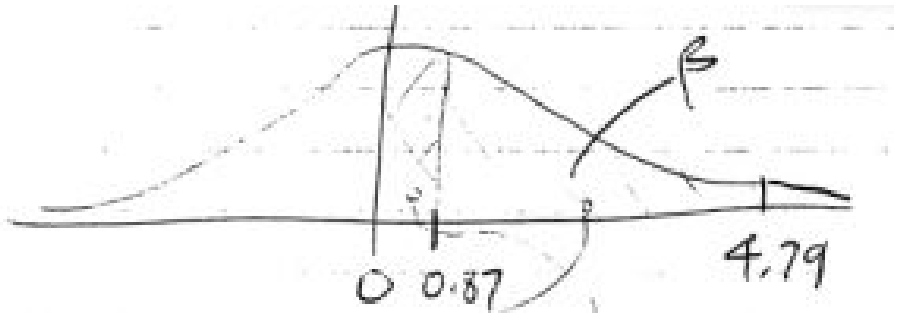
Step 4 : Compute the z-transform of $\overline{x}_{0,L}$ and $\overline{x}_{0,R}$ relative to H_a:

Rendered by QuickLaTeX.com

and

Rendered by QuickLaTeX.com

Step 5 : The two values, $z_{a,L}$ and $z_{a,R}$ appear on the z-distribution as :



So using the areas $A(z)$ from the **Standard Normal Distribution Table** we find

$$\beta = A(4.79) - A(0.87) = 0.5 - 0.3078 = 0.1922 \tag{13.1.5}$$

Notice that $z = 4.79$ is way the heck out there, it is higher than any z given in the **Standard Normal Distribution Table**. So $A(7.79)$ is essentially 0.5; the tail area past $z = 4.79$ is essentially zero. So the effect of going to from a one-tail to a two-tail test is only felt by the size of the H_0 critical region on the side where the test statistic ($\overline{x}$ here) is, which is half the size of the critical region in a one-tail test for a fixed α. In this case, then, the **power** = $1 - \beta =$ **0.8078** which is smaller than the value found in Example 13.5.

□

Using observed power

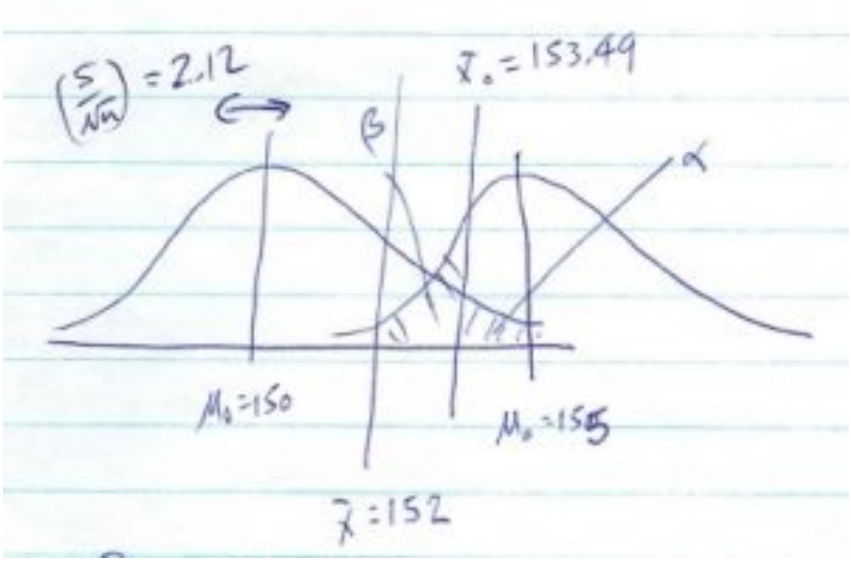
As mentioned earlier, almost no one is interested in observed power because we must reject H_0 to compute it. People are interested in β and power only when you report a failure to reject H_0.

Suppose in the situation of Example 13.1 we wanted to find evidence that $\mu_a = 155$ but measured $\overline{x} = 152$ (fail to reject H_0). Then, with our given information of

$H_0 : \mu \leq 150$, *image* 150" title="Rendered by QuickLaTeX.com" height="16" width="73" style="vertical-align: -3px;">

$n = 50$, $\alpha = 0.05$, $s = 15$, $\overline{x} = 152$ and $\mu_a = 155$

we have



Based on the calculation we did in Example 13.1 we would report that we had a power of 0.7611 to detect an effect of $\mu_a = 155$ but with $\overline{x} = 152$ we were unable to detect μ_a.

---

1. And a corollary of this will be that one-tailed tests are more powerful than two-tailed tests. ↵
2. Effect size as defined in the Green and Salkind SPSS book would be $\mu_a - \mu_0/\sigma$. But that quantity is not useful here, so we define effect size as the difference of the means for the purpose of this discussion on power. Reference: Green SB, Salkind NJ. *Using*

SPSS for Windows and Macintosh: Analyzing and Understanding Data, new edition pretty much every year, Pearson, Toronto, circa 2005. ↵

This page titled 13.1: Power is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14: Correlation and Regression

This page titled 14: Correlation and Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.1: Scatter Plots

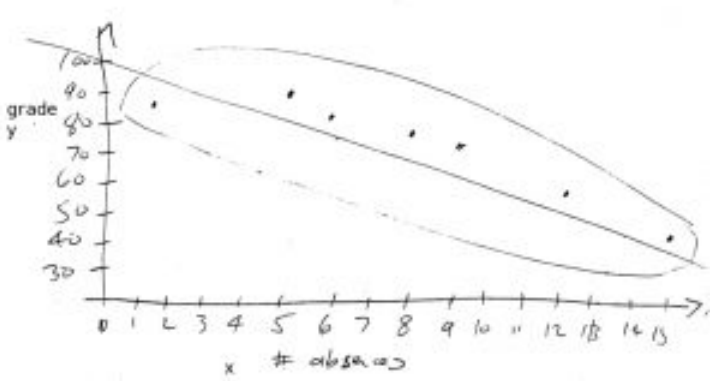
You can make a scatter plot of your data when you have values for two or more variables for each subject. Here we will only be interested in the case where we have a pair of variables (2D plot).

Of the two variables, for application to regression, one will be an independent variable (IV) and the other a dependent variable (DV). The IV is usually a variable that is known with a high degree of precision (like age). The idea with regression (when we get to it) is to come up with a formula that allows you to predict what the DV will be if you know the IV. We will use the symbol x for the IV and y for the DV.

The best way to see what a scatter plot is is to plot one. With the data:

Student	No. of absences, x	grade, y
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

the scatterplot is:



A couple of things to notice in the plot are: 1. An eyeball best line fit has been drawn through the scatterplot points. With regression we will calculate exactly what that best fit line is. 2. If x and y are ***linearly related*** then the points will fall inside an ellipse. If the ellipse is long and skinny, x and y are said to to be highly correlated. If the ellipse is more like a circle the x and y are not correlated. By looking at a scatter plot you can judge if x and y are linearly related. If your scatterplot looks like:



then you could conclude that x and y are not linearly related and it will not make much sense to try and fit a line through the data or to compute a correlation coefficient.

This page titled 14.1: Scatter Plots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

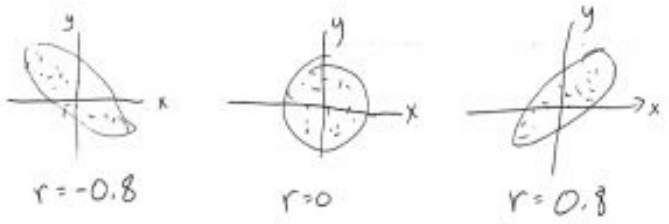
14.2: Correlation

The correlation coefficient we will use here is called the “Pearson product moment correlation coefficient” and will be represented by the following symbols :

ρ — population correlation

r — sample correlation

The correlation is always a number between -1 and $+1$: $-1 \leq r \leq +1$ and $-1 \leq \rho \leq +1$. If r (or ρ) equals 0 then that means there is *no correlation* between x and y. A minus sign means a minus slope, a plus sign means a positive slope.



The formula for r is[1] :

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Example 14.1 : Compute the correlation between x and y for the data on Section 14.1 used for the scatter plot.

Solution : To compute r, first make a table, fill in the data columns (on the right of the double vertical line below), fill in the other computed columns, sum the columns and finally plug the sums into the formula for r :

Subject	x	y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	86	172	4	7396
C	15	43	645	225	1849
D	9	74	666	81	5476
E	12	58	696	144	3364
F	5	90	450	25	8100
G	8	78	624	64	6084
$n = 7$	$\sum x = 57$	$\sum y = 511$	$\sum xy = 3745$	$\sum x^2 = 579$	$\sum y^2 = 38993$

Plug in the numbers :

$$\begin{aligned} r &= \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{7(3745) - (57)(511)}{\sqrt{[7(579) - (57)^2][7(38993) - (511)^2]}} \\ &= -0.944 \end{aligned}$$

Here there is a strong negative relationship between x and y. That is, as x goes up, y goes down with a fair degree of certainty. Note the r is **not** the slope. All we know here, from the correlation coefficient, is that the slope is negative and the scatterplot ellipse is long and skinny.

▢

Standard warning about correlation and causation : If you find that x and y are highly correlated (i.e. r is close to $+1$ or -1) then you cannot say that x causes y or that y causes x or that there is and causal relationship between x and y at all. In other words, it is true that if x causes y or that y causes x then x will be correlated with y but the reverse implication does not logically follow. So beware of looking for relations between variables by looking at correlation alone. Simply finding correlations by themselves doesn't prove anything.

The significance of r is assessed by a ***hypothesis test*** of

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0 \tag{14.2.1}$$

To test this hypothesis, you need to convert r to t via:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{14.2.2}$$

and use $\nu = n - 2$ to find t_{crit}. The **Pearson Correlation Coefficient Critical Values Table** offers a shortcut and lists critical r values that correspond to the critical t values.

Example 14.2 : Given $r = 0.897$, $n = 6$ and $\alpha = 0.05$, test if r is significant.

