

# Introduction to Statistics

## Online Edition

Primary author and editor:

David M. Lane<sup>1</sup>

Other authors:

David Scott<sup>1</sup>, Mikki Hebl<sup>1</sup>, Rudy Guerra<sup>1</sup>, Dan Osherson<sup>1</sup>, and Heidi Zimmer<sup>2</sup>

<sup>1</sup>Rice University; <sup>2</sup>University of Houston, Downtown Campus

Section authors specified on each section.

This work is in the public domain. Therefore, it can be copied and reproduced without limitation.

1. Introduction .....	10
What Are Statistics .....	11
Importance of Statistics.....	13
Descriptive Statistics .....	15
Inferential Statistics.....	20
Variables .....	26
Percentiles .....	29
Levels of Measurement .....	34
Distributions .....	40
Summation Notation .....	52
Linear Transformations.....	55

Logarithms.....	58
Statistical Literacy .....	61
Exercises .....	62
<b>2. Graphing Distributions .....</b>	<b>65</b>
Graphing Qualitative Variables.....	66
Graphing Quantitative Variables .....	75
Stem and Leaf Displays.....	76
Histograms.....	82
Frequency Polygons .....	86
Box Plots .....	92
Bar Charts .....	101
Line Graphs.....	105
Dot Plots .....	109
Statistical Literacy .....	113
References.....	115
Exercises .....	116
<b>3. Summarizing Distributions .....</b>	<b>123</b>
What is Central Tendency? .....	124
Measures of Central Tendency .....	131
Median and Mean .....	134
Additional Measures of Central Tendency .....	136
Comparing Measures of Central Tendency.....	140
Measures of Variability .....	144

Shapes of Distributions .....	152
Effects of Linear Transformations.....	154
Variance Sum Law I.....	156
Statistical Literacy .....	158
Exercises .....	159
<b>4. Describing Bivariate Data.....</b>	<b>164</b>
Introduction to Bivariate Data .....	165
Values of the Pearson Correlation.....	170
Properties of Pearson's r .....	175
Computing Pearson's r.....	176
Variance Sum Law II.....	178
Statistical Literacy .....	180
Exercises .....	181
<b>5. Probability.....</b>	<b>185</b>
Remarks on the Concept of “Probability” .....	186
Basic Concepts.....	189
Permutations and Combinations .....	198
Binomial Distribution.....	203
Poisson Distribution.....	207
Multinomial Distribution.....	208
Hypergeometric Distribution .....	210
Base Rates.....	212
Statistical Literacy .....	215

Exercises .....	216
<b>6. Research Design.....</b>	<b>222</b>
Scientific Method.....	223
Measurement.....	225
Basics of Data Collection .....	231
Sampling Bias .....	235
Experimental Designs.....	238
Causation.....	242
Statistical Literacy .....	245
References.....	246
Exercises .....	247
<b>7. Normal Distributions .....</b>	<b>248</b>
Introduction to Normal Distributions .....	249
History of the Normal Distribution .....	252
Areas Under Normal Distributions .....	256
Standard Normal Distribution .....	259
Normal Approximation to the Binomial .....	263
Statistical Literacy .....	266
Exercises .....	267
<b>8. Advanced Graphs .....</b>	<b>272</b>
Quantile-Quantile (q-q) Plots .....	273
Contour Plots.....	289
3D Plots .....	292

Statistical Literacy .....	297
Exercises .....	298
<b>9. Sampling Distributions .....</b>	<b>299</b>
Introduction to Sampling Distributions.....	300
Sampling Distribution of the Mean .....	307
Sampling Distribution of Difference Between Means.....	311
Sampling Distribution of Pearson's r.....	316
Figure 2. The sampling distribution of $r$ for $N = 12$ and $\rho = 0.90$ . ....	318
Sampling Distribution of $p$ .....	319
Statistical Literacy .....	322
Exercises .....	323
<b>10. Estimation .....</b>	<b>328</b>
Introduction to Estimation .....	329
Degrees of Freedom .....	330
Characteristics of Estimators.....	333
Confidence Intervals.....	336
Introduction to Confidence Intervals .....	337
t Distribution.....	339
Confidence Interval for the Mean .....	343
Difference between Means .....	349
Correlation .....	356
Proportion.....	358
Statistical Literacy .....	360

Exercises .....	362
<b>11. Logic of Hypothesis Testing .....</b>	<b>369</b>
Introduction .....	370
Significance Testing .....	375
Type I and II Errors .....	377
One- and Two-Tailed Tests .....	379
Interpreting Significant Results .....	383
Interpreting Non-Significant Results .....	385
Steps in Hypothesis Testing .....	388
Significance Testing and Confidence Intervals .....	389
Misconceptions .....	391
Statistical Literacy .....	392
References .....	393
Exercises .....	394
<b>12. Testing Means .....</b>	<b>398</b>
Testing a Single Mean .....	399
Differences between Two Means (Independent Groups) .....	406
All Pairwise Comparisons Among Means .....	412
Specific Comparisons (Independent Groups) .....	418
Difference Between Two Means (Correlated Pairs) .....	428
Specific Comparisons (Correlated Observations) .....	432
Pairwise Comparisons (Correlated Observations) .....	436
Statistical Literacy .....	438

References.....	439
Exercises .....	440
<b>13. Power.....</b>	<b>447</b>
Introduction to Power.....	448
Example Calculations.....	450
Factors Affecting Power .....	454
Statistical Literacy .....	458
Exercises .....	459
<b>14. Regression .....</b>	<b>461</b>
Introduction to Linear Regression.....	462
Partitioning the Sums of Squares .....	468
Standard Error of the Estimate.....	473
Inferential Statistics for b and r .....	476
Influential Observations .....	482
Regression Toward the Mean.....	487
Introduction to Multiple Regression.....	495
Statistical Literacy .....	507
References.....	508
Exercises .....	509
<b>15. Analysis of Variance .....</b>	<b>515</b>
Introduction.....	516
Analysis of Variance Designs.....	518
Between- and Within-Subjects Factors .....	519

One-Factor ANOVA (Between Subjects).....	521
Multi-Factor Between-Subjects Designs .....	532
Unequal Sample Sizes .....	544
Tests Supplementing ANOVA.....	553
Within-Subjects ANOVA.....	562
Statistical Literacy .....	569
Exercises .....	570
<b>16. Transformations.....</b>	<b>576</b>
Log Transformations .....	577
Tukey Ladder of Powers.....	580
Box-Cox Transformations .....	588
Statistical Literacy .....	594
References.....	595
Exercises .....	596
<b>17. Chi Square .....</b>	<b>597</b>
Chi Square Distribution .....	598
One-Way Tables (Testing Goodness of Fit).....	601
Contingency Tables .....	605
Statistical Literacy .....	608
References.....	609
Exercises .....	610
<b>18. Distribution-Free Tests .....</b>	<b>616</b>
Benefits .....	617

Randomization Tests: Two Conditions .....	618
Randomization Tests: Two or More Conditions .....	620
Randomization Tests: Association (Pearson's $r$ ).....	622
Randomization Tests: Contingency Tables: (Fisher's Exact Test) .....	624
Rank Randomization: Two Conditions (Mann-Whitney U, Wilcoxon Rank Sum).....	626
Rank Randomization: Two or More Conditions (Kruskal-Wallis).....	631
Rank Randomization for Association (Spearman's $\rho$ ) .....	633
Statistical Literacy .....	636
Exercises .....	637
<b>19. Effect Size.....</b>	<b>639</b>
Proportions .....	640
Difference Between Two Means.....	643
Proportion of Variance Explained .....	647
References.....	653
Statistical Literacy .....	654
Exercises .....	655
<b>20. Case Studies .....</b>	<b>657</b>
<b>21. Glossary.....</b>	<b>659</b>

# 1. Introduction

This chapter begins by discussing what statistics are and why the study of statistics is important. Subsequent sections cover a variety of topics all basic to the study of statistics. The only theme common to all of these sections is that they cover concepts and ideas important for other chapters in the book.

- A. What are Statistics?
- B. Importance of Statistics
- C. Descriptive Statistics
- D. Inferential Statistics
- E. Variables
- F. Percentiles
- G. Measurement
- H. Levels of Measurement
- I. Distributions
- J. Summation Notation
- K. Linear Transformations
- L. Logarithms
- M. Exercises

# What Are Statistics

by Mikki Hebl

## *Learning Objectives*

1. Describe the range of applications of statistics
2. Identify situations in which statistics can be misleading
3. Define “Statistics”

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following three scenarios and the interpretations based upon the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong. Try to identify a major flaw with each interpretation before we describe it.

- 1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.

- 2) The more churches in a city, the more crime there is. Thus, churches lead to crime.

A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem, which we will discuss in more detail in Chapter 6, refers to the third-variable problem.

Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

3) 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

A major flaw is that we don't have the information that we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). But this latter number is hardly evidence suggesting the acceptability of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics.

As a whole, these examples show that statistics are *not only facts and figures*; they are something more than that. In the broadest sense, “statistics” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

# Importance of Statistics

by Mikki Hebl

## *Learning Objectives*

1. Give examples of statistics encountered in everyday life
2. Give examples of how statistics can lend credibility to an argument

Like most people, you probably feel that it is important to “take control of your life.” But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- Native Americans are significantly more likely to be hit crossing the street than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data and data interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) The present electronic textbook is designed to help you learn statistical essentials. **It will make you into an intelligent consumer of statistical claims.**

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to **question** the statistics that you encounter. The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, “There are three kinds of lies -- lies, damned lies, and statistics.” This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have put the emphasis on defending ourselves against fraudulent claims wrapped up as statistics. We close this section on a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

Now let us get to work!

# Descriptive Statistics

by Mikki Hebl

## *Prerequisites*

- none

## *Learning Objectives*

1. Define “descriptive statistics”
2. Distinguish between descriptive statistics and inferential statistics

**Descriptive statistics** are numbers that are used to summarize and describe data.

The word “data” refers to the information that has been collected from an experiment, a survey, an historical record, etc. (By the way, “data” is plural. One piece of information is called a “datum.”) If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time to give a full picture of the data.

Descriptive statistics are just descriptive. They do not involve **generalizing** beyond the data at hand. Generalizing from our data to another set of cases is the business of **inferential statistics**, which you'll be studying in another section. Here we focus on (mere) descriptive statistics.

Some descriptive statistics are shown in Table 1. The table shows the average salaries for various occupations in the United States in 1999.

Table 1. Average salaries for various occupations in 1999.

\$112,760	pediatricians
\$106,130	dentists
\$100,090	podiatrists
\$76,140	physicists
\$53,410	architects,
\$49,720	school, clinical, and counseling psychologists
\$47,910	flight attendants
\$39,560	elementary school teachers
\$38,710	police officers
\$18,980	floral designers

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table 2. It shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table 2. Number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990.

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69

4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70
6. State College, PA	113	6. Gadsden, AL	70
7. Clarksville-Hopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-Seaside-Monterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

*NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.*

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) The following table shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Table 3. Winning Olympic marathon times.

Women			
Year	Winner	Country	Time
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40

1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20

**Men**

Year	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50
1900	Michel Theato	FRA	2:59:45
1904	Thomas Hicks	USA	3:28:53
1906	Billy Sherring	CAN	2:51:23
1908	Johnny Hayes	USA	2:55:18
1912	Kenneth McArthur	S. Afr.	2:36:54
1920	Hannes Kolehmainen	FIN	2:32:35
1924	Albin Stenroos	FIN	2:41:22
1928	Boughra El Ouafi	FRA	2:32:57
1932	Juan Carlos Zabala	ARG	2:31:36
1936	Sohn Kee-Chung	JPN	2:29:19
1948	Delfo Cabrera	ARG	2:34:51
1952	Emil Ztopek	CZE	2:23:03
1956	Alain Mimoun	FRA	2:25:00
1960	Abebe Bikila	ETH	2:15:16
1964	Abebe Bikila	ETH	2:12:11
1968	Mamo Wolde	ETH	2:20:26
1972	Frank Shorter	USA	2:12:19
1976	Waldemar Cierpinski	E.Ger	2:09:55
1980	Waldemar Cierpinski	E.Ger	2:11:03
1984	Carlos Lopes	POR	2:09:21
1988	Gelindo Bordin	ITA	2:10:32

1992	Hwang Young-Cho	S. Kor	2:13:23
1996	Josia Thugwane	S. Afr.	2:12:36
2000	Gezahenge Abera	ETH	2:10.10
2004	Stefano Baldini	ITA	2:10:55

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written 2:44:22). The mean winning time for the second 13 races is 2:13:18. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can't answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table 3 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how far they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

# Inferential Statistics

by Mikki Hebl

## *Prerequisites*

- Chapter 1: Descriptive Statistics

## *Learning Objectives*

1. Distinguish between a sample and a population
2. Define inferential statistics
3. Identify biased samples
4. Distinguish between simple random sampling and stratified sampling
5. Distinguish between random sampling and random assignment

## **Populations and samples**

In statistics, we often rely on a sample --- that is, a small subset of a larger set of data --- to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

Example #1: You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Who will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. The mathematical procedures whereby we convert information about the sample into intelligent guesses about the population fall under the rubric of inferential statistics.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer

the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferential statistics are based on the assumption that sampling is random. We trust a random sample to represent different segments of society in close to the appropriate proportions (provided the sample is large enough; see below).

Example #2: We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school. Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

Example #3: A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example #3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example #4: A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example #4, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

## Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check

your understanding of simple random sampling, consider the following example. What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

Example #5: A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the **National Twin Registry**, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

In Example #5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the Z's may bias the sample. Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the B's is that the "every-other-one" procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

## Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the **sampling procedure** rather than the **results** of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small,

are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. (To see how to obtain this probability, see the section on the binomial distribution in Chapter 5.) Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

## More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competing to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

## Random assignment

In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a placebo, there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (drug) and the other group is assigned to the control condition (placebo). This random division of the sample into two groups is called **random assignment**. Random assignment is

critical for the validity of an experiment. For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

In experimental research of this kind, failure to assign subjects randomly to groups is generally more serious than having a non-random sample. Failure to randomize (the former error) invalidates the experimental findings. A non-random sample (the latter error) simply restricts the generalizability of the results.

## **Stratified Sampling**

Since simple random sampling often does not ensure a representative sample, a sampling method called stratified random sampling is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct “strata” or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

# Variables

by Heidi Ziemer

## *Prerequisites*

- none

## *Learning Objectives*

1. Define and distinguish between independent and dependent variables
2. Define and distinguish between discrete and continuous variables
3. Define and distinguish between qualitative and quantitative variables

## **Independent and dependent variables**

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as  $\pi$  that do not vary). When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is “type of antidepressant.” When a variable is manipulated by an experimenter, it is called an independent variable. The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a dependent variable. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Example #1: Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

1. What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)

2. What are the dependent variables? (memory test and motor skills test)

Example #2: Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

1. What is the independent variable? (supplements: beta-carotene or placebo)
2. What is the dependent variable? (occurrence of cancer)

Example #3: How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

1. What is the independent variable? (brightness of brake lights)
2. What is the dependent variable? (time to hit brakes)

## **Levels of an Independent Variable**

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

## Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on.

The values of a qualitative variable do not imply a numerical ordering. Values of the variable “religion” differ qualitatively; no ordering of religions is implied.

Qualitative variables are sometimes referred to as categorical variables.

Quantitative variables are those variables that are measured in terms of numbers.

Some examples of quantitative variables are height, weight, and shoe size.

In the study on the effect of diet discussed previously, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable “type of supplement” is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable “memory test” is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

## Discrete and Continuous Variables

Variables such as number of children in a household are called discrete variables since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as “time to respond to a question” are continuous variables since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

# Percentiles

by David Lane

## *Prerequisites*

- none

## *Learning Objectives*

1. Define percentiles
2. Use three formulas for computing percentiles

A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65th percentile.

## Two Simple Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this “Definition 1.” The 65th percentile can also be defined as the smallest score that is greater than or equal to 65% of the scores. This we will call “Definition 2.” Unfortunately, these two definitions can lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5% of the scores? A third way to compute percentiles (presented below) is a weighted average of the percentiles computed according to the first two definitions. This third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50th percentile.

## A Third Definition

Unless otherwise specified, when we refer to “percentile,” we will be referring to this third definition of percentiles. Let's begin with an example. Consider the 25th percentile for the 8 numbers in Table 1. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = \frac{P}{100} \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = \frac{25}{100} \times (8 + 1) = \frac{9}{4} = 2.25$$

If R is an integer, the Pth percentile is be the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, IR = 2.
2. Define FR as the fractional portion of R. For this example, FR = 0.25.

3. Find the scores with Rank  $I_R$  and with Rank  $I_R + 1$ . For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.

4. Interpolate by multiplying the difference between the scores by  $F_R$  and add the result to the lower score. For these data, this is  $(0.25)(7 - 5) + 5 = 5.5$ .

Therefore, the 25th percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25th percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25th percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table 2.

Table 2. 20 Quiz Scores.

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

We will compute the 25th and the 85th percentiles. For the 25th,

$$R = \frac{25}{100} \times (20 + 1) = \frac{21}{4} = 5.25$$

*IR = 5 and FR = 0.25.*

Since the score with a rank of IR (which is 5) and the score with a rank of IR + 1 (which is 6) are both equal to 5, the 25th percentile is 5. In terms of the formula:

$$25th \ percentile = (.25) \times (5 - 5) + 5 = 5.$$

For the 85th percentile,

$$R = \frac{85}{100} \times (20 + 1) = 17.85$$

*IR = 17 and FR = 0.85*

**Caution: FR does not generally equal the percentile to be computed as it does here.**

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85th percentile is:

$$(0.85)(10 - 9) + 9 = 9.85$$

Consider the 50th percentile of the numbers 2, 3, 5, 9.

$$R = \frac{50}{100} \times (4 + 1) = 2.5$$

*IR = 2 and FR = 0.5.*

The score with a rank of IR is 3 and the score with a rank of IR + 1 is 5. Therefore, the 50th percentile is:

$$(0.5)(5 - 3) + 3 = 4.$$

Finally, consider the 50th percentile of the numbers 2, 3, 5, 9, 11.

$$R = \frac{50}{100} \times (5 + 1) = 3$$

*IR = 3 and FR = 0.*

Whenever  $FR = 0$ , you simply find the number with rank  $IR$ . In this case, the third number is equal to 5, so the 50th percentile is 5. You will also get the right answer if you apply the general formula:

$$50th \text{ percentile} = (0.00) (9 - 5) + 5 = 5.$$

# Levels of Measurement

by Dan Osherson and David M. Lane

## *Prerequisites*

- Chapter 1: Variables

## *Learning Objectives*

1. Define and distinguish among nominal, ordinal, interval, and ratio scales
2. Identify a scale type
3. Discuss the type of scale used in psychological measurement
4. Give examples of errors that can be made by failing to understand the proper use of measurement scales

## Types of Scales

Before we can conduct a statistical analysis, we need to measure our dependent variable. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a dependent variable such as "favorite color," you can simply note the color-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are described in this section.

### Nominal scales

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which

green is placed “ahead of” blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

## Ordinal scales

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either “very dissatisfied,” “somewhat dissatisfied,” “somewhat satisfied,” or “very satisfied.” The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses “very dissatisfied” and “somewhat dissatisfied” is probably not equivalent to the difference between “somewhat dissatisfied” and “somewhat satisfied.” Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the mental step from 1 to 2 (for example) is the same as the mental step from 3 to 4.

## Interval scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name “zero.” The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label “zero” is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the “zero” label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is “twice as hot” as 40 degrees. Such a claim would depend on an arbitrary decision about where to “start” the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

## Ratio scales

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This

makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).

## What level of measurement is used for psychological variables?

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point; some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

Subject	Easy Items					Difficult Items					Score
A	0	0	1	1	0	0	0	0	0	0	2
B	1	0	1	1	0	0	0	0	0	0	3
C	1	1	1	1	1	1	1	0	0	0	7
D	1	1	1	1	1	0	1	1	0	1	8

Let's compare (i) the difference between Subject A's score of 2 and Subject B's score of 3 and (ii) the difference between Subject C's score of 7 and Subject D's score of 8. The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond "No" to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio.

### Consequences of level of measurement

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

This means that if a child said her favorite color was "Red," then the choice was coded as "2," if the child said her favorite color was "Purple," then the response was coded as 5, and so forth. Consider the following hypothetical data:

Subject	Color	Code
1	Blue	1
2	Blue	1
3	Green	4
4	Green	4
5	Purple	5

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favorite color is yellow (the color with a code of 3). Such nonsense arises because favorite color is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic. However, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

# Distributions

by David M. Lane and Heidi Ziemer

## *Prerequisites*

- Chapter 1: Variables

## *Learning Objectives*

1. Define “distribution”
2. Interpret a frequency distribution
3. Distinguish between a frequency distribution and a probability distribution
4. Construct a grouped frequency distribution for a continuous variable
5. Identify the skew of a distribution
6. Identify bimodal, leptokurtic, and platykurtic distributions

## Distributions of Discrete Variables

I recently purchased a bag of Plain M&M's. The M&M's were in six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table 1.

Table 1. Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a frequency distribution. Often a frequency distribution is shown graphically as in Figure 1.

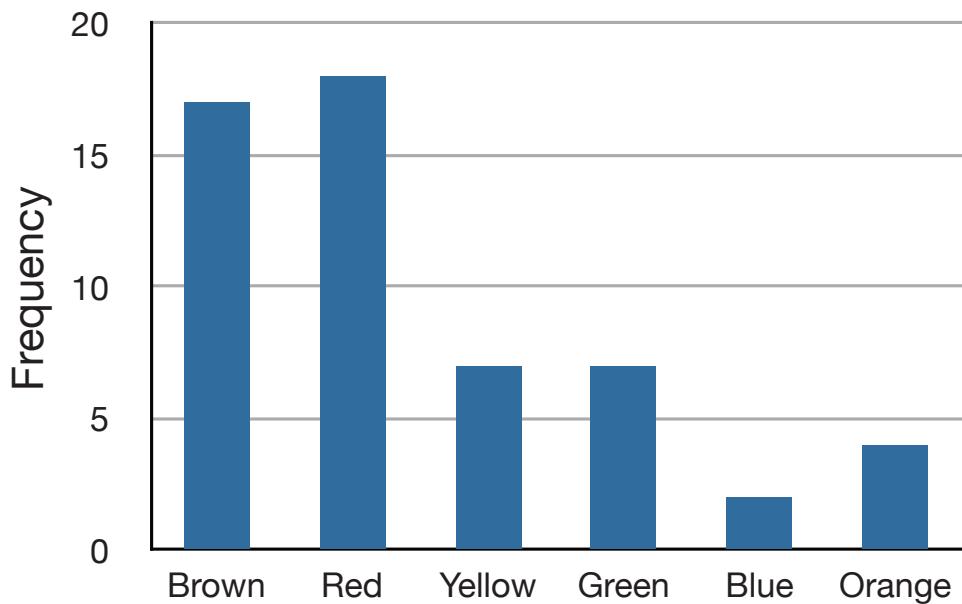


Figure 1. Distribution of 55 M&M's.

The distribution shown in Figure 1 concerns just my one bag of M&M's. You might be wondering about the distribution of colors for all M&M's. The manufacturer of M&M's provides some information about this matter, but they do not tell us exactly how many M&M's of each color they have ever produced. Instead, they report proportions rather than frequencies. Figure 2 shows these proportions. Since every M&M is one of the six familiar colors, the six proportions shown in the figure add to one. We call Figure 2 a probability distribution because if you choose an M&M at random, the probability of getting, say, a brown M&M is equal to the proportion of M&M's that are brown (0.30).

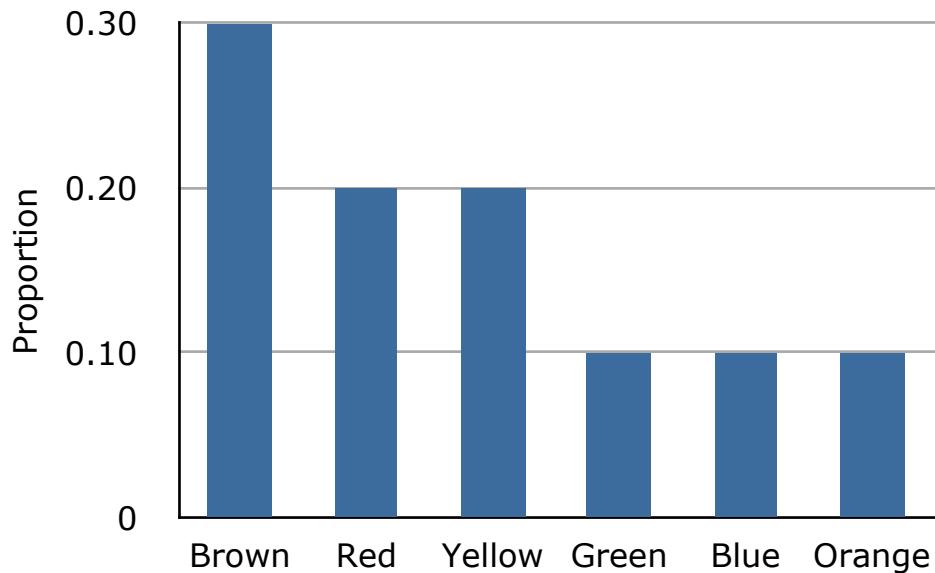


Figure 2. Distribution of all M&M's.

Notice that the distributions in Figures 1 and 2 are not identical. Figure 1 portrays the distribution in a sample of 55 M&M's. Figure 2 shows the proportions for all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colors that is close to Figure 2; others will be further away.

## Continuous Variables

The variable “color of M&M” used in this example is a **discrete variable**, and its distribution is also called **discrete**. Let us now extend the concept of a distribution to **continuous variables**.

The data shown in Table 2 are the times it took one of us (DL) to move the cursor over a small target in a series of 20 trials. The times are sorted from shortest to longest. The variable “time to respond” is a **continuous variable**. With time measured accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in psychology. As you can see in Table 2, measuring DL's responses this way produced times no two of which were the same. As a result, a frequency distribution would be **uninformative**: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 2. Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

The solution to this problem is to create a grouped frequency distribution. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 3 shows a grouped frequency distribution for these 20 times.

Table 3. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Grouped frequency distributions can be portrayed graphically. Figure 3 shows a graphical representation of the frequency distribution in Table 3. This kind of graph is called a histogram. Chapter 2 contains an entire section devoted to histograms.

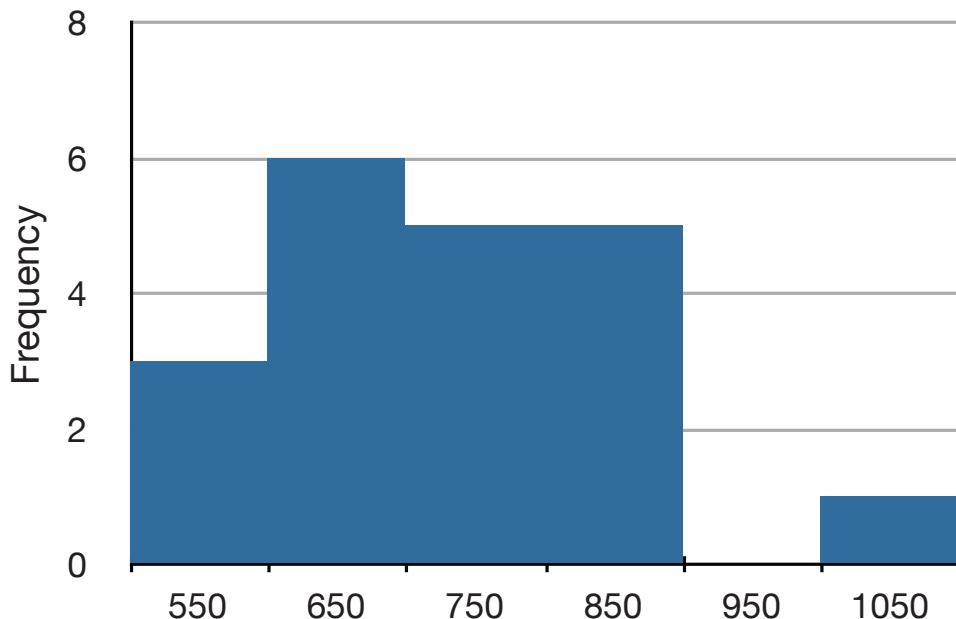


Figure 3. A histogram of the grouped frequency distribution shown in Table 3. The labels on the X-axis are the middle values of the range they represent.

## Probability Densities

The histogram in Figure 3 portrays just DL's 20 times in the one experiment he performed. To represent the probability associated with an arbitrary movement (which can take any positive amount of time), we must represent all these potential times at once. For this purpose, we plot the distribution for the continuous variable of time. Distributions for continuous variables are called continuous distributions. They also carry the fancier name probability density. Some probability densities have particular importance in statistics. A very important one is shaped like a bell, and called the normal distribution. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

An example of a normal distribution is shown in Figure 4. Do you see the “bell”? The normal distribution doesn't represent a real bell, however, since the left and right tips extend indefinitely (we can't draw them any further so they look like they've stopped in our diagram). The Y-axis in the normal distribution represents the “density of probability.” Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis. In Figure 4, for example, the probability of an observation with value near 40 is about half of the probability of an

observation with value near 50. (For more information, see Chapter 7.)

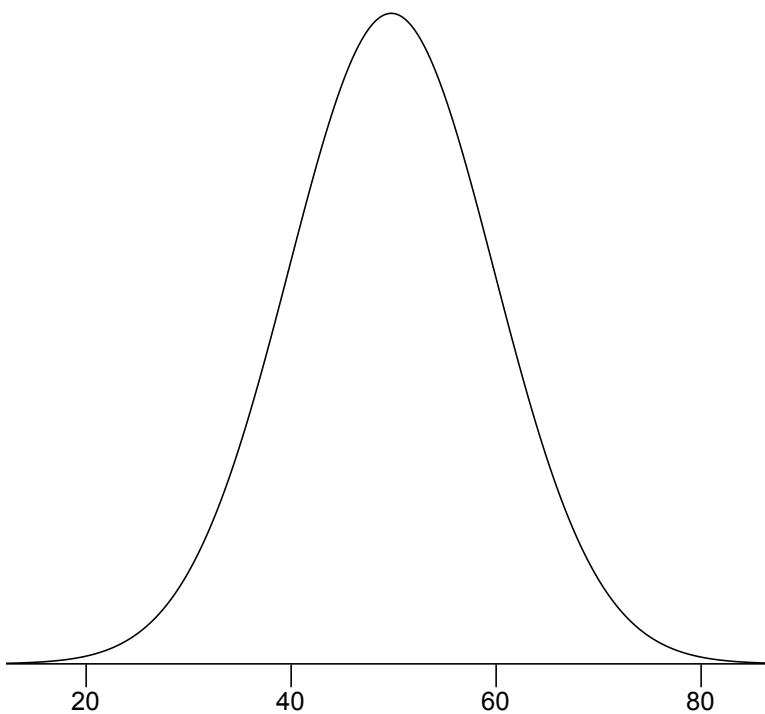


Figure 4. A normal distribution.

Although this text does not discuss the concept of probability density in detail, you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution). First, the area under the curve equals 1. Second, the probability of any exact value of  $X$  is 0. Finally, the area under the curve and bounded between two given points on the  $X$ -axis is the probability that a number chosen at random will fall between the two points. Let us illustrate with DL's hand movements. First, the probability that his movement takes some amount of time is one! (We exclude the possibility of him never finishing his gesture.) Second, the probability that his movement takes exactly 598.956432342346576 milliseconds is essentially zero. (We can make the probability as close as we like to zero by making the time measurement more and more precise.) Finally, suppose that the probability of DL's movement taking between 600 and 700 milliseconds is one tenth. Then the continuous distribution for DL's possible times would have a shape that places 10% of the area below the curve in the region bounded by 600 and 700 on the  $X$ -axis.

## Shapes of Distributions

Distributions have different shapes; they don't all look like the normal distribution in Figure 4. For example, the normal probability density is higher in the middle compared to its two tails. Other distributions need not have this feature. There is even variation among the distributions that we call "normal." For example, some normal distributions are more spread out than the one shown in Figure 4 (their tails begin to hit the X-axis further from the middle of the curve --for example, at 10 and 90 if drawn in place of Figure 4). Others are less spread out (their tails might approach the X-axis at 30 and 70). More information on the normal distribution can be found in a later chapter completely devoted to them.

The distribution shown in Figure 4 is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure 5 shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in the negative direction. A distribution with the longer tail extending in the positive direction is said to have a positive skew. It is also described as "skewed to the right."

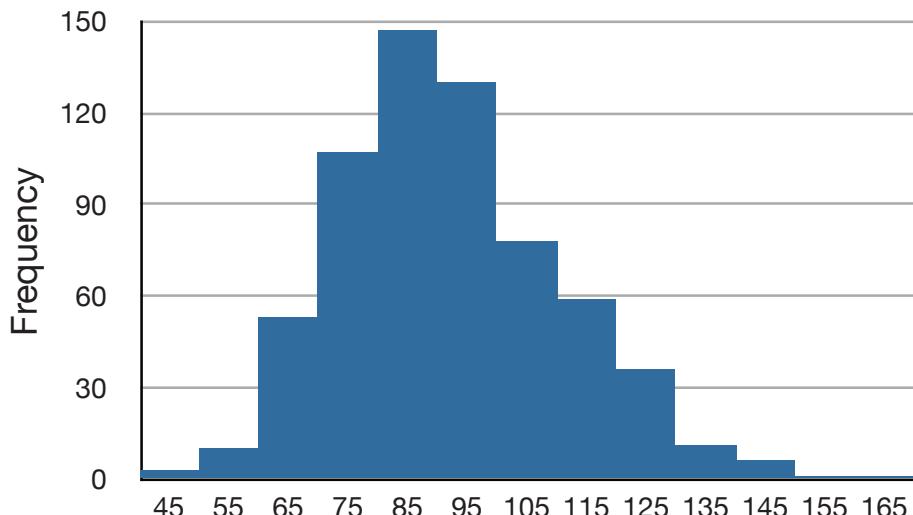


Figure 5. A distribution with a positive skew.

Figure 6 shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

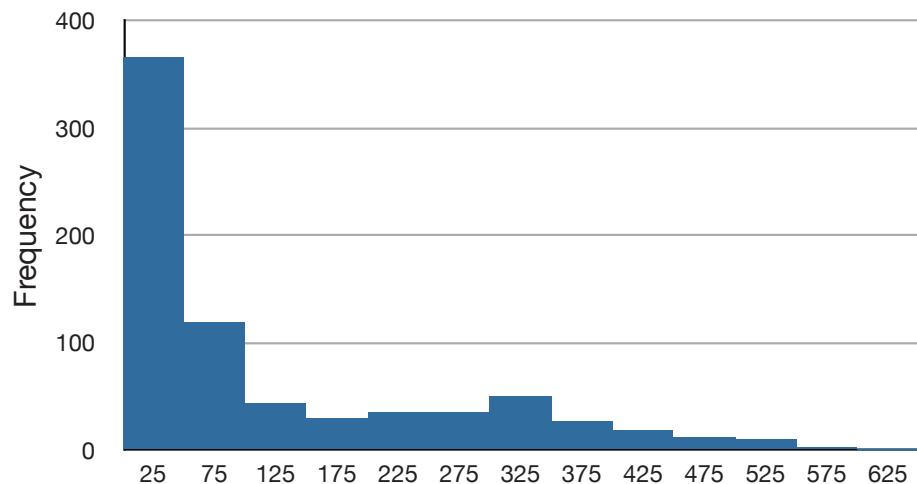


Figure 6. A distribution with a very large positive skew.

A continuous distribution with a positive skew is shown in Figure 7.

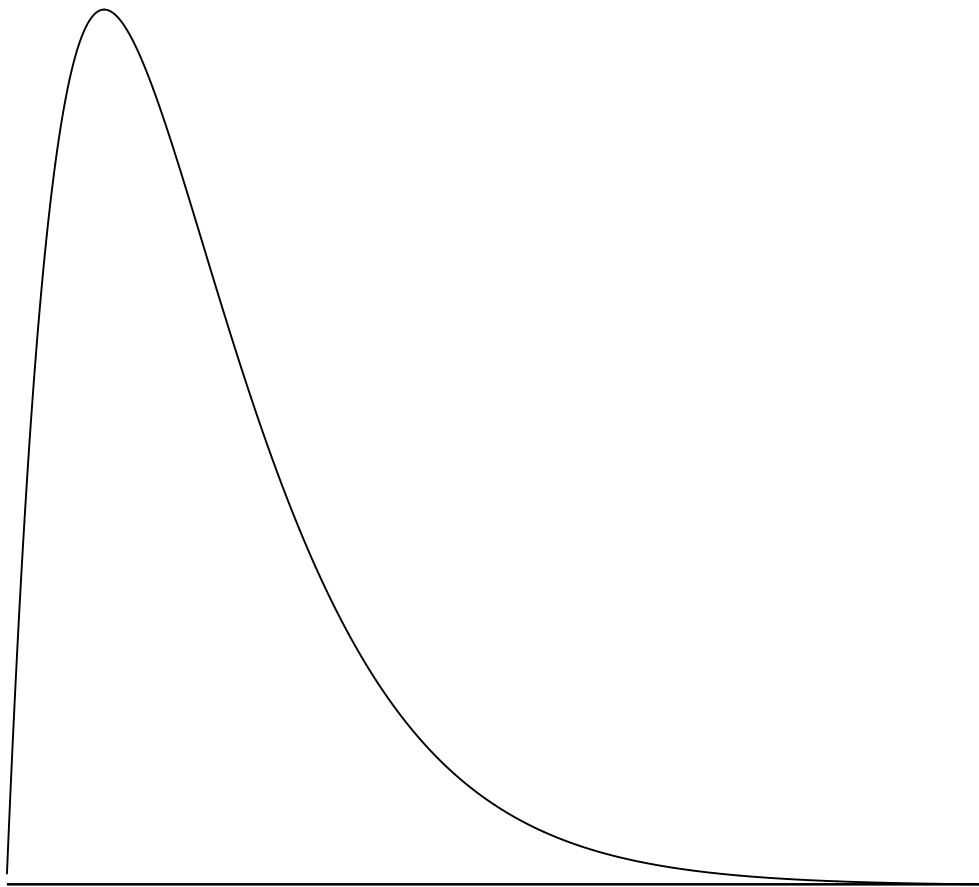


Figure 7. A continuous distribution with a positive skew.

Although less common, some distributions have a negative skew. Figure 8 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the left, this distribution is skewed to the left.

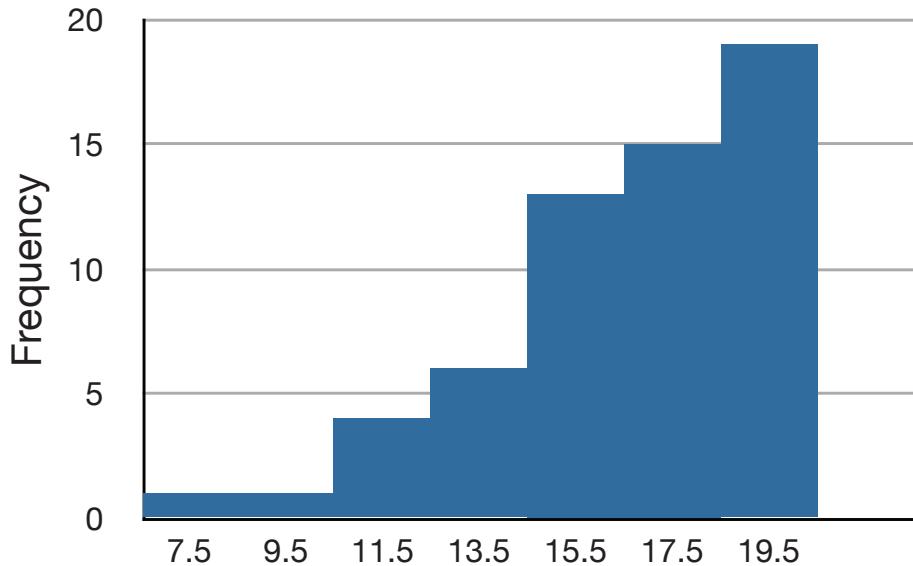


Figure 8. A distribution with negative skew. This histogram shows the frequencies of various scores on a 20-point question on a statistics test.

A continuous distribution with a negative skew is shown in Figure 9.

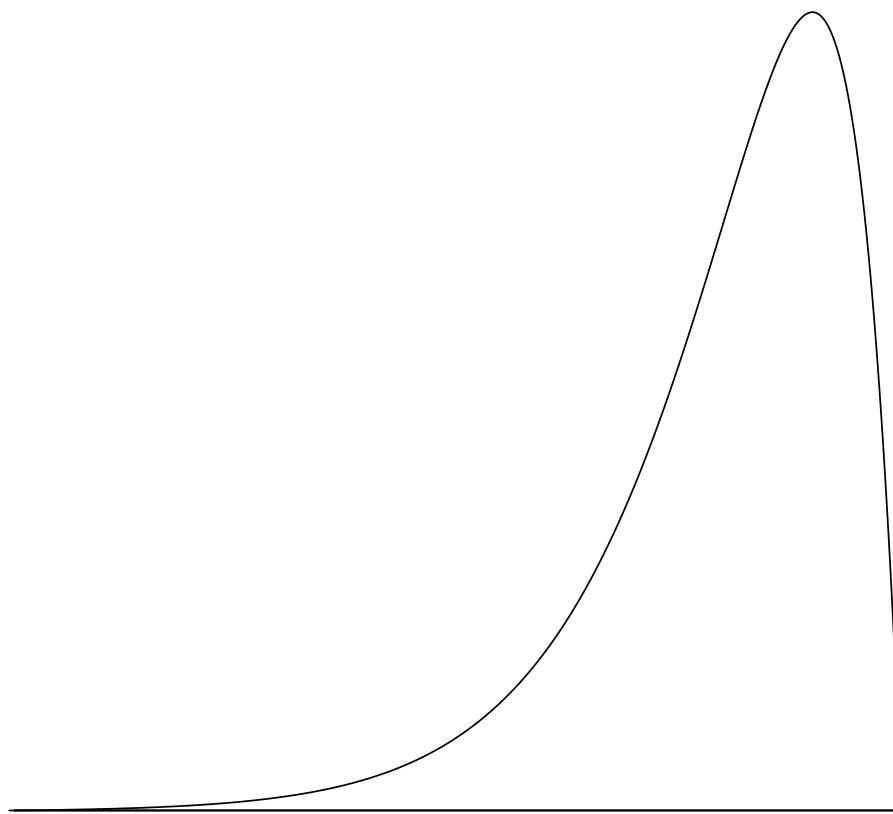


Figure 9. A continuous distribution with a negative skew.

The distributions shown so far all have one distinct high point or peak. The distribution in Figure 10 has two distinct peaks. A distribution with two peaks is called a **bimodal distribution**.

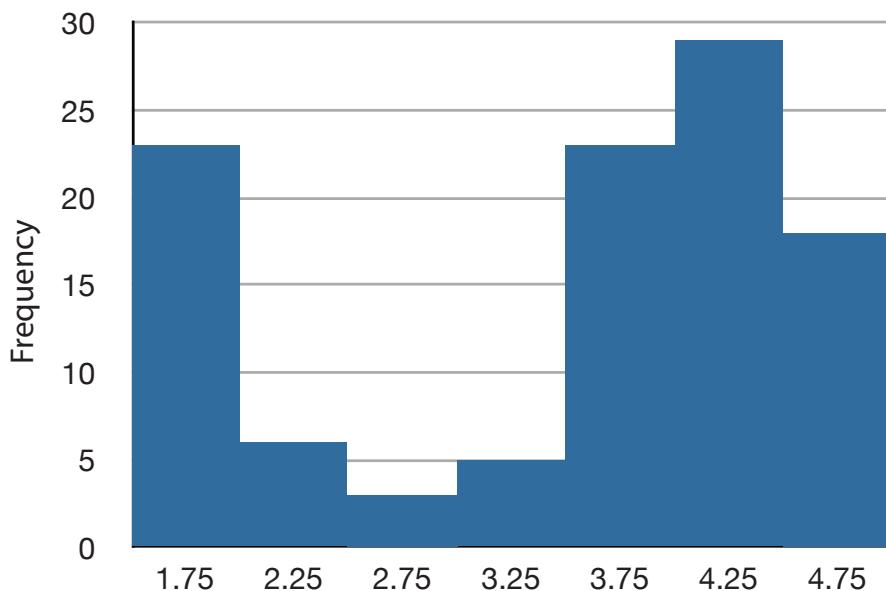


Figure 10. Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

Distributions also differ from each other in terms of how large or “fat” their tails are. Figure 11 shows two distributions that differ in this respect. The upper distribution has relatively more scores in its tails; its shape is called leptokurtic. The lower distribution has relatively fewer scores in its tails; its shape is called platykurtic.

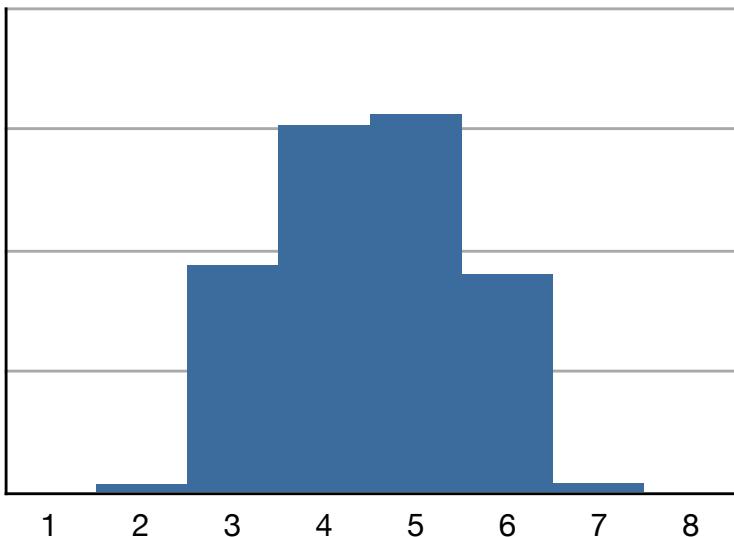
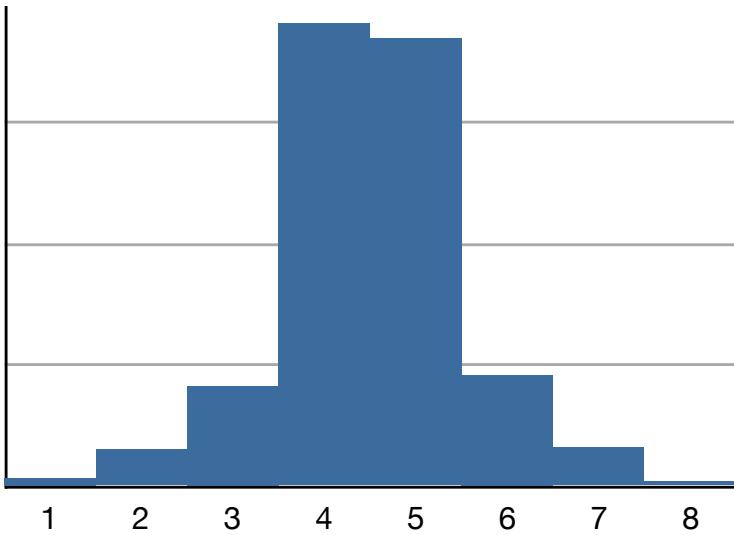


Figure 11. Distributions differing in kurtosis. The top distribution has long tails. It is called “leptokurtic.” The bottom distribution has short tails. It is called “platykurtic.”

# Summation Notation

by David M. Lane

## Prerequisites

- None

## Learning Objectives

1. Use summation notation to express the sum of all numbers
2. Use summation notation to express the sum of a subset of numbers
3. Use summation notation to express the sum of squares

Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable  $X$  that represents the weights (in grams) of 4 grapes. The data are shown in Table 1.

Table 1. Weights of 4 grapes.

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

We label Grape 1's weight  $X_1$ , Grape 2's weight  $X_2$ , etc. The following formula means to sum up the weights of the four grapes:

$$\sum_{i=1}^4 X_i$$

The Greek letter  $\Sigma$  indicates summation. The “ $i = 1$ ” at the bottom indicates that the summation is to start with  $X_1$  and the 4 at the top indicates that the summation will end with  $X_4$ . The “ $X_i$ ” indicates that  $X$  is the variable to be summed as  $i$  goes from 1 to 4. Therefore,

$$\sum_{i=1}^4 X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

The symbol

$$\sum_{i=1}^3 X_i$$

indicates that only the first 3 scores are to be summed. The index variable  $i$  goes from 1 to 3.

When all the scores of a variable (such as  $X$ ) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X$$

Thus, when no values of  $i$  are shown, it means to sum all the values of  $X$ .

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\begin{aligned} \sum X^2 &= 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2 \\ &= 21.16 + 26.01 + 24.01 + 19.36 = 90.54. \end{aligned}$$

Notice that:

$$\left(\sum X\right)^2 \neq \sum X^2$$

because the expression on the left means to sum up all the values of  $X$  and then square the sum ( $19^2 = 361$ ), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Table 2 shows the data for variables  $X$  and  $Y$ . The cross products ( $XY$ ) are shown in the third column. The sum of the cross products is  $3 + 4 + 21 = 28$ .

Table 2. Cross Products.

X	Y	XY
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28$$

# Linear Transformations

by David M. Lane

## *Prerequisites*

- None

## *Learning Objectives*

1. Give the formula for a linear transformation
2. Determine whether a transformation is linear
3. Describe what is linear about a linear transformation

Often it is necessary to transform data from one measurement scale to another. For example, you might want to convert height measured in feet to height measured in inches. Table 1 shows the heights of four people measured in both feet and inches. To transform feet to inches, you simply multiply by 12. Similarly, to transform inches to feet, you divide by 12.

Table 1. Converting between feet and inches.

Feet	Inches
5.00	60
6.25	75
5.50	66
5.75	69

Some conversions require that you multiply by a number and then add a second number. A good example of this is the transformation between degrees Centigrade and degrees Fahrenheit. Table 2 shows the temperatures of 5 US cities in the early afternoon of November 16, 2002.

Table 2. Temperatures in 5 cities on 11/16/2002.

City	Degrees Fahrenheit	Degrees Centigrade
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11

The formula to transform Centigrade to Fahrenheit is:

$$F = 1.8C + 32$$

The formula for converting from Fahrenheit to Centigrade is

$$C = 0.5556F - 17.778$$

The transformation consists of multiplying by a constant and then adding a second constant. For the conversion from Centigrade to Fahrenheit, the first constant is 1.8 and the second is 32.

Figure 1 shows a plot of degrees Centigrade as a function of degrees Fahrenheit. Notice that the points form a straight line. This will always be the case if the transformation from one scale to another consists of multiplying by one constant and then adding a second constant. Such transformations are therefore called linear transformations.

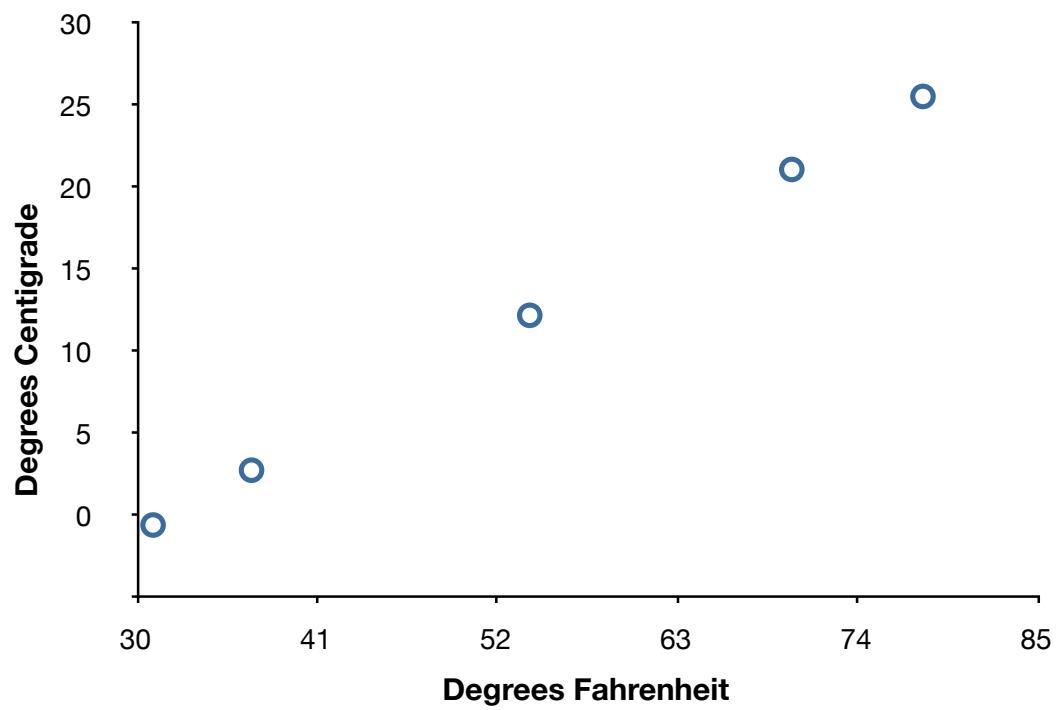


Figure 1. Degrees Centigrade as a function of degrees Fahrenheit

# Logarithms

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions

## *Learning Objectives*

1. Compute logs using different bases
2. Convert between bases
3. State the relationship between logs and proportional change

The log transformation reduces positive skew. This can be valuable both for making the data more interpretable and for helping to meet the assumptions of inferential statistics.

## **Basics of Logarithms (Logs)**

Logs are, in a sense, the opposite of exponents. Consider the following simple expression:

$$10^2 = 100$$

Here we can say the base of 10 is raised to the second power. Here is an example of a log:

$$\log_{10}(100) = 2$$

This can be read as: The log base ten of 100 equals 2. The result is the power that the base of 10 has to be raised to in order to equal the value (100). Similarly,

$$\log_{10}(1000) = 3$$

since 10 has to be raised to the third power in order to equal 1,000.

These examples all used base 10, but any base could have been used. There is a base which results in “natural logarithms” and that is called e and equals approximately 2.718. It is beyond the scope of this book to explain what is “natural” about it. Natural logarithms can be indicated either as:  $\ln(x)$  or  $\log_e(x)$

Changing the base of the log changes the result by a multiplicative constant. To convert from  $\log_{10}$  to natural logs, you multiply by 2.303. Analogously, to convert in the other direction, you divide by 2.303.

Taking the antilog of a number undoes the operation of taking the log. Therefore, since  $\text{Log}_{10}(1000) = 3$ , the antilog<sub>10</sub> of 3 is 1,000. Taking the antilog of a number simply raises the base of the logarithm in question to that number.

## Logs and Proportional Change

A series of numbers that increases proportionally will increase in equal amounts when converted to logs. For example, the numbers in the first column of Table 1 increase by a factor of 1.5 so that each row is 1.5 times as high as the preceding row. The  $\text{Log}_{10}$  transformed numbers increase in equal steps of 0.176.

Table 1. Proportional raw changes are equal in log units.

Raw	Log
4.0	0.602
6.0	0.778
9.0	0.954
13.5	1.130

As another example, if one student increased their score from 100 to 200 while a second student increased their's from 150 to 300, the percentage change (100%) is the same for both students. The log difference is also the same, as shown below.

$$\text{Log}_{10}(100) = 2.000$$

$$\text{Log}_{10}(200) = 2.301$$

$$\text{Difference: } 0.301$$

$$\text{Log}_{10}(150) = 2.176$$

$$\text{Log}_{10}(300) = 2.477$$

$$\text{Difference: } 0.301$$

## Arithmetic Operations

Rules for logs of products and quotients are shown below.

$$\text{Log}(AB) = \text{Log}(A) + \text{Log}(B)$$

$$\text{Log}(A/B) = \text{Log}(A) - \text{Log}(B)$$

For example,

$$\text{Log}_{10}(10 \times 100) = \text{Log}_{10}(10) + \text{Log}_{10}(100) = 1 + 2 = 3.$$

Similarly,

$$\log_{10}(100/10) = \log_{10}(100) - \log_{10}(10) = 2 - 1 = 1.$$

# Statistical Literacy

by Denise Harvey and David M. Lane

## *Prerequisites*

- Chapter 1: Levels of Measurement

The Board of Trustees at a university commissioned a top management-consulting firm to address the admission processes for academic and athletic programs. The consulting firm wrote a report discussing the trade-off between maintaining academic and athletic excellence. One of their key findings was:

The standard for an athlete's admission, as reflected in SAT scores alone, is lower than the standard for non-athletes by as much as 20 percent, with the weight of this difference being carried by the so-called "revenue sports" of football and basketball. Athletes are also admitted through a different process than the one used to admit non-athlete students.

## **What do you think?**

Based on what you have learned in this chapter about measurement scales, does it make sense to compare SAT scores using percentages? Why or why not?

Think about this before continuing:

As you may know, the SAT has an arbitrarily-determined lower limit on test scores of 200. Therefore, SAT is measured on either an ordinal scale or, at most, an interval scale. However, it is clearly not measured on a ratio scale. Therefore, it is not meaningful to report SAT score differences in terms of percentages. For example, consider the effect of subtracting 200 from every student's score so that the lowest possible score is 0. How would that affect the difference as expressed in percentages?

## Exercises

### *Prerequisites*

- All material presented in Chapter: “Introduction”

1. A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics?
2. A cognitive psychologist is interested in comparing two ways of presenting stimuli on sub- sequent memory. Twelve subjects are presented with each method and a memory test is given. What would be the roles of descriptive and inferential statistics in the analysis of these data?
3. If you are told only that you scored in the 80th percentile, do you know from that description exactly how it was calculated? Explain.
4. A study is conducted to determine whether people learn better with spaced or massed practice. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the massed-practice condition. At the end of the semester 12 subjects volunteer and are assigned to the spaced-practice condition. This experiment involves two kinds of non-random sampling: (1) Subjects are not randomly sampled from some specified population and (2) subjects are not randomly assigned to conditions. Which of the problems relates to the generality of the results? Which of the problems relates to the validity of the results? Which problem is more serious?
5. Give an example of an independent and a dependent variable.
6. Categorize the following variables as being qualitative or quantitative:  
Rating of the quality of a movie on a 7-point scale  
Age  
Country you were born in  
Favorite Color  
Time to respond to a question

7. Specify the level of measurement used for the items in Question 6.
8. Which of the following are linear transformations?
- Converting from meters to kilometers
  - Squaring each side to find the area
  - Converting from ounces to pounds
  - Taking the square root of each person's height.
  - Multiplying all numbers by 2 and then adding 5
  - Converting temperature from Fahrenheit to Centigrade
9. The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows:  $g = 16 + 3s$

Is this a linear transformation?

If a student got a raw score of 20, what is his test grade?

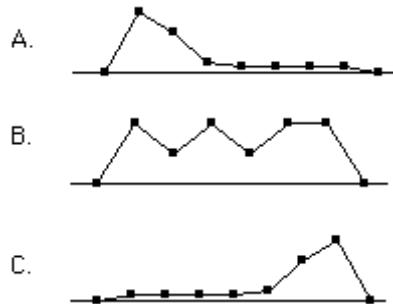
10. For the numbers 1, 2, 4, 16, compute the following:

$$\Sigma X$$

$$\Sigma X^2$$

$$(\Sigma X)^2$$

11. Which of the frequency polygons has a large positive skew? Which has a large negative skew?



12. What is more likely to have a skewed distribution: time to solve an anagram problem (where the letters of a word or phrase are rearranged into another

word or phrase like “dear” and “read” or “funeral” and “real fun”) or scores on a vocabulary test?

*Questions from Case Studies*

Angry Moods (AM) case study

13. (AM) Which variables are the participant variables? (They act as independent variables in this study.)

14. (AM) What are the dependent variables?

15. (AM) Is Anger-Out a quantitative or qualitative variable?

Teacher Ratings (TR) case study

16. (TR) What is the independent variable in this study?

ADHD Treatment (AT) case study

17. (AT) What is the independent variable of this experiment? How many levels does it have?

18. (AT) What is the dependent variable? On what scale (nominal, ordinal, interval, ratio) was it measured?

# 2. Graphing Distributions

- A. Qualitative Variables
- B. Quantitative Variables
  - 1. Stem and Leaf Displays
  - 2. Histograms
  - 3. Frequency Polygons
  - 4. Box Plots
  - 5. Bar Charts
  - 6. Line Graphs
  - 7. Dot Plots
- C. Exercises

Graphing data is the first and often most important step in data analysis. In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves. This is all the more unfortunate because computers can create many types of graphs quickly and easily.

This chapter covers some classic types of graphs such bar charts that were invented by William Playfair in the 18th century as well as graphs such as box plots invented by John Tukey in the 20th century.

# Graphing Qualitative Variables

by David M. Lane

## *Prerequisites*

- Chapter 1: Variables

## *Learning Objectives*

1. Create a frequency table
2. Determine when pie charts are valuable and when they are not
3. Create and interpret bar charts
4. Identify common graphical mistakes

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

## Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative

frequencies, which are the proportion of responses in each category. For example, the relative frequency for “none” of  $0.17 = 85/500$ .

Table 1. Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1

## Pie Charts

The pie chart in Figure 1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

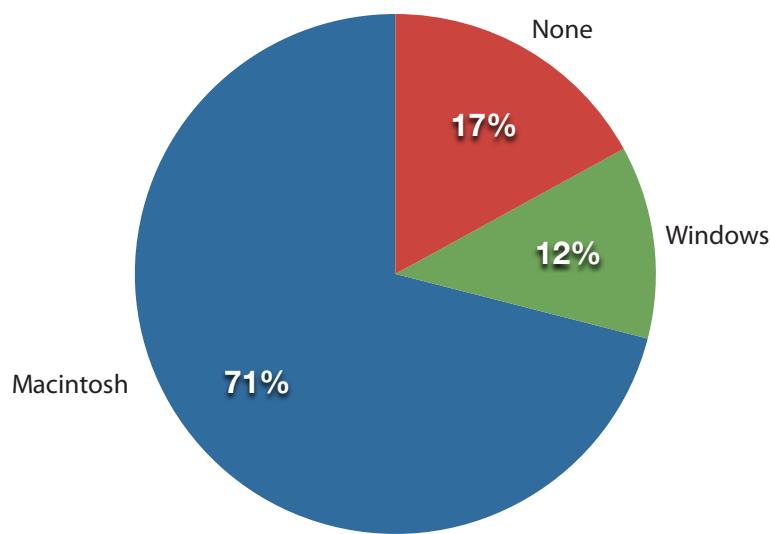


Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted “The only worse design than a pie chart is several of them.”

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

## Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 2. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations in each category as is typical in pie charts.

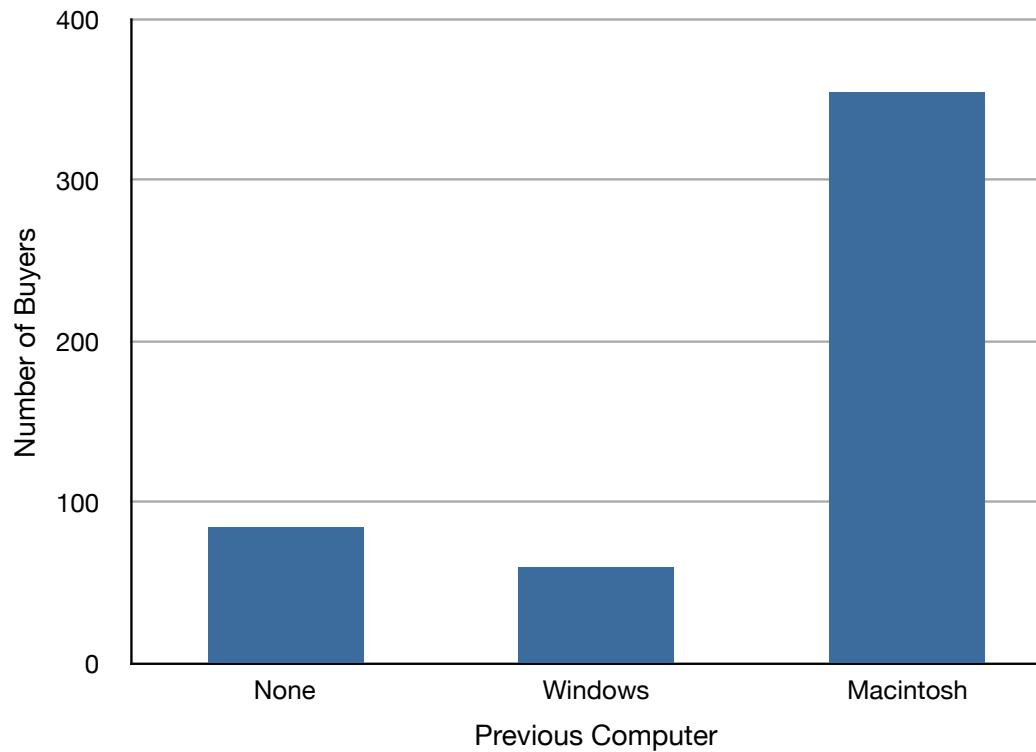


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

## Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the “distributions” of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 3 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

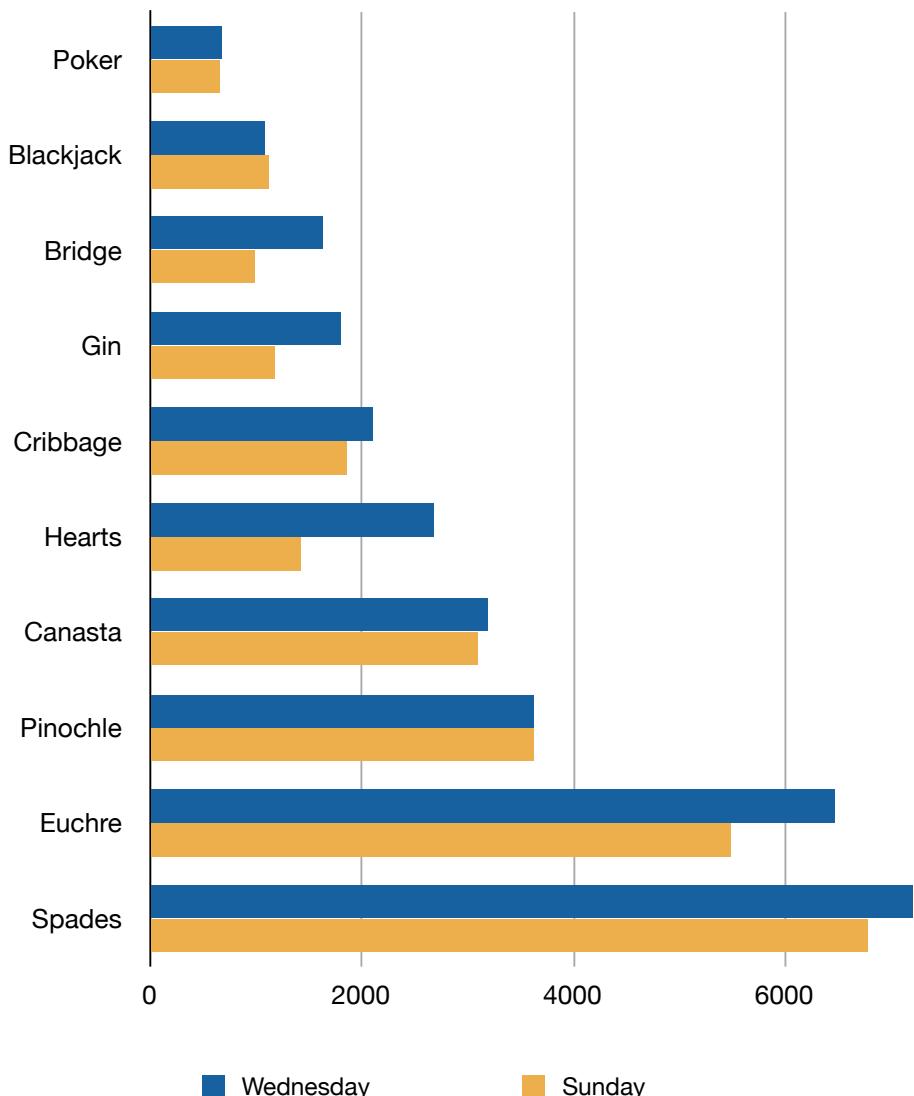


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.

The bars in Figure 3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in this chapter.

### Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 4 are usually not as effective as their two-dimensional counterparts.