Solution :

1. Hypothesis. $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$

2. Critical statistic.

From the ***t* Distribution Table** with $\nu = n - 2 = 6 - 2 = 4$ and $\alpha = 0.05$ for a two-tailed test find

$$t_{\text{crit}} = \pm 2.776 \tag{14.2.3}$$

As a short cut, you can also look in the **Pearson Correlation Coefficient Critical Values Table** for $\alpha = 0.05$, $\nu = 4$ to find the corresponding

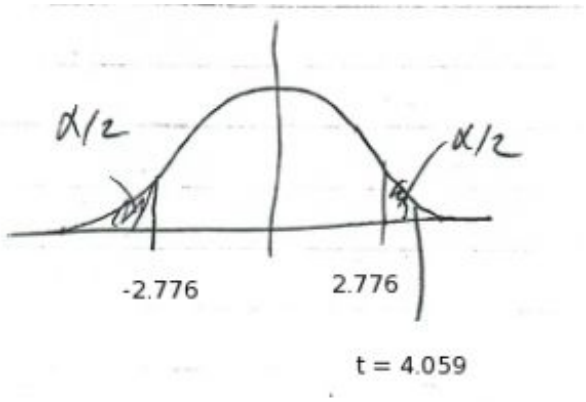
$$r_{\text{crit}} = \pm 0.811 \tag{14.2.4}$$

3. Test statistic.

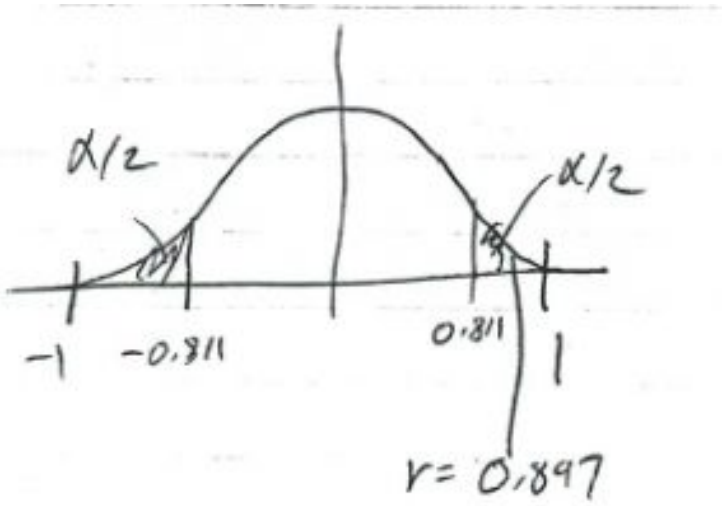
$$t_{\text{test}} = r\sqrt{\frac{n-2}{1-r^2}} = 0.897\sqrt{\frac{6-2}{1-(0.897)^2}} = 4.059 \tag{14.2.5}$$

4. Decision.

Using the t :



or using the **Pearson Correlation Coefficient Critical Values Table** short cut :



we conclude that we can reject H_0.

5. Interpretation. The correlation is statistically significant at $\alpha = 0.05$.

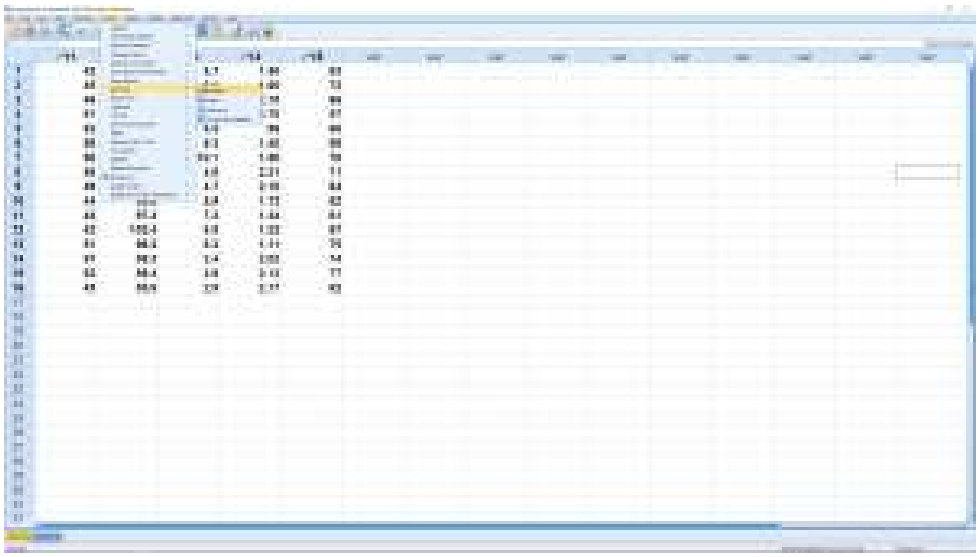
□

1. The formula for ρ is the same with all x and y in the population used. ↵

This page titled 14.2: Correlation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

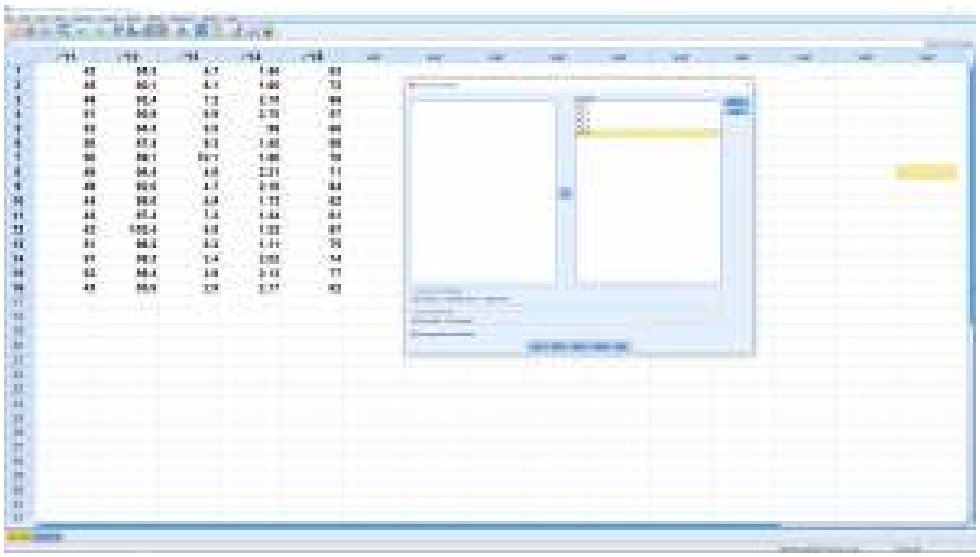
14.3: SPSS Lesson 10- Scatterplots and Correlation

Open “Hypertension.sav” from the Data Sets and pick Analyze → Correlate → Bivariate:



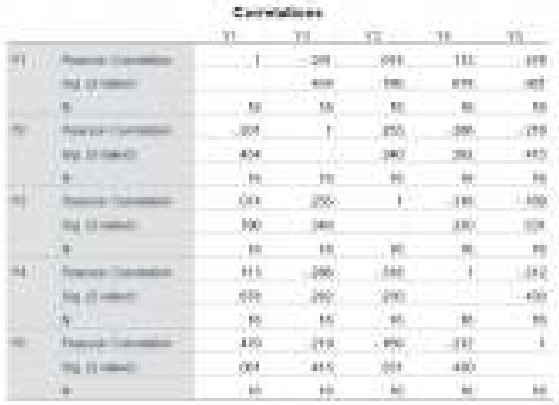
SPSS screenshot © International Business Machines Corporation.

In the menu that pops up, move all the variables over:



SPSS screenshot © International Business Machines Corporation.

and hit OK to get the following output:



SPSS screenshot © International Business Machines Corporation.

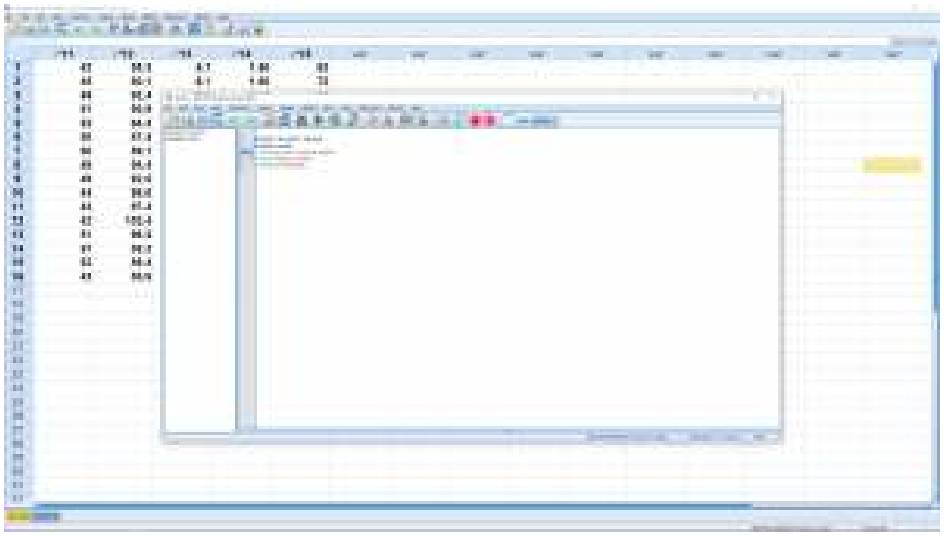
This result, when you just look at the Pearson correlation coefficients, is a correlation matrix. Specifically, the correlation matrix for these four variables, looking at the SPSS output is:

$$\begin{bmatrix} 1 & -0.201 & 0.074 & 0.113 & 0.479 \\ -0.201 & 1 & 0.255 & -0.286 & -0.219 \\ 0.074 & 0.255 & 1 & -0.318 & -0.169 \\ 0.113 & -0.286 & -0.318 & 1 & -0.212 \\ 0.479 & -0.219 & -0.169 & -0.212 & 1 \end{bmatrix} \tag{14.3.1}$$

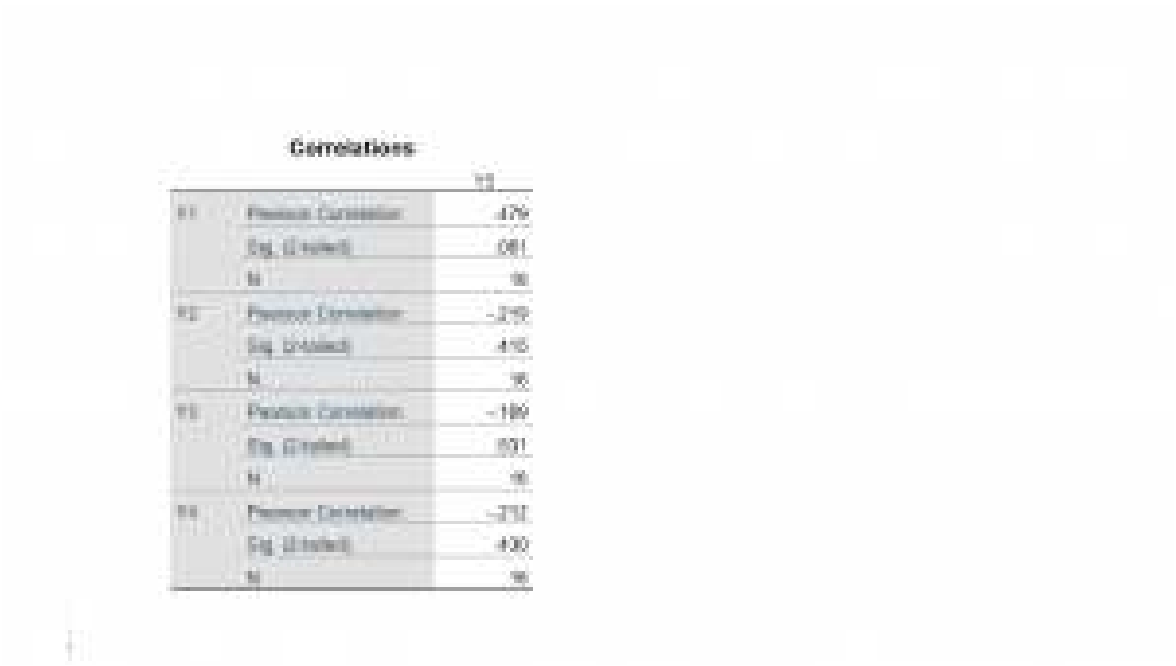
Note that the correlation matrix has ones on the diagonal — a variable is perfectly correlated with itself. The matrix is also symmetric which means that the numbers above the ones are the same as the ones directly across below the ones — the correlation between a and b is the same as the correlation between b and a. We'll be introduced to matrices more systematically in Chapter 17. The correlation matrix is at the heart of multivariate statistics in a way that standard deviation is at the heart of univariate statistics.