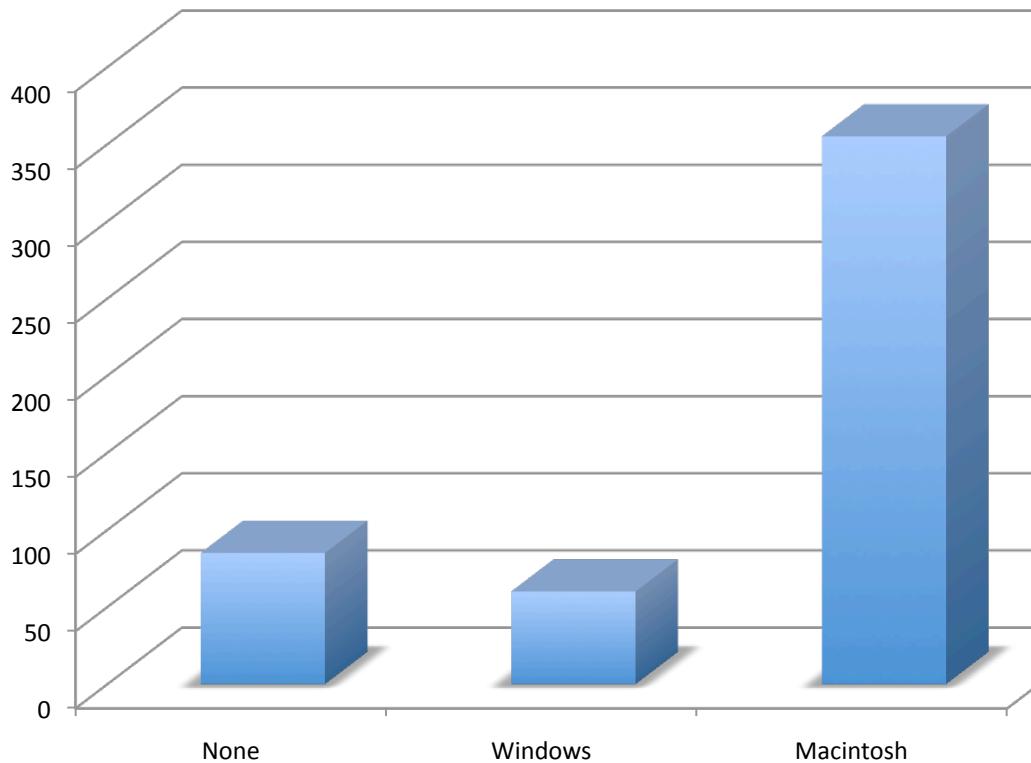


Figure 4. A three-dimensional version of Figure 2.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 5 instead of Figure 2! Edward Tufte coined the term “lie factor” to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

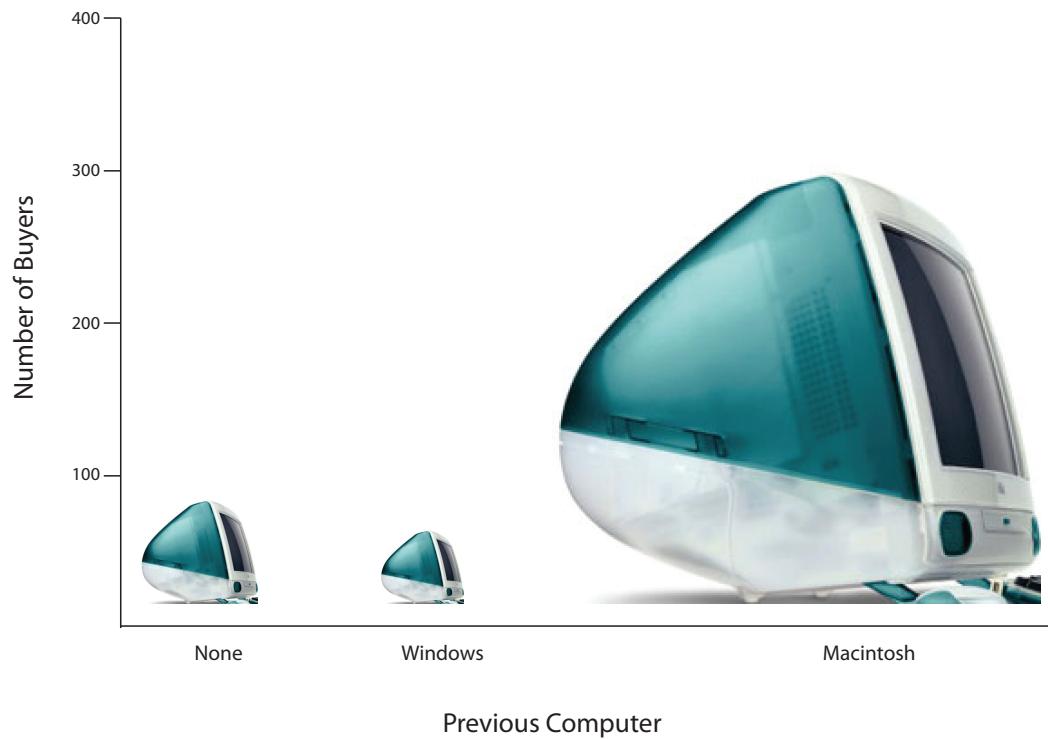


Figure 5. A redrawing of Figure 2 with a lie factor greater than 8.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

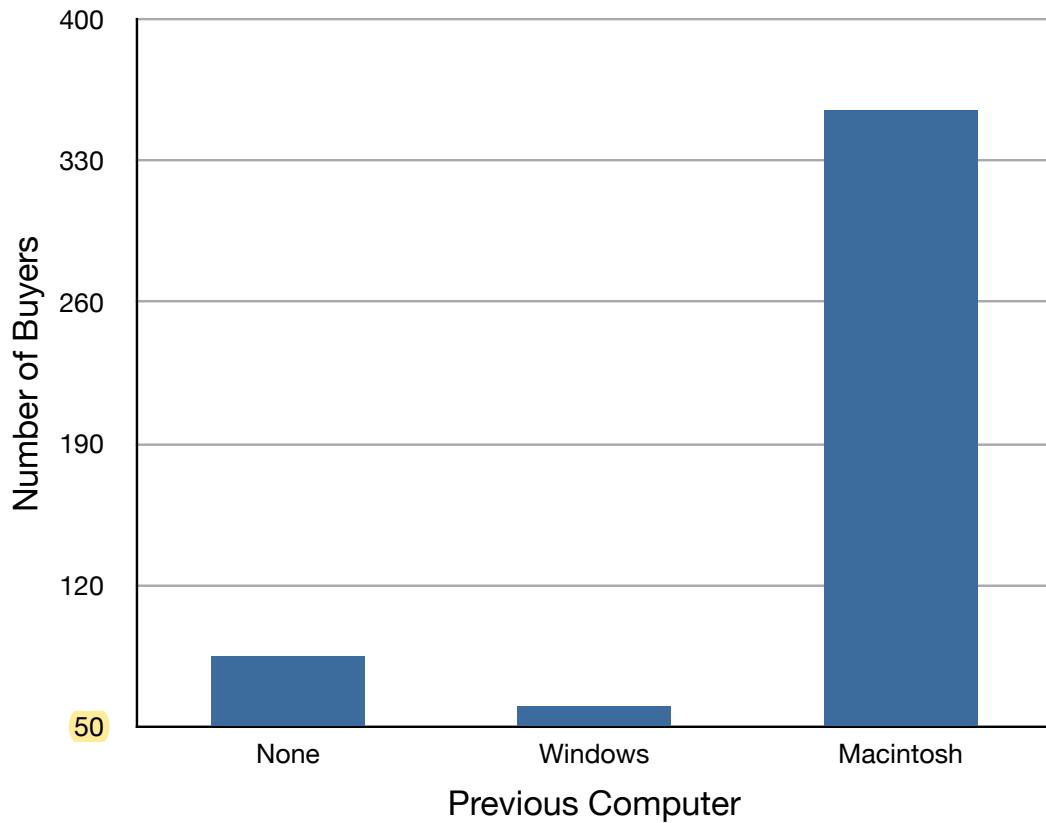


Figure 6. A redrawing of Figure 2 with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

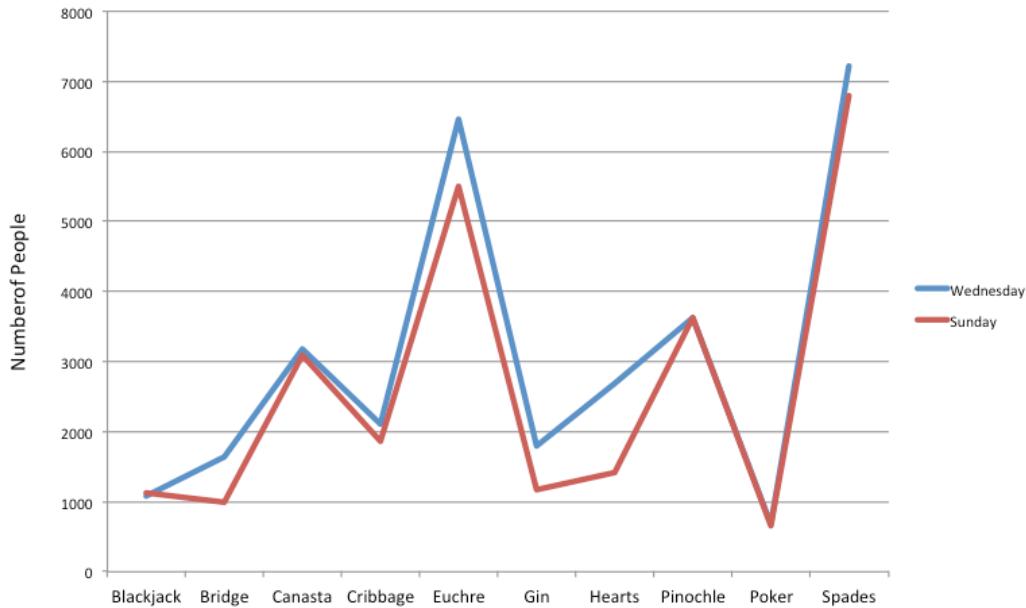


Figure 7. A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

## Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

# Graphing Quantitative Variables

1. Stem and Leaf Displays
2. Histograms
3. Frequency Polygons
4. Box Plots
5. Bar Charts
6. Line Graphs
7. Dot Plots

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs: (1) stem and leaf displays, (2) histograms, (3) frequency polygons, (4) box plots, (5) bar charts, (6) line graphs, (7) dot plots, and (8) scatter plots (discussed in a different chapter). Some graph types such as stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best-suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

# Stem and Leaf Displays

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions

## *Learning Objectives*

1. Create and interpret basic stem and leaf displays
2. Create and interpret back-to-back stem and leaf displays
3. Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, we will start with an example. Consider Table 1 that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6
--

A stem and leaf display of the data is shown in Figure 1. The left portion of Figure 1 contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits. A stem of 3, for example, can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

3 2337
2 001112223889
1 2244456888899
0 69

Figure 1. Stem and leaf display of the number of touchdown passes.

To make this clear, let us examine Figure 1 more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD passes for the first four teams in Table 1. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table 1, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 1 than in Table 1. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TD's, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure 2 shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table 1. The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

3 7
3 233
2 889
2 001112223
1 56888899
1 22444
0 69

Figure 2. Stem and leaf display with the stems split in two.

Figure 2 is more revealing than Figure 1 because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a “back-to-back stem and leaf display.” Figure 3 shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with 11, 12, and 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Figure 3. Back-to-back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

Figure 3 helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table 2 shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such

as “bug.” In the second condition, the same words were preceded by a weapon word such as “gun” or “knife.” The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

Table 2. The effects of priming (thousandths of a second).

43.2, 42.9, 35.6, 25.6, 25.4, 23.6, 20.5, 19.9, 14.4, 12.7, 11.3, 10.2, 10.0, 9.1, 7.5, 5.4, 4.7, 3.8, 2.1, 1.2, -0.2, -6.3, -6.7, -8.8, -10.4, -10.5, -14.9, -14.9, -15.0, -18.5, -27.4
--

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure 4. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf) the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number -27. The second-to-last row represents the numbers -10, -10, -15, etc. Once again, we have rounded the original values from Table 2.

4   33
3   6
2   00456
1   00134
0   1245589
-0   0679
-1   005559
-2   7

Figure 4. Stem and leaf display with negative numbers and rounding.

Observe that the figure contains a row headed by “0” and another headed by “-0.” The stem of 0 is for numbers between 0 and 9, whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, -6, -7, and -9. Values that are exactly 0 before rounding should be split as evenly as possible between the “0” and “-0” rows. In Table 2, none of the values are 0 before rounding. The “0” that appears in the “-0” row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure 5 portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.

Figure 5. Stem and leaf display of populations of 185 US cities with populations between 100,000 and 500,000 in 1988.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example the largest number (493,559) was

rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: 0-1, 2-3, 4-5, 6-7, and 8-9.

Whether your data can be suitably represented by a stem and leaf display depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure 5 fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

# Histograms

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 2: Graphing Qualitative Data

## *Learning Objectives*

1. Create a grouped frequency distribution
2. Create a histogram based on a grouped frequency distribution
3. Determine an appropriate bin width

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items each graded as “correct” or “incorrect.” The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 1.

Table 1. Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36

129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too “choppy.” More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 1.

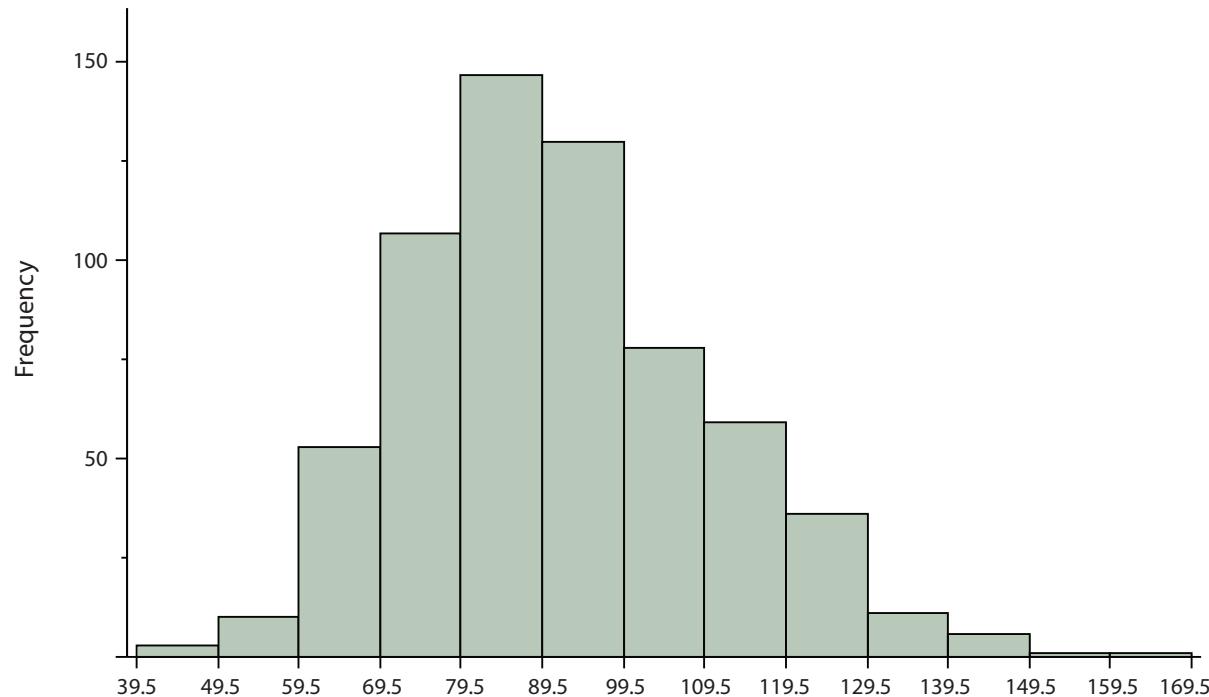


Figure 1. Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in Chapter 3.)

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some “rules of thumb” that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) Sturges’ rule is to set the number of intervals as close as possible to  $1 + \text{Log}_2(N)$ , where  $\text{Log}_2(N)$  is the base 2 log of the number of observations. The formula can also be written as  $1 + 3.3 \text{ Log}_{10}(N)$  where  $\text{Log}_{10}(N)$  is the log base 10 of the number of observations. According to Sturges’ rule, 1000 observations

would be graphed with 11 class intervals since 10 is the closest integer to  $\log_2(1000)$ . We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by Sturges' rule. For the psychology test example used above, Sturges' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure 1 to create a histogram that seemed clearest. **The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.**

To provide experience in constructing histograms, we have developed an interactive demonstration ([external link](#); Java required). The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

# Frequency Polygons

by David M. Lane

## *Prerequisites*

- Chapter 2: Histograms

## *Learning Objectives*

1. Create and interpret frequency polygons
2. Create and interpret cumulative frequency polygons
3. Create and interpret overlaid frequency polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

A frequency polygon for 642 psychology test scores shown in Figure 1 was constructed from the frequency table shown in Table 1.

Table 1. Frequency Distribution of Psychology Test Scores

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173

79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	170.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 1. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

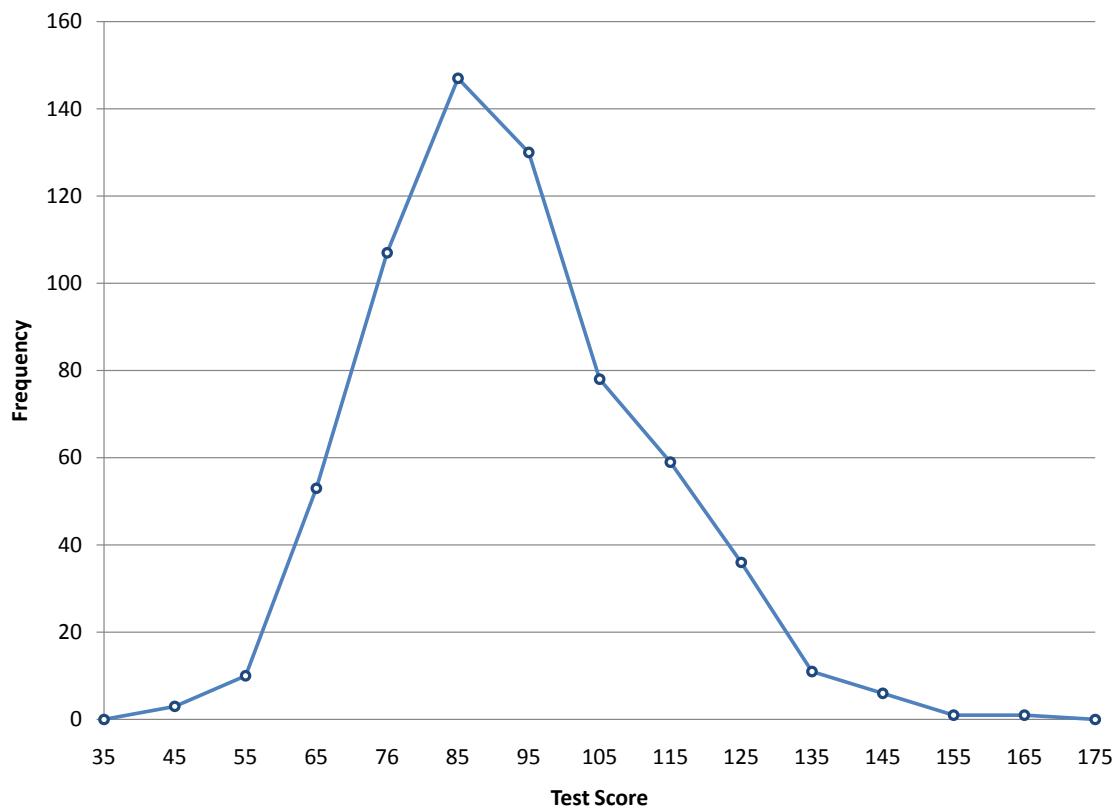


Figure 1. Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in Figure 2. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled “35,” three in the interval “45,” and 10 in the interval “55.” Therefore, the Y value corresponding to “55” is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

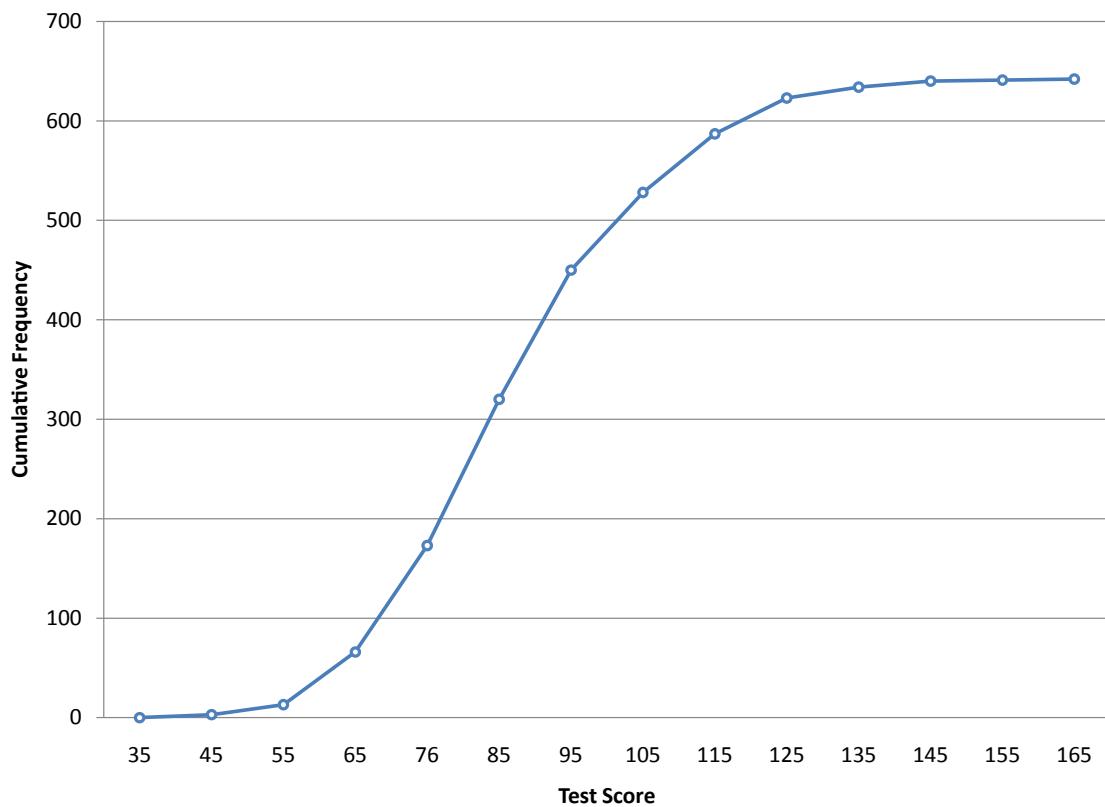


Figure 2. Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 3. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

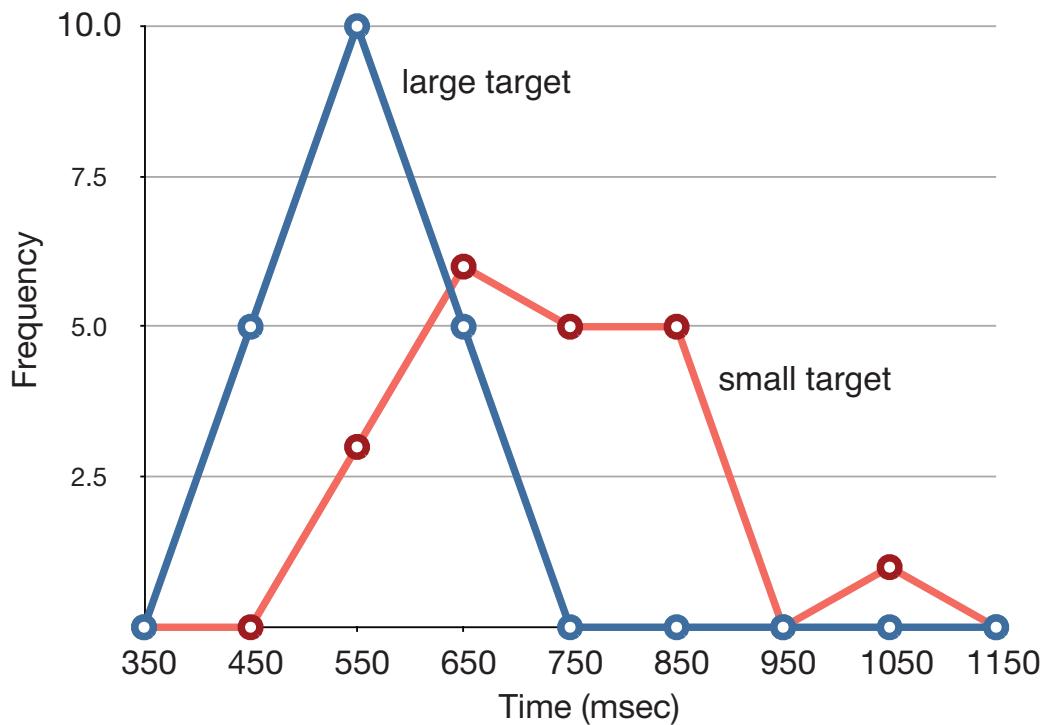


Figure 3. Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4 using the same data from the cursor task. The

difference in distributions for the two targets is again evident.

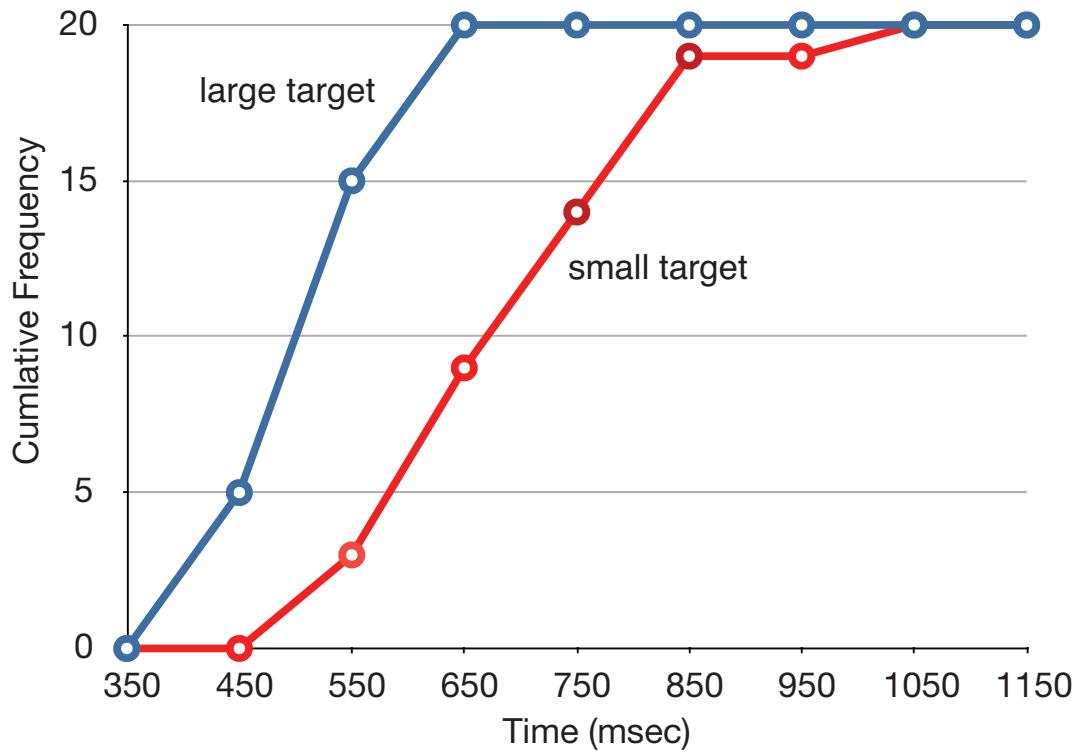


Figure 4. Overlaid cumulative frequency polygons.

# Box Plots

by David M. Lane

## *Prerequisites*

- Chapter 1: Percentiles
- Chapter 2: Histograms
- Chapter 2: Frequency Polygons

## *Learning Objectives*

1. Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
2. Create a box plot
3. Create parallel box plots
4. Determine whether a box plot is appropriate for a given data set

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section we present another important graph, called a box plot. Box plots are useful for identifying outliers and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. Students in Introductory Statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25th, 50th, and 75th percentiles in the distribution of scores. Figure 1 shows how these three statistics are used. For each gender we draw a box extending from the 25th percentile to the 75th percentile. The 50th percentile is drawn inside the box. Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile.

The data for the women in our sample are shown in Table 1.

Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

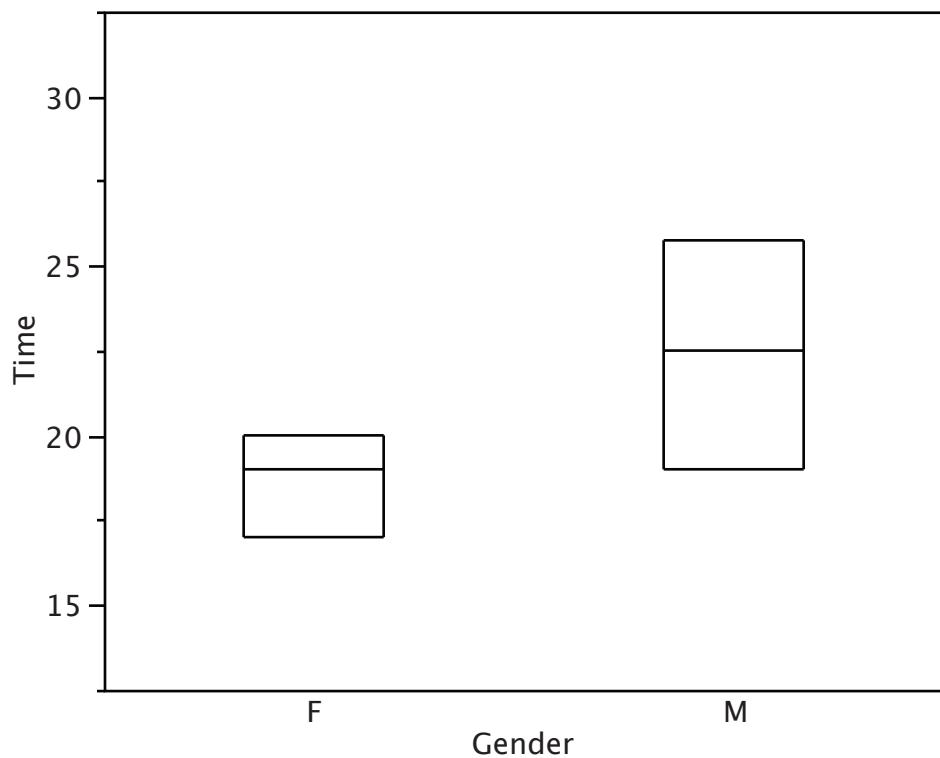


Figure 1. The first step in creating box plots.

Before proceeding, the terminology in Table 2 is helpful.

Table 2. Box plot terms and values for women's times.

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17

H-Spread	Upper Hinge - Lower Hinge	3
Step	$1.5 \times \text{H-Spread}$	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put “whiskers” above and below each box to give additional information about the spread of data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).

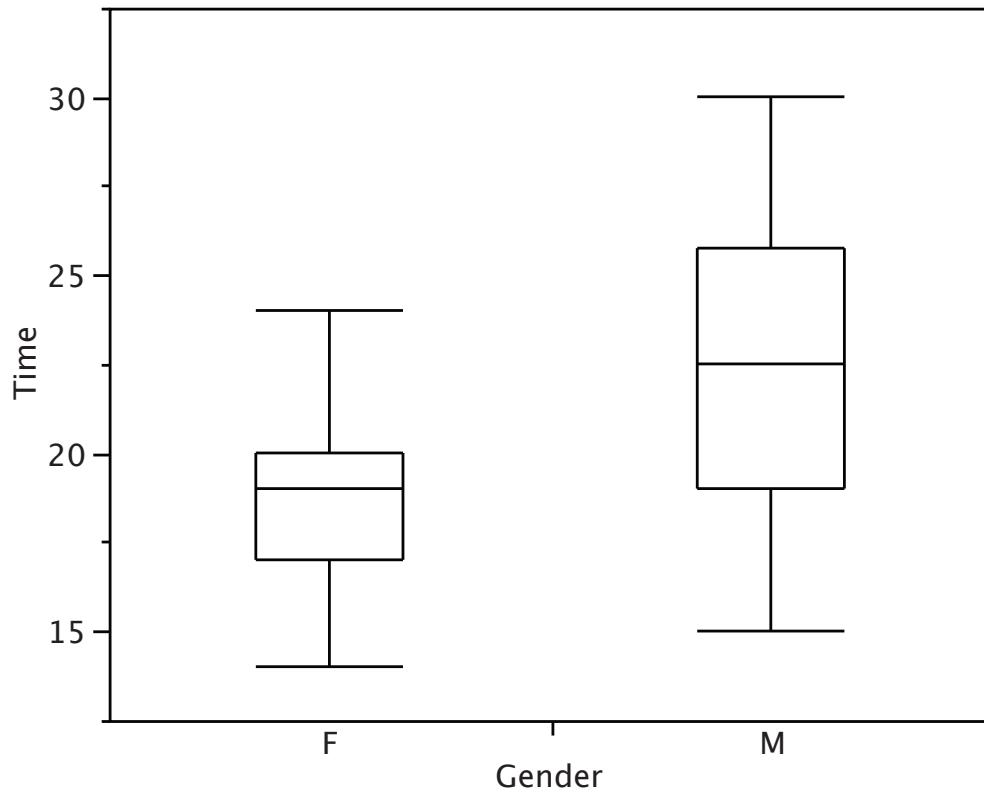


Figure 2. The box plots with the whiskers drawn.

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "o's" and far out values are indicated by asterisks (\*). In our data, there are no far-out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 3.

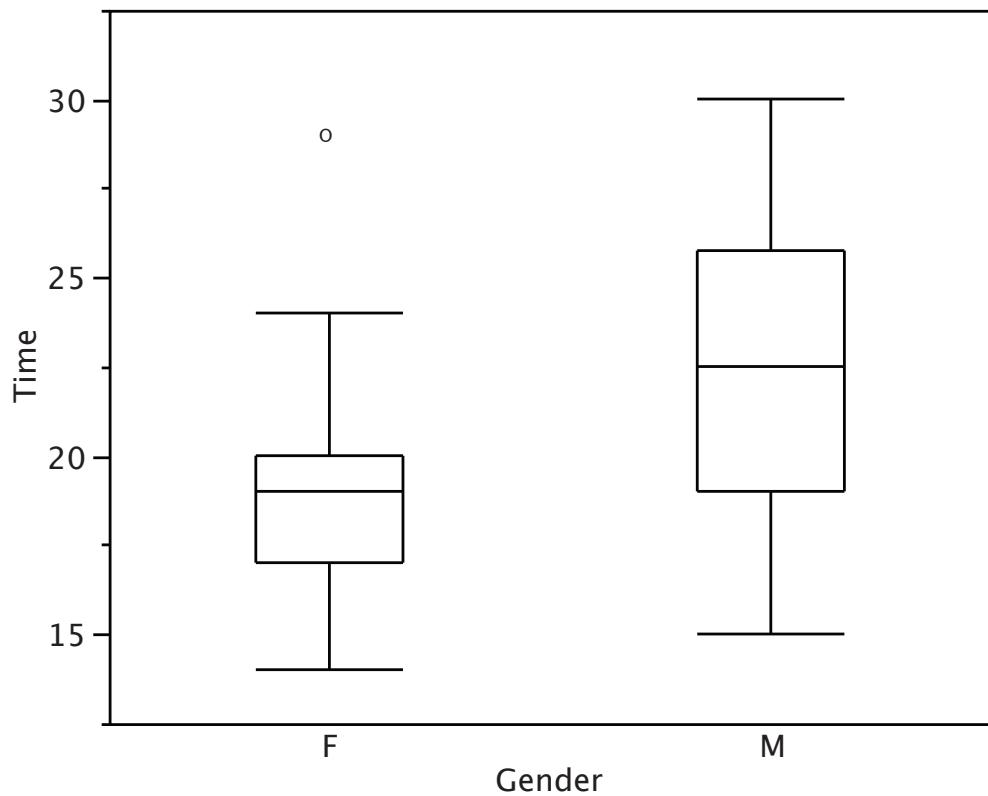


Figure 3. The box plots with the outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 4 shows the result of adding means to our box plots.

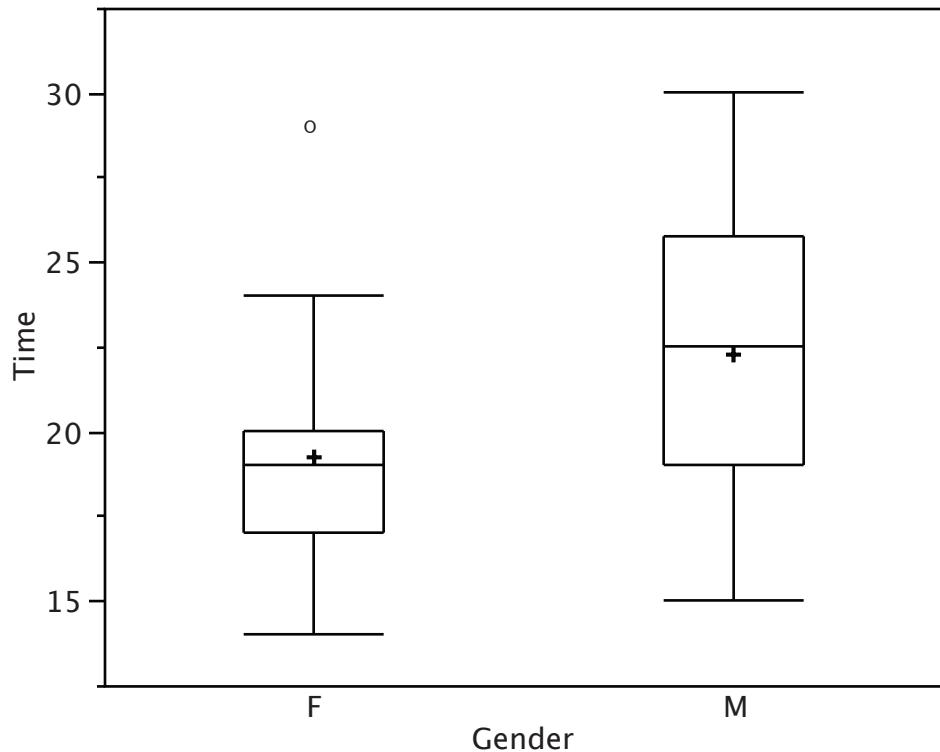


Figure 4. The completed box plots.

Figure 4 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women's times are between 17 and 20 seconds whereas half the men's times are between 19 and 25.5 seconds. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 5 shows the box plot for the women's data with detailed labels.

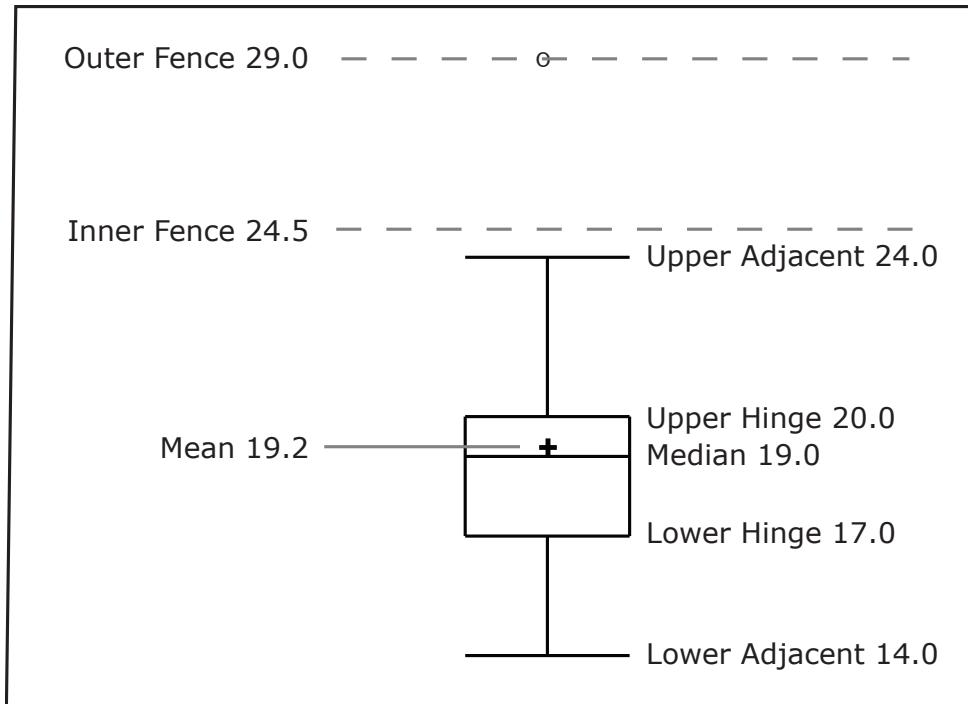


Figure 5. The box plots for the women's data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot and to examine these details one should use create a histogram and/or a stem and leaf display.

### Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 6 are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.

5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

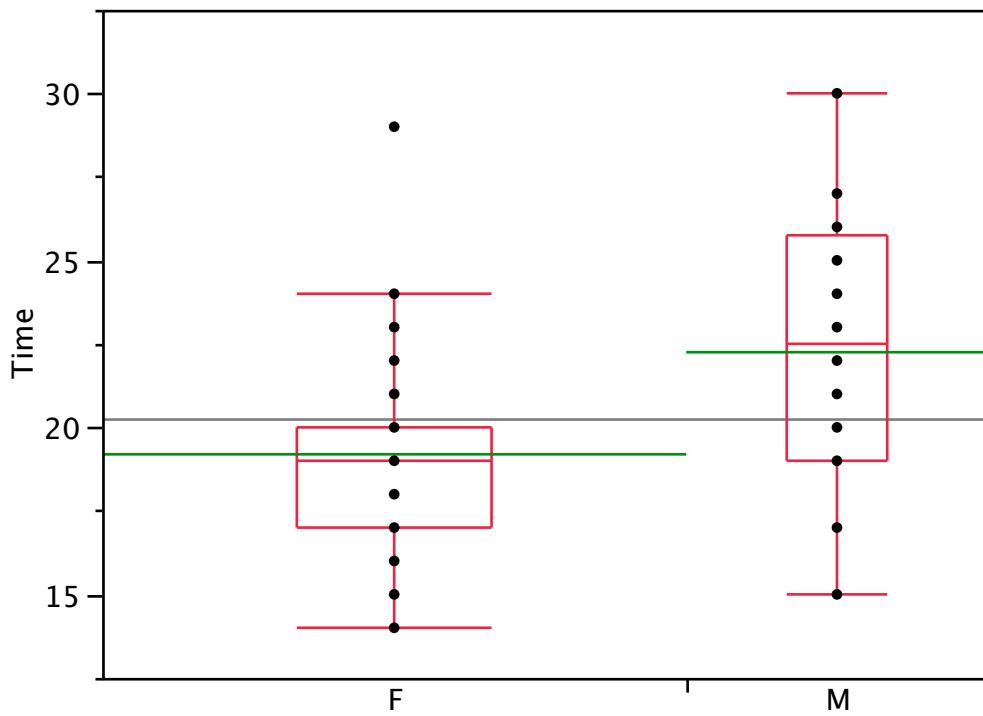


Figure 6. Box plots showing the individual scores and the means.

Each dot in Figure 6 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter the points. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittered. Figure 7 shows what jittering looks like.

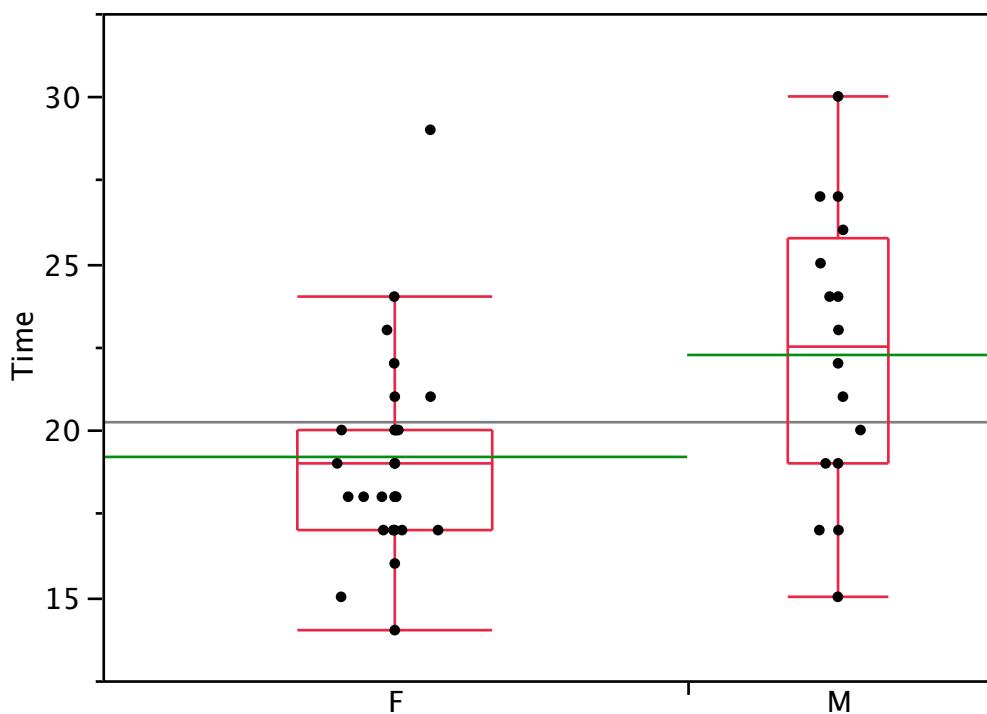


Figure 7. Box plots with the individual scores jittered.

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

# Bar Charts

by David M. Lane

## *Prerequisites*

- Chapter 2: Graphing Qualitative Variables

## *Learning Objectives*

1. Create and interpret bar charts
2. Judge whether a bar chart or another graph such as a box plot would be more appropriate

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 1 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

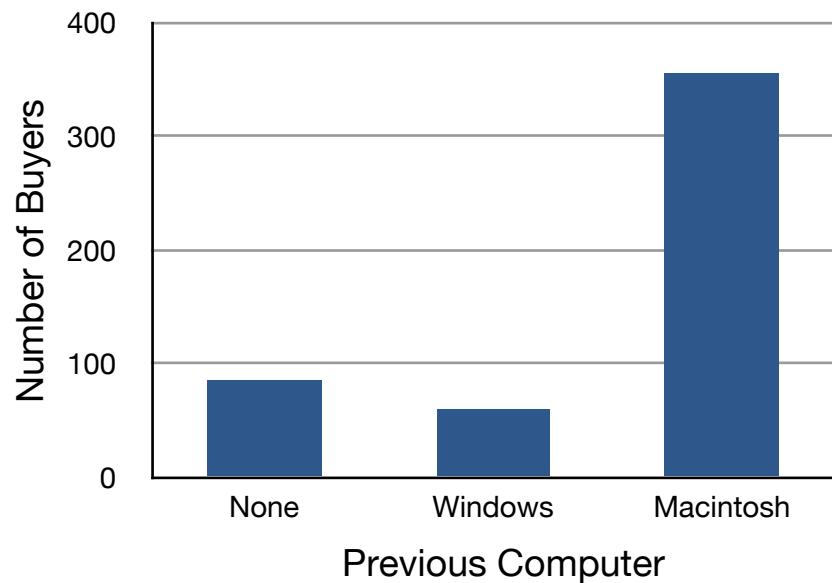


Figure 1. iMac buyers as a function of previous computer ownership.

In this section we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 2 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 2001. Notice that both the S & P and the Nasdaq had “negative increases” which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity *percentage increase*.

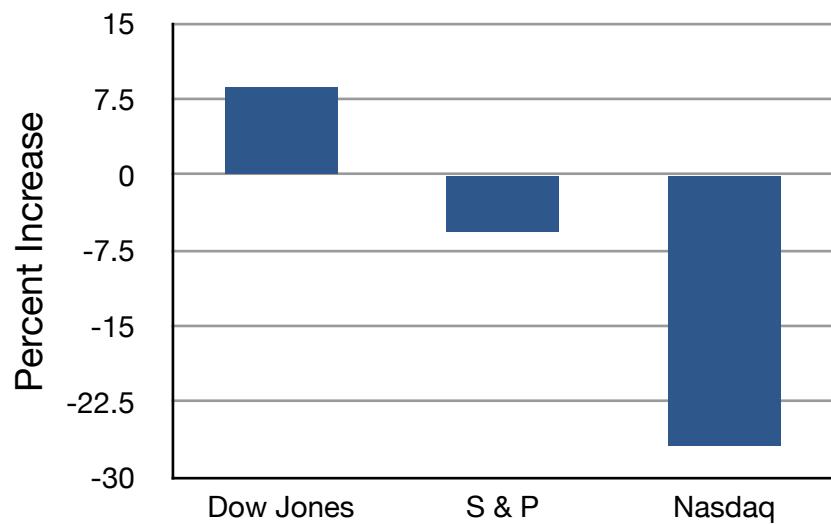


Figure 2. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Bar charts are particularly effective for showing change over time. Figure 3, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

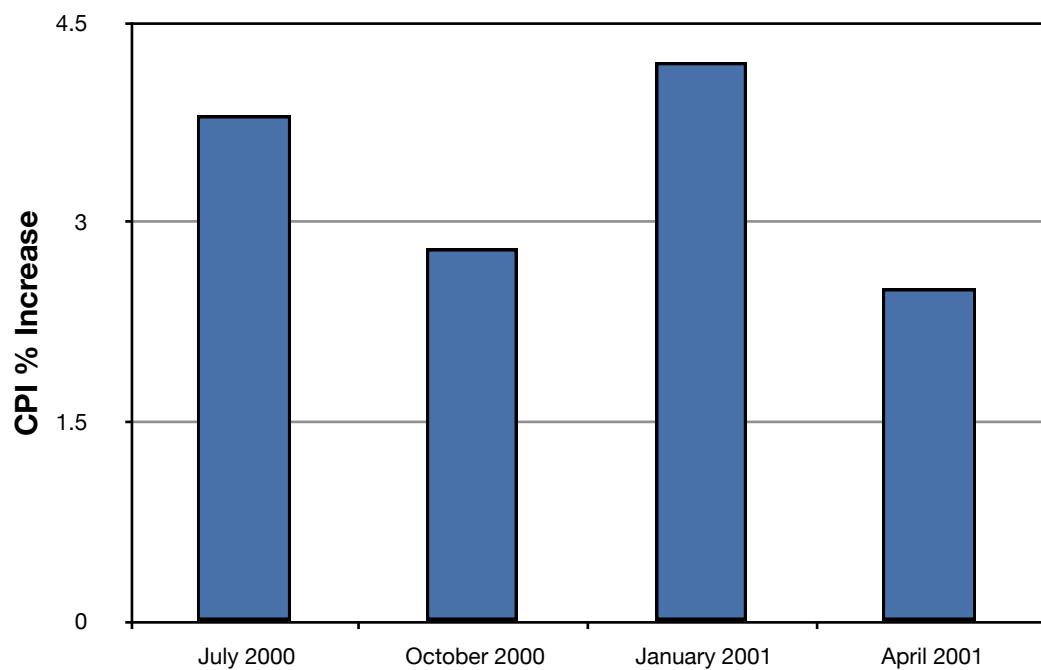


Figure 3. Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 4 shows the mean time it took one of us (DL) to move the cursor to either a small target or a large target. On average, more time was required for small targets than for large ones.

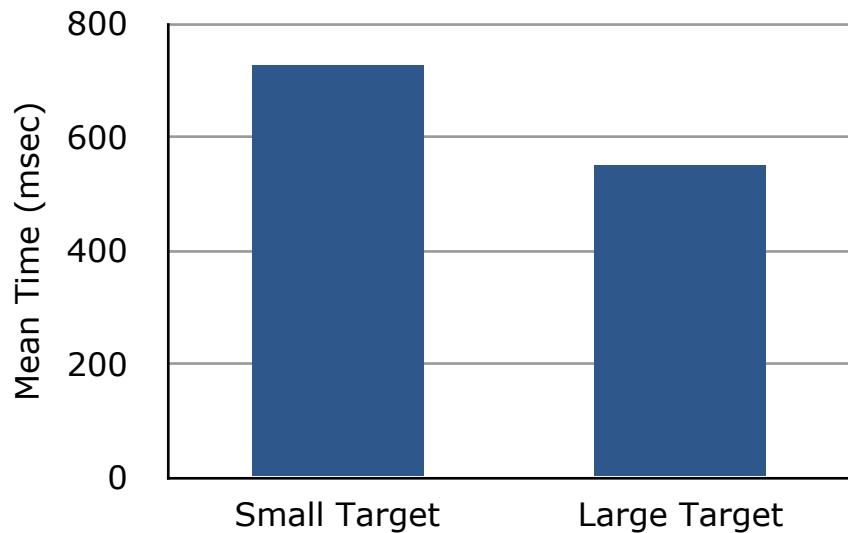


Figure 4. Bar chart showing the means for the two conditions.

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the cursor-movement data is shown in Figure 5. You can see that Figure 5 reveals more about the distribution of movement times than does Figure 4.

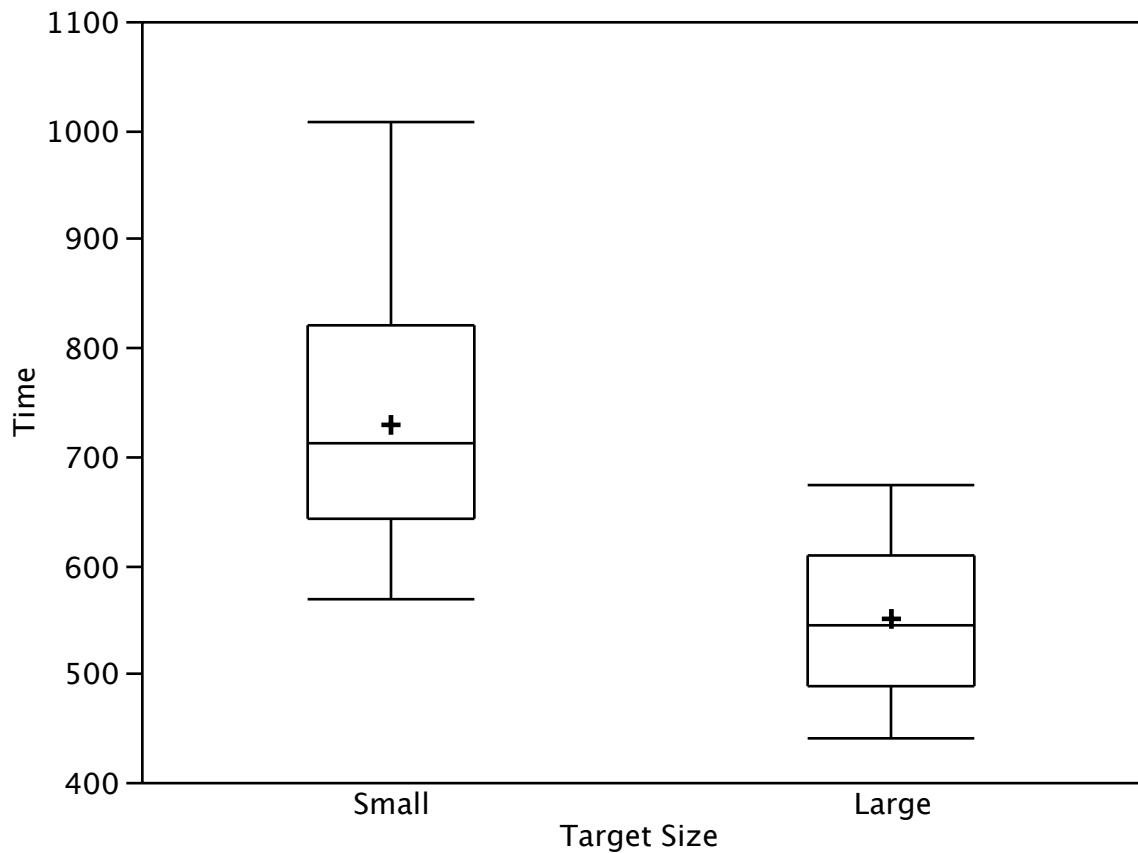


Figure 5. Box plots of times to move the cursor to the small and large targets.

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

# Line Graphs

by David M. Lane

## *Prerequisites*

- Chapter 2: Bar Charts

## *Learning Objectives*

1. Create and interpret line graphs
2. Judge whether a line graph would be appropriate for a given data set

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure 1 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

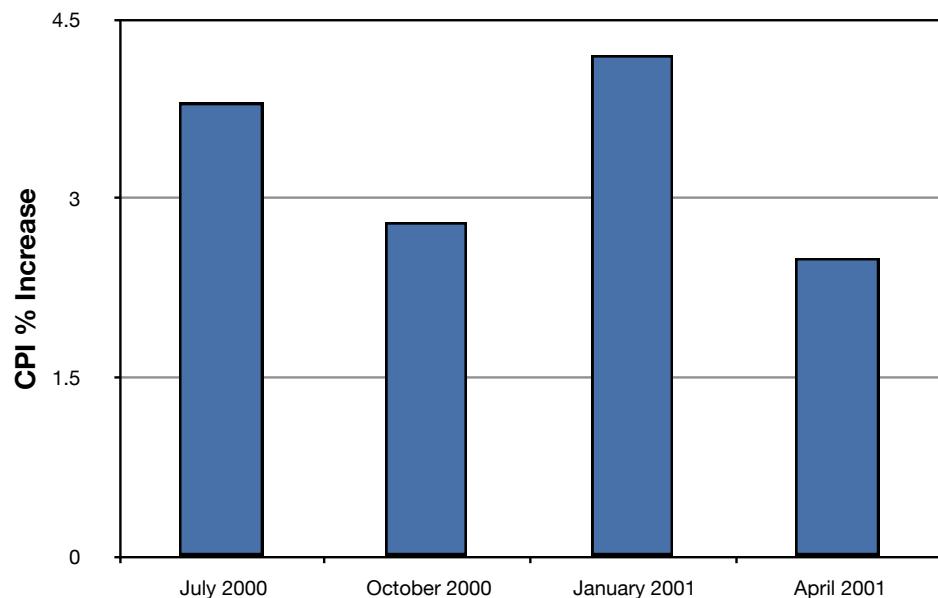


Figure 1. A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure 2. Although the figures are similar, the line graph emphasizes the change from period to period.

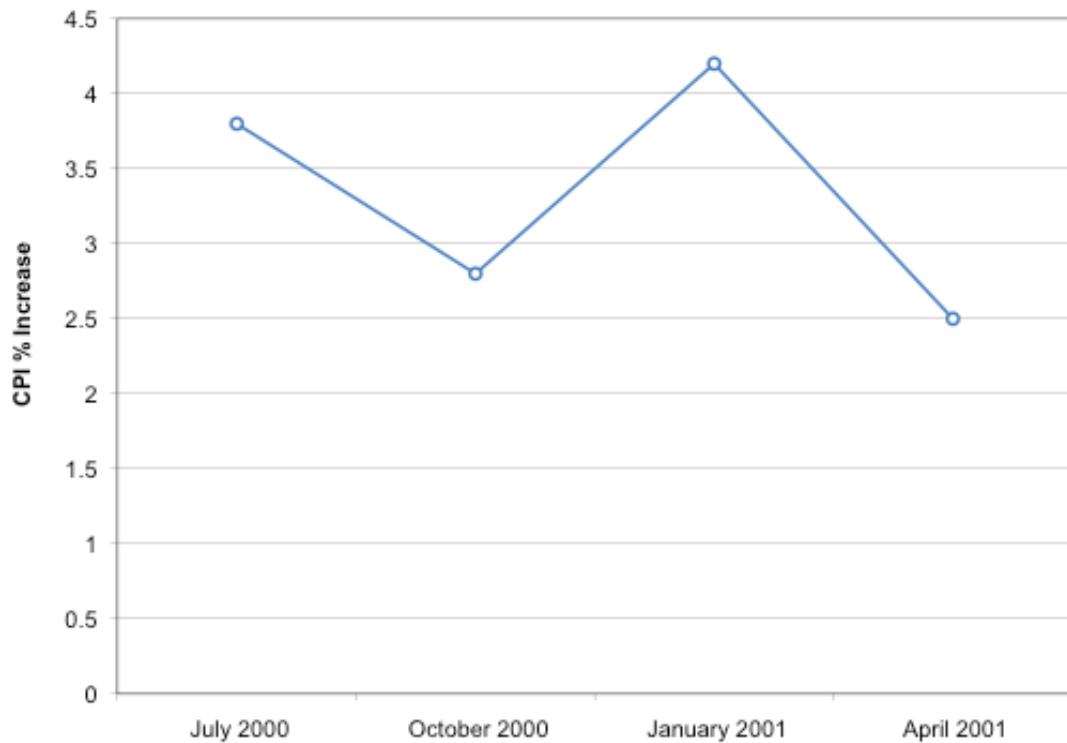


Figure 2. A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 3, for example, shows percent increases and decreases in five components of the CPI. The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its

interpretation would not be as easy.

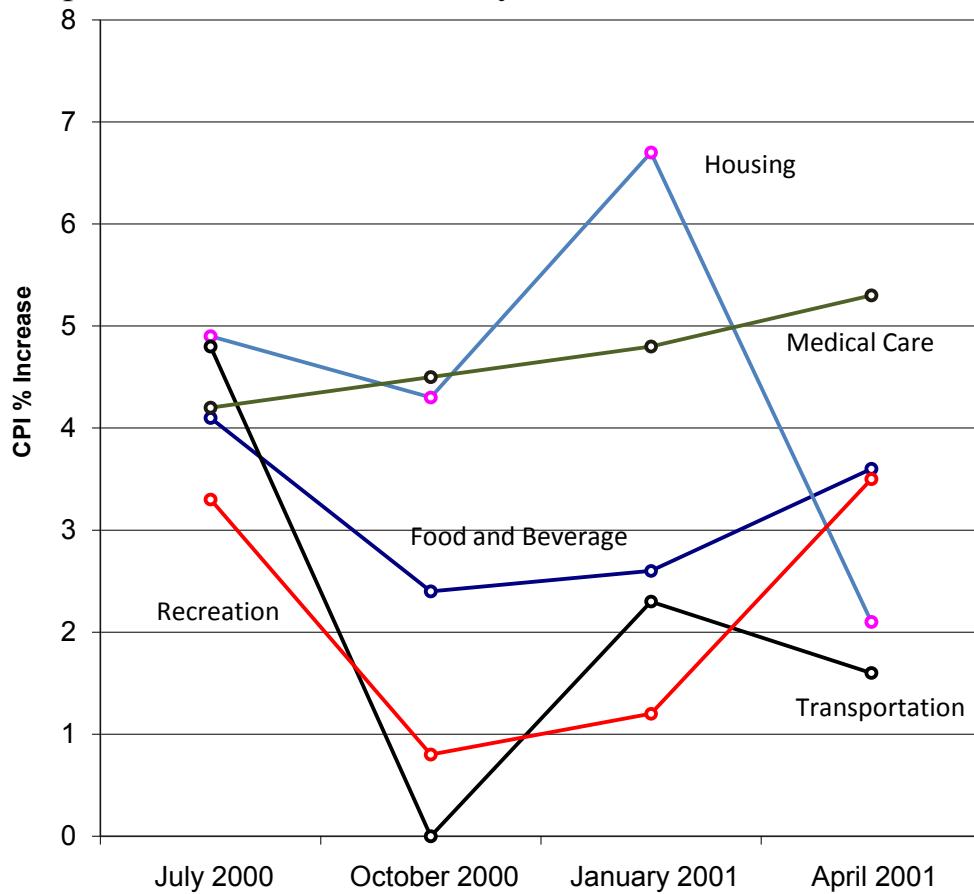


Figure 3. A line graph of the percent change in five components of the CPI over time.

Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables. Figure 4 inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 4 is that it gives the false impression that the games are naturally ordered in a numerical way.

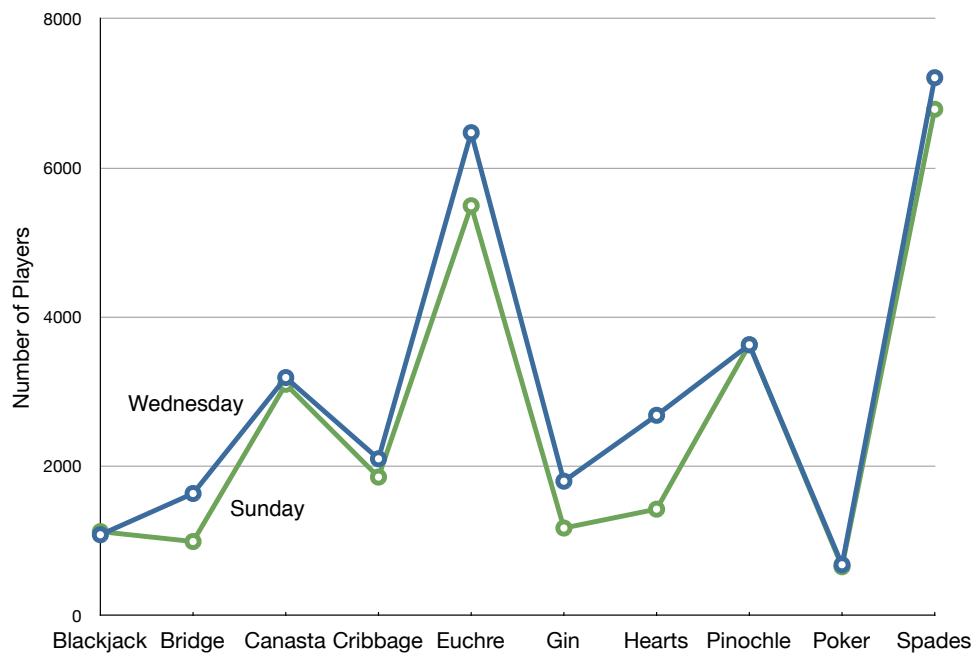


Figure 4. A line graph, inappropriately used, depicting the number of people playing different card games on Wednesday and Sunday.

# Dot Plots

by David M. Lane

## *Prerequisites*

- Chapter 2: Bar Charts

## *Learning Objectives*

1. Create and interpret dot plots
2. Judge whether a dot plot would be appropriate for a given data set

Dot plots can be used to display various types of information. Figure 1 uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. Each dot represents a single M & M. From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc.

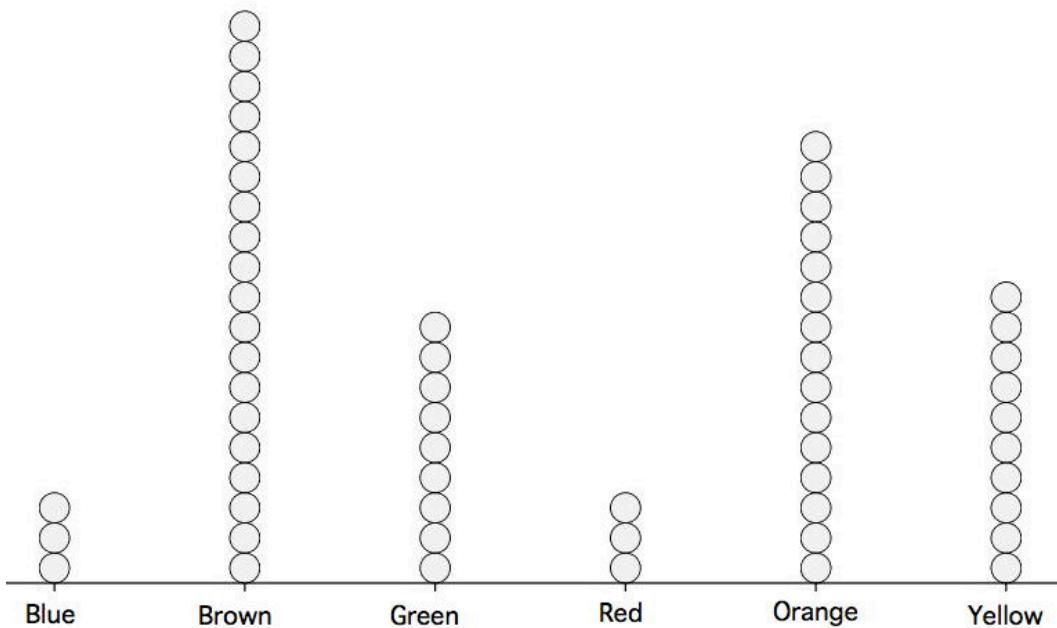


Figure 1. A dot plot showing the number of M & M's of various colors in a bag of M & M's.

The dot plot in Figure 2 shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure 1, the location rather than the number of dots represents the frequency.

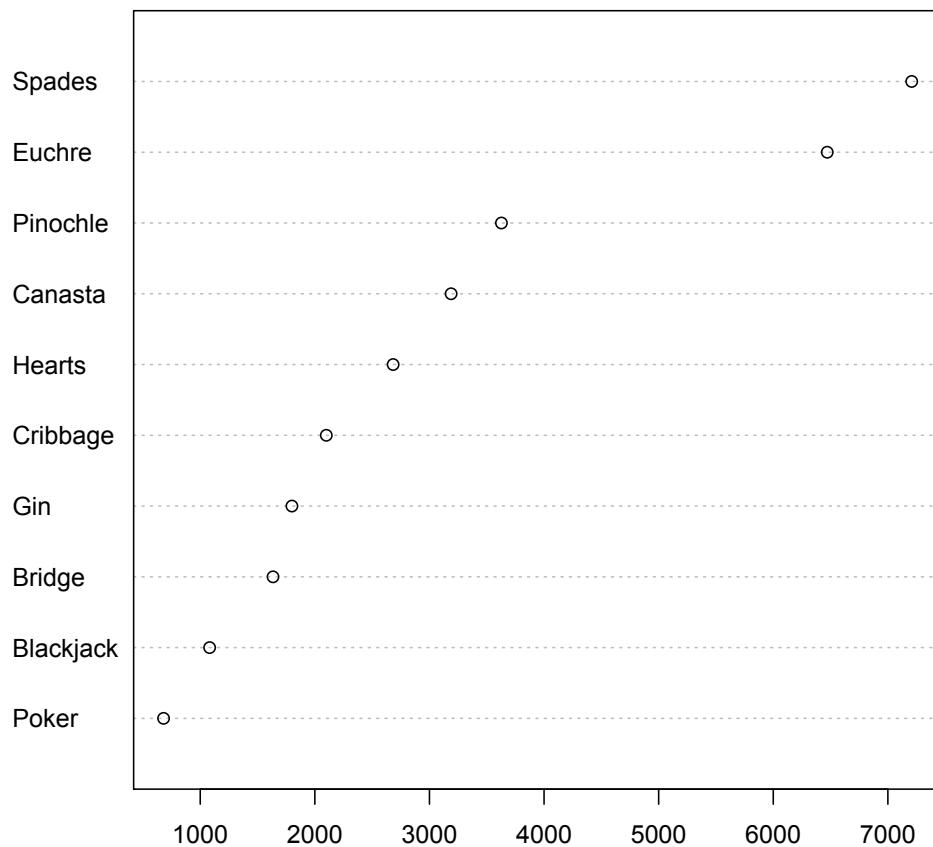


Figure 2. A dot plot showing the number of people playing various card games on a Wednesday.

The dot plot in Figure 3 shows the number of people playing on a Sunday and on a Wednesday. This graph makes it easy to compare the popularity of the games separately for the two days, but does not make it easy to compare the popularity of a given game on the two days.

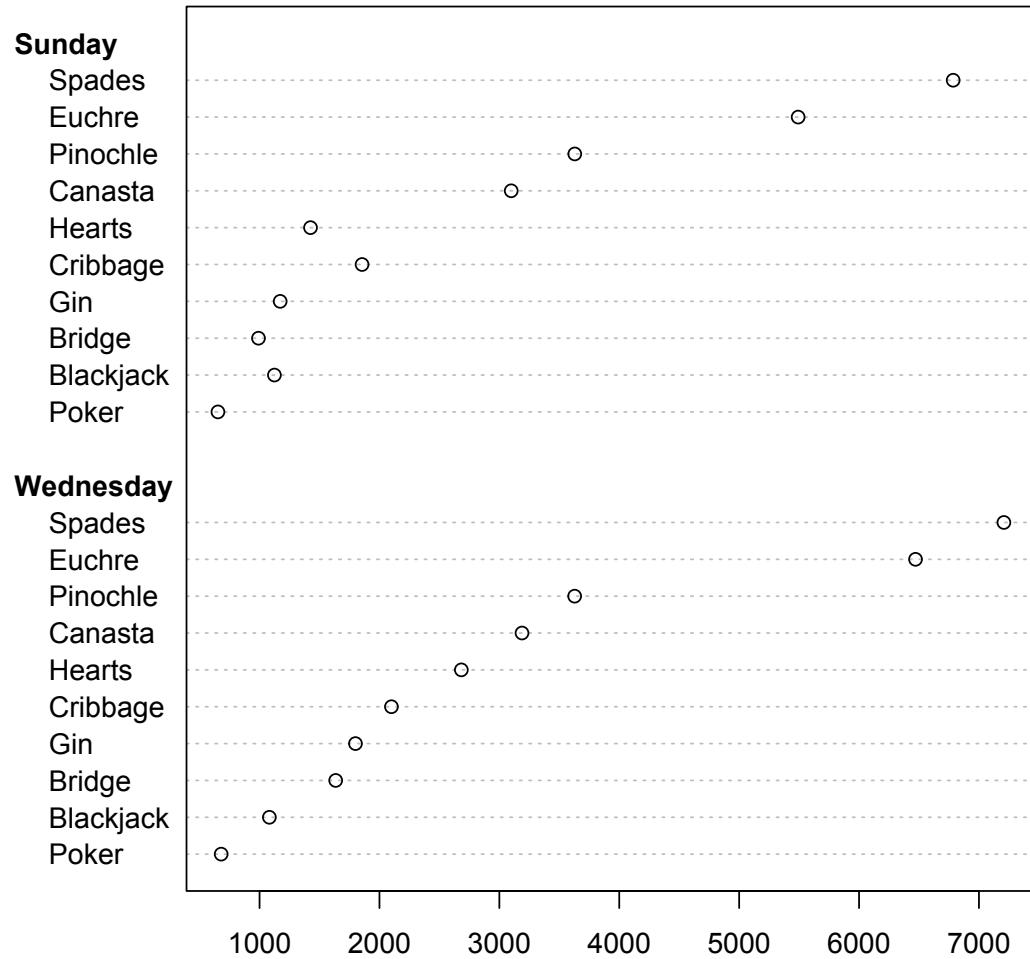


Figure 3. A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday.

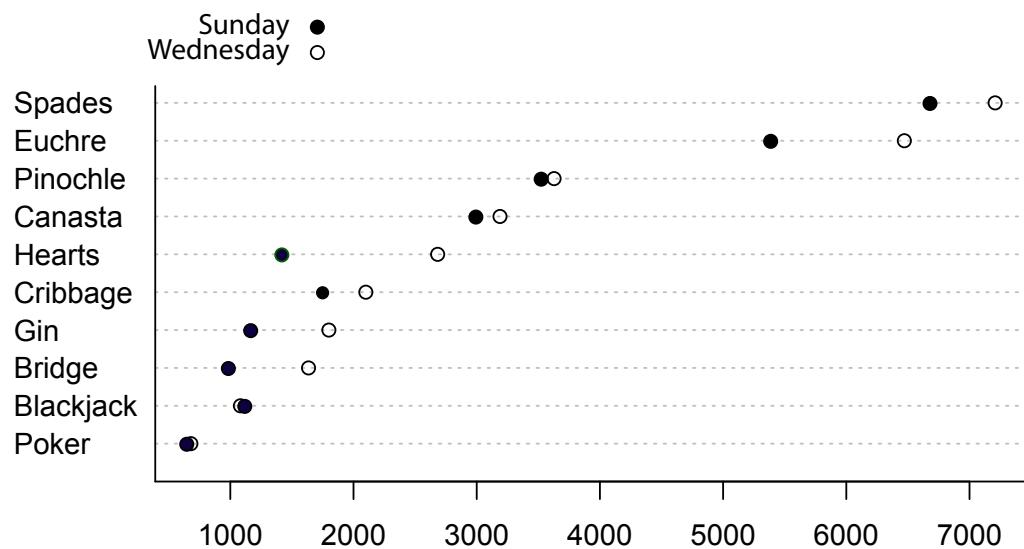


Figure 4. An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday.

The dot plot in Figure 4 makes it easy to compare the days of the week for specific games while still portraying differences among games.

# Statistical Literacy

by Seyd Ercan and David Lane

## *Prerequisites*

- Chapter 2: Graphing Distributions

Fox News aired the line graph below showing the number unemployed during four quarters between 2007 and 2010.



## **What do you think?**

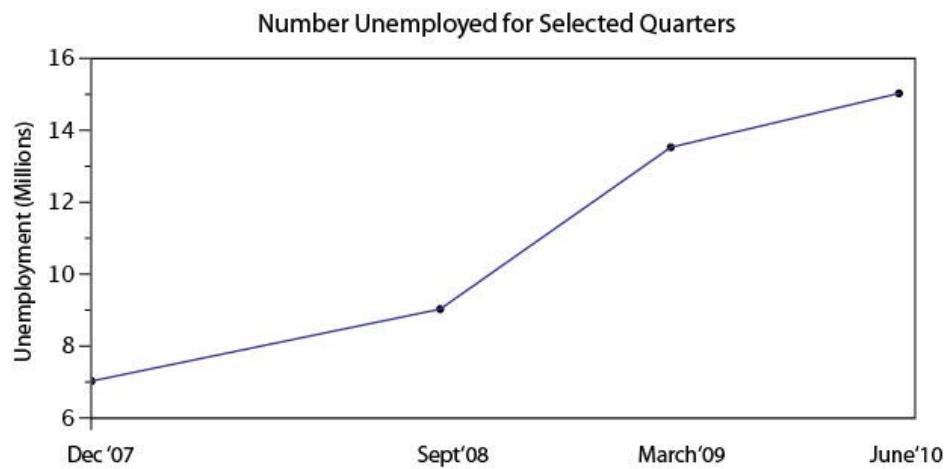
Does Fox News' line graph provide misleading information? Why or Why not?