Other thing to notice in the SPSS output is the significance of the correlation coefficients. This significance is determined using the t statistic given in Section 14.2. SPSS puts ** by r values that have $p < 0.01$ and a * by those correlations with $p < 0.05$. The p-values themselves are also given in the SPSS output.

Sometimes you will not be interested in the complete correlation matrix but only in the correlations of one group of variables with another group. For example here we may want to lump the variables academic common friend and intimate together and see what their correlations are with the general variable. To get the associated partial correlation matrix, open the Analyze → Correlate → Bivariate dialog again, move all the variables over (if they are not already there) and hit Paste instead of OK. That will bring up the syntax editor. In the /VARIABLES line, add the word "with" between academic and general as shown:



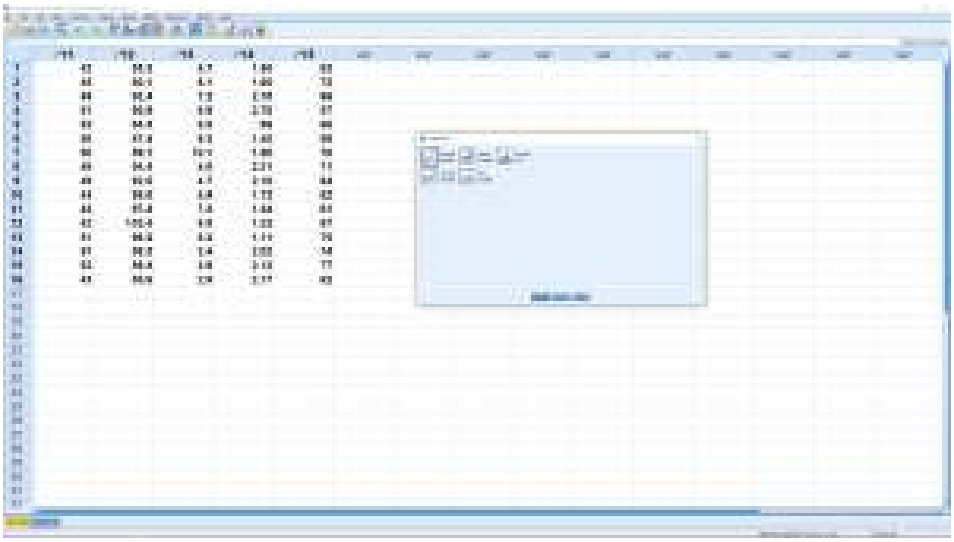
SPSS screenshot © International Business Machines Corporation.



SPSS screenshot © International Business Machines Corporation.

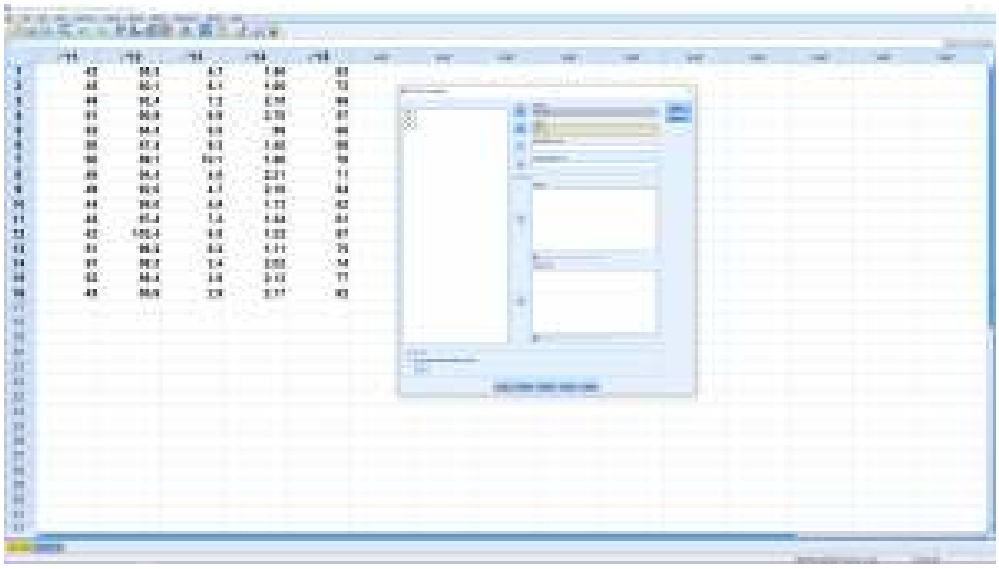
Then hit the big green triangle (“run”) to get:

Next, let’s do some scatterplots. First, a simple scatterplot of two variables. Pick Graphs → Legacy dialogs → Scatter/Dot to get:



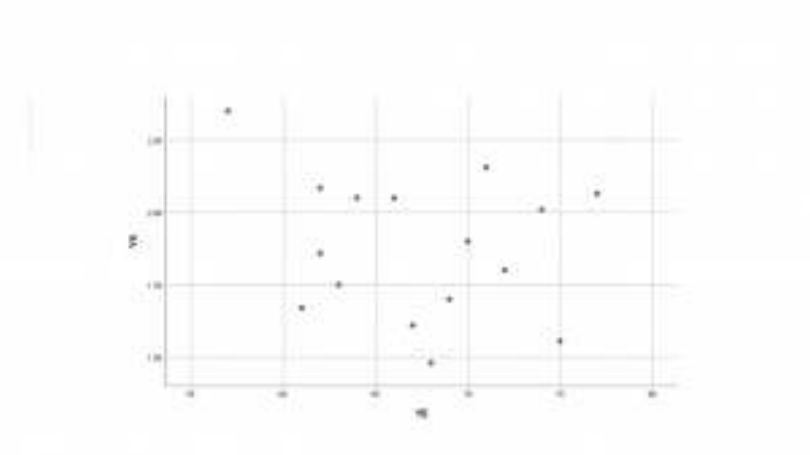
SPSS screenshot © International Business Machines Corporation.

where we pick Simple. This gives:



SPSS screenshot © International Business Machines Corporation.

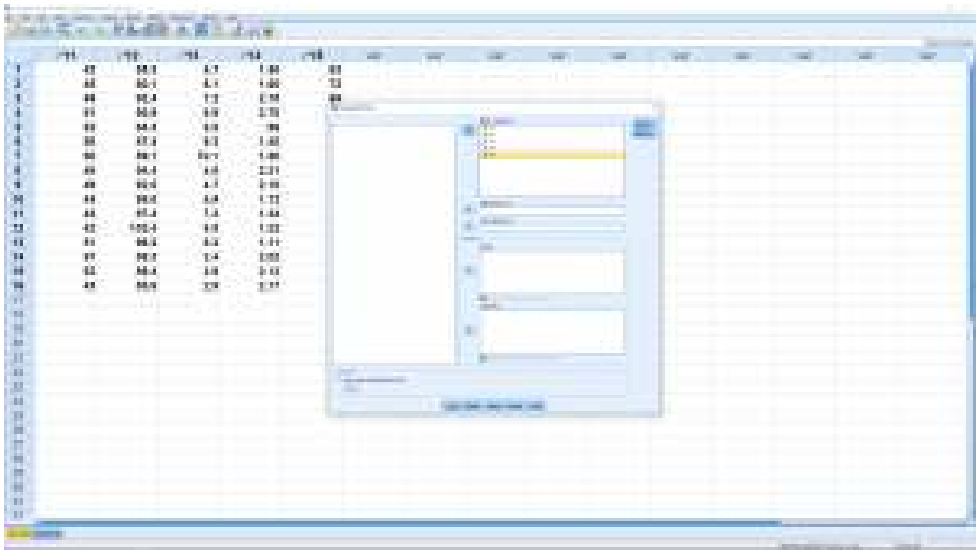
where two variables have been picked for plotting. (Note that if we had a variable with subject names, we could move that variable into the labels slot and get scatter plots with each point labeled by the subject name.) The result, after hitting OK, is:



SPSS screenshot © International Business Machines Corporation.

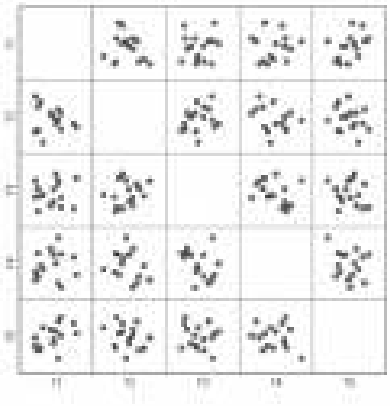
The correlation matrix output above reported that the correlation between these two variables was not significant and it does not appear that the points in the scatterplot are contained in a longish ellipse.

Instead of picking Simple in the graph pop up menu, pick Matrix Scatter and move all of the variables over for analysis in the menu that pops up after that:



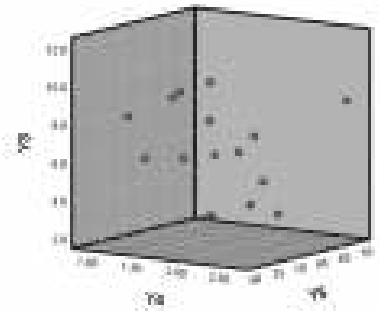
SPSS screenshot © International Business Machines Corporation.

The result is a matrix of scatterplots:

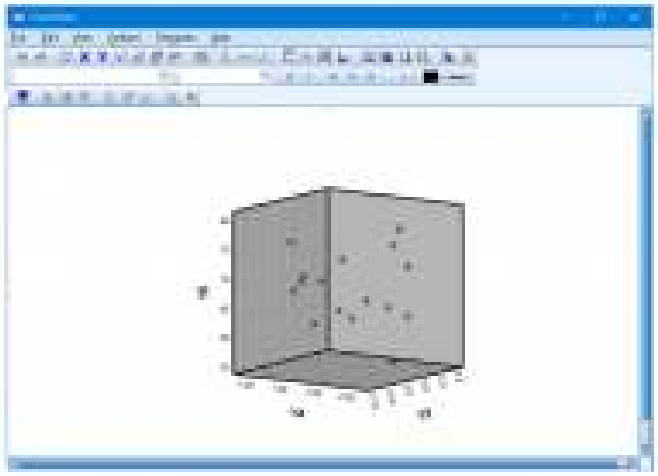


SPSS screenshot © International Business Machines Corporation.

Finally, for fun, pick 3D scatter in the first pop up menu and then pick three variables to get:



SPSS screenshot © International Business Machines Corporation.



SPSS screenshot © International Business Machines Corporation.

If you double click on the graphic, it will pop out into a Chart Editor and you can select a 3D rotation icon at the top (the little helper pop ups can help you find it):

After you click the 3D rotation icon at the top, you can grab the 3D plot with the mouse and rotate it around.