Think about this before continuing:

There are major flaws with the Fox News graph. First, the title of the graph is misleading. Although the data show the number unemployed, Fox News' graph is titled "Job Loss by Quarter." Second, the intervals on the X-axis are misleading. Although there are 6 months between September 2008 and March 2009 and 15 months between March 2009 and June 2010, the intervals are represented in the graph by very similar lengths. This gives the false impression that unemployment increased steadily.

The graph presented below is corrected so that distances on the X-axis are proportional to the number of days between the

dates. This graph shows clearly that the rate of increase in the number unemployed is greater between September 2008 and March 2009 than it is between March 2009 and June 2010.



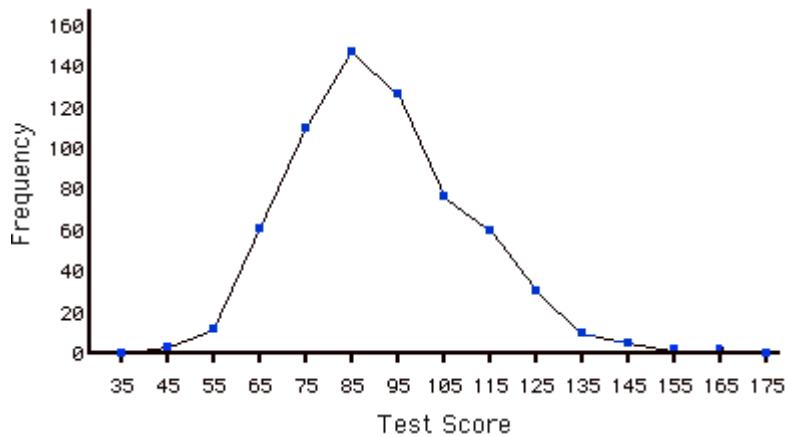
## References

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.) (p. 178). Cheshire, CT: Graphics Press.

## Exercises

### Prerequisites

- All material presented in the Graphing Distributions chapter
1. Name some ways to graph quantitative variables and some ways to graph qualitative variables.
  2. Based on the frequency polygon displayed below, the most common test grade was around what score? Explain.



3. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?

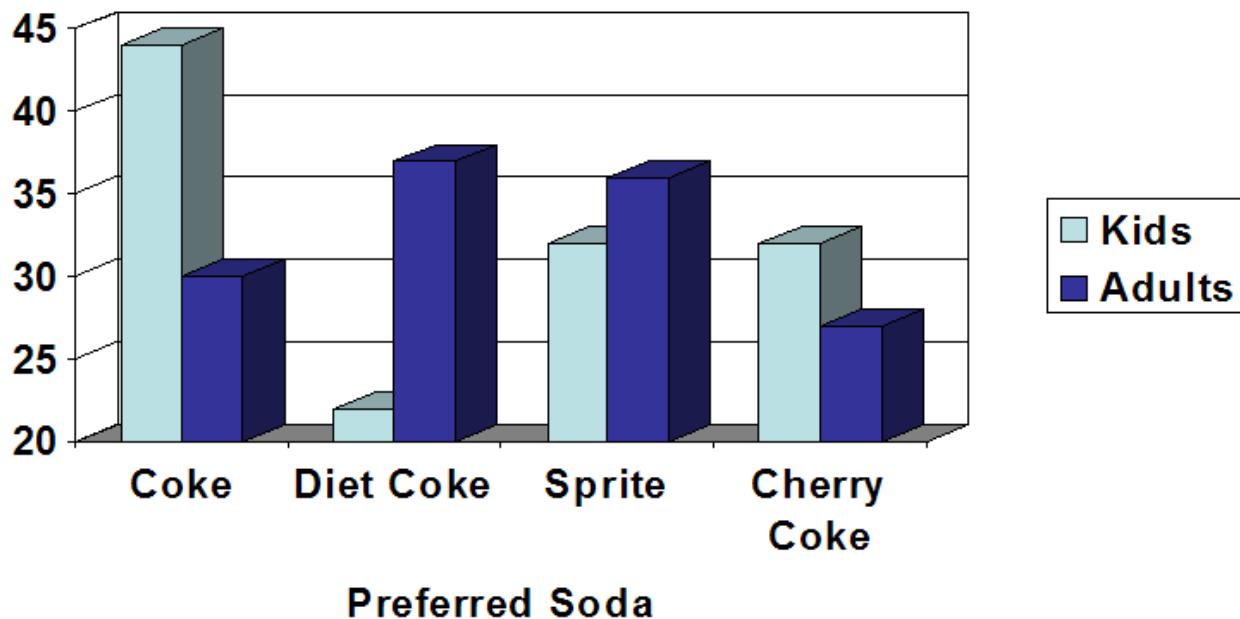
<b>Non-players</b>	<b>Beginners</b>	<b>Tournament players</b>
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

4. You have to decide between displaying your data with a histogram or with a stem and leaf display. What factor(s) would affect your choice?
5. In a box plot, what percent of the scores are between the lower and upper hinges?
6. A student has decided to display the results of his project on the number of hours people in various countries slept per night. He compared the sleeping patterns of people from the US, Brazil, France, Turkey, China, Egypt, Canada, Norway, and Spain. He was planning on using a line graph to display this data. Is a line graph appropriate? What might be a better choice for a graph?
7. For the data from the 1977 Stat. and Biom. 200 class for eye color, construct:
  - a. pie graph
  - b. horizontal bar graph
  - c. vertical bar graph
  - d. a frequency table with the relative frequency of each eye color

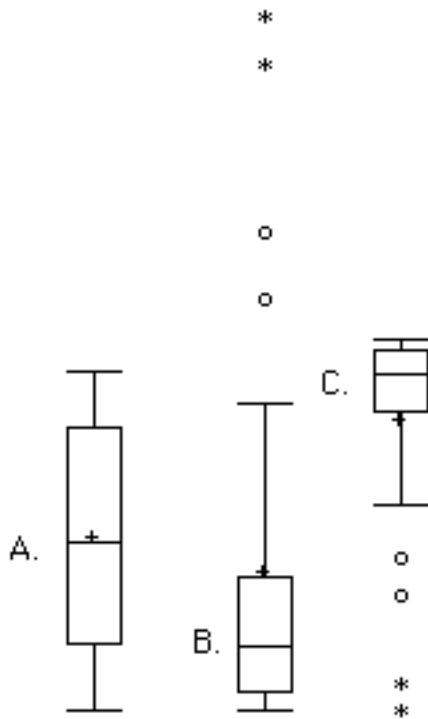
Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

(Question submitted by J. Warren, UNH)

8. A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improved.



9. Which of the box plots on the graph has a large positive skew? Which has a large negative skew?



### *Question from Case Studies*

#### Angry Moods (AM) case study

10. (AM) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the “Anger-Out” scores:
  - a. Create parallel box plots.
  - b. Create a back to back stem and leaf displays (You may have trouble finding a computer to do this so you may have to do it by hand. Use a fixed-width font such as Courier.)
11. (AM) Create parallel box plots for the Anger-In scores by sports participation.
12. (AM) Plot a histogram of the distribution of the Control-Out scores.
13. (AM) Create a bar graph comparing the mean Control-In score for the athletes and the non- athletes. What would be a better way to display this data?

14. (AM) Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger?

Flatulence (F) case study

15. (F) Plot a histogram of the variable “per day.”

16. (F) Create parallel box plots of “how long” as a function gender. Why is the 25th percentile not showing? What can you say about the results?

17. (F) Create a stem and leaf plot of the variable “how long.” What can you say about the shape of the distribution?

Physicians’ Reactions (PR) case study

18. (PR) Create box plots comparing the time expected to be spent with the average-weight and overweight patients.

19. (PR) Plot histograms of the time spent with the average-weight and overweight patients.

20. (PR) To which group does the patient with the highest expected time belong?

Smiles and Leniency (SL) case study

21. (SL) Create parallel box plots for the four conditions.

22. (SL) Create back to back stem and leaf displays for the false smile and neutral conditions. (It may be hard to find a computer program to do this for you, so be prepared to do it by hand).

ADHD Treatment (AT) case study

23. (AT) Create a line graph of the data. Do certain dosages appear to be more effective than others?

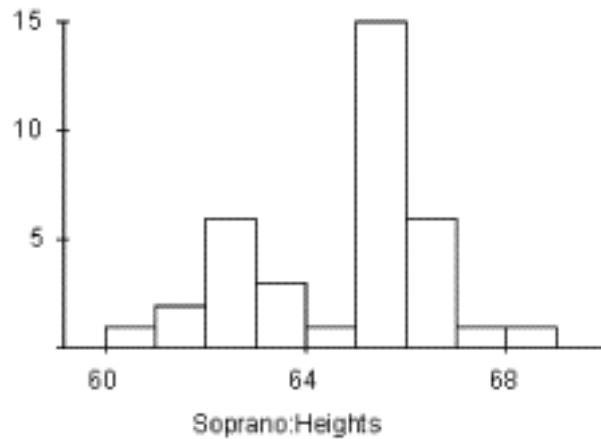
24. (AT) Create a stem and leaf plot of the number of correct responses of the participants after taking the placebo (d0 variable). What can you say about the shape of the distribution?
25. (AT) Create box plots for the four conditions. You may have to rearrange the data to get a computer program to create the box plots.

### SAT and College GPA (SG) case study

26. (SG) Create histograms and stem and leaf displays of both high-school grade point average and university grade point average. In what way(s) do the distributions differ?
27. The April 10th issue of the Journal of the American Medical Association reports a study on the effects of anti-depressants. The study involved 340 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (Pfizer's cousin of Lilly's Prozac) or placebo for an 8-week period. The following are the mean scores (approximately) for the three groups of subjects over the eight-week experiment. The first column is the baseline. Lower scores mean less depression. Create a graph to display these means.

<b>Placebo</b>	22.5	19.1	17.9	17.1	16.2	15.1	12.1	12.3
<b>Wort</b>	23.0	20.2	18.2	18.0	16.5	16.1	14.2	13.0
<b>Zoloft</b>	22.4	19.2	16.6	15.5	14.2	13.1	11.8	10.5

28. For the graph below, of heights of singers in a large chorus. What word starting with the letter "B" best describes the distribution?



29. Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired. (a) What is on the Y-axis? Explain. (b) What is on the X-axis? Explain. (c) What would be the probable shape of the salary distribution? Explain why.

# 3. Summarizing Distributions

## A. Central Tendency

1. What is Central Tendency
2. Measures of Central Tendency
3. Median and Mean
4. Additional Measures
5. Comparing measures

## B. Variability

1. Measures of Variability

## C. Shape

1. Effects of Transformations
2. Variance Sum Law I

## D. Exercises

Descriptive statistics often involves using a few numbers to summarize a distribution. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the numbers in the distribution vary from one another. The second section describes measures of variability. Distributions can differ in shape. Some distributions are symmetric whereas others have long tails in just one direction. The third section describes measures of the shape of distributions. The final two sections concern (1) how transformations affect measures summarizing distributions and (2) the variance sum law, an important relationship involving a measure of variability.

# What is Central Tendency?

by David M. Lane and Heidi Ziemer

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 2: Stem and Leaf Displays

## *Learning Objectives*

1. Identify situations in which knowing the center of a distribution would be valuable
2. Give three different ways the center of a distribution can be defined
3. Describe how the balance is different for symmetric distributions than it is for asymmetric distributions.

What is “central tendency,” and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is “3/5.” How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, “Whad'ja get?” and then ask the instructor, “How did the class do?” In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table 1. They are labeled “Dataset A,” “Dataset B,” and “Dataset C.” Which of

the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Table 1. Three possible datasets for the 5-point make-up quiz.

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the **center of the distribution**.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. Figure 1 shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

8	05
7	156
6	233
5	168
330	4 06
9420	3
622	2

Figure 1. Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

## Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

## Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure 2 shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and is placed at its

position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8.

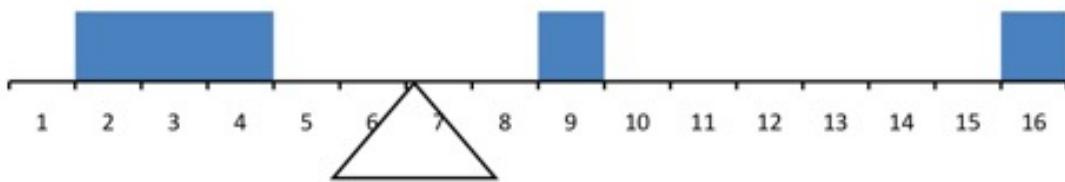


Figure 2. A balance scale.

For another example, consider the distribution shown in Figure 3. It is balanced by placing the fulcrum in the geometric middle.

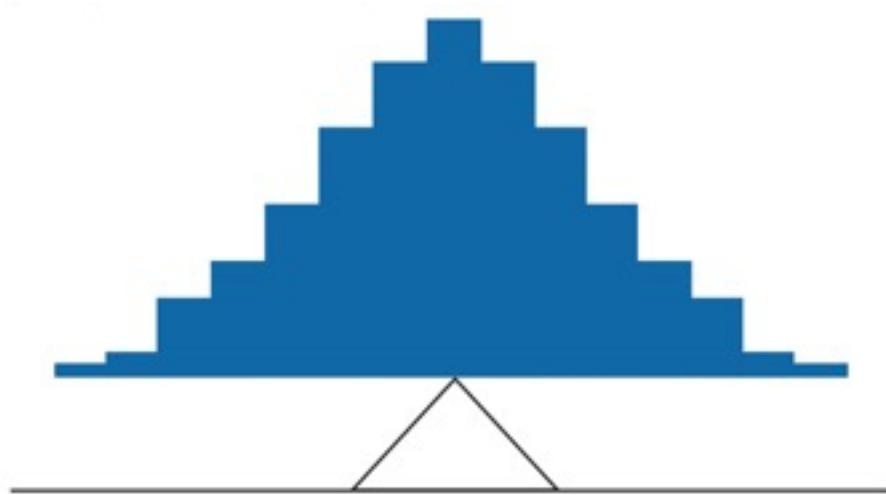


Figure 3. A distribution balanced on the tip of a triangle.

Figure 4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

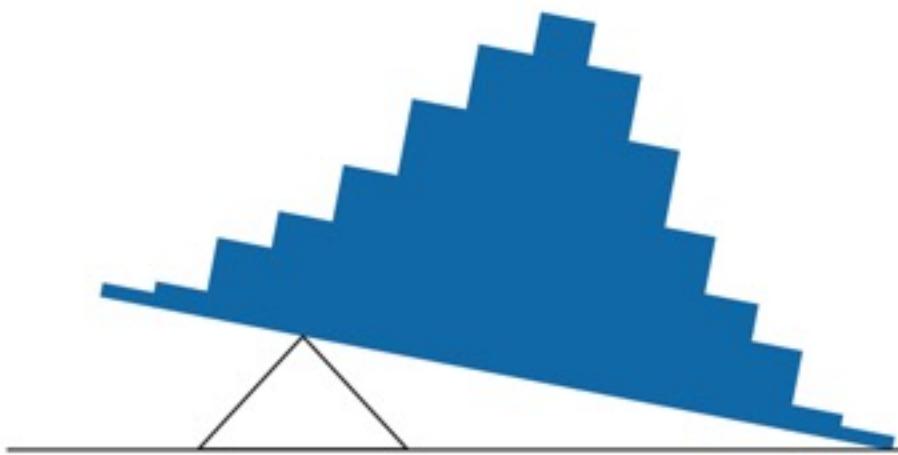


Figure 4. The distribution is not balanced.

Figure 5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3). Placing the fulcrum at the “half way” point would cause it to tip towards the left.

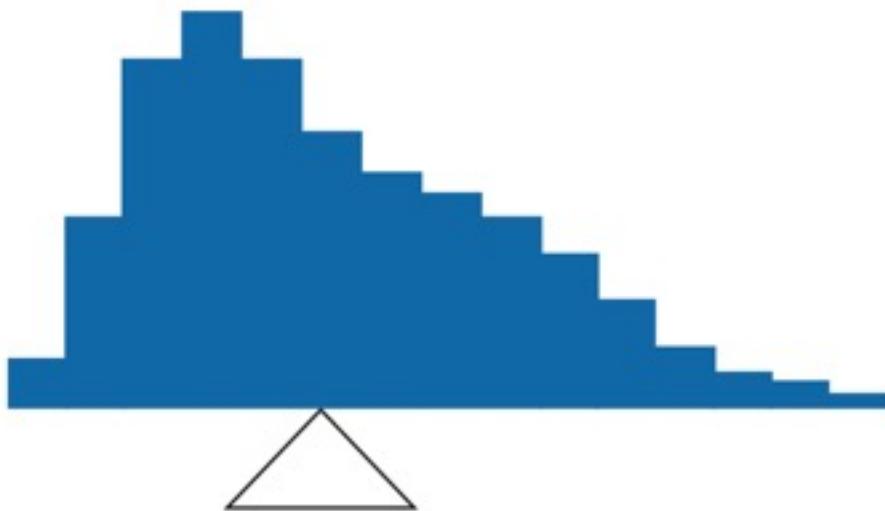


Figure 5. An asymmetric distribution balanced on the tip of a triangle.

The balance point defines one sense of a distribution's center.

### **Smallest Absolute Deviation**

Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers 2, 3, 4, 9, 16. Let's see how far the distribution is from 10

(picking a number arbitrarily). Table 2 shows the sum of the absolute deviations of these numbers from the number 10.

Table 2. An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
<b>Sum</b>	<b>28</b>

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations, we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals  $3 + 2 + 1 + 4 + 11 = 21$ . So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which the sum of absolute deviations is only 20. See if you can find it.

### **Smallest Squared Deviation**

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers 2, 3, 4, 9, 16. Table 3 shows the sum of the squared deviations of these numbers from the number 10.

Table 3. An example of the sum of squared deviations.

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
<b>Sum</b>	<b>186</b>

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186.

Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as  $9 + 4 + 1 + 16 + 121 = 151$ . So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum.

# Measures of Central Tendency

by David M. Lane

## *Prerequisites*

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Central Tendency

## *Learning Objectives*

1. Compute mean
2. Compute median
3. Compute mode

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

## **Arithmetic Mean**

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol “ $\mu$ ” is used for the mean of a population. The symbol “ $M$ ” is used for the mean of a sample.

The formula for  $\mu$  is shown below:

$$\mu = \frac{\sum X}{N}$$

where  $\sum X$  is the sum of all the numbers in the population and  $N$  is the number of numbers in the population.

The formula for  $M$  is essentially identical:

$$M = \frac{\sum X}{N}$$

where  $\Sigma X$  is the sum of all the numbers in the sample and  $N$  is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is  $20/5 = 4$  regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$$

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6
--

Although the arithmetic mean is not the only “mean” (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term “mean” is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

## Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15

scores above the 16th score. The median can also be thought of as the 50th percentile.

## Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is:

$$\frac{(4 + 7)}{2} = 5.5$$

When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

## Mode

The mode is the most frequently occurring value. For the data in Table 1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

# Median and Mean

by David M. Lane

## *Prerequisites*

- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency

## *Learning Objectives*

1. State when the mean and median are the same
2. State whether it is the mean or median that minimizes the mean absolute deviation
3. State whether it is the mean or median that minimizes the mean squared deviation
4. State whether it is the mean or median that is the balance point on a balance scale

In the section “What is central tendency,” we saw that the center of a distribution could be defined three ways: (1) the point on which a distribution would balance, (2) the value whose average absolute deviation from all the other values is minimized, and (3) the value whose squared difference from all the other values is minimized. The mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations.

Table 1 shows the absolute and squared deviations of the numbers 2, 3, 4, 9, and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Table 1. Absolute and squared deviations from the median of 4 and the mean of 6.8.

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
<b>Total</b>	<b>20</b>	<b>22.8</b>	<b>174</b>	<b>134.8</b>

Figure 1 shows that the distribution balances at the mean of 6.8 and not at the median of 4.0. The relative advantages and disadvantages of the mean and median are discussed in the section “Comparing Measures” later in this chapter.

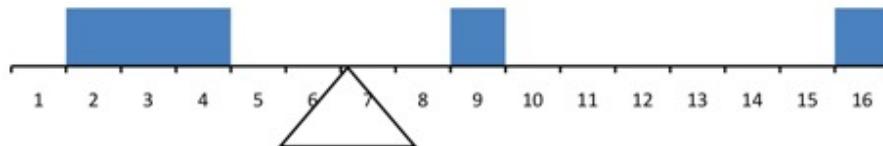


Figure 1. The distribution balances at the mean of 6.8 and not at the median of 4.0.

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: 1, 3, 4, 5, 6, 7, 9. The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

# Additional Measures of Central Tendency

by David M. Lane

## *Prerequisites*

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

## *Learning Objectives*

1. Compute the trimean
2. Compute the geometric mean directly
3. Compute the geometric mean using logs
4. Use the geometric to compute annual portfolio returns
5. Compute a trimmed mean

Although the mean, median, and mode are by far the most commonly used measures of central tendency, they are by no means the only measures. This section defines three additional measures of central tendency: the trimean, the geometric mean, and the trimmed mean. These measures will be discussed again in the section “Comparing Measures of Central Tendency.”

## **Trimean**

The trimean is a weighted average of the 25th percentile, the 50th percentile, and the 75th percentile. Letting P25 be the 25th percentile, P50 be the 50th and P75 be the 75th percentile, the formula for the trimean is:

$$\text{Trimean} = \frac{(P25 + 2P50 + P75)}{4}$$

Consider the data in Table 2. The 25th percentile is 15, the 50th is 20 and the 75th percentile is 23.

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6
---

Table 2. Percentiles.

Percentile	Value
25	15
50	20
75	23

The trimean is therefore :

$$\frac{(15 + 2 \times 20 + 23)}{4} = \frac{78}{4} = 19.5$$

### Geometric Mean

The geometric mean is computed by multiplying all the numbers together and then taking the nth root of the product. For example, for the numbers 1, 10, and 100, the product of all the numbers is:  $1 \times 10 \times 100 = 1,000$ . Since there are three numbers, we take the cubed root of the product (1,000) which is equal to 10. The formula for the geometric mean is therefore

$$\left( \prod X \right)^{\frac{1}{N}}$$

where the symbol  $\prod$  means to multiply. Therefore, the equation says to multiply all the values of  $X$  and then raise the result to the  $1/N$ th power. Raising a value to the  $1/N$ th power is, of course, the same as taking the  $N$ th root of the value. In this case,  $1000^{1/3}$  is the cube root of 1,000.

The geometric mean has a close relationship with logarithms. Table 3 shows the logs (base 10) of these three numbers. The arithmetic mean of the three logs is 1. The anti-log of this arithmetic mean of 1 is the geometric mean. The anti-log of 1 is  $10^1 = 10$ . Note that the geometric mean only makes sense if all the numbers are positive.

Table 3. Logarithms.

X	Log10(X)
1	0
10	1
100	2

The geometric mean is an appropriate measure to use for averaging rates. For example, consider a stock portfolio that began with a value of \$1,000 and had annual returns of 13%, 22%, 12%, -5%, and -13%. Table 4 shows the value after each of the five years.

Table 4. Portfolio Returns

Year	Return	Value
1	13%	1,130
2	22%	1,379
3	12%	1,544
4	-5%	1,467
5	-13%	1,276

The question is how to compute average annual rate of return. The answer is to compute the geometric mean of the returns. Instead of using the percents, each return is represented as a multiplier indicating how much higher the value is after the year. This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss. The multipliers for this example are 1.13, 1.22, 1.12, 0.95, and 0.87. The geometric mean of these multipliers is 1.05. Therefore, the average annual rate of return is 5%. Table 5 shows how a portfolio gaining 5% a year would end up with the same value (\$1,276) as shown in Table 4.

Table 5. Portfolio Returns

Year	Return	Value

1	5%	1,050
2	5%	1,103
3	5%	1,158
4	5%	1,216
5	5%	1,276

### Trimmed Mean

To compute a *trimmed mean*, you remove some of the higher and lower scores and compute the mean of the remaining scores. A mean trimmed 10% is a mean computed with 10% of the scores trimmed off: 5% from the bottom and 5% from the top. A mean trimmed 50% is computed by trimming the upper 25% of the scores and the lower 25% of the scores and computing the mean of the remaining scores. The trimmed mean is similar to the median which, in essence, trims the upper 49+% and the lower 49+% of the scores. Therefore the trimmed mean is a hybrid of the mean and the median. To compute the mean trimmed 20% for the touchdown pass data shown in Table 1, you remove the lower 10% of the scores (6, 9, and 12) as well as the upper 10% of the scores (33, 33, and 37) and compute the mean of the remaining 25 scores. This mean is 20.16.

# Comparing Measures of Central Tendency

by David M. Lane

## *Prerequisites*

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

## *Learning Objectives*

1. Understand how the difference between the mean and median is affected by skew
2. State how the measures differ in symmetric distributions
3. State which measure(s) should be used to describe the center of a skewed distribution

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean, median, trimean, and trimmed mean are equal, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 1 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

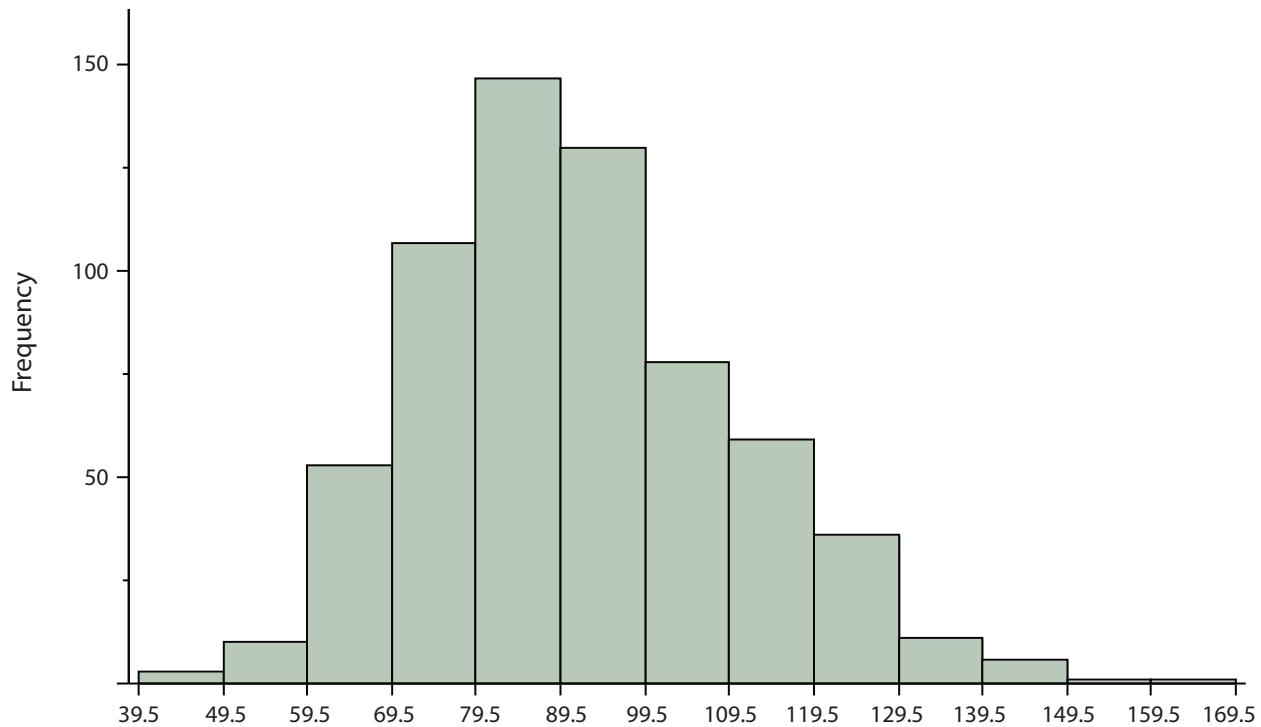


Figure 1. A distribution with a positive skew.

Measures of central tendency are shown in Table 1. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90. Typically the trimean and trimmed mean will fall between the median and the mean, although in this case, the trimmed mean is slightly lower than the median. The geometric mean is lower than all measures except the mode.

Table 1. Measures of central tendency for the test scores.

Measure	Value
Mode	84.00
Median	90.00
Geometric Mean	89.70
Trimean	90.25
Mean trimmed 50%	89.81
Mean	91.58

The distribution of baseball salaries (in 1994) shown in Figure 2 has a much more pronounced skew than the distribution in Figure 1.

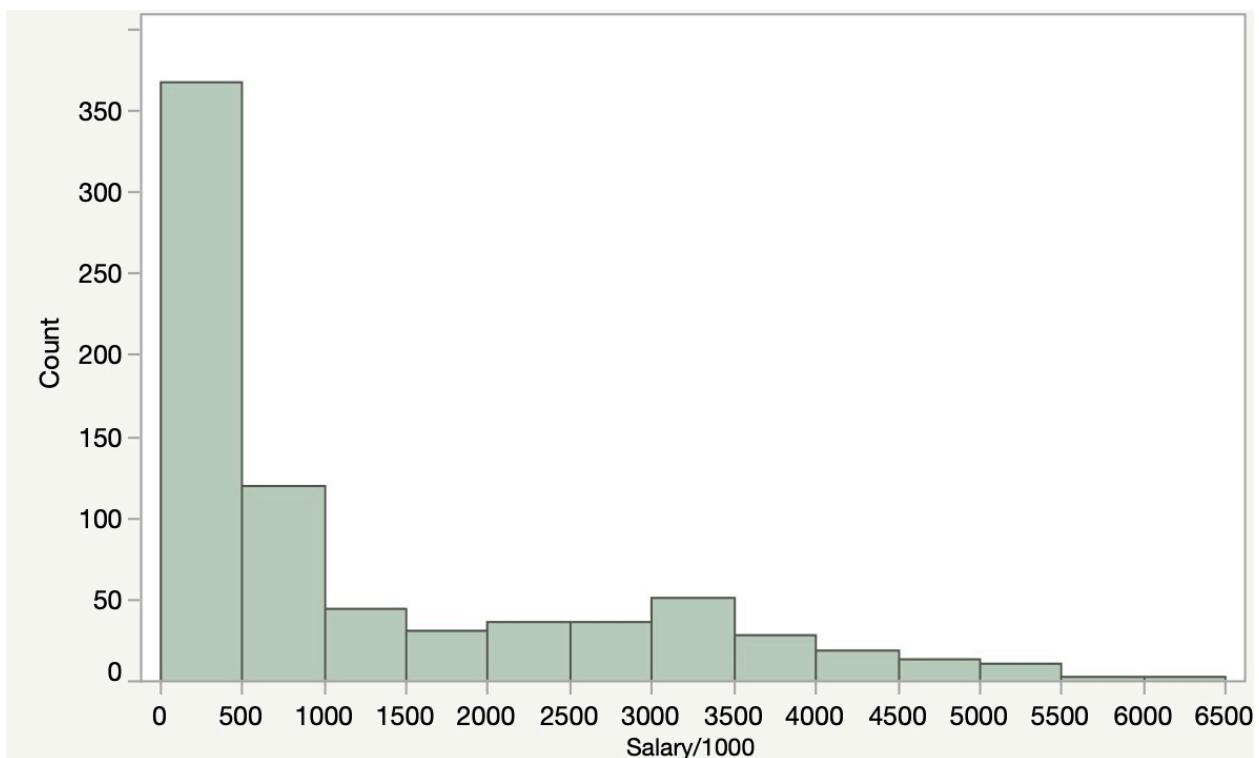


Figure 2. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars: 250 equals 250,000).

Table 2 shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central

tendency is sufficient for data such as these. If you were asked the very general question: “So, what do baseball players make?” and answered with the mean of \$1,183,000, you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of \$250,000 or the median of \$500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean, median, and either the trimean or the mean trimmed 50%. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table 2. Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	Value
Mode	250
Median	500
Geometric Mean	555
Trimean	792
Mean trimmed 50%	619
Mean	1,183

# Measures of Variability

by David M. Lane

## *Prerequisites*

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency

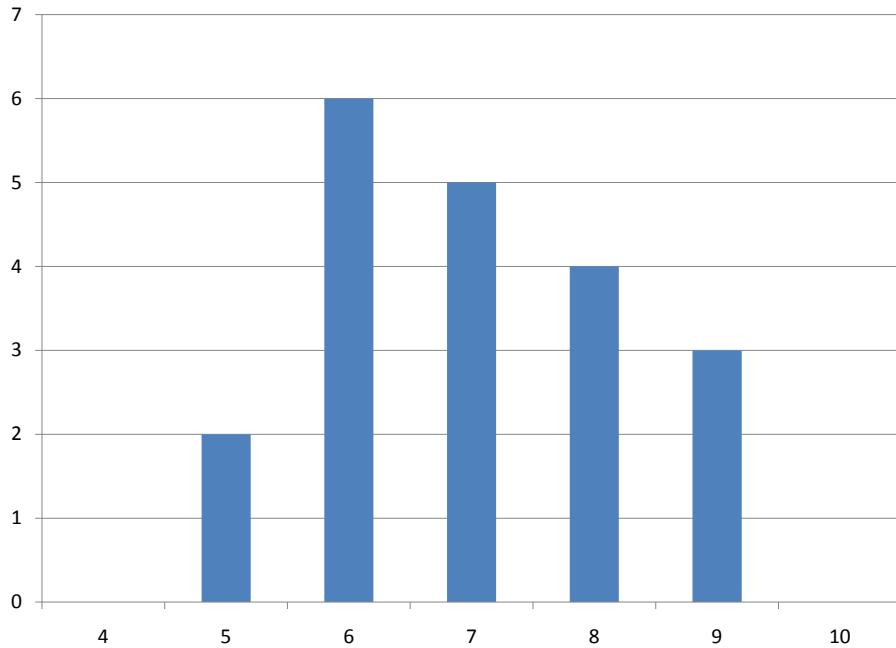
## *Learning Objectives*

1. Determine the relative variability of two distributions
2. Compute the range
3. Compute the inter-quartile range
4. Compute the variance in the population
5. Estimate the variance from a sample
6. Compute the standard deviation from the variance

## **What is Variability?**

Variability refers to how “spread out” a group of scores is. To see what we mean by spread out, consider graphs in Figure 1. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

## Quiz 1



## Quiz 2

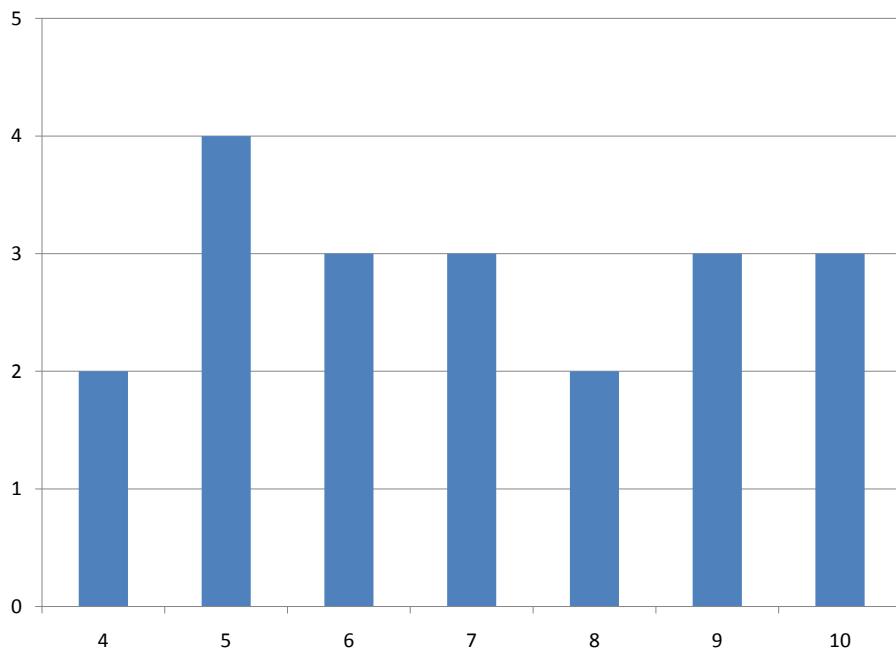


Figure 1. Bar charts of two quizzes.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will

discuss measures of the variability of a distribution. There are four frequently used measures of variability: range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

## Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so  $10 - 2 = 8$ . The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so  $99 - 23 = 76$ ; the range is 76. Now consider the two quizzes shown in Figure 1. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

## Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

$$\text{IQR} = \text{75th percentile} - \text{25th percentile}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

A related measure of variability is called the semi-interquartile range. The semi-interquartile range is defined simply as the interquartile range divided by 2. If a distribution is symmetric, the median plus or minus the semi-interquartile range contains half the scores in the distribution.