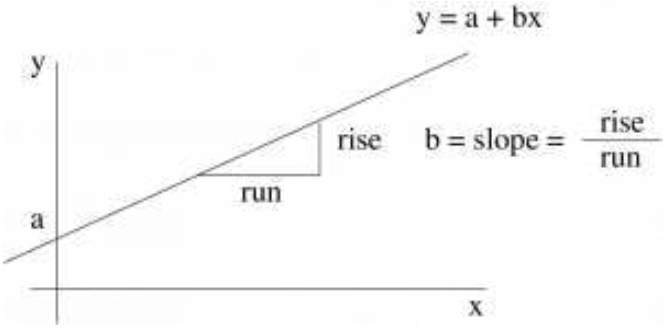
This page titled 14.3: SPSS Lesson 10- Scatterplots and Correlation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.4: Linear Regression

Linear regression gives us the best equation of a line through the scatter plot data in terms of *least squares*. Let's begin with the equation of a line:

$$y = a + bx \tag{14.4.1}$$

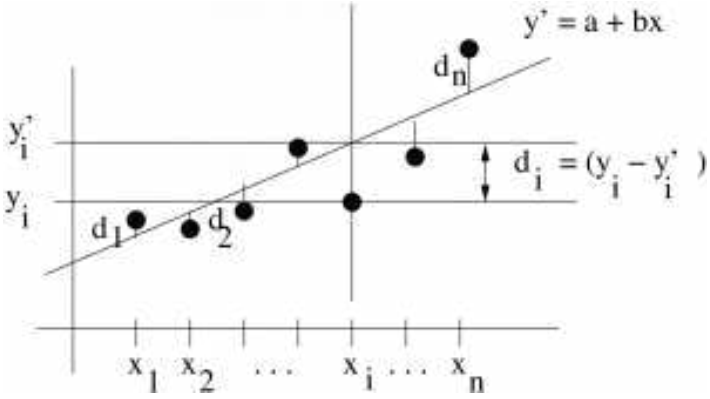
where a is the intercept and b is the slope.



The data, the collection of (x, y) points, rarely lie on a perfect straight line in a scatter plot. So we write

$$y' = a + bx \tag{14.4.2}$$

as the equation of the best fit line. The quantity y' is the predicted value of y (predicted from the value of x) and y is the measured value of y. Now consider :



The difference between the measured and predicted value at data point i, $d_i = y_i - y_i'$, is the *deviation*. The quantity

$$d_i^2 = (y_i - y_i')^2 = (y_i - (a + bx_i))^2 \tag{14.4.3}$$

is the *squared deviation*. The *sum of the squared deviations* is

$$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \tag{14.4.4}$$

The least squares solution for a and b is the solution that minimizes E, the sum of squares, over all possible selections of a and b. Minimization problems are easily handled with differential calculus by solving the differential equations:

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 \tag{14.4.5}$$

The solution to those two differential equations is

$$a = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{14.4.6}$$

and

$$b = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{14.4.7}$$

Example 14.3 : Continue with the data from Example 14.1 and find the best fit line. The data again are:

Subject	x	y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	86	172	4	7396
C	15	43	645	225	1849
D	9	74	666	81	5476
E	12	58	696	144	3364
F	5	90	450	25	8100
G	8	78	624	64	6084
$n = 7$	$\sum x = 57$	$\sum y = 511$	$\sum xy = 3745$	$\sum x^2 = 579$	$\sum y^2 = 38993$

Using the sums of the columns, compute:

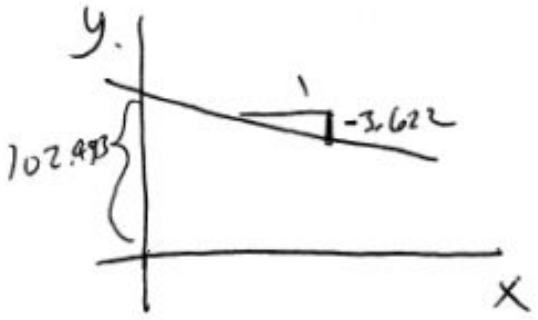
$$\begin{aligned} a &= \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \\ &= \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} \\ &= 102.493 \end{aligned}$$

and

$$\begin{aligned} b &= \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \\ &= \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} \\ &= -3.622 \end{aligned}$$

So

$$\begin{aligned} y' &= a + bx \\ y' &= 102.493 - 3.622x \end{aligned}$$



□

14.5.1: Relationship between correlation and slope

The relationship is

$$r = \frac{bs_x}{} \tag{14.4.8}$$

$$r = \frac{}{s_y} \tag{14.4.6}$$

where

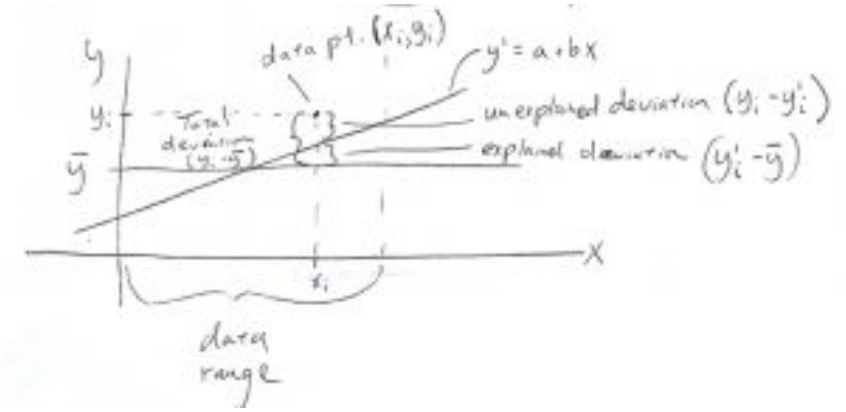
$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$
$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

are the standard deviations of the x and y datasets considered separately.

This page titled 14.4: Linear Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.5: r² and the Standard Error of the Estimate of y′

Consider the deviations :



Looking at the picture we see that

$$\begin{aligned} \text{total deviation} &= \text{explained deviation} + \text{unexplained deviation} \\ (y_i - \overline{y}_i) &= (y_i' - \overline{y}_i) + (y_i - y_i') \end{aligned}$$

Remember that variance is the sum of the squared deviations (divided by degrees of freedom), so squaring the above and summing gives:

$$\sum_{i=1}^{n} (y_i - \overline{y}_i)^2 = \sum_{i=1}^{n} (y_i' - \overline{y}_i)^2 + \sum_{i=1}^{n} (y_i - y_i')^2 \tag{14.5.1}$$

(the cross terms all cancel because y' is the least square solution and $a = \overline{y} - b\overline{x}$, see Section 14.6.1, below, for details). This is also a sum of squares statement:

$$\text{SS}_T = \text{SS}_R + \text{SS}_E \tag{14.5.2}$$

where $\text{SS}_E = \sum (y_i - y_i')^2$, $\text{SS}_T = \sum (y_i - \overline{y})^2$ and $\text{SS}_R = \sum (y_i' - \overline{y})^2$ are the sum of squares — error, sum of squares — total and sum of squares — regression (explained) respectively.

Dividing by the degrees of freedom, which is $n - 2$ in this {\em bivariate} situation, we get:

$$\begin{aligned} \frac{\sum (y_i - \overline{y}_i)^2}{n-2} &= \frac{\sum (y_i' - \overline{y}_i)^2}{n-2} + \frac{\sum (y_i - y_i')^2}{n-2} \\ \text{total variance} &= \text{explained variance} + \text{unexplained variance} \\ &= \text{signal (or model)} + \text{noise} \end{aligned}$$

It turns out that

$$r^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{\text{SS}_R}{\text{SS}_T} \tag{14.5.3}$$

The quantity r^2 is called the *coefficient of determination* and gives the ***the fraction of variance explained by the model*** (here the model is the equation of a line). The quantity r^2 appears with many statistical models. For example with ANOVA it turns out that the “effect size” eta-squared is the fraction of variance explained by the ANOVA model[1], $\eta^2 = r^2$.

The *standard error of the estimate* is the standard deviation of the noise (the square root of the unexplained variance) and is given by

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n-2}} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} \tag{14.5.4}$$

Example 14.4: Continuing with the data of Example 14.3, we had

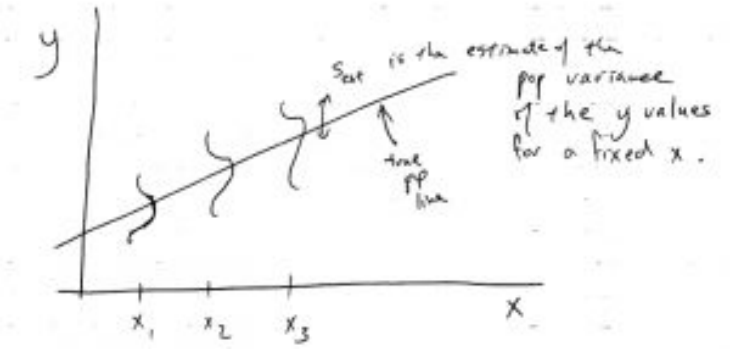
$$\sum y = 511 \quad \sum y^2 = 38993 \quad \sum xy = 3745 \quad a = 102.493 \quad b = -3.622 \quad n = 7 \tag{14.5.5}$$

so

$$s_{\text{est}} = \sqrt{\frac{(38993) - (102.493)(511) - (-3.622)(3745)}{5}}$$
$$s_{\text{est}} = \sqrt{\frac{38993 - 52373.923 + 13564.39}{5}}$$
$$s_{\text{est}} = 6.06$$

□

Here is a graphical interpretation of s_{est} :



The assumption for computing confidence intervals for is that s_{est} is independent of x. This is the assumption of homoscedasticity. You can think of the regression situation as a generalized one-way ANOVA where instead of having a finite number of discrete populations for the IV, we have an infinite number of (continuous) populations. All the populations have the same variance σ^2 (and they are assumed to be normal) and s_{est^2} is the pooled estimate of that variance.

14.6.1: **Details: from deviations to variances

Squaring both sides of

$$(y_i - \overline{y}_i) = (y'_i - \overline{y}_i) + (y_i - y'_i) \tag{14.5.6}$$

and summing gives

$$\sum (y_i - \overline{y}_i)^2 = \sum (y'_i - \overline{y}_i)^2 + \sum (y_i - y'_i)^2 + \sum 2(y'_i - \overline{y})(y_i - y'_i) \tag{14.5.7}$$

Working on that cross term, using $a = \overline{y} - b\overline{x}$, we get

$$\begin{aligned}
\sum 2(y'_i - \overline{y})(y_i - y'_i) &= \sum 2((\overline{y} - b\overline{x} + bx_i) - \overline{y})(y_i - y'_i) \\
&= \sum 2((\overline{y} + b(x_i - \overline{x})) - \overline{y})(y_i - y'_i) \\
&= \sum 2(b(x_i - \overline{x}))(y_i - y'_i) \\
&= \sum 2b(x_i - \overline{x})(y_i - (\overline{y} + b(x_i - \overline{x}))) \\
&= \sum 2b((y_i - \overline{y})(x_i - \overline{x}) - b(x_i - \overline{x})^2) \\
&= 2b \sum ((y_i - \overline{y})(x_i - \overline{x}) - (y_i - \overline{y})(x_i - \overline{x})) = 0
\end{aligned}$$

where

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \tag{14.5.8}$$

was used in the last line.

---

1. In ANOVA the ``model" is the difference of means between the groups. We will see more about this aspect of ANOVA in Chapter 17. ↵

This page titled 14.5: r² and the Standard Error of the Estimate of y′ is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.6: Confidence Interval for y′ at a Given x

At a fixed x (that is important to remember) the confidence interval for y is

$$y' - E < y < y' + E \tag{14.6.1}$$

where

$$E = t_{\mathcal{C}}\ s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x-\overline{x})^2}{n(\sum x^2) - (\sum x)^2}} \tag{14.6.2}$$

where, as usual, $t_{\mathcal{C}}$ comes from the ***t* Distribution Table** with $\nu = n - 2$.