## **Variance**

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 1. The mean score is 7.0. Therefore, the column “Deviation from Mean” contains the score minus 7. The column “Squared Deviation” is simply the previous column squared.

Table 1. Calculation of Variance for Quiz 1 scores.

<b>Scores</b>	<b>Deviation from Mean</b>	<b>Squared Deviation</b>
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
Means		
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that its variance is 6.7. The formula for the variance is:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

where  $\sigma^2$  is the variance,  $\mu$  is the mean, and  $N$  is the number of numbers. For Quiz 1,  $\mu = 7$  and  $N = 20$ .

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X - M)^2}{N - 1}$$

where  $s^2$  is the estimate of the variance and  $M$  is the sample mean. Note that  $M$  is the mean of a sample taken from a population with a mean of  $\mu$ . Since, in practice, the variance is usually computed in a sample, this formula is most often used.

Let's take a concrete example. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute  $s^2$  as follows:

$$M = \frac{1 + 2 + 4 + 5}{4} = \frac{12}{4} = 3$$

$$s^2 = \frac{(1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2}{4 - 1} = \frac{4 + 1 + 1 + 4}{3} = \frac{10}{3} = 3.333$$

There are alternate formulas that can be easier to use if you are doing your calculations with a hand calculator:

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

and

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

For this example,

$$(\sum X)^2 = \frac{(1 + 2 + 4 + 5)^2}{4} = \frac{144}{4} = 36$$

$$\sigma^2 = \frac{(46 - 36)}{4} = 2.5$$

$$s^2 = \frac{(46 - 36)}{3} = 3.333$$

as with the other formula.

## Standard Deviation

The standard deviation is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal (see Chapter 7) because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between  $50 - 10 = 40$  and  $50 + 10 = 60$ . Similarly, about 95% of the distribution would be between  $50 - 2 \times 10 = 30$  and  $50 + 2 \times 10 = 70$ . The symbol for the population standard deviation is  $\sigma$ ; the symbol for an estimate computed in a sample is  $s$ .

Figure 2 shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 45 and 55; for the blue distribution, 68% is between 50 and 70.

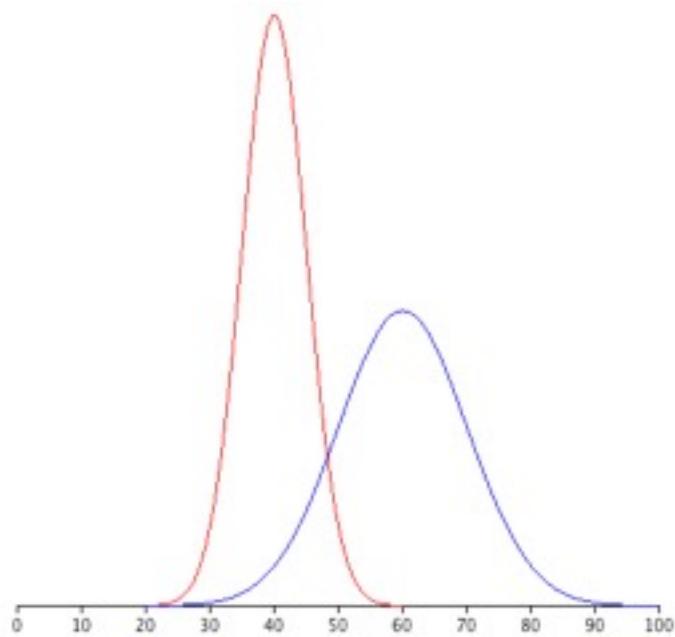


Figure 2. Normal distributions with standard deviations of 5 and 10.

# Shapes of Distributions

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability

## *Learning Objectives*

1. Compute skew using two different formulas
2. Compute kurtosis

We saw in the section on distributions in Chapter 1 that shapes of distributions can differ in skew and/or kurtosis. This section presents numerical indexes of these two measures of shape.

## **Skew**

Figure 1 shows a distribution with a very large positive skew. Recall that distributions with positive skew have tails that extend to the right.

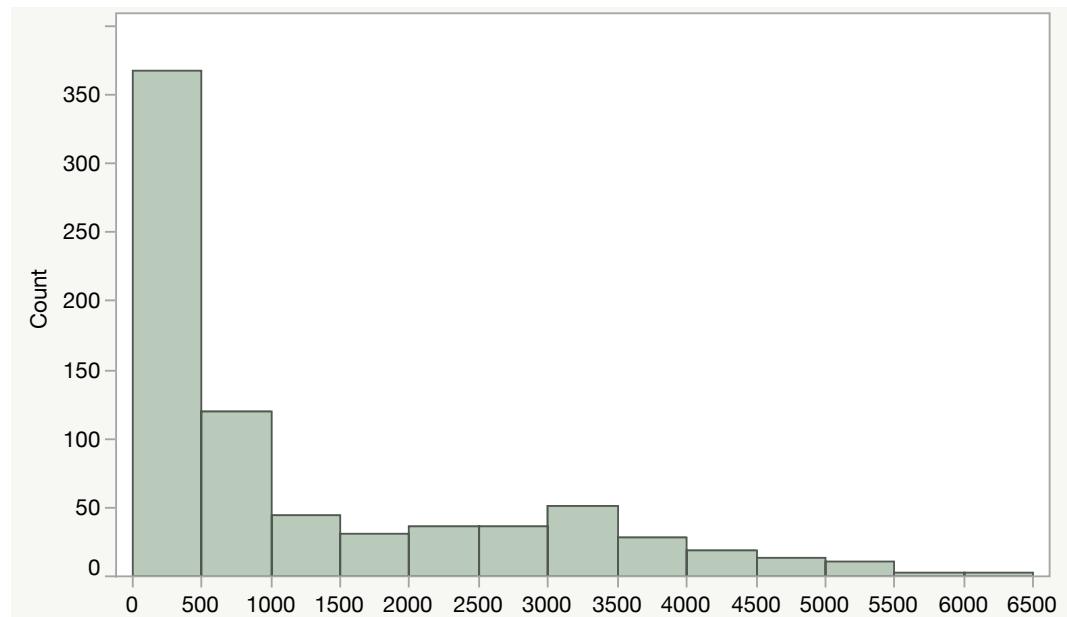


Figure 1. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Distributions with positive skew normally have larger means than medians. The mean and median of the baseball salaries shown in Figure 1 are \$1,183,417 and \$500,000 respectively. Thus, for this highly-skewed distribution, the mean is more than twice as high as the median. The relationship between skew and the relative size of the mean and median lead the statistician Pearson to propose the following simple and convenient numerical index of skew:

$$\frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The standard deviation of the baseball salaries is 1,390,922. Therefore, Pearson's measure of skew for this distribution is  $3(1,183,417 - 500,000)/1,390,922 = 1.47$ .

Just as there are several measures of central tendency, there is more than one measure of skew. Although Pearson's measure is a good one, the following measure is more commonly used. It is sometimes referred to as the third moment about the mean.

$$\sum \frac{(X - \mu)^3}{\sigma^3}$$

## Kurtosis

The following measure of kurtosis is similar to the definition of skew. The value “3” is subtracted to define “no kurtosis” as the kurtosis of a normal distribution. Otherwise, a normal distribution would have a kurtosis of 3.

$$\sum \frac{(X - \mu)^4}{\sigma^4} - 3$$

# Effects of Linear Transformations

by David M. Lane

## *Prerequisites*

- Chapter 1: Linear Transformations

## *Learning Objectives*

1. Define a linear transformation
2. Compute the mean of a transformed variable
3. Compute the variance of a transformed variable

This section covers the effects of linear transformations on measures of central tendency and variability. Let's start with an example we saw before in the section that defined linear transformation: temperatures of cities. Table 1 shows the temperatures of 5 cities.

Table 1. Temperatures in 5 cities on 11/16/2002.

City	Degrees Fahrenheit	Degrees Centigrade
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11
Mean	54.000	12.220
Median	54.000	12.220
Variance	330	101.852
SD	18.166	10.092

Recall that to transform the degrees Fahrenheit to degrees Centigrade, we use the formula

$$C = 0.55556F - 17.7778$$

which means we multiply each temperature Fahrenheit by 0.556 and then subtract 17.7778. As you might have expected, you multiply the mean temperature in Fahrenheit by 0.556 and then subtract 17.778 to get the mean in Centigrade. That is,  $(0.556)(54) - 17.7778 = 12.22$ . The same is true for the median. Note that this

relationship holds even if the mean and median are not identical as they are in Table 1.

The formula for the standard deviation is just as simple: the standard deviation in degrees Centigrade is equal to the standard deviation in degrees Fahrenheit times 0.556. Since the variance is the standard deviation squared, the variance in degrees Centigrade is equal to  $0.556^2$  times the variance in degrees Fahrenheit.

To sum up, if a variable  $X$  has a mean of  $\mu$ , a standard deviation of  $\sigma$ , and a variance of  $\sigma^2$ , then a new variable  $Y$  created using the linear transformation

$$Y = bX + A$$

will have a mean of  $b\mu+A$ , a standard deviation of  $b\sigma$ , and a variance of  $b^2\sigma^2$ .

It should be noted that the term “linear transformation” is defined differently in the field of linear algebra. For details, follow [this link](#).

# Variance Sum Law I

by David M. Lane

## *Prerequisites*

- Chapter 3: Variance

## *Learning Objectives*

1. Compute the variance of the sum of two uncorrelated variables
2. Compute the variance of the difference between two uncorrelated variables

As you will see in later sections, there are many occasions in which it is important to know the variance of the sum of two variables. Consider the following situation: (a) you have two populations, (b) you sample one number from each population, and (c) you add the two numbers together. The question is, “What is the variance of this sum?” For example, suppose the two populations are the populations of 8-year old males and 8-year-old females in Houston, Texas, and that the variable of interest is memory span. You repeat the following steps thousands of times: (1) sample one male and one female, (2) measure the memory span of each, and (3) sum the two memory spans. After you have done this thousands of times, you compute the variance of the sum. It turns out that the variance of this sum can be computed according to the following formula:

$$\sigma_{sum}^2 = \sigma_M^2 + \sigma_F^2$$

where the first term is the variance of the sum, the second term is the variance of the males and the third term is the variance of the females. Therefore, if the variances on the memory span test for the males and females respectively were 0.9 and 0.8, respectively, then the variance of the sum would be 1.7.

The formula for the variance of the difference between the two variables (memory span in this example) is shown below. Notice that the expression for the difference is the same as the formula for the sum.

$$\sigma_{difference}^2 = \sigma_M^2 + \sigma_F^2$$

More generally, the variance sum law can be written as follows:

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$$

which is read: “The variance of X plus or minus Y is equal to the variance of X plus the variance of Y.”

These formulas for the sum and difference of variables given above only apply when the variables are independent.

In this example, we have thousands of randomly-paired scores. Since the scores are paired randomly, there is no relationship between the memory span of one member of the pair and the memory span of the other. Therefore the two scores are independent. Contrast this situation with one in which thousands of people are sampled and two measures (such as verbal and quantitative SAT) are taken from each. In this case, there would be a relationship between the two variables since higher scores on the verbal SAT are associated with higher scores on the quantitative SAT (although there are many examples of people who score high on one test and low on the other). Thus the two variables are not independent and the variance of the total SAT score would not be the sum of the variances of the verbal SAT and the quantitative SAT. The general form of the variance sum law is presented in a section in the chapter on correlation.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 3: Median and Mean

The playbill for the Alley Theatre in Houston wants to appeal to advertisers. They reported the mean household income and the median age of theatergoers.

## **What do you think?**

What might have guided their choice of the mean or median?

It is likely that they wanted to emphasize that theatergoers had high income but de-emphasize how old they are. The distributions of income and age of theatergoers probably have positive skew. Therefore the mean is probably higher than the median, which results in higher income and lower age than if the median household income and mean age had been presented.

## Exercises

### Prerequisites

- All material presented in the Summarizing Distributions chapter

1. Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?
2. Repeat Problem 1 only this time make the dataset have a negative skew.
3. Make up three data sets with 5 numbers each that have:
  - (a) the same mean but different standard deviations.
  - (b) the same mean but different medians.
  - (c) the same median but different means.
4. Find the mean and median for the following three variables:

A	B	C
8	4	6
5	4	2
7	6	3
1	3	4
3	4	1

5. A sample of 30 distance scores measured in yards has a mean of 10, a variance of 9, and a standard deviation of 3 (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation? (b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?
6. You recorded the time in seconds it took for 8 participants to solve a puzzle. These times appear below. However, when the data was entered into the statistical program, the score that was supposed to be 22.1 was entered as 21.2.

You had calculated the following measures of central tendency: the mean, the median, and the mean trimmed 25%. Which of these measures of central tendency will change when you correct the recording error?

Time (seconds)
15.2
18.8
19.3
19.7
20.2
21.8
22.1
29.4

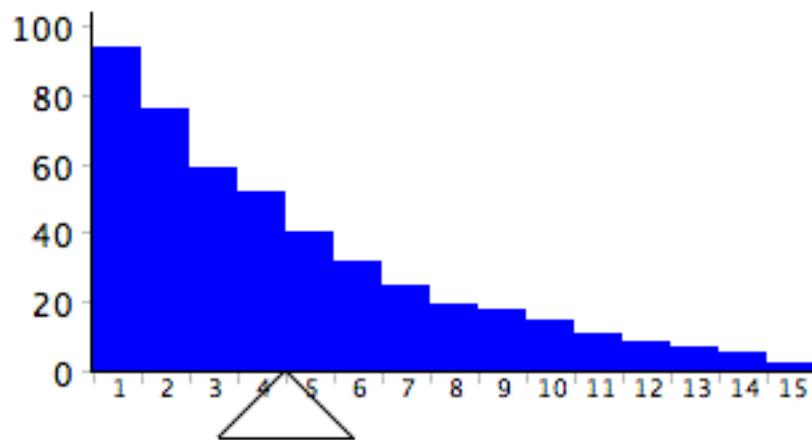
7. For the test scores in question #6, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?
8. You know the minimum, the maximum, and the 25th, 50th, and 75th percentiles of a distribution. Which of the following measures of central tendency or variability can you determine?  
mean, median, mode, trimean, geometric mean, range, interquartile range, variance, standard deviation
9. For the numbers 1, 3, 4, 6, and 12:  
Find the value (v) for which  $\Sigma(X-v)^2$  is minimized.  
Find the value (v) for which  $\Sigma|x-v|$  is minimized.
10. Your younger brother comes home one day after taking a science test. He says that someone at school told him that “60% of the students in the class scored above the median test grade.” What is wrong with this statement? What if he had said “60% of the students scored below the mean?”
11. An experiment compared the ability of three groups of participants to remember briefly- presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess

positions. Compare the performance of each group. Consider spread as well as central tendency.

<b>Non-players</b>	<b>Beginners</b>	<b>Tournament players</b>
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

12. True/False: A bimodal distribution has two modes and two medians.
13. True/False: The best way to describe a skewed distribution is to report the mean.
14. True/False: When plotted on the same graph, a distribution with a mean of 50 and a standard deviation of 10 will look more spread out than will a distribution with a mean of 60 and a standard deviation of 5.
15. Compare the mean, median, trimean in terms of their sensitivity to extreme scores.
16. If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?
17. A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?
18. Which measure of central tendency is most often used for returns on investment?

19. The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?



### *Questions from Case Studies*

#### Angry Moods (AM) case study

20. (AM) Does Anger-Out have a positive skew, a negative skew, or no skew?

21. (AM) What is the range of the Anger-In scores? What is the interquartile range?

22. (AM) What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes?

23. (AM) What is the variance of the Control-In scores for the athletes? What is the variance of the Control-In scores for the non-athletes?

#### Flatulence (F) case study

24. (F) Based on a histogram of the variable “perday”, do you think the mean or median of this variable is larger? Calculate the mean and median to see if you are right.

#### Stroop (S) case study

25.(S) Compute the mean for “words”.

26. (S#2) Compute the mean and standard deviation for “colors”.

Physicians’ Reactions (PR) case study

27.(PR) What is the mean expected time spent for the average-weight patients?

What is the mean expected time spent for the overweight patients?

28.(PR) What is the difference in means between the groups? By approximately how many standard deviations do the means differ?

Smiles and Leniency (SL) case study

29.(SL) Find the mean, median, standard deviation, and interquartile range for the leniency scores of each of the four groups.

ADHD Treatment (AT) case study

30.(AT) What is the mean number of correct responses of the participants after taking the placebo (0 mg/kg)?

31.(AT) What are the standard deviation and the interquartile range of the d0 condition?

# 4. Describing Bivariate Data

- A. Introduction to Bivariate Data
- B. Values of the Pearson Correlation
- C. Properties of Pearson's  $r$
- D. Computing Pearson's  $r$
- E. Variance Sum Law II
- F. Exercises

A dataset with two variables contains what is called bivariate data. This chapter discusses ways to describe the relationship between two variables. For example, you may wish to describe the relationship between the heights and weights of people to determine the extent to which taller people weigh more.

The introductory section gives more examples of bivariate relationships and presents the most common way of portraying these relationships graphically. The next five sections discuss Pearson's correlation, the most common index of the relationship between two variables. The final section, “Variance Sum Law II,” makes use of Pearson's correlation to generalize this law to bivariate data.

# Introduction to Bivariate Data

by Rudy Guerra and David M. Lane

## *Prerequisites*

- Chapter 1: Variables
- Chapter 1: Distributions
- Chapter 2: Histograms
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability
- Chapter 3: Shapes of Distributions

## *Learning Objectives*

1. Define “bivariate data”
2. Define “scatter plot”
3. Distinguish between a linear and a nonlinear relationship
4. Identify positive and negative associations from a scatter plot

Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two quantitative variables for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. Let's begin by asking if people tend to marry other people of about the same age. Our experience tells us “yes,” but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table 1 below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big

surprise, but at least the data bear out our experiences, which is not always the case.

Table 1. Sample of spousal ages of 10 White American Couples.

<b>Husband</b>	36	72	37	36	51	50	47	50	37	41
<b>Wife</b>	35	67	33	35	50	46	47	42	36	41

The pairs of ages in Table 1 are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a histogram (see Figure 1) and by a mean and standard deviation (See Table 2).

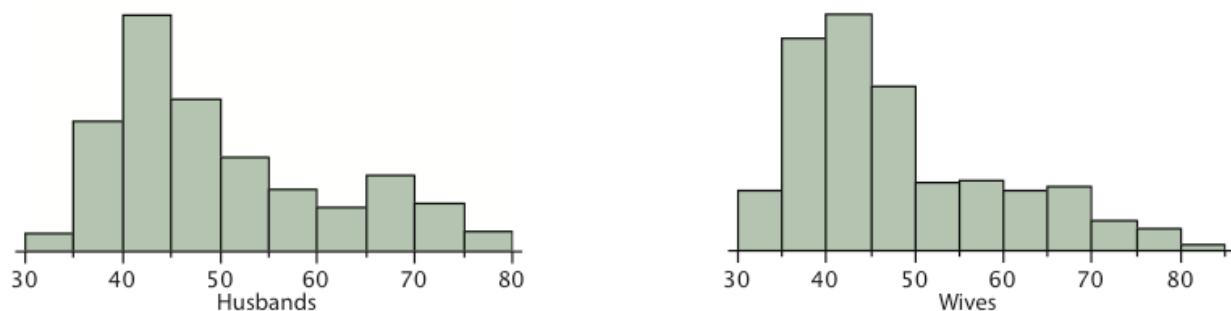


Figure 1. Histograms of spousal ages.

Table 2. Means and standard deviations of spousal ages.

	Mean	Standard Deviation
Husbands	49	11
Wives	47	11

Each distribution is fairly skewed with a long right tail. From Table 1 we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples has younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives

of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

We can learn much more by displaying the bivariate data in a graphical form that maintains the pairing. Figure 2 shows a scatter plot of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

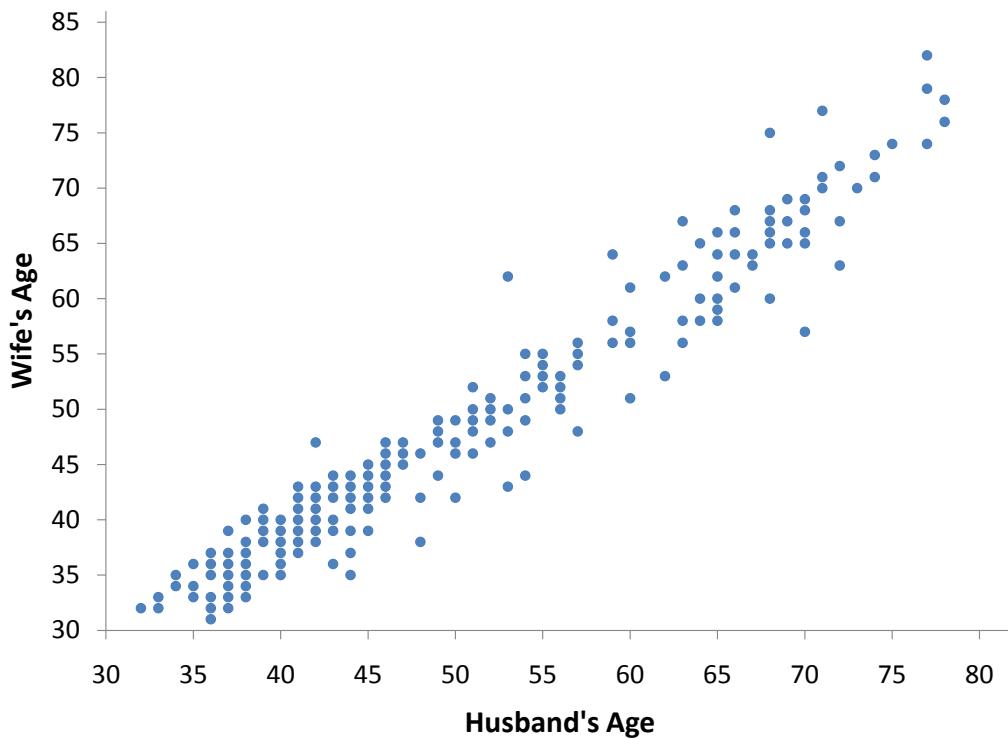


Figure 2. Scatter plot showing wife's age as a function of husband's age.

There are two important characteristics of the data revealed by Figure 2. First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (Y) increases with the second variable (X), we say that X and Y have a positive association. Conversely, when Y decreases as X increases, we say that they have a negative association.

Second, the points cluster along a straight line. When this occurs, the relationship is called a linear relationship.

Figure 3 shows a scatter plot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a

positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.

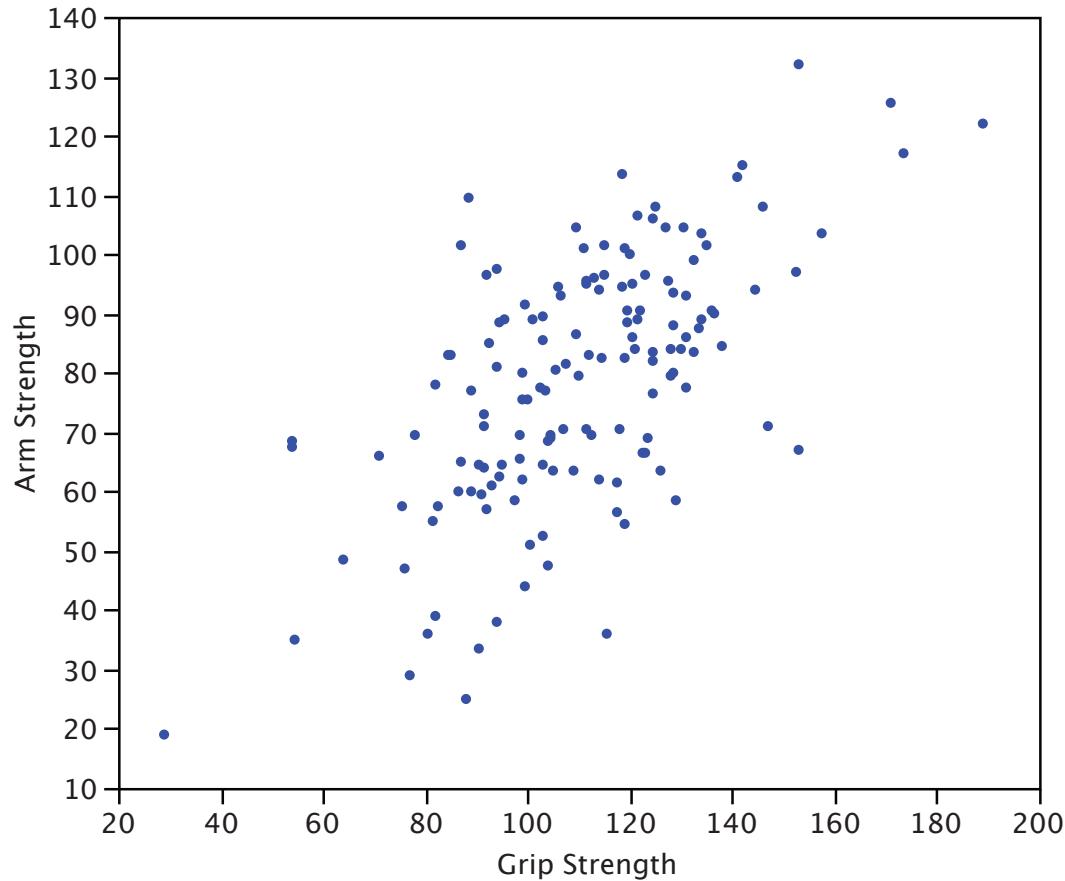


Figure 3. Scatter plot of Grip Strength and Arm Strength.

Not all scatter plots show linear relationships. Figure 4 shows the results of an experiment conducted by Galileo on projectile motion. In the experiment, Galileo rolled balls down an incline and measured how far they traveled as a function of the release height. It is clear from Figure 4 that the relationship between “Release Height” and “Distance Traveled” is not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola.

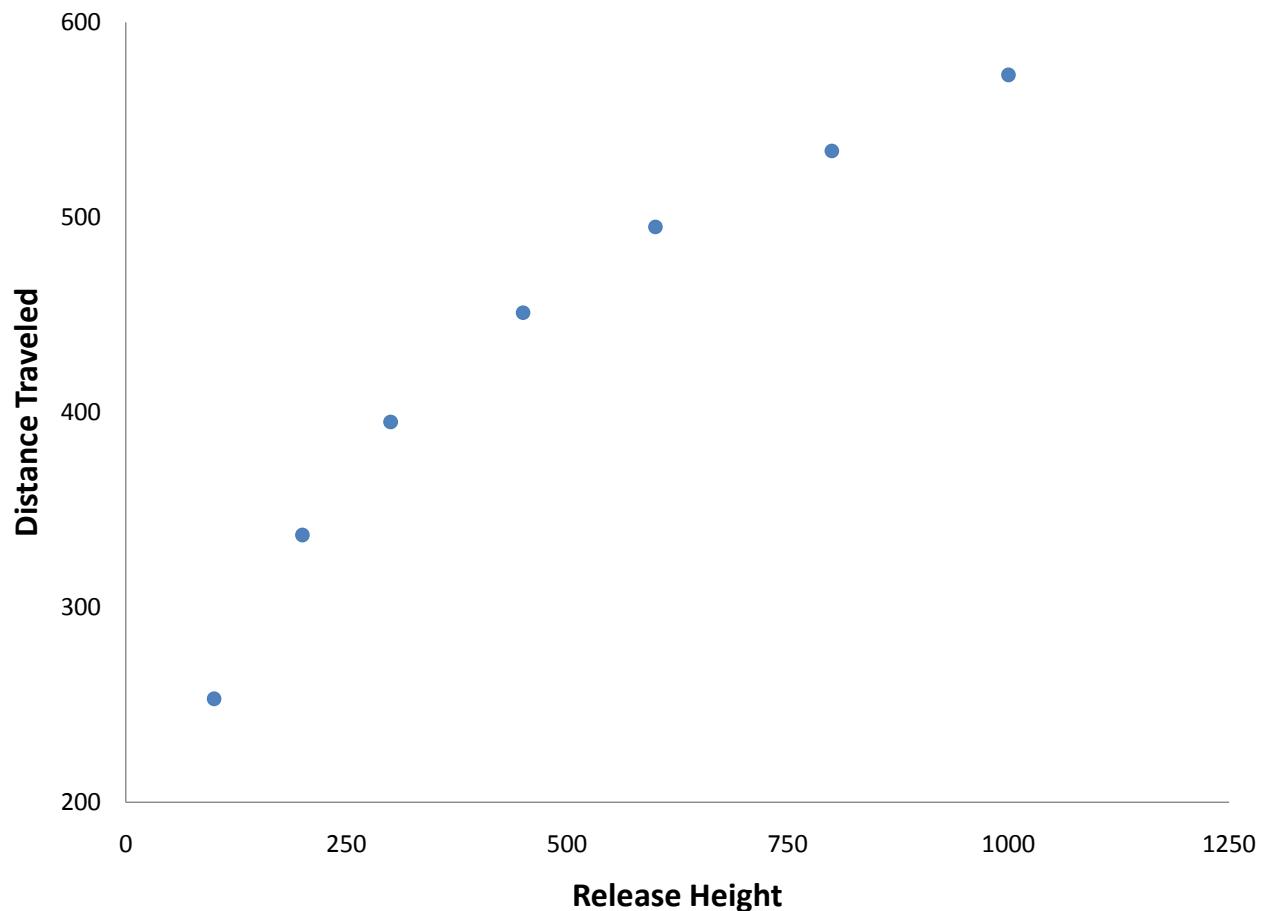


Figure 4. Galileo's data showing a non-linear relationship.

Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line. A statistical measure of the strength of the relationship between two quantitative variables that takes these factors into account is the subject of the next section.

# Values of the Pearson Correlation

by David M. Lane

## *Prerequisites*

- Chapter 4: Introduction to Bivariate Data

## *Learning Objectives*

1. Describe what Pearson's correlation measures
2. Give the symbols for Pearson's correlation in the sample and in the population
3. State the possible range for Pearson's correlation
4. Identify a perfect linear relationship

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is “ $\rho$ ” when it is measured in the population and “ $r$ ” when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use  $r$  to represent Pearson's correlation unless otherwise noted.

Pearson's  $r$  can range from -1 to 1. An  $r$  of -1 indicates a perfect negative linear relationship between variables, an  $r$  of 0 indicates no linear relationship between variables, and an  $r$  of 1 indicates a perfect positive linear relationship between variables. Figure 1 shows a scatter plot for which  $r = 1$ .

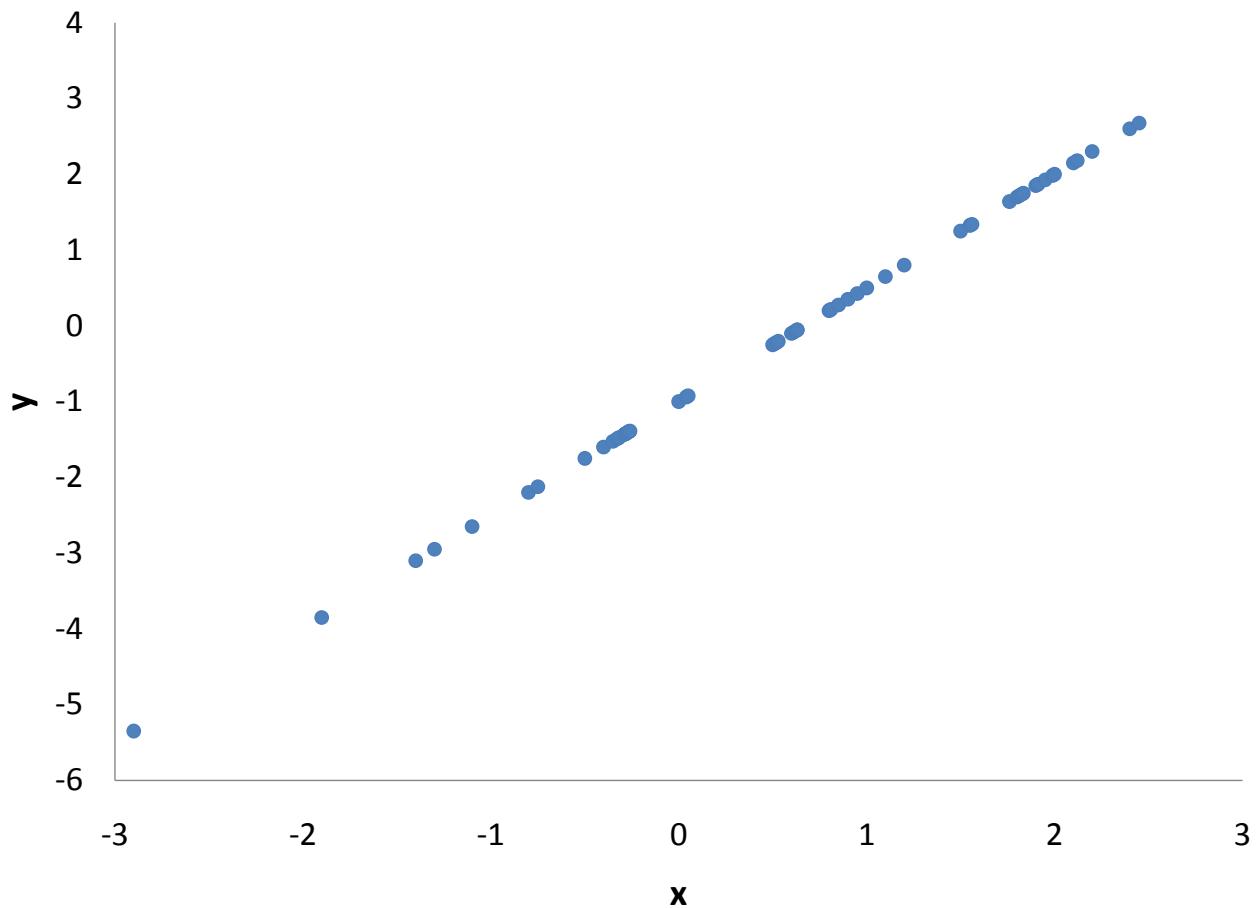


Figure 1. A perfect linear relationship,  $r = 1$ .

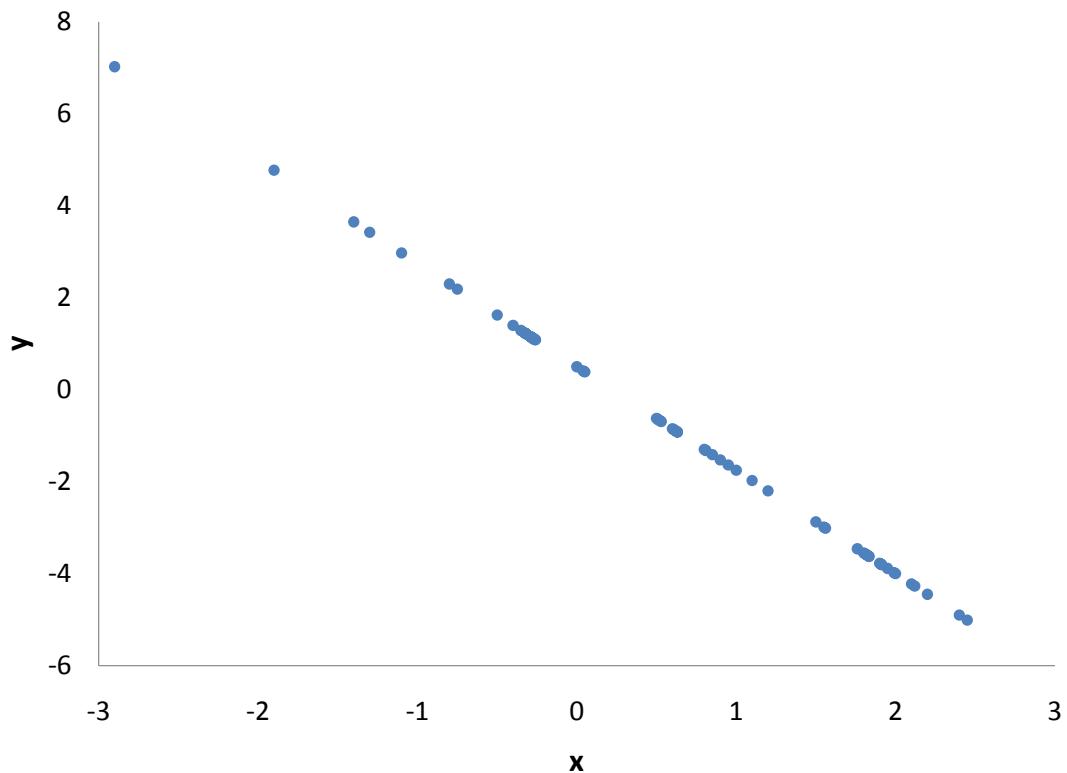


Figure 2. A perfect negative linear relationship,  $r = -1$ .