Example 14.5 : Continuing from Example 14.4 (so you can see how an exam will go), say we want to predict the grade (y) in terms of a 95% confidence interval for the number of absences (x) equal to 10.

First, find the value predicted from the regression line, which we previously found to be :

$$y' = 102.493 - 3.622x \tag{14.6.3}$$

at $x = 10$. The result is

$$y' = 102.493 - 3.622(10) = 66.273 \tag{14.6.4}$$

Furthermore, from the last example, we found

$$s_{\text{est}} = 6.06 \tag{14.6.5}$$

and, from the completed data table (Example 14.3)

$$\sum x = 57 \quad \sum x^2 = 579 \tag{14.6.6}$$

We still need $t\mathcal{C}$ and $\overline{x}$. Using our sums:

$$\overline{x} = \frac{\sum x}{n} = \frac{57}{7} = 8.143 \tag{14.6.7}$$

and from ***t* Distribution Table** for the 95% confidence interval, $\nu = 7 - 2 = 5$ we get

$$t\mathcal{C} = 2.571 \tag{14.6.8}$$

Now we compute E :

$$E = t_{\mathcal{C}}\ s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x-\overline{x})^2}{n(\sum x^2) - (\sum x)^2}}$$
$$E = (2.571)\ (6.06)\sqrt{1 + \frac{1}{7} + \frac{7(10-8.143)^2}{7(579) - (52)^2}}$$
$$E = 15.58026\sqrt{1 + 0.1428571 + \frac{24.139}{804}}$$
$$E = 16.77$$

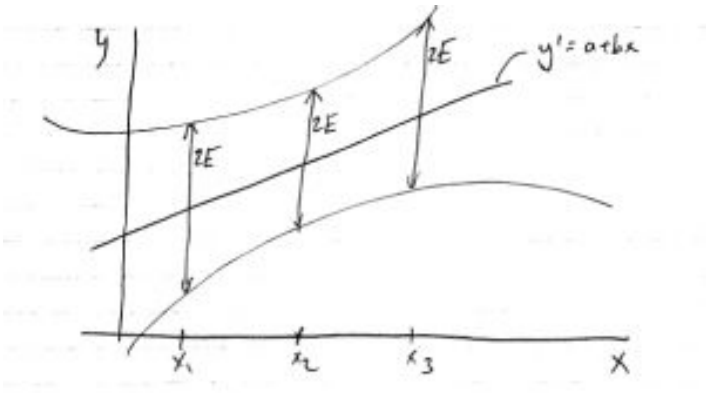
So

$$y' - E < y < y' + E$$
$$66.273 - 16.77 < y < 66.273 + 16.77$$
$$49.5 < y < 83.0$$

This is the 95% confidence interval for predicting the mark of a person who was absent for 10 days.

□

Important: s_{set} is independent of x but E is not. So confidence intervals look like :

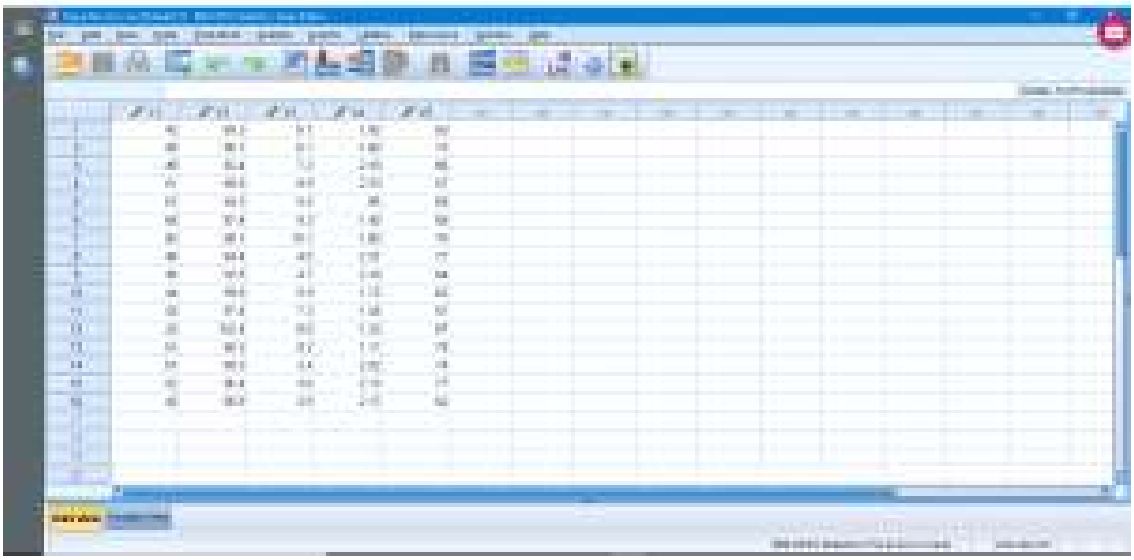


The reason for this variance of the width of the confidence interval comes from the uncertainty in the slope b. You can make plots like the one above in SPSS.

This page titled 14.6: Confidence Interval for y′ at a Given x is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

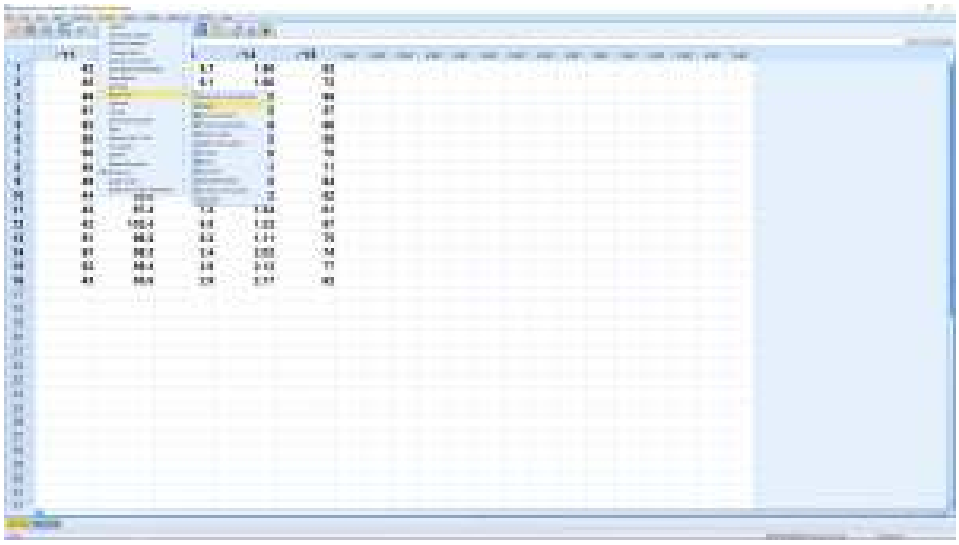
14.7: SPSS Lesson 11- Linear Regression

Open “Hypertension.sav” from the Data Sets:



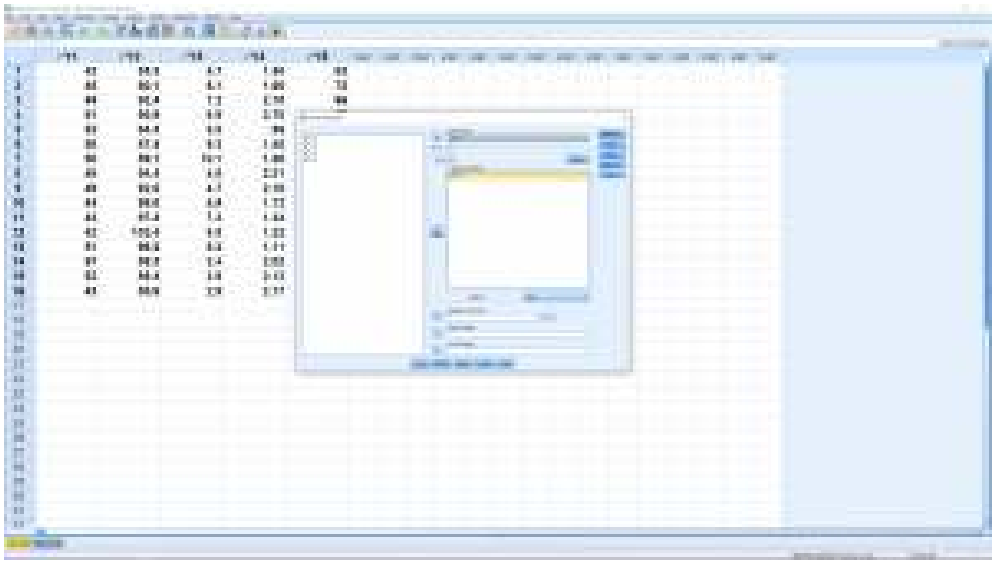
SPSS screenshot © International Business Machines Corporation.

This dataset has a number of variables having to do with a study that is looking for a way to predict injury on the basis of strength. So y_1 is the dependent (y) variable. To get one independent variable, we’ll arbitrarily pick y_2 as our independent variable x. Next pick Analyze → Regression → Linear,



SPSS screenshot © International Business Machines Corporation.

and move the independent and dependent variables into the right slots :



SPSS screenshot © International Business Machines Corporation.



SPSS screenshot © International Business Machines Corporation.

You can look through the submenus if you like but they primarily give options for multiple regression and require the consideration of the independent variable as a vector instead of as a number — this elevation of data from a number to a vector is the basis of multivariate statistics so we'll leave that for now. Running the analysis produces four output tables. You can ignore the "Variables Entered/Removed" table (it is for advanced multiple regression analysis). The other tables show :

The "Model Summary" gives r and s_{est} plus r^2 and r^2_{adj} that we'll discuss when we look at multiple regression. The ANOVA table gives information about the significance of r (and therefore of the overall significance of the regression). We used t to test the significance of r. You can recover the t test statistic from F in the ANOVA table; $t = \sqrt{F}$. Here $p = 0.454$ so the model fit is not significant (do not reject H_0). Even though the fit is not significant, the regression can still be done and this is reported in the last output table. The coefficients are reported in the B column. They are called Unstandardized Coefficients because the data, x and y, have not been z-transformed. The first line gives the intercept (a or b_0), the second line the slope (b or b_1) so

$$y = a + bx$$
$$y = b_0 + b_1 x$$
$$y = 64.309 - 0.173x$$

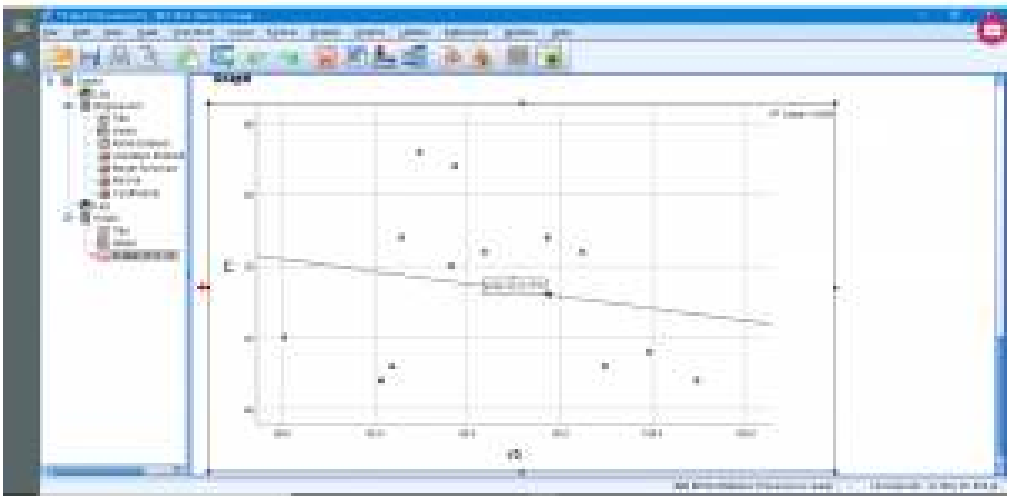
For each of the two regression coefficients, a standard error can be computed, along with confidence intervals for the coefficients, and the significance of the coefficients (H_0: $b_i = 0$) tested with a t test statistic. We haven't covered that aspect of linear regression but we can see the standard errors, t test statistics and associated p values in the "Coefficients" output table. Here b_0, the intercept,

is significant while b_1 the slope, is not. The last thing to notice is Beta in the Standardized Coefficients column. Imagine that we z-transform our variables x and y to z_x and z_y and then did a linear regression on the z-transformed variables. Then the result would be

$$z_y = \beta z_x$$
$$z_y = -0.201 z_x$$

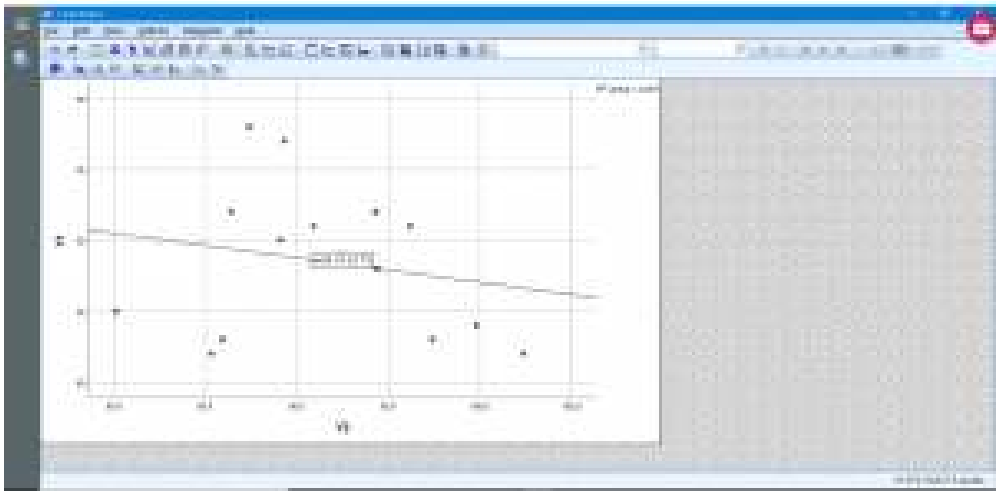
In this case the regression is still insignificant, z-transforming can't change that. There is no intercept in this case because the average of each of z-transformed variables is zero and this leads to an intercept of zero.

Finally, let's see how we can plot the regression line. Generate a scatterplot and then double click on the plot and then click on the little icon that shows a line through scatterplot data :



SPSS screenshot © International Business Machines Corporation.

and



SPSS screenshot © International Business Machines Corporation.

The equation of the regression line is computed instantly and is plotted.

This page titled 14.7: SPSS Lesson 11- Linear Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.8: Multiple Regression

Multiple regression is to the linear regression we just covered as one-way ANOVA is to m-way ANOVA. In m-way ANOVA we have one DV and m discrete IVs. With multiple regression we have one DV (univariate) and k continuous IVs. We will label the DV with y and the IVs with $x_1, x_2, \ldots, x_k$. The idea is to predict y with y' via

$$y' = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{14.8.1}$$

or, using summation notation

$$y' = a + \sum_{j=1}^{k} b_j x_j \tag{14.8.2}$$

Sometimes we (and SPSS) write $a = b_0$. The explicit formula for the coefficients a and b_j are long so we won't give them here but, instead, we will rely on SPSS to compute the coefficients for us. Just the same, we should remember that the coefficients are computed using the least squares method, where the sum of the squared deviations is minimized. That is, a and the b_j are such that

$$\begin{aligned} E &= \sum_{i=1}^{n} (y_i - y_i')^2 \\ &= \sum_{i=1}^{n} (y_i - [a + \sum_{j=1}^{k} b_j x_{ji}])^2 \end{aligned}$$

is minimized. (Here we are using $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$ to represent data point i.) If you like calculus and have a few minutes to spare, the equations for a and the b_j can be found by solving:

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b_1} = 0, \quad \cdots \quad \frac{\partial E}{\partial b_k} = 0 \tag{14.8.3}$$

for a and the b_j. The result will contain al the familiar terms like $\sum y$, $\sum y x_j$, etc. It also turns out that the "normal equations" for a and the b_j that result have a pattern that can be captured with a simple linear algebra equation that we will see in Chapter 17.

Some terminology: the b_j (including b_0) are known as *partial regression coefficients*.

14.10.1: Multiple regression coefficient, *r*

An overall correlation coefficient, r, can be computed using pairwise bivariate correlation coefficients as defined in the previous Section 14.2. This overall correlation is defined as $r = r_{y'y}$, the bivariate correlation coefficient of the predicted values y' versus the data y. For the case of 2 IVs, the formula is

$$r = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}} \tag{14.8.4}$$

where r_{yx_1} is the bivariate correlation coefficient between y and x_1, etc. It is true that $-1 \le r \le 1$ as with the bivariate r.

Example 14.6 : Suppose that you have used SPSS to obtain the regression equation

$$y' = -44.572 + 87.679 x_1 + 14.519 x_2 \tag{14.8.5}$$

for the following data :

Student	GPA, x_1	Age, x_2	Score, y	x_1^2	x_2^2	y^2	$x_1 y$	$x_2 y$	$x_1 x_2$
A	3.2	22	550	10.24	484	302500	1760	12100	70.4
B	2.7	27	570	7.29	729	324900	1539	15390	72.9
C	2.5	24	525	6.25	576	275625	1312.5	12600	60
D	3.4	28	670	11.56	784	448900	2278	18760	95.2
E	2.2	23	490	4.84	529	240100	1078	11270	50.6

$n = 5$	$\sum x_1 =$ 14	$\sum x_2 =$ 124	$\sum y =$ 2805	$\sum x_1^2 =$ 40.18	$\sum x_2^2 =$ 3102	$\sum y^2 =$ 1592025	$\sum x_1 y =$ 7967.5	$\sum x_2 y =$ 70120	$\sum x_1 x_2 =$ 349.1

Compute the multiple correlation coefficient.

Solution :

First we need to compute the pairwise correlations r_{x_1y}, r_{x_2y}, and $r_{x_1x_2}$. (Note that $r_{x_1y} = r_{yx_1}$, etc. because the correlation matrix is symmetric.)

$$\begin{aligned} r_{x_1y} &= \frac{n(\sum x_1 y) - (\sum x_1)(\sum y)}{\sqrt{[n(\sum x_1^2) - (\sum x_1)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{5(7967.5) - (14)(2805)}{\sqrt{[5(40.18) - (14)^2][5(1592025) - (2805)^2]}} \\ &= 0.845 \end{aligned}$$

$$\begin{aligned} r_{x_2y} &= \frac{n(\sum x_2 y) - (\sum x_2)(\sum y)}{\sqrt{[n(\sum x_2^2) - (\sum x_2)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{5(70120) - (124)(2805)}{\sqrt{[5(3102) - (124)^2][5(1592025) - (2805)^2]}} \\ &= 0.791 \end{aligned}$$

$$\begin{aligned} r_{x_1x_2} &= \frac{n(\sum x_1 x_2) - (\sum x_1)(\sum x_2)}{\sqrt{[n(\sum x_1^2) - (\sum x_1)^2][n(\sum x_2^2) - (\sum x_2)^2]}} \\ &= \frac{5(349.1) - (14)(124)}{\sqrt{[5(40.18) - (14)^2][5(3102) - (124)^2]}} \\ &= 0.371 \end{aligned}$$

Now use these in :

$$\begin{aligned} r &= \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}} \\ &= \sqrt{\frac{(0.845)^2 + (0.791)^2 - (2)(0.845)(0.791)(0.371)}{1 - (0.371)^2}} \\ &= 0.989 \end{aligned}$$

□

14.10.2: Significance of *r*

Here we want to test the hypotheses :

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

where ρ is the population multiple regression correlation coefficient.

To test the hypothesis we use

Rendered by QuickLaTeX.com

with

$$\nu_1 = n - k \text{ (d.f.N.) and } \nu_2 = n - k - 1 \text{ (d.f.D.)} \tag{14.8.6}$$

here:

$$
\begin{aligned}
n &= \text{ sample size} \\
k &= \text{ number of IVs} \\
r &= \text{ multiple correlation coefficient}
\end{aligned}
$$

(Note: This "F-test" is similar to but not the same as the "ANOVA" output given by SPSS when you run a regression.)

Example 14.7 : Continuing with Example 14.6, test the significance of r.

Solution :

1. Hypotheses.

$$
\begin{aligned}
H_0 &: \rho = 0 \\
H_1 &: \rho \neq 0
\end{aligned}
$$

2. Critical statistic. From the **Rank Correlation Coefficient Critical Values Table** (i.e., the critical values for the Spearman correlation) with

$$
\begin{aligned}
\nu_1 &= n - k = 5 - 2 = 3 \\
\nu_2 &= n - k - 1 = 5 - 2 - 1 = 2 \\
\alpha &= 0.05
\end{aligned}
$$

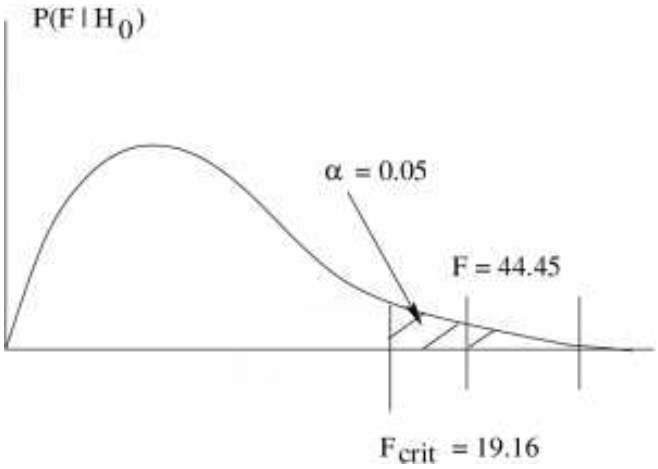
find

$$
F_{\text{crit}} = 19.16 \tag{14.8.7}
$$

3. Test statistic.

Rendered by QuickLaTeX.com

4. Decision.



Reject H_0.

5. Interpretation.

$r = 0.989$ is significant.

□

14.10.3: Other descriptions of correlation

1. Coefficient of multiple determination: r^2. This quantity still has the interpretation as fraction of variance explained by the (multiple regression) model.
2. Adjusted r^2:

$$
r^2_{\text{adj}} = 1 - \left[\frac{(1 - r^2)(n - 1)}{n - k - 1} \right] \tag{14.8.8}
$$

r^2_{adj} gives a better (unbiased) estimate of the population value for ρ^2 by correcting for degrees of freedom just as the sample s^2 with its degrees of freedom equal to $n-1$ gives an unbiased estimate of the population σ^2.