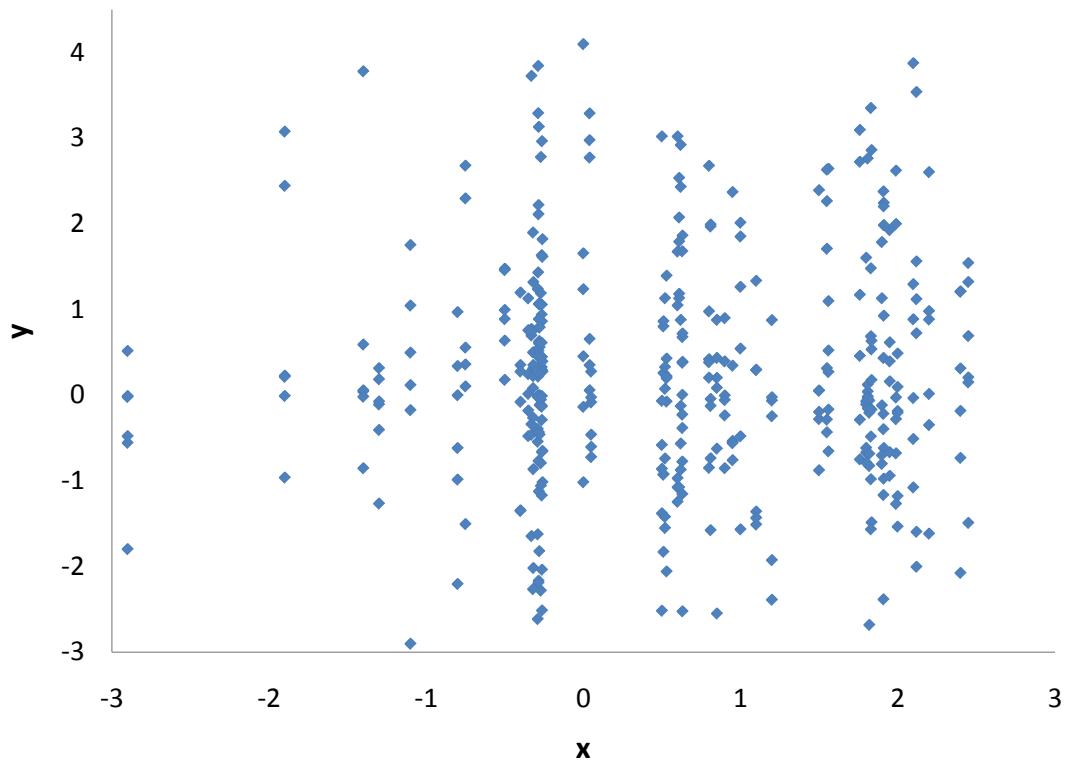


Figure 3. A scatter plot for which  $r = 0$ . Notice that there is no relationship between X and Y.

With real data, you would not expect to get values of  $r$  of exactly  $-1$ ,  $0$ , or  $1$ . The data for spousal ages shown in Figure 4 and described in the introductory section has an  $r$  of  $0.97$ .

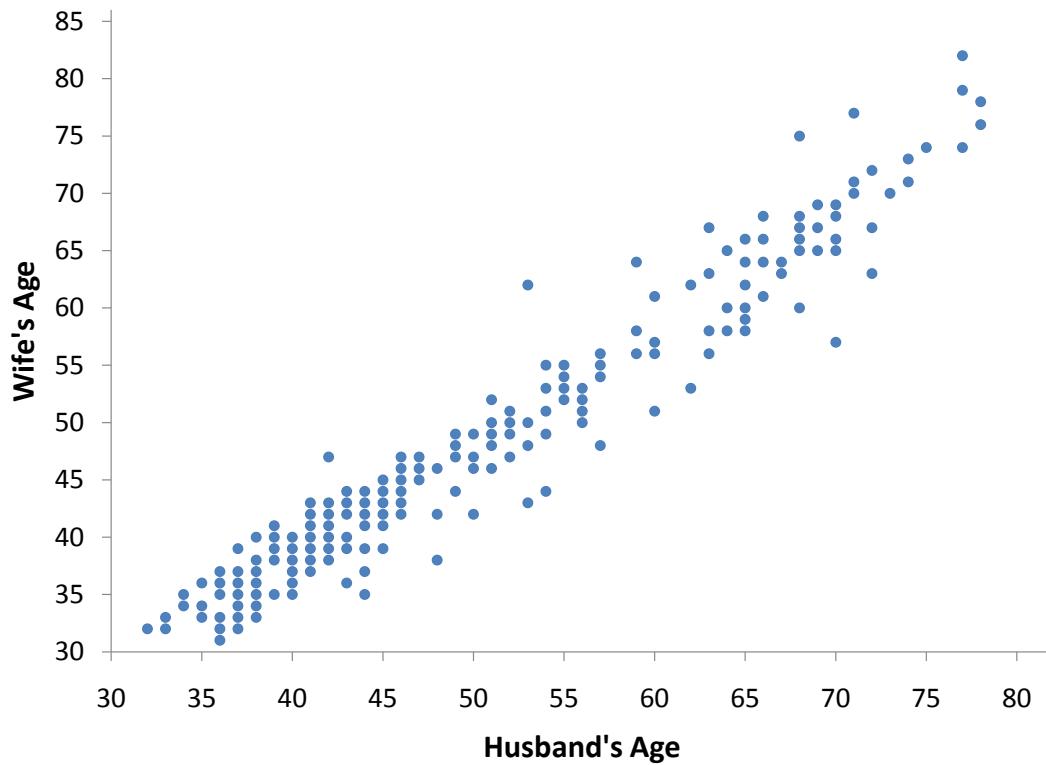


Figure 4. Scatter plot of spousal ages,  $r = 0.97$ .

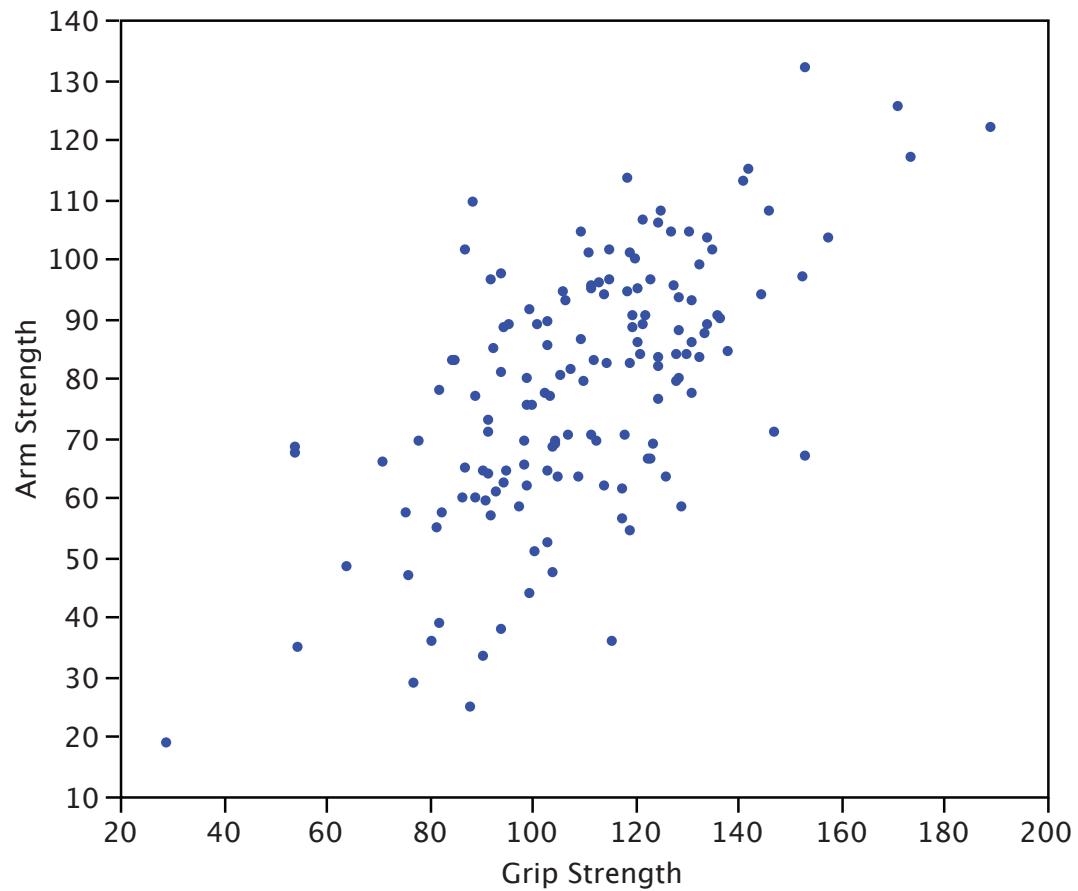


Figure 5. Scatter plot of Grip Strength and Arm Strength,  $r = 0.63$ .

The relationship between grip strength and arm strength depicted in Figure 5 (also described in the introductory section) is 0.63.

# Properties of Pearson's $r$

by David M. Lane

## *Prerequisites*

- Chapter 1: Linear Transformations
- Chapter 4: Introduction to Bivariate Data

## *Learning Objectives*

1. State the range of values for Pearson's correlation
2. State the values that represent perfect linear relationships
3. State the relationship between the correlation of  $Y$  with  $X$  and the correlation of  $X$  with  $Y$
4. State the effect of linear transformations on Pearson's correlation

A basic property of Pearson's  $r$  is that its possible range is from -1 to 1. A correlation of -1 means a perfect negative linear relationship, a correlation of 0 means no linear relationship, and a correlation of 1 means a perfect positive linear relationship.

Pearson's correlation is symmetric in the sense that the correlation of  $X$  with  $Y$  is the same as the correlation of  $Y$  with  $X$ . For example, the correlation of Weight with Height is the same as the correlation of Height with Weight.

A critical property of Pearson's  $r$  is that it is unaffected by linear transformations. This means that multiplying a variable by a constant and/or adding a constant does not change the correlation of that variable with other variables. For instance, the correlation of Weight and Height does not depend on whether Height is measured in inches, feet, or even miles. Similarly, adding five points to every student's test score would not change the correlation of the test score with other variables such as GPA.

# Computing Pearson's $r$

by David M. Lane

## *Prerequisites*

- Chapter 1: Summation Notation
- Chapter 4: Introduction to Bivariate Data

## *Learning Objectives*

1. Define  $X$  and  $x$
2. State why  $\Sigma xy = 0$  when there is no relationship
3. Calculate  $r$

There are several formulas that can be used to compute Pearson's correlation. Some formulas make more conceptual sense whereas others are easier to actually compute. We are going to begin with a formula that makes more conceptual sense.

We are going to compute the correlation between the variables  $X$  and  $Y$  shown in Table 1. We begin by computing the mean for  $X$  and subtracting this mean from all values of  $X$ . The new variable is called “ $x$ .” The variable “ $y$ ” is computed similarly. The variables  $x$  and  $y$  are said to be deviation scores because each score is a deviation from the mean. Notice that the means of  $x$  and  $y$  are both 0. Next we create a new column by multiplying  $x$  and  $y$ .

Before proceeding with the calculations, let's consider why the sum of the  $xy$  column reveals the relationship between  $X$  and  $Y$ . If there were no relationship between  $X$  and  $Y$ , then positive values of  $x$  would be just as likely to be paired with negative values of  $y$  as with positive values. This would make negative values of  $xy$  as likely as positive values and the sum would be small. On the other hand, consider Table 1 in which high values of  $X$  are associated with high values of  $Y$  and low values of  $X$  are associated with low values of  $Y$ . You can see that positive values of  $x$  are associated with positive values of  $y$  and negative values of  $x$  are associated with negative values of  $y$ . In all cases, the product of  $x$  and  $y$  is positive, resulting in a high total for the  $xy$  column. Finally, if there were a negative relationship then positive values of  $x$  would be associated with negative values of  $y$  and negative values of  $x$  would be associated with positive values of  $y$ . This would lead to negative values for  $xy$ .

Table 1. Calculation of r.

	X	Y	x	y	xy	x <sup>2</sup>	y <sup>2</sup>
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

Pearson's r is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet. To achieve this property, Pearson's correlation is computed by dividing the sum of the xy column ( $\Sigma xy$ ) by the square root of the product of the sum of the  $x^2$  column ( $\Sigma x^2$ ) and the sum of the  $y^2$  column ( $\Sigma y^2$ ). The resulting formula is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

and therefore

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

An alternative computational formula that avoids the step of computing deviation scores is:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{N}\right)} \sqrt{\left(\sum y^2 - \frac{(\sum y)^2}{N}\right)}}$$

## Variance Sum Law II

by David M. Lane

### *Prerequisites*

- Chapter 1: Variance Sum Law I
- Chapter 4: Values of Pearson's Correlation

### *Learning Objectives*

1. State the variance sum law when X and Y are not assumed to be independent
2. Compute the variance of the sum of two variables if the variance of each and their correlation is known
3. Compute the variance of the difference between two variables if the variance of each and their correlation is known

Recall that when the variables X and Y are independent, the variance of the sum or difference between X and Y can be written as follows:

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$$

which is read: “The variance of X plus or minus Y is equal to the variance of X plus the variance of Y.”

When X and Y are correlated, the following formula should be used:

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y$$

where  $\rho$  is the correlation between X and Y in the population. For example, if the variance of verbal SAT were 10,000, the variance of quantitative SAT were 11,000 and the correlation between these two tests were 0.50, then the variance of total SAT (verbal + quantitative) would be:

$$\sigma_{verbal+quant}^2 = 10,000 + 11,000 + (2)(0.5)\sqrt{10,000}\sqrt{11,000}$$

which is equal to 31,488. The variance of the difference is:

$$\sigma_{verbal-quant}^2 = 10,000 + 11,000 - (2)(0.5)\sqrt{10,000}\sqrt{11,000}$$

which is equal to 10,512.

If the variances and the correlation are computed in a sample, then the following notation is used to express the variance sum law:

$$s_{X \pm Y}^2 = s_X^2 + s_Y^2 \pm 2rs_Xs_y$$

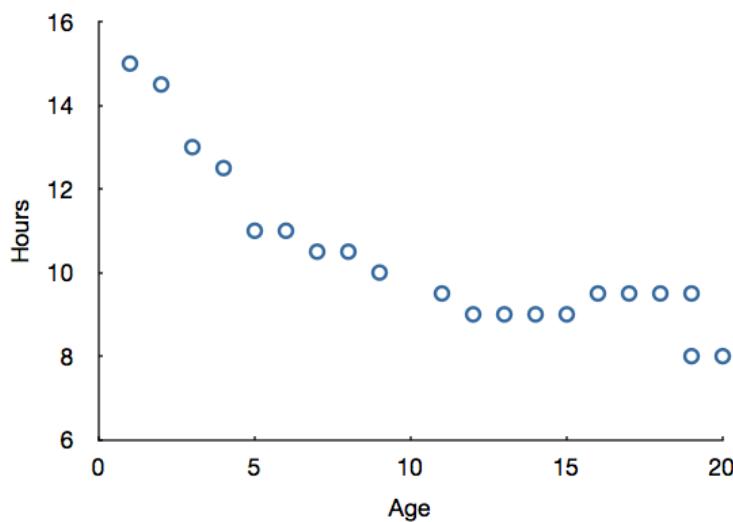
# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 4: Values of Pearson's Correlation

The graph below showing the relationship between age and sleep is based on a graph that appears on [this web page](#).



## What do you think?

Why might Pearson's correlation not be a good way to describe the relationship?

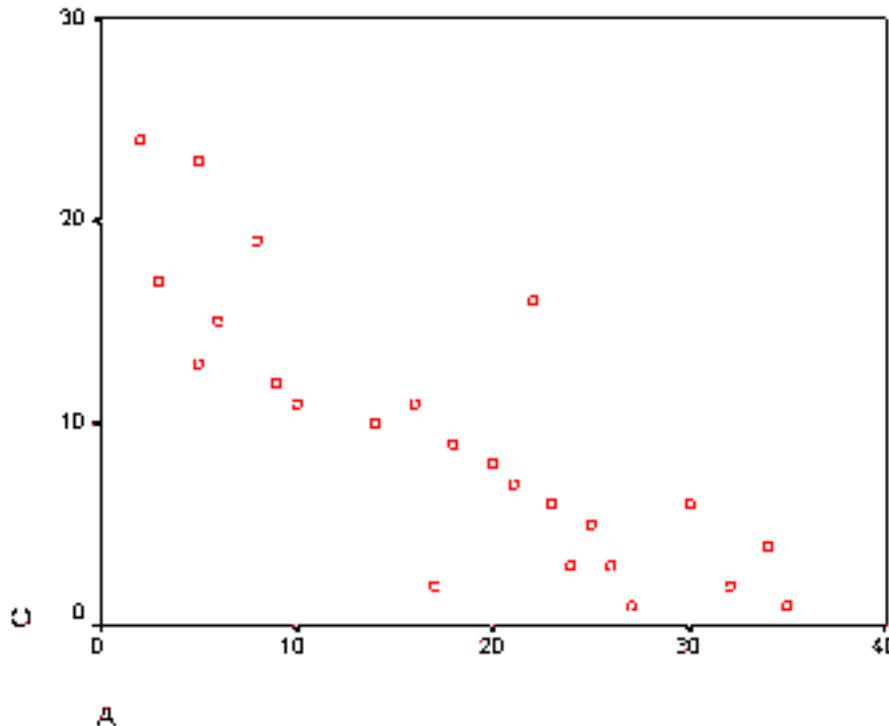
Pearson's correlation measures the strength of the linear relationship between two variables. The relationship here is not linear. As age increases, hours slept decreases rapidly at first but then levels off.

## Exercises

### Prerequisites

- All material presented in the Describing Bivariate Data chapter

1. Describe the relationship between variables A and C. Think of things these variables could represent in real life.



2. Make up a data set with 10 numbers that has a positive correlation.
3. Make up a data set with 10 numbers that has a negative correlation.
4. If the correlation between weight (in pounds) and height (in feet) is 0.58, find:
  - (a) the correlation between weight (in pounds) and height (in yards)
  - (b) the correlation between weight (in kilograms) and height (in meters).
5. Would you expect the correlation between High School GPA and College GPA to be higher when taken from your entire high school class or when taken from only the top 20 students? Why?

6. For a certain class, the relationship between the amount of time spent studying and the test grade earned was examined. It was determined that as the amount of time they studied increased, so did their grades. Is this a positive or negative association?
7. For this same class, the relationship between the amount of time spent studying and the amount of time spent socializing per week was also examined. It was determined that the more hours they spent studying, the fewer hours they spent socializing. Is this a positive or negative association?
8. For the following data:
- Find the deviation scores for Variable A that correspond to the raw scores of 2 and 8.
  - Find the deviation scores for Variable B that correspond to the raw scores of 5 and 4.
  - Just from looking at these scores, do you think these variables are positively or negatively correlated? Why?
  - Now calculate the correlation. Were you right?

A	B
2	8
5	5
6	2
8	4
9	1

9. Students took two parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 49. The correlation between the test scores is 0.6. (a) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be? (b) What would the variance of Part A - Part B be?
10. True/False: The correlation in real life between height and weight is  $r=1$ .

11. True/False: It is possible for variables to have  $r=0$  but still have a strong association.
12. True/False: Two variables with a correlation of 0.3 have a stronger linear relationship than two variables with a correlation of -0.7.
13. True/False: After polling a certain group of people, researchers found a 0.5 correlation between the number of car accidents per year and the driver's age. This means that older people get in more accidents.
14. True/False: The correlation between R and T is the same as the correlation between T and R.
15. True/False: To examine bivariate data graphically, the best choice is two side by side histograms.
16. True/False: A correlation of  $r=1.2$  is not possible.

*Questions from Case Studies*

Angry Moods (AM) case study

17. (AM) What is the correlation between the Control-In and Control-Out scores?
18. (AM) Would you expect the correlation between the Anger-Out and Control-Out scores to be positive or negative? Compute this correlation.

Flatulence (F) case study

19. (F) Is there a relationship between the number of male siblings and embarrassment in front of romantic interests? Create a scatterplot and compute  $r$ .

Stroop (S) case study

20. (S) Create a scatterplot showing "words" on the X-axis and "colors" on the Y-axis.

21. (S) Compute the correlation between “colors” and “words.”
22. (S) Sort the data by color-naming time. Choose only the 23 fastest color-namers.
  - (a) What is the new correlation?
  - (b) What is the technical term for the finding that this correlation is smaller than the correlation for the full dataset?

#### Animal Research (AR) case study

23. (AR) What is the overall correlation between the belief that animal research is wrong and belief that animal research is necessary?

#### ADHD Treatment (AT) case study

24. (AT) What is the correlation between the participants’ correct number of responses after taking the placebo and their correct number of responses after taking 0.60 mg/kg of MPH?

# 5. Probability

- A. Introduction
- B. Basic Concepts
- C. Permutations and Combinations
- D. Poisson Distribution
- E. Multinomial Distribution
- F. Hypergeometric Distribution
- G. Base Rates
- H. Exercises

Probability is an important and complex field of study. Fortunately, only a few basic issues in probability theory are essential for understanding statistics at the level covered in this book. These basic issues are covered in this chapter.

The introductory section discusses the definitions of probability. This is not as simple as it may seem. The section on basic concepts covers how to compute probabilities in a variety of simple situations. The section on base rates discusses an important but often-ignored factor in determining probabilities.

# Remarks on the Concept of “Probability”

by Dan Osherson

## *Prerequisites*

- None

## *Learning Objectives*

1. Define symmetrical outcomes
2. Distinguish between frequentist and subjective approaches
3. Determine whether the frequentist or subjective approach is better suited for a given situation

Inferential statistics is built on the foundation of probability theory, and has been remarkably successful in guiding opinion about the conclusions to be drawn from data. Yet (paradoxically) the very idea of probability has been plagued by controversy from the beginning of the subject to the present day. In this section we provide a glimpse of the debate about the interpretation of the probability concept.

One conception of probability is drawn from the idea of **symmetrical outcomes**. For example, the two possible outcomes of tossing a fair coin seem not to be distinguishable in any way that affects which side will land up or down. Therefore the probability of heads is taken to be  $1/2$ , as is the probability of tails. In general, if there are  $N$  symmetrical outcomes, the probability of any given one of them occurring is taken to be  $1/N$ . Thus, if a six-sided die is rolled, the probability of any one of the six sides coming up is  $1/6$ .

Probabilities can also be thought of in terms of **relative frequencies**. If we tossed a coin millions of times, we would expect the proportion of tosses that came up heads to be pretty close to  $1/2$ . As the number of tosses increases, the proportion of heads approaches  $1/2$ . Therefore, we can say that the probability of a head is  $1/2$ .

If it has rained in Seattle on 62% of the last 100,000 days, then the probability of it raining tomorrow might be taken to be 0.62. This is a natural idea but nonetheless unreasonable if we have further information relevant to whether it will rain tomorrow. For example, if tomorrow is August 1, a day of the year on which it seldom rains in Seattle, we should only consider the percentage of the time it rained on August 1. But even this is not enough since the probability of rain on the next August 1 depends on the humidity. (The chances are higher in the presence of high humidity.) So, we should consult only the prior occurrences of

August 1 that had the same humidity as the next occurrence of August 1. Of course, wind direction also affects probability. You can see that our sample of prior cases will soon be reduced to the empty set. Anyway, past meteorological history is misleading if the climate is changing.

For some purposes, probability is best thought of as subjective. Questions such as “What is the probability that Ms. Garcia will defeat Mr. Smith in an upcoming congressional election?” do not conveniently fit into either the symmetry or frequency approaches to probability. Rather, assigning probability 0.7 (say) to this event seems to reflect the speaker's personal opinion --- perhaps his willingness to bet according to certain odds. Such an approach to probability, however, seems to lose the objective content of the idea of chance; probability becomes mere opinion.

Two people might attach different probabilities to the election outcome, yet there would be no criterion for calling one “right” and the other “wrong.” We cannot call one of the two people right simply because she assigned higher probability to the outcome that actually transpires. After all, you would be right to attribute probability 1/6 to throwing a six with a fair die, and your friend who attributes 2/3 to this event would be wrong. And you are still right (and your friend is still wrong) even if the die ends up showing a six! The lack of objective criteria for adjudicating claims about probabilities in the subjective perspective is an unattractive feature of it for many scholars.

Like most work in the field, the present text adopts the frequentist approach to probability in most cases. Moreover, almost all the probabilities we shall encounter will be nondogmatic, that is, neither zero nor one. An event with probability 0 has no chance of occurring; an event of probability 1 is certain to occur. It is hard to think of any examples of interest to statistics in which the probability is either 0 or 1. (Even the probability that the Sun will come up tomorrow is less than 1.)

The following example illustrates our attitude about probabilities. Suppose you wish to know what the weather will be like next Saturday because you are planning a picnic. You turn on your radio, and the weather person says, “There is a 10% chance of rain.” You decide to have the picnic outdoors and, lo and behold, it rains. You are furious with the weather person. But was she wrong? No, she did not say it would not rain, only that rain was unlikely. She would have been flatly wrong only if she said that the probability is 0 and it subsequently rained. However, if you kept track of her weather predictions over a long period of time

and found that it rained on 50% of the days that the weather person said the probability was 0.10, you could say her probability assessments are wrong.

So when is it accurate to say that the probability of rain is 0.10? According to our frequency interpretation, it means that it will rain 10% of the days on which rain is forecast with this probability.

# Basic Concepts

by David M. Lane

## *Prerequisites*

- Chapter 5: Introduction to Probability

## *Learning Objectives*

1. Compute probability in a situation where there are equally-likely outcomes
2. Apply concepts to cards and dice
3. Compute the probability of two independent events both occurring
4. Compute the probability of either of two independent events occurring
5. Do problems that involve conditional probabilities
6. Compute the probability that in a room of  $N$  people, at least two share a birthday
7. Describe the gambler's fallacy

## Probability of a Single Event

If you roll a six-sided die, there are six possible outcomes, and each of these outcomes is equally likely. A six is as likely to come up as a three, and likewise for the other four sides of the die. What, then, is the probability that a one will come up? Since there are six possible outcomes, the probability is  $1/6$ . What is the probability that either a one or a six will come up? The two outcomes about which we are concerned (a one or a six coming up) are called favorable outcomes. Given that all outcomes are equally likely, we can compute the probability of a one or a six using the formula:

$$\text{probability} = \frac{\text{Number of favorable outcomes}}{\text{Number of possible equally-likely outcomes}}$$

In this case there are two favorable outcomes and six possible outcomes. So the probability of throwing either a one or six is  $1/3$ . Don't be misled by our use of the term "favorable," by the way. You should understand it in the sense of "favorable to the event in question happening." That event might not be favorable to your well-being. You might be betting on a three, for example.

The above formula applies to many games of chance. For example, what is the probability that a card drawn at random from a deck of playing cards will be an ace? Since the deck has four aces, there are four favorable outcomes; since the deck has 52 cards, there are 52 possible outcomes. The probability is therefore  $4/52 = 1/13$ . What about the probability that the card will be a club? Since there are 13 clubs, the probability is  $13/52 = 1/4$ .

Let's say you have a bag with 20 cherries: 14 sweet and 6 sour. If you pick a cherry at random, what is the probability that it will be sweet? There are 20 possible cherries that could be picked, so the number of possible outcomes is 20. Of these 20 possible outcomes, 14 are favorable (sweet), so the probability that the cherry will be sweet is  $14/20 = 7/10$ . There is one potential complication to this example, however. It must be assumed that the probability of picking any of the cherries is the same as the probability of picking any other. This wouldn't be true if (let us imagine) the sweet cherries are smaller than the sour ones. (The sour cherries would come to hand more readily when you sampled from the bag.) Let us keep in mind, therefore, that when we assess probabilities in terms of the ratio of favorable to all potential cases, we rely heavily on the assumption of equal probability for all outcomes.

Here is a more complex example. You throw 2 dice. What is the probability that the sum of the two dice will be 6? To solve this problem, list all the possible outcomes. There are 36 of them since each die can come up one of six ways. The 36 possibilities are shown in Table 1.

Table 1. 36 possible outcomes.

Die 1	Die 2	Total	Die 1	Die 2	Total	Die 1	Die 2	Total
1	1	2	3	1	4	5	1	6
1	2	3	3	2	5	5	2	7
1	3	4	3	3	6	5	3	8
1	4	5	3	4	7	5	4	9
1	5	6	3	5	8	5	5	10
1	6	7	3	6	9	5	6	11
2	1	3	4	1	5	6	1	7
2	2	4	4	2	6	6	2	8
2	3	5	4	3	7	6	3	9
2	4	6	4	4	8	6	4	10
2	5	7	4	5	9	6	5	11
2	6	8	4	6	10	6	6	12

You can see that 5 of the 36 possibilities total 6. Therefore, the probability is 5/36.

If you know the probability of an event occurring, it is easy to compute the probability that the event does not occur. If  $P(A)$  is the probability of Event A, then  $1 - P(A)$  is the probability that the event does not occur. For the last example, the probability that the total is 6 is 5/36. Therefore, the probability that the total is not 6 is  $1 - 5/36 = 31/36$ .

## Probability of Two (or more) Independent Events

Events A and B are independent events if the probability of Event B occurring is the same whether or not Event A occurs. Let's take a simple example. A fair coin is tossed two times. The probability that a head comes up on the second toss is 1/2 regardless of whether or not a head came up on the first toss. The two events are (1) first toss is a head and (2) second toss is a head. So these events are independent. Consider the two events (1) "It will rain tomorrow in Houston" and (2) "It will rain tomorrow in Galveston" (a city near Houston). These events are not independent because it is more likely that it will rain in Galveston on days it rains in Houston than on days it does not.

## Probability of A and B

When two events are independent, the probability of both occurring is the product of the probabilities of the individual events. More formally, if events A and B are independent, then the probability of both A and B occurring is:

$$P(A \text{ and } B) = P(A) \times P(B)$$

where  $P(A \text{ and } B)$  is the probability of events A and B both occurring,  $P(A)$  is the probability of event A occurring, and  $P(B)$  is the probability of event B occurring.

If you flip a coin twice, what is the probability that it will come up heads both times? Event A is that the coin comes up heads on the first flip and Event B is that the coin comes up heads on the second flip. Since both  $P(A)$  and  $P(B)$  equal  $1/2$ , the probability that both events occur is

$$1/2 \times 1/2 = 1/4$$

Let's take another example. If you flip a coin and roll a six-sided die, what is the probability that the coin comes up heads and the die comes up 1? Since the two events are independent, the probability is simply the probability of a head (which is  $1/2$ ) times the probability of the die coming up 1 (which is  $1/6$ ). Therefore, the probability of both events occurring is  $1/2 \times 1/6 = 1/12$ .

One final example: You draw a card from a deck of cards, put it back, and then draw another card. What is the probability that the first card is a heart and the second card is black? Since there are 52 cards in a deck and 13 of them are hearts, the probability that the first card is a heart is  $13/52 = 1/4$ . Since there are 26 black cards in the deck, the probability that the second card is black is  $26/52 = 1/2$ . The probability of both events occurring is therefore  $1/4 \times 1/2 = 1/8$ .

See the discussion on conditional probabilities later in this section to see how to compute  $P(A \text{ and } B)$  when A and B are not independent.

## Probability of A or B

If Events A and B are independent, the probability that either Event A or Event B occurs is:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

In this discussion, when we say “A or B occurs” we include three possibilities:

1. A occurs and B does not occur
2. B occurs and A does not occur
3. Both A and B occur

This use of the word “or” is technically called inclusive or because it includes the case in which both A and B occur. If we included only the first two cases, then we would be using an exclusive or.

**(Optional)** We can derive the law for  $P(A\text{-or-}B)$  from our law about  $P(A\text{-and-}B)$ .

The event “A-or-B” can happen in any of the following ways:

1. A-and-B happens
2. A-and-not-B happens
3. not-A-and-B happens.

The simple event A can happen if either A-and-B happens or A-and-not-B happens. Similarly, the simple event B happens if either A-and-B happens or not-A-and-B happens.  $P(A) + P(B)$  is therefore  $P(A\text{-and-}B) + P(A\text{-and-not-}B) + P(A\text{-and-}B) + P(\text{not-}A\text{-and-}B)$ , whereas  $P(A\text{-or-}B)$  is  $P(A\text{-and-}B) + P(A\text{-and-not-}B) + P(\text{not-}A\text{-and-}B)$ . We can make these two sums equal by subtracting one occurrence of  $P(A\text{-and-}B)$  from the first. Hence,  $P(A\text{-or-}B) = P(A) + P(B) - P(A\text{-and-}B)$ .

Now for some examples. If you flip a coin two times, what is the probability that you will get a head on the first flip or a head on the second flip (or both)? Letting Event A be a head on the first flip and Event B be a head on the second flip, then  $P(A) = 1/2$ ,  $P(B) = 1/2$ , and  $P(A \text{ and } B) = 1/4$ . Therefore,

$$P(A \text{ or } B) = 1/2 + 1/2 - 1/4 = 3/4.$$

If you throw a six-sided die and then flip a coin, what is the probability that you will get either a 6 on the die or a head on the coin flip (or both)? Using the formula,

$$\begin{aligned} P(6 \text{ or head}) &= P(6) + P(\text{head}) - P(6 \text{ and head}) \\ &= (1/6) + (1/2) - (1/6)(1/2) \\ &= 7/12 \end{aligned}$$

An alternate approach to computing this value is to start by computing the probability of not getting either a 6 or a head. Then subtract this value from 1 to compute the probability of getting a 6 or a head. Although this is a complicated

method, it has the advantage of being applicable to problems with more than two events. Here is the calculation in the present case. The probability of not getting either a 6 or a head can be recast as the probability of

(not getting a 6) AND (not getting a head).

This follows because if you did not get a 6 and you did not get a head, then you did not get a 6 or a head. The probability of not getting a six is  $1 - 1/6 = 5/6$ . The probability of not getting a head is  $1 - 1/2 = 1/2$ . The probability of not getting a six and not getting a head is  $5/6 \times 1/2 = 5/12$ . This is therefore the probability of not getting a 6 or a head. The probability of getting a six or a head is therefore (once again)  $1 - 5/12 = 7/12$ .

If you throw a die three times, what is the probability that one or more of your throws will come up with a 1? That is, what is the probability of getting a 1 on the first throw OR a 1 on the second throw OR a 1 on the third throw? The easiest way to approach this problem is to compute the probability of

NOT getting a 1 on the first throw  
AND not getting a 1 on the second throw  
AND not getting a 1 on the third throw.

The answer will be 1 minus this probability. The probability of not getting a 1 on any of the three throws is  $5/6 \times 5/6 \times 5/6 = 125/216$ . Therefore, the probability of getting a 1 on at least one of the throws is  $1 - 125/216 = 91/216$ .

## Conditional Probabilities

Often it is required to compute the probability of an event given that another event has occurred. For example, what is the probability that two cards drawn at random from a deck of playing cards will both be aces? It might seem that you could use the formula for the probability of two independent events and simply multiply  $4/52 \times 4/52 = 1/169$ . This would be **incorrect**, however, because the two events are not independent. If the first card drawn is an ace, then the probability that the second card is also an ace would be lower because there would only be three aces left in the deck.

Once the first card chosen is an ace, the probability that the second card chosen is also an ace is called the conditional probability of drawing an ace. In this case, the “condition” is that the first card is an ace. Symbolically, we write this as:

$$P(\text{ace on second draw} \mid \text{an ace on the first draw})$$

The vertical bar “|” is read as “given,” so the above expression is short for: “The probability that an ace is drawn on the second draw given that an ace was drawn on the first draw.” What is this probability? Since after an ace is drawn on the first draw, there are 3 aces out of 51 total cards left. This means that the probability that one of these aces will be drawn is  $3/51 = 1/17$ .

If Events A and B are not independent, then  
 $P(A \text{ and } B) = P(A) \times P(B|A)$ .

Applying this to the problem of two aces, the probability of drawing two aces from a deck is  $4/52 \times 3/51 = 1/221$ .