Example 14.8 : Continuing Example 14.6, we had $r = 0.989$ so

$$r^2 = 0.978 \tag{14.8.9}$$

and

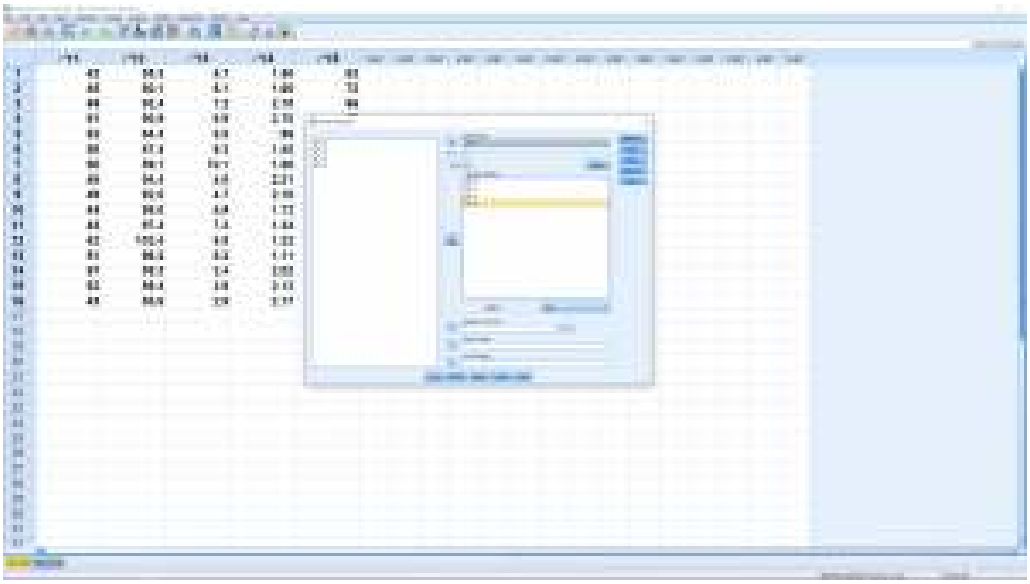
$$\begin{aligned} r^2_{\text{adj}} &= 1 - \left[\frac{(1-r^2)(n-1)}{n-k-1}\right] \\ r^2_{\text{adj}} &= 1 - \left[\frac{(1-0.978)(5-1)}{5-2-1}\right] \\ &= 0.956 \end{aligned}$$

□

This page titled 14.8: Multiple Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

14.9: SPSS Lesson 12- Multiple Regression

Open “Hypertension.sav” from the Data Sets: It is very similar to the data file we used for demonstrating simple linear regression in SPSS but now we have more variables to choose from for independent variables. As before, we really should combine the strength variables but we’ll pick y_2 and y_3. Let’s pick age as a second independent variable, y_1. Pick Analyze → Regression → Linear and enter the independent and dependent variables :



SPSS screenshot © International Business Machines Corporation.

We will again ignore the submenus but note this time that they are to set up what is known as step-up and step-down analysis where independent variables are added or removed in an attempt to get a better fitting model by removing independent variables that are correlated with each other. The relevant output is (ignoring the table meant for step-up and step-down analysis) :



SPSS screenshot © International Business Machines Corporation.

The “Model Summary” table gives r, r^2 (here the model explains 5.7% of the variance of y), r^2_{adj} and s_{est} for multiple regression which we did not look at explicitly for multiple regression. The “ANOVA” table gives the test statistic F for the significance of r along with its p value, which is not significant here. Again, note that this is not the F we looked at in Section 14.10.2, notice the drastic difference in the degrees of freedom between for the two F values. But both do test the significance of the overall r. The models given by the “Coefficients” table are :

$$y = b_0 + b_1 y_1 + b_2 y_2$$
$$y = 65.118 - 0.202 y_1 + 0.295 y_2$$

Note that the intercept is significant but the two slopes are not. If the variables were z-transformed first then we'd have:

$$z_y = \beta_1 z_{y1} + \beta_2 z_{y2}$$
$$z_y = -0.236 z_{y1} + 0.134 z_{y2}$$

There is no way to get SPSS to plot the best fit plane through 3D scatterplot data.

This page titled 14.9: SPSS Lesson 12- Multiple Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

15: Chi Squared- Goodness of Fit and Contingency Tables

Recall that the χ^2 is essentially the distribution of sample variances s^2 from a normal population. It has three important applications (there are others) :

1. Hypothesis test of population variance (covered in Section 9.5).
2. Model fitting through $\chi^2 = \mathrm{SS}_{\mathrm{error}}$ (not covered in this course).
3. Hypothesis test of frequencies :
   a) Goodness of fit
   b) contingency tables.

Here we focus on the last application. We will use the χ^2 statistic to compare the measured (or observed) statistic with expected (H_0) frequencies. The difference of observed and expected frequencies squared represents a variance. If the difference between observed and expected frequencies is due to noise, which will have some sort of binomial distribution, then we expect the χ^2 statistic to be low. If the difference between observed and expected frequencies is large then there must be an effect other than noise that is causing that difference.

This page titled 15: Chi Squared- Goodness of Fit and Contingency Tables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

15.1: Goodness of Fit

Page created for new attachment

This page titled 15.1: Goodness of Fit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gordon E. Sarty via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

16: Non-parametric Tests

16.1: New Page
16.2: New Page
16.3: New Page
16.4: New Page
16.5: New Page
16.6: New Page
16.7: New Page
16.8: New Page
16.9: New Page

16: Non-parametric Tests is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.1: New Page

16.1: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.2: New Page

16.2: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.3: New Page

16.3: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.4: New Page

16.4: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.5: New Page

16.5: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.6: New Page

16.6: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.7: New Page

16.7: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.8: New Page

16.8: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.9: New Page

16.9: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

17: Overview of the General Linear Model

17.1: New Page
17.2: New Page

17: Overview of the General Linear Model is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.1: New Page

17.1: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.2: New Page

17.2: New Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

Index

D

dire

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Introduction to Applied Statistics for Psychology Students (Sarty)

Webpages: 107

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 81.3% (87 pages)
- Undeclared: 18.7% (20 pages)

By Page

- Introduction to Applied Statistics for Psychology Students (Sarty) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *CC BY-NC-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-NC-SA 4.0*
    - About This Book - *Undeclared*
    - Acknowledgements - *Undeclared*
    - Statistical Software Used in this Book - *Undeclared*
    - Data Sets - *Undeclared*
  - 1: Background and Motivation - *CC BY-NC-SA 4.0*
    - 1.1: Overview - *Undeclared*
    - 1.2: Basic Definitions - *CC BY-NC-SA 4.0*
    - 1.3: Summation Convention - *CC BY-NC-SA 4.0*
  - 2: Descriptive Statistics - Frequency Data (Counting) - *CC BY-NC-SA 4.0*
    - 2.1: Frequency Tables - *CC BY-NC-SA 4.0*
    - 2.2: Plotting Frequency Data - *CC BY-NC-SA 4.0*
    - 2.3: SPSS Lesson 1- Getting Started with SPSS - *CC BY-NC-SA 4.0*
  - 3: Descriptive Statistics- Central Tendency and Dispersion - *CC BY-NC-SA 4.0*
    - 3.1: Central Tendency- Mean, Median, Mode - *CC BY-NC-SA 4.0*
    - 3.2: Dispersion- Variance and Standard Deviation - *CC BY-NC-SA 4.0*
    - 3.3: z-score / z-transformation - *CC BY-NC-SA 4.0*
    - 3.4: SPSS Lesson 2- Combining variables and recoding - *CC BY-NC-SA 4.0*
  - 4: Probability and the Binomial Distributions - *CC BY-NC-SA 4.0*
    - 4.1: Probability - *CC BY-NC-SA 4.0*
    - 4.2: Binomial Distribution - *CC BY-NC-SA 4.0*
    - 4.3: SPSS Lesson 3- Combining variables - advanced - *CC BY-NC-SA 4.0*
  - 5: The Normal Distributions - *CC BY-NC-SA 4.0*
    - 5.1: Discrete versus Continuous Distributions - *CC BY-NC-SA 4.0*
    - 5.2: **The Normal Distribution as a Limit of Binomial Distributions - *CC BY-NC-SA 4.0*
    - 5.3: Normal Distribution - *CC BY-NC-SA 4.0*
  - 6: Percentiles and Quartiles - *CC BY-NC-SA 4.0*
    - 6.1: Discrete Data Percentiles and Quartiles - *CC BY-NC-SA 4.0*
    - 6.2: Finding Outliers Using Quartiles - *CC BY-NC-SA 4.0*
    - 6.3: Box Plots - *CC BY-NC-SA 4.0*
    - 6.4: Robust Statistics - *CC BY-NC-SA 4.0*
    - 6.5: SPSS Lesson 4- Percentiles - *CC BY-NC-SA 4.0*
  - 7: The Central Limit Theorem - *CC BY-NC-SA 4.0*
    - 7.1: Using the Normal Distribution to Approximate the Binomial Distribution - *CC BY-NC-SA 4.0*
    - 7.2: The Central Limit Theorem - *CC BY-NC-SA 4.0*
  - 8: Confidence Intervals - *CC BY-NC-SA 4.0*
    - 8.1: Confidence Intervals Using the z-Distribution - *CC BY-NC-SA 4.0*
    - 8.2: **Bayesian Statistics - *CC BY-NC-SA 4.0*
    - 8.3: The t-Distributions - *CC BY-NC-SA 4.0*
    - 8.4: Proportions and Confidence Intervals for Proportions - *CC BY-NC-SA 4.0*
    - 8.5: Chi Squared Distribution - *CC BY-NC-SA 4.0*
  - 9: Hypothesis Testing - *CC BY-NC-SA 4.0*
    - 9.1: Hypothesis Testing Problem Solving Steps - *CC BY-NC-SA 4.0*
    - 9.2: z-Test for a Mean - *CC BY-NC-SA 4.0*
    - 9.3: t-Test for Means - *CC BY-NC-SA 4.0*
    - 9.4: z-Test for Proportions - *CC BY-NC-SA 4.0*
    - 9.5: Chi Squared Test for Variance or Standard Deviation - *CC BY-NC-SA 4.0*
    - 9.6: SPSS Lesson 5- Single Sample t-Test - *CC BY-NC-SA 4.0*

- 10: Comparing Two Population Means - *CC BY-NC-SA 4.0*
  - 10.1: Unpaired z-Test - *CC BY-NC-SA 4.0*
  - 10.2: Confidence Interval for Difference of Means (Large Samples) - *CC BY-NC-SA 4.0*
  - 10.3: Difference between Two Variances - the F Distributions - *CC BY-NC-SA 4.0*
  - 10.4: Unpaired or Independent Sample t-Test - *CC BY-NC-SA 4.0*
  - 10.5: Confidence Intervals for the Difference of Two Means - *CC BY-NC-SA 4.0*
  - 10.6: SPSS Lesson 6- Independent Sample t-Test - *CC BY-NC-SA 4.0*
  - 10.7: Paired t-Test - *CC BY-NC-SA 4.0*
  - 10.8: Confidence Intervals for Paired t-Tests - *CC BY-NC-SA 4.0*
  - 10.9: SPSS Lesson 7- Paired Sample t-Test - *CC BY-NC-SA 4.0*
- 11: Comparing Proportions - *CC BY-NC-SA 4.0*
  - 11.1: z-Test for Comparing Proportions - *CC BY-NC-SA 4.0*
  - 11.2: Confidence Interval for the Difference between Two Proportions - *CC BY-NC-SA 4.0*
- 12: ANOVA - *CC BY-NC-SA 4.0*
  - 12.1: One-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.2: Post hoc Comparisons - *CC BY-NC-SA 4.0*
  - 12.3: SPSS Lesson 8- One-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.4: Two-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.5: SPSS Lesson 9- Two-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.6: Higher Factorial ANOVA - *CC BY-NC-SA 4.0*
  - 12.7: Between and Within Factors - *CC BY-NC-SA 4.0*
  - 12.8: *Contrasts - *CC BY-NC-SA 4.0*
- 13: Power - *CC BY-NC-SA 4.0*
  - 13.1: Power - *CC BY-NC-SA 4.0*
- 14: Correlation and Regression - *CC BY-NC-SA 4.0*
  - 14.1: Scatter Plots - *CC BY-NC-SA 4.0*
  - 14.2: Correlation - *CC BY-NC-SA 4.0*
  - 14.3: SPSS Lesson 10- Scatterplots and Correlation - *CC BY-NC-SA 4.0*
  - 14.4: Linear Regression - *CC BY-NC-SA 4.0*
  - 14.5: r^2 and the Standard Error of the Estimate of y' - *CC BY-NC-SA 4.0*
  - 14.6: Confidence Interval for y' at a Given x - *CC BY-NC-SA 4.0*
  - 14.7: SPSS Lesson 11- Linear Regression - *CC BY-NC-SA 4.0*
  - 14.8: Multiple Regression - *CC BY-NC-SA 4.0*
  - 14.9: SPSS Lesson 12- Multiple Regression - *CC BY-NC-SA 4.0*
- 15: Chi Squared- Goodness of Fit and Contingency Tables - *CC BY-NC-SA 4.0*
  - 15.1: Goodness of Fit - *CC BY-NC-SA 4.0*
- 16: Non-parametric Tests - *Undeclared*
  - 16.1: New Page - *Undeclared*
  - 16.2: New Page - *Undeclared*
  - 16.3: New Page - *Undeclared*
  - 16.4: New Page - *Undeclared*
  - 16.5: New Page - *Undeclared*
  - 16.6: New Page - *Undeclared*
  - 16.7: New Page - *Undeclared*
  - 16.8: New Page - *Undeclared*
  - 16.9: New Page - *Undeclared*
- 17: Overview of the General Linear Model - *Undeclared*
  - 17.1: New Page - *Undeclared*
  - 17.2: New Page - *Undeclared*
- Back Matter - *CC BY-NC-SA 4.0*
  - Index - *CC BY-NC-SA 4.0*
  - Glossary - *CC BY-NC-SA 4.0*
  - Detailed Licensing - *CC BY-NC-SA 4.0*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Introduction to Applied Statistics for Psychology Students (Sarty)

Webpages: 107

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 81.3% (87 pages)
- Undeclared: 18.7% (20 pages)

By Page

- Introduction to Applied Statistics for Psychology Students (Sarty) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *CC BY-NC-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-NC-SA 4.0*
    - About This Book - *Undeclared*
    - Acknowledgements - *Undeclared*
    - Statistical Software Used in this Book - *Undeclared*
    - Data Sets - *Undeclared*
  - 1: Background and Motivation - *CC BY-NC-SA 4.0*
    - 1.1: Overview - *Undeclared*
    - 1.2: Basic Definitions - *CC BY-NC-SA 4.0*
    - 1.3: Summation Convention - *CC BY-NC-SA 4.0*
  - 2: Descriptive Statistics - Frequency Data (Counting) - *CC BY-NC-SA 4.0*
    - 2.1: Frequency Tables - *CC BY-NC-SA 4.0*
    - 2.2: Plotting Frequency Data - *CC BY-NC-SA 4.0*
    - 2.3: SPSS Lesson 1- Getting Started with SPSS - *CC BY-NC-SA 4.0*
  - 3: Descriptive Statistics- Central Tendency and Dispersion - *CC BY-NC-SA 4.0*
    - 3.1: Central Tendency- Mean, Median, Mode - *CC BY-NC-SA 4.0*
    - 3.2: Dispersion- Variance and Standard Deviation - *CC BY-NC-SA 4.0*
    - 3.3: z-score / z-transformation - *CC BY-NC-SA 4.0*
    - 3.4: SPSS Lesson 2- Combining variables and recoding - *CC BY-NC-SA 4.0*
  - 4: Probability and the Binomial Distributions - *CC BY-NC-SA 4.0*
    - 4.1: Probability - *CC BY-NC-SA 4.0*
    - 4.2: Binomial Distribution - *CC BY-NC-SA 4.0*
    - 4.3: SPSS Lesson 3- Combining variables - advanced - *CC BY-NC-SA 4.0*
  - 5: The Normal Distributions - *CC BY-NC-SA 4.0*
    - 5.1: Discrete versus Continuous Distributions - *CC BY-NC-SA 4.0*
    - 5.2: **The Normal Distribution as a Limit of Binomial Distributions - *CC BY-NC-SA 4.0*
    - 5.3: Normal Distribution - *CC BY-NC-SA 4.0*
  - 6: Percentiles and Quartiles - *CC BY-NC-SA 4.0*
    - 6.1: Discrete Data Percentiles and Quartiles - *CC BY-NC-SA 4.0*
    - 6.2: Finding Outliers Using Quartiles - *CC BY-NC-SA 4.0*
    - 6.3: Box Plots - *CC BY-NC-SA 4.0*
    - 6.4: Robust Statistics - *CC BY-NC-SA 4.0*
    - 6.5: SPSS Lesson 4- Percentiles - *CC BY-NC-SA 4.0*
  - 7: The Central Limit Theorem - *CC BY-NC-SA 4.0*
    - 7.1: Using the Normal Distribution to Approximate the Binomial Distribution - *CC BY-NC-SA 4.0*
    - 7.2: The Central Limit Theorem - *CC BY-NC-SA 4.0*
  - 8: Confidence Intervals - *CC BY-NC-SA 4.0*
    - 8.1: Confidence Intervals Using the z-Distribution - *CC BY-NC-SA 4.0*
    - 8.2: **Bayesian Statistics - *CC BY-NC-SA 4.0*
    - 8.3: The t-Distributions - *CC BY-NC-SA 4.0*
    - 8.4: Proportions and Confidence Intervals for Proportions - *CC BY-NC-SA 4.0*
    - 8.5: Chi Squared Distribution - *CC BY-NC-SA 4.0*
  - 9: Hypothesis Testing - *CC BY-NC-SA 4.0*
    - 9.1: Hypothesis Testing Problem Solving Steps - *CC BY-NC-SA 4.0*
    - 9.2: z-Test for a Mean - *CC BY-NC-SA 4.0*
    - 9.3: t-Test for Means - *CC BY-NC-SA 4.0*
    - 9.4: z-Test for Proportions - *CC BY-NC-SA 4.0*
    - 9.5: Chi Squared Test for Variance or Standard Deviation - *CC BY-NC-SA 4.0*
    - 9.6: SPSS Lesson 5- Single Sample t-Test - *CC BY-NC-SA 4.0*

- 10: Comparing Two Population Means - *CC BY-NC-SA 4.0*
  - 10.1: Unpaired z-Test - *CC BY-NC-SA 4.0*
  - 10.2: Confidence Interval for Difference of Means (Large Samples) - *CC BY-NC-SA 4.0*
  - 10.3: Difference between Two Variances - the F Distributions - *CC BY-NC-SA 4.0*
  - 10.4: Unpaired or Independent Sample t-Test - *CC BY-NC-SA 4.0*
  - 10.5: Confidence Intervals for the Difference of Two Means - *CC BY-NC-SA 4.0*
  - 10.6: SPSS Lesson 6- Independent Sample t-Test - *CC BY-NC-SA 4.0*
  - 10.7: Paired t-Test - *CC BY-NC-SA 4.0*
  - 10.8: Confidence Intervals for Paired t-Tests - *CC BY-NC-SA 4.0*
  - 10.9: SPSS Lesson 7- Paired Sample t-Test - *CC BY-NC-SA 4.0*
- 11: Comparing Proportions - *CC BY-NC-SA 4.0*
  - 11.1: z-Test for Comparing Proportions - *CC BY-NC-SA 4.0*
  - 11.2: Confidence Interval for the Difference between Two Proportions - *CC BY-NC-SA 4.0*
- 12: ANOVA - *CC BY-NC-SA 4.0*
  - 12.1: One-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.2: Post hoc Comparisons - *CC BY-NC-SA 4.0*
  - 12.3: SPSS Lesson 8- One-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.4: Two-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.5: SPSS Lesson 9- Two-way ANOVA - *CC BY-NC-SA 4.0*
  - 12.6: Higher Factorial ANOVA - *CC BY-NC-SA 4.0*
  - 12.7: Between and Within Factors - *CC BY-NC-SA 4.0*
  - 12.8: *Contrasts - *CC BY-NC-SA 4.0*
- 13: Power - *CC BY-NC-SA 4.0*
  - 13.1: Power - *CC BY-NC-SA 4.0*
- 14: Correlation and Regression - *CC BY-NC-SA 4.0*
  - 14.1: Scatter Plots - *CC BY-NC-SA 4.0*
  - 14.2: Correlation - *CC BY-NC-SA 4.0*
  - 14.3: SPSS Lesson 10- Scatterplots and Correlation - *CC BY-NC-SA 4.0*
  - 14.4: Linear Regression - *CC BY-NC-SA 4.0*
  - 14.5: r^2 and the Standard Error of the Estimate of y' - *CC BY-NC-SA 4.0*
  - 14.6: Confidence Interval for y' at a Given x - *CC BY-NC-SA 4.0*
  - 14.7: SPSS Lesson 11- Linear Regression - *CC BY-NC-SA 4.0*
  - 14.8: Multiple Regression - *CC BY-NC-SA 4.0*
  - 14.9: SPSS Lesson 12- Multiple Regression - *CC BY-NC-SA 4.0*
- 15: Chi Squared- Goodness of Fit and Contingency Tables - *CC BY-NC-SA 4.0*
  - 15.1: Goodness of Fit - *CC BY-NC-SA 4.0*
- 16: Non-parametric Tests - *Undeclared*
  - 16.1: New Page - *Undeclared*
  - 16.2: New Page - *Undeclared*
  - 16.3: New Page - *Undeclared*
  - 16.4: New Page - *Undeclared*
  - 16.5: New Page - *Undeclared*
  - 16.6: New Page - *Undeclared*
  - 16.7: New Page - *Undeclared*
  - 16.8: New Page - *Undeclared*
  - 16.9: New Page - *Undeclared*
- 17: Overview of the General Linear Model - *Undeclared*
  - 17.1: New Page - *Undeclared*
  - 17.2: New Page - *Undeclared*
- Back Matter - *CC BY-NC-SA 4.0*
  - Index - *CC BY-NC-SA 4.0*
  - Glossary - *CC BY-NC-SA 4.0*
  - Detailed Licensing - *CC BY-NC-SA 4.0*
  - Detailed Licensing - *Undeclared*