One more example: If you draw two cards from a deck, what is the probability that you will get the Ace of Diamonds and a black card? There are two ways you can satisfy this condition: (1) You can get the Ace of Diamonds first and then a black card or (2) you can get a black card first and then the Ace of Diamonds. Let's calculate Case 1. The probability that the first card is the Ace of Diamonds is  $1/52$ . The probability that the second card is black given that the first card is the Ace of Diamonds is  $26/51$  because 26 of the remaining 51 cards are black. The probability is therefore  $1/52 \times 26/51 = 1/102$ . Now for Case 2: the probability that the first card is black is  $26/52 = 1/2$ . The probability that the second card is the Ace of Diamonds given that the first card is black is  $1/51$ . The probability of Case 2 is therefore  $1/2 \times 1/51 = 1/102$ , the same as the probability of Case 1. Recall that the probability of A or B is  $P(A) + P(B) - P(A \text{ and } B)$ . In this problem,  $P(A \text{ and } B) = 0$  since a card cannot be the Ace of Diamonds and be a black card. Therefore, the probability of Case 1 or Case 2 is  $1/102 + 1/102 = 2/102 = 1/51$ . So,  $1/51$  is the probability that you will get the Ace of Diamonds and a black card when drawing two cards from a deck.

## Birthday Problem

If there are 25 people in a room, what is the probability that at least two of them share the same birthday. If your first thought is that it is  $25/365 = 0.068$ , you will be surprised to learn it is much higher than that. This problem requires the application of the sections on  $P(A \text{ and } B)$  and conditional probability.

This problem is best approached by asking what is the probability that no two people have the same birthday. Once we know this probability, we can simply subtract it from 1 to find the probability that two people share a birthday.

If we choose two people at random, what is the probability that they do not share a birthday? Of the 365 days on which the second person could have a birthday, 364 of them are different from the first person's birthday. Therefore the probability is  $364/365$ . Let's define  $P_2$  as the probability that the second person drawn does not share a birthday with the person drawn previously.  $P_2$  is therefore  $364/365$ . Now define  $P_3$  as the probability that the third person drawn does not share a birthday with anyone drawn previously **given** that there are no previous birthday matches.  $P_3$  is therefore a conditional probability. If there are no previous birthday matches, then two of the 365 days have been "used up," leaving 363 non-matching days. Therefore  $P_3 = 363/365$ . In like manner,  $P_4 = 362/365$ ,  $P_5 = 361/365$ , and so on up to  $P_{25} = 341/365$ .

In order for there to be no matches, the second person must not match any previous person **and** the third person must not match any previous person, and the fourth person must not match any previous person, etc. Since  $P(A \text{ and } B) = P(A)P(B)$ , all we have to do is multiply  $P_2, P_3, P_4 \dots P_{25}$  together. The result is 0.431. Therefore the probability of at least one match is 0.569.

## Gambler's Fallacy

A fair coin is flipped five times and comes up heads each time. What is the probability that it will come up heads on the sixth flip? The correct answer is, of course,  $1/2$ . But many people believe that a tail is more likely to occur after throwing five heads. Their **faulty reasoning** may go something like this: "In the long run, the number of heads and tails will be the same, so the tails have some catching up to do."

The error in this reasoning is that the proportion of heads approaches 0.5 but the number of heads does not approach the number of tails. The results of a simulation ([external link](#); requires Java) are shown in Figure 1. (The quality of the image is somewhat low because it was captured from the screen.)

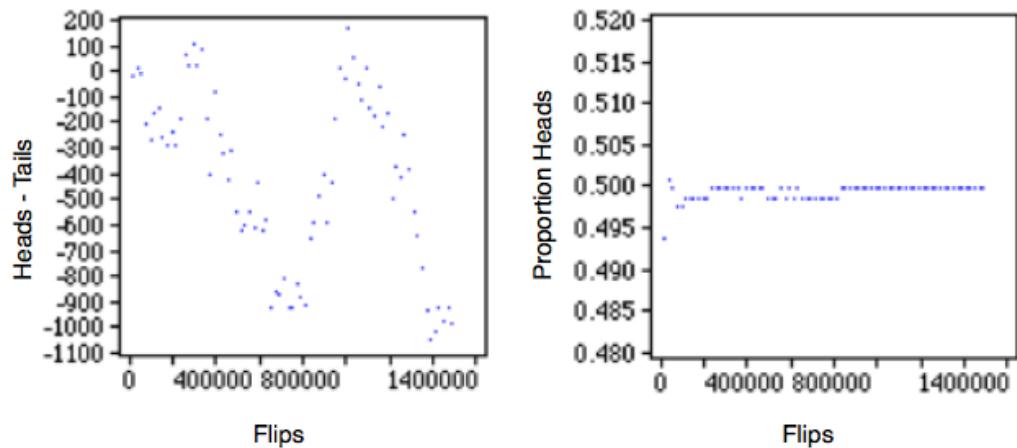


Figure 1. The results of simulating 1,500,000 coin flips. The graph on the left shows the difference between the number of heads and the number of tails as a function of the number of flips. You can see that there is no consistent pattern. After the final flip, there are 968 more tails than heads. The graph on the right shows the proportion of heads. This value goes up and down at the beginning, but converges to 0.5 (rounded to 3 decimal places) before 1,000,000 flips.

# Permutations and Combinations

by David M. Lane

## *Prerequisites*

none

## *Learning Objectives*

1. Calculate the probability of two independent events occurring
2. Define permutations and combinations
3. List all permutations and combinations
4. Apply formulas for permutations and combinations

This section covers basic formulas for determining the number of various possible types of outcomes. The topics covered are: (1) counting the number of possible orders, (2) counting using the multiplication rule, (3) counting the number of permutations, and (4) counting the number of combinations.

## Possible Orders

Suppose you had a plate with three pieces of candy on it: one green, one yellow, and one red. You are going to pick up these three pieces one at a time. The question is: In how many different orders can you pick up the pieces? Table 1 lists all the possible orders. There are two orders in which red is first: red, yellow, green and red, green, yellow. Similarly, there are two orders in which yellow is first and two orders in which green is first. This makes six possible orders in which the pieces can be picked up.

Table 1. Six Possible Orders.

Number	First	Second	Third
1	red	yellow	green
2	red	green	yellow
3	yellow	red	green
4	yellow	green	red
5	green	red	yellow
6	green	yellow	red

The formula for the number of orders is shown below.

$$\text{Number of orders} = n!$$

where  $n$  is the number of pieces to be picked up. The symbol “!” stands for factorial. Some examples are:

$$\begin{aligned}3! &= 3 \times 2 \times 1 = 6 \\4! &= 4 \times 3 \times 2 \times 1 = 24 \\5! &= 5 \times 4 \times 3 \times 2 \times 1 = 120\end{aligned}$$

This means that if there were 5 pieces of candy to be picked up, they could be picked up in any of  $5! = 120$  orders.

## Multiplication Rule

Imagine a small restaurant whose menu has 3 soups, 6 entrées, and 4 desserts. How many possible meals are there? The answer is calculated by multiplying the numbers to get  $3 \times 6 \times 4 = 72$ . You can think of it as first there is a choice among 3 soups. Then, for each of these choices there is a choice among 6 entrées resulting in  $3 \times 6 = 18$  possibilities. Then, for each of these 18 possibilities there are 4 possible desserts yielding  $18 \times 4 = 72$  total possibilities.

## Permutations

Suppose that there were four pieces of candy (red, yellow, green, and brown) and you were only going to pick up exactly two pieces. How many ways are there of

picking up two pieces? Table 2 lists all the possibilities. The first choice can be any of the four colors. For each of these 4 first choices there are 3 second choices. Therefore there are  $4 \times 3 = 12$  possibilities.

Table 2. Twelve Possible Orders.

Number	First	Second
1	red	yellow
2	red	green
3	red	brown
4	yellow	red
5	yellow	green
6	yellow	brown
7	green	red
8	green	yellow
9	green	brown
10	brown	red
11	brown	yellow
12	brown	green

More formally, this question is asking for the number of permutations of four things taken two at a time. The general formula is:

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

where  ${}_n P_r$  is the number of permutations of  $n$  things taken  $r$  at a time. In other words, it is the number of ways  $r$  things can be selected from a group of  $n$  things. In this case,

$${}_4 P_2 = \frac{4!}{(4 - 2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 12$$

It is important to note that order counts in permutations. That is, choosing red and then yellow is counted separately from choosing yellow and then red. Therefore permutations refer to the number of ways of choosing rather than the number of

possible outcomes. When order of choice is not considered, the formula for combinations is used.

## Combinations

Now suppose that you were not concerned with the way the pieces of candy were chosen but only in the final choices. In other words, how many different combinations of two pieces could you end up with? In counting combinations, choosing red and then yellow is the same as choosing yellow and then red because in both cases you end up with one red piece and one yellow piece. Unlike permutations, order does not count. Table 3 is based on Table 2 but is modified so that repeated combinations are given an “x” instead of a number. For example, “yellow then red” has an “x” because the combination of red and yellow was already included as choice number 1. As you can see, there are six combinations of the three colors.

Table 3. Six Combinations.

Number	First	Second
1	red	yellow
2	red	green
3	red	brown
x	yellow	red
4	yellow	green
5	yellow	brown
x	green	red
x	green	yellow
6	green	brown
x	brown	red
x	brown	yellow
x	brown	green

The formula for the number of combinations is shown below where  $nCr$  is the number of combinations for  $n$  things taken  $r$  at a time.

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

For our example,

$${}_4C_2 = \frac{4!}{(4 - 2)! 2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = 6$$

which is consistent with Table 3.

As an example application, suppose there were six kinds of toppings that one could order for a pizza. How many combinations of exactly 3 toppings could be ordered? Here  $n = 6$  since there are 6 toppings and  $r = 3$  since we are taking 3 at a time. The formula is then:

$${}_6C_3 = \frac{6!}{(6 - 3)! 3!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(3 \times 2 \times 1)} = 20.$$

# Binomial Distribution

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Variability
- Chapter 5: Basic Probability

## *Learning Objectives*

1. Define binomial outcomes
2. Compute the probability of getting X successes in N trials
3. Compute cumulative binomial probabilities
4. Find the mean and standard deviation of a binomial distribution

When you flip a coin, there are two possible outcomes: heads and tails. Each outcome has a fixed probability, the same from trial to trial. In the case of coins, heads and tails each have the same probability of 1/2. More generally, there are situations in which the coin is biased, so that heads and tails have different probabilities. In the present section, we consider probability distributions for which there are just two possible outcomes with fixed probabilities summing to one. These distributions are called binomial distributions.

## A Simple Example

The four possible outcomes that could occur if you flipped a coin twice are listed below in Table 1. Note that the four outcomes are equally likely: each has probability 1/4. To see this, note that the tosses of the coin are independent (neither affects the other). Hence, the probability of a head on Flip 1 and a head on Flip 2 is the product of  $P(H)$  and  $P(H)$ , which is  $1/2 \times 1/2 = 1/4$ . The same calculation applies to the probability of a head on Flip 1 and a tail on Flip 2. Each is  $1/2 \times 1/2 = 1/4$ .

Table 1. Four Possible Outcomes.

Outcome	First Flip	Second Flip
1	Heads	Heads
2	Heads	Tails

3	Tails	Heads
4	Tails	Tails

The four possible outcomes can be classified in terms of the number of heads that come up. The number could be two (Outcome 1), one (Outcomes 2 and 3) or 0 (Outcome 4). The probabilities of these possibilities are shown in Table 2 and in Figure 1. Since two of the outcomes represent the case in which just one head appears in the two tosses, the probability of this event is equal to  $1/4 + 1/4 = 1/2$ . Table 2 summarizes the situation.

Table 2. Probabilities of Getting 0, 1, or 2 Heads.

Number of Heads	Probability
0	1/4
1	1/2
2	1/4

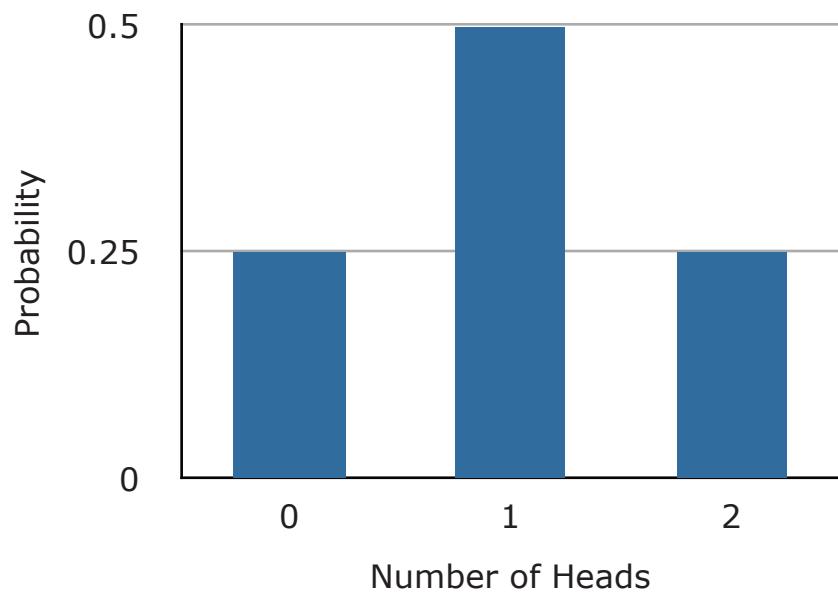


Figure 1. Probabilities of 0, 1, and 2 heads.

Figure 1 is a discrete probability distribution: It shows the probability for each of the values on the X-axis. Defining a head as a “success,” Figure 1 shows the probability of 0, 1, and 2 successes for two trials (flips) for an event that has a

probability of 0.5 of being a success on each trial. This makes Figure 1 an example of a binomial distribution.

## The Formula for Binomial Probabilities

The binomial distribution consists of the probabilities of each of the possible numbers of successes on  $N$  trials for independent events that each have a probability of  $\pi$  (the Greek letter pi) of occurring. For the coin flip example,  $N = 2$  and  $\pi = 0.5$ . The formula for the binomial distribution is shown below:

$$P(x) = \frac{N!}{x!(N-x)!} \pi^x (1-\pi)^{N-x}$$

where  $P(x)$  is the probability of  $x$  successes out of  $N$  trials,  $N$  is the number of trials, and  $\pi$  is the probability of success on a given trial. Applying this to the coin flip example,

$$P(0) = \frac{2!}{0!(2-0)!} (.5^0)(1-.5)^{2-0} = \frac{2}{2} (1)(.25) = 0.25$$

$$P(1) = \frac{2!}{1!(2-1)!} (.5^1)(1-.5)^{2-1} = \frac{2}{1} (.5)(.5) = 0.50$$

$$P(2) = \frac{2!}{2!(2-2)!} (.5^2)(1-.5)^{2-2} = \frac{2}{2} (.25)(1) = 0.25$$

If you flip a coin twice, what is the probability of getting one or more heads? Since the probability of getting exactly one head is 0.50 and the probability of getting exactly two heads is 0.25, the probability of getting one or more heads is  $0.50 + 0.25 = 0.75$ .

Now suppose that the coin is biased. The probability of heads is only 0.4. What is the probability of getting heads at least once in two tosses? Substituting into the general formula above, you should obtain the answer .64.

## Cumulative Probabilities

We toss a coin 12 times. What is the probability that we get from 0 to 3 heads? The answer is found by computing the probability of exactly 0 heads, exactly 1 head, exactly 2 heads, and exactly 3 heads. The probability of getting from 0 to 3 heads

is then the sum of these probabilities. The probabilities are: 0.0002, 0.0029, 0.0161, and 0.0537. The sum of the probabilities is 0.073. The calculation of cumulative binomial probabilities can be quite tedious. Therefore we have provided a binomial calculator ([external link](#); requires Java) to make it easy to calculate these probabilities.

## Mean and Standard Deviation of Binomial Distributions

Consider a coin-tossing experiment in which you tossed a coin 12 times and recorded the number of heads. If you performed this experiment over and over again, what would the mean number of heads be? On average, you would expect half the coin tosses to come up heads. Therefore the mean number of heads would be 6. In general, the mean of a binomial distribution with parameters  $N$  (the number of trials) and  $\pi$  (the probability of success on each trial) is:

$$\mu = N\pi$$

where  $\mu$  is the mean of the binomial distribution. The variance of the binomial distribution is:

$$\sigma^2 = N\pi(1-\pi)$$

where  $\sigma^2$  is the variance of the binomial distribution.

Let's return to the coin-tossing experiment. The coin was tossed 12 times, so  $N = 12$ . A coin has a probability of 0.5 of coming up heads. Therefore,  $\pi = 0.5$ . The mean and variance can therefore be computed as follows:

$$\mu = N\pi = (12)(0.5) = 6$$

$$\sigma^2 = N\pi(1-\pi) = (12)(0.5)(1.0 - 0.5) = 3.0.$$

Naturally, the standard deviation ( $\sigma$ ) is the square root of the variance ( $\sigma^2$ ).

$$\sigma = \sqrt{N\pi(1 - \pi)}$$

# Poisson Distribution

by David M. Lane

## *Prerequisites*

- Chapter 1: Logarithms

The Poisson distribution can be used to calculate the probabilities of various numbers of “successes” based on the mean number of successes. In order to apply the Poisson distribution, the various events must be independent. Keep in mind that the term “success” does not really mean success in the traditional positive sense. It just means that the outcome in question occurs.

Suppose you knew that the mean number of calls to a fire station on a weekday is 8. What is the probability that on a given weekday there would be 11 calls? This problem can be solved using the following formula based on the Poisson distribution:

$$p = \frac{e^{-\mu} \mu^x}{x!}$$

$e$  is the base of natural logarithms (2.7183)

$\mu$  is the mean number of “successes”

$x$  is the number of “successes” in question

For this example,

$$p = \frac{e^{-8} 8^{11}}{11!} = 0.072$$

since the mean is 8 and the question pertains to 11 fires.

The mean of the Poisson distribution is  $\mu$ . The variance is also equal to  $\mu$ . Thus, for this example, both the mean and the variance are equal to 8.

# Multinomial Distribution

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Variability
- Chapter 5: Basic Probability
- Chapter 5: Binomial Distribution

## *Learning Objectives*

1. Define multinomial outcomes
2. Compute probabilities using the multinomial distribution

The binomial distribution allows one to compute the probability of obtaining a given number of binary outcomes. For example, it can be used to compute the probability of getting 6 heads out of 10 coin flips. The flip of a coin is a binary outcome because it has only two possible outcomes: heads and tails. The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes. For example, suppose that two chess players had played numerous games and it was determined that the probability that Player A would win is 0.40, the probability that Player B would win is 0.35, and the probability that the game would end in a draw is 0.25. The multinomial distribution can be used to answer questions such as: "If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?" The following formula gives the probability of obtaining a specific set of outcomes when there are three possible outcomes for each event:

$$p = \frac{n!}{(n_1!)(n_2!)(n_3!)} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

where

p is the probability,  
n is the total number of events

$n_1$  is the number of times Outcome 1 occurs,  
 $n_2$  is the number of times Outcome 2 occurs,  
 $n_3$  is the number of times Outcome 3 occurs,  
 $p_1$  is the probability of Outcome 1  
 $p_2$  is the probability of Outcome 2, and  
 $p_3$  is the probability of Outcome 3.

For the chess example,

$n = 12$  (12 games are played),  
 $n_1 = 7$  (number won by Player A),  
 $n_2 = 2$  (number won by Player B),  
 $n_3 = 3$  (the number drawn),  
 $p_1 = 0.40$  (probability Player A wins)  
 $p_2 = 0.35$  (probability Player B wins)  
 $p_3 = 0.25$  (probability of a draw)

$$p = \frac{12!}{(7!)(2!)(3!)} \cdot 40^7 \cdot 35^2 \cdot 25^3 = 0.0248$$

The formula for  $k$  outcomes is:

$$p = \frac{n!}{(n_1!)(n_2!)\dots(n_k!)} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

# Hypergeometric Distribution

by David M. Lane

## *Prerequisites*

- Chapter 5: Binomial Distribution
- Chapter 5: Permutations and Combinations

The hypergeometric distribution is used to calculate probabilities when sampling without replacement. For example, suppose you first randomly sample one card from a deck of 52. Then, without putting the card back in the deck you sample a second and then (again without replacing cards) a third. Given this sampling procedure, what is the probability that exactly two of the sampled cards will be aces (4 of the 52 cards in the deck are aces). You can calculate this probability using the following formula based on the hypergeometric distribution:

$$p = \frac{kC_x (N - k)C_{(n - x)}}{nC_n}$$

where

$k$  is the number of "successes" in the population

$x$  is the number of "successes" in the sample

$N$  is the size of the population

$n$  is the number sampled

$p$  is the probability of obtaining exactly  $x$  successes

$kC_x$  is the number of combinations of  $k$  things taken  $x$  at a time

In this example,  $k = 4$  because there are four aces in the deck,  $x = 2$  because the problem asks about the probability of getting two aces,  $N = 52$  because there are 52 cards in a deck, and  $n = 3$  because 3 cards were sampled. Therefore,

$$p = \frac{^4C_2 (52 - 4)C_{(3 - 2)}}{52C_3}$$

$$p = \frac{\frac{4!}{2!2!} \frac{48!}{47!1!}}{\frac{52!}{49!3!}} = 0.013$$

The mean and standard deviation of the hypergeometric distribution are:

$$\text{mean} = \frac{(n)(k)}{N}$$

$$\text{sd} = \sqrt{\frac{(n)(k)(N - k)(N - n)}{N^2(N - 1)}}$$

# Base Rates

by David M. Lane

## *Prerequisites*

- Chapter 5: Basic Concepts

## *Learning Objectives*

1. Compute the probability of a condition from hits, false alarms, and base rates using a tree diagram
2. Compute the probability of a condition from hits, false alarms, and base rates using Bayes' Theorem

Suppose that at your regular physical exam you test positive for Disease X. Although Disease X has only mild symptoms, you are concerned and ask your doctor about the accuracy of the test. It turns out that the test is 95% accurate. It would appear that the probability that you have Disease X is therefore 0.95. However, the situation is not that simple.

For one thing, more information about the accuracy of the test is needed because there are two kinds of errors the test can make: misses and false positives. If you actually have Disease X and the test failed to detect it, that would be a miss. If you did not have Disease X and the test indicated you did, that would be a false positive. The miss and false positive rates are not necessarily the same. For example, suppose that the test accurately indicates the disease in 99% of the people who have it and accurately indicates no disease in 91% of the people who do not have it. In other words, the test has a miss rate of 0.01 and a false positive rate of 0.09. This might lead you to revise your judgment and conclude that your chance of having the disease is 0.91. This would not be correct since the probability depends on the proportion of people having the disease. This proportion is called the base rate.

Assume that Disease X is a rare disease, and only 2% of people in your situation have it. How does that affect the probability that you have it? Or, more generally, what is the probability that someone who tests positive actually has the disease? Let's consider what would happen if one million people were tested. Out of these one million people, 2% or 20,000 people would have the disease. Of these 20,000 with the disease, the test would accurately detect it in 99% of them. This means that 19,800 cases would be accurately identified. Now let's consider the

98% of the one million people (980,000) who do not have the disease. Since the false positive rate is 0.09, 9% of these 980,000 people will test positive for the disease. This is a total of 88,200 people incorrectly diagnosed.

To sum up, 19,800 people who tested positive would actually have the disease and 88,200 people who tested positive would not have the disease. This means that of all those who tested positive, only

$$19,800 / (19,800 + 88,200) = 0.1833$$

of them would actually have the disease. So the probability that you have the disease is not 0.95, or 0.91, but only 0.1833.

These results are summarized in Table 1. The numbers of people diagnosed with the disease are shown in red. Of the one million people tested, the test was correct for 891,800 of those without the disease and for 19,800 with the disease; the test was correct 91% of the time. However, if you look only at the people testing positive (shown in red), only 19,800 (0.1833) of the  $88,200 + 19,800 = 108,000$  testing positive actually have the disease.

Table 1. Diagnosing Disease X.

True Condition			
		Test Result	
		Positive	Negative
No Disease	980,000	Disease	20,000
Positive	88,200	Negative	19,800
	891,800		200

## Bayes' Theorem

This same result can be obtained using Bayes' theorem. Bayes' theorem considers both the prior probability of an event and the diagnostic value of a test to determine the posterior probability of the event. For the current example, the event is that you have Disease X. Let's call this Event D. Since only 2% of people in your situation have Disease X, the prior probability of Event D is 0.02. Or, more formally,  $P(D) = 0.02$ . If  $D'$  represents the probability that Event D is false, then  $P(D') = 1 - P(D) = 0.98$ .

To define the diagnostic value of the test, we need to define another event: that you test positive for Disease X. Let's call this Event T. The diagnostic value of the test depends on the probability you will test positive given that you actually have the disease, written as  $P(T|D)$ , and the probability you test positive given that you do not have the disease, written as  $P(T|D')$ . Bayes' theorem shown below allows you to calculate  $P(D|T)$ , the probability that you have the disease given that you test positive for it.

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')}$$

The various terms are:

$$\begin{aligned} P(T|D) &= 0.99 \\ P(T|D') &= 0.09 \\ P(D) &= 0.02 \\ P(D') &= 0.98 \end{aligned}$$

Therefore,

$$P(D|T) = \frac{(0.99)(0.02)}{(0.99)(0.02) + (0.09)(0.98)} = 0.1833$$

which is the same value computed previously.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 5: Base Rates

This [webpage](#) gives the FBI list of warning signs for school shooters.

## **What do you think?**

Do you think it is likely that someone showing a majority of these signs would actually shoot people in school?

Fortunately the vast majority of students do not become shooters. It is necessary to take this base rate information into account in order to compute the probability that any given student will be a shooter. The warning signs are unlikely to be sufficiently predictive to warrant the conclusion that a student will become a shooter. If an action is taken on the basis of these warning signs, it is likely that the student involved would never have become a shooter even without the action.

## Exercises

### *Prerequisites*

- All material presented in the Probability Chapter

1. (a) What is the probability of rolling a pair of dice and obtaining a total score of 9 or more? (b) What is the probability of rolling a pair of dice and obtaining a total score of 7?
2. A box contains four black pieces of cloth, two striped pieces, and six dotted pieces. A piece is selected randomly and then placed back in the box. A second piece is selected randomly. What is the probability that:
  - a. both pieces are dotted?
  - b. the first piece is black and the second piece is dotted?
  - c. one piece is black and one piece is striped?
3. A card is drawn at random from a deck. (a) What is the probability that it is an ace or a king? (b) What is the probability that it is either a red card or a black card?
4. The probability that you will win a game is 0.45. (a) If you play the game 80 times, what is the most likely number of wins? (b) What are the mean and variance of a binomial distribution with  $p = 0.45$  and  $N = 80$ ?
5. A fair coin is flipped 9 times. What is the probability of getting exactly 6 heads?
6. When Susan and Jessica play a card game, Susan wins 60% of the time. If they play 9 games, what is the probability that Jessica will have won more games than Susan?
7. You flip a coin three times. (a) What is the probability of getting heads on only one of your flips? (b) What is the probability of getting heads on at least one flip?
8. A test correctly identifies a disease in 95% of people who have it. It correctly identifies no disease in 94% of people who do not have it. In the population, 3% of the people have the disease. What is the probability that you have the disease if you tested positive?

9. A jar contains 10 blue marbles, 5 red marbles, 4 green marbles, and 1 yellow marble. Two marbles are chosen (without replacement). (a) What is the probability that one will be green and the other red? (b) What is the probability that one will be blue and the other yellow?
10. You roll a fair die five times, and you get a 6 each time. What is the probability that you get a 6 on the next roll?
11. You win a game if you roll a die and get a 2 or a 5. You play this game 60 times.
- What is the probability that you win between 5 and 10 times (inclusive)?
  - What is the probability that you will win the game at least 15 times?
  - What is the probability that you will win the game at least 40 times?
  - What is the most likely number of wins.
  - What is the probability of obtaining the number of wins in d?
- Explain how you got each answer or show your work.
12. In a baseball game, Tommy gets a hit 30% of the time when facing this pitcher. Joey gets a hit 25% of the time. They are both coming up to bat this inning.
- What is the probability that Joey or Tommy will get a hit?
  - What is the probability that neither player gets a hit?
  - What is the probability that they both get a hit?
13. An unfair coin has a probability of coming up heads of 0.65. The coin is flipped 50 times. What is the probability it will come up heads 25 or fewer times? (Give answer to at least 3 decimal places).
14. You draw two cards from a deck, what is the probability that:
- both of them are face cards (king, queen, or jack)?
  - you draw two cards from a deck and both of them are hearts?
15. True/False: You are more likely to get a pattern of HTHHHTHTTH than HHHHHHHHTT when you flip a coin 10 times.

16. True/False: Suppose that at your regular physical exam you test positive for a relatively rare disease. You will need to start taking medicine if you have the disease, so you ask your doctor about the accuracy of the test. It turns out that the test is 98% accurate. The probability that you have Disease X is therefore 0.98 and the probability that you do not have it is .02. Explain your answer.

*Questions from Case Studies*

Diet and Health (DH) case study

17. (DH)

- a. What percentage of people on the AHA diet had some sort of illness or death?
- b. What is the probability that if you randomly selected a person on the AHA diet, he or she would have some sort of illness or death?
- c. If 3 people on the AHA diet are chosen at random, what is the probability that they will all be healthy?

18. (DH)

- a. What percentage of people on the Mediterranean diet had some sort of illness or death?
- b. What is the probability that if you randomly selected a person on the Mediterranean diet, he or she would have some sort of illness or death?
- c. What is the probability that if you randomly selected a person on the Mediterranean diet, he or she would have cancer?
- d. If you randomly select five people from the Mediterranean diet, what is the probability that they would all be healthy?

The following questions are from ARTIST (reproduced with permission)



19. Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times. Which of the following results is more likely?

  - Black side up on five of the rolls; white side up on the other roll
  - Black side up on all six rolls
  - a and b are equally likely

20. One of the items on the student survey for an introductory statistics course was “Rate your intelligence on a scale of 1 to 10.” The distribution of this variable for the 100 women in the class is presented below. What is the probability of randomly selecting a woman from the class who has an intelligence rating that is LESS than seven (7)?

Intelligence Rating	Count
5	12
6	24
7	38
8	23
9	2
10	1

  - $(12 + 24)/100 = .36$
  - $(12 + 24 + 38)/100 = .74$
  - $38/100 = .38$
  - $(23 + 2 + 1)/100 = .26$
  - None of the above.

21. You roll 2 fair six-sided dice. Which of the following outcomes is most likely to occur on the next roll? A. Getting double 3. B. Getting a 3 and a 4. C. They are equally likely. Explain your choice.

22. If Tahnee flips a coin 10 times, and records the results (Heads or Tails), which outcome below is more likely to occur, A or B? Explain your choice.

Throw Number	1	2	3	4	5	6	7	8	9	10
A	H	T	T	H	T	H	H	T	T	T
B	H	T	H	T	H	T	H	T	H	T

23. A bowl has 100 wrapped hard candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Jenny pulls out a handful of 10 candies, counts the number of reds, and tells her teacher. The teacher writes the number of red candies on a list. Then, Jenny puts the candies back into the bowl, and mixes them all up again. Four of Jenny's classmates, Jack, Julie, Jason, and Jerry do the same thing. They each pick ten candies, count the reds, and the teacher writes down the number of reds. Then they put the candies back and mix them up again each time. The teacher's list for the number of reds is most likely to be (please select one):
- a. 8,9,7,10,9
  - b. 3,7,5,8,5
  - c. 5,5,5,5,5
  - d. 2,4,3,4,3
  - e. 3,0,9,2,8
24. An insurance company writes policies for a large number of newly-licensed drivers each year. Suppose 40% of these are low-risk drivers, 40% are moderate risk, and 20% are high risk. The company has no way to know which group any individual driver falls in when it writes the policies. None of the low-risk drivers will have an at-fault accident in the next year, but 10% of the moderate-risk and 20% of the high-risk drivers will have such an accident. If a driver has an at-fault accident in the next year, what is the probability that he or she is high-risk?
25. You are to participate in an exam for which you had no chance to study, and for that reason cannot do anything but guess for each question (all questions being of the multiple choice type, so the chance of guessing the correct answer for each question is  $1/d$ ,  $d$  being the number of options (distractors) per question;

so in case of a 4-choice question, your guess chance is 0.25). Your instructor offers you the opportunity to choose amongst the following exam formats: I. 6 questions of the 4-choice type; you pass when 5 or more answers are correct; II. 5 questions of the 5-choice type; you pass when 4 or more answers are correct; III. 4 questions of the 10-choice type; you pass when 3 or more answers are correct. Rank the three exam formats according to their attractiveness. It should be clear that the format with the highest probability to pass is the most attractive format. Which would you choose and why?

26. Consider the question of whether the home team wins more than half of its games in the National Basketball Association. Suppose that you study a simple random sample of 80 professional basketball games and find that 52 of them are won by the home team.
  - a. Assuming that there is no home court advantage and that the home team therefore wins 50% of its games in the long run, determine the probability that the home team would win 65% or more of its games in a simple random sample of 80 games.
  - b. Does the sample information (that 52 of a random sample of 80 games are won by the home team) provide strong evidence that the home team wins more than half of its games in the long run? Explain.
27. A refrigerator contains 6 apples, 5 oranges, 10 bananas, 3 pears, 7 peaches, 11 plums, and 2 mangos.
  - a. Imagine you stick your hand in this refrigerator and pull out a piece of fruit at random. What is the probability that you will pull out a pear?
  - b. Imagine now that you put your hand in the refrigerator and pull out a piece of fruit. You decide you do not want to eat that fruit so you put it back into the refrigerator and pull out another piece of fruit. What is the probability that the first piece of fruit you pull out is a banana and the second piece you pull out is an apple?
  - c. What is the probability that you stick your hand in the refrigerator one time and pull out a mango or an orange?

# 6. Research Design

- A. Scientific Method
- B. Measurement
- C. Basics of Data Collection
- D. Sampling Bias
- E. Experimental Designs
- F. Causation
- G. Exercises

# Scientific Method

by David M. Lane

## *Prerequisites*

- none

This section contains a brief discussion of the most important principles of the scientific method. A thorough treatment of the philosophy of science is beyond the scope of this work.

One of the hallmarks of the scientific method is that it depends on empirical data. To be a proper scientific investigation, the data must be collected systematically. However, scientific investigation does not necessarily require experimentation in the sense of manipulating variables and observing the results. Observational studies in the fields of astronomy, developmental psychology, and ethology are common and provide valuable scientific information.

Theories and explanations are very important in science. Theories in science can never be proved since one can never be 100% certain that a new empirical finding inconsistent with the theory will never be found.

Scientific theories must be potentially disconfirmable. If a theory can accommodate all possible results then it is not a scientific theory. Therefore, a scientific theory should lead to testable hypotheses. If a hypothesis is disconfirmed, then the theory from which the hypothesis was deduced is incorrect. For example, the secondary reinforcement theory of attachment states that an infant becomes attached to its parent by means of a pairing of the parent with a primary reinforcer (food). It is through this “secondary reinforcement” that the child-parent bond forms. The secondary reinforcement theory has been disconfirmed by numerous experiments. Perhaps the most notable is one in which infant monkeys were fed by a surrogate wire mother while a surrogate cloth mother was available. The infant monkeys formed no attachment to the wire monkeys and frequently clung to the cloth surrogate mothers (Harlow, 1958).

If a hypothesis derived from a theory is confirmed, then the theory has survived a test and it becomes more useful and better thought of by the researchers in the field. A theory is not confirmed when correct hypotheses are derived from it.

A key difference between scientific explanations and faith-based explanations is simply that faith-based explanations are based on faith and do not

need to be testable. This does not mean that an explanation that cannot be tested is incorrect in some cosmic sense. It just means that it is not a scientific explanation.

The method of investigation in which a hypothesis is developed from a theory and then confirmed or disconfirmed involves deductive reasoning. However, deductive reasoning does not explain where the theory came from in the first place. In general, a theory is developed by a scientist who is aware of many empirical findings on a topic of interest. Then, through a generally poorly understood process called “induction,” the scientist develops a way to explain all or most of the findings within a relatively simple framework or theory.

An important attribute of a good scientific theory is that it is parsimonious. That is, that it is simple in the sense that it uses relatively few constructs to explain many empirical findings. A theory that is so complex that it has as many assumptions as it has predictions is not very valuable.

Although strictly speaking, disconfirming an hypothesis deduced from a theory disconfirms the theory, it rarely leads to the abandonment of the theory. Instead, the theory will probably be modified to accommodate the inconsistent finding. If the theory has to be modified over and over to accommodate new findings, the theory generally becomes less and less parsimonious. This can lead to discontent with the theory and the search for a new theory. If a new theory is developed that can explain the same facts in a more parsimonious way, then the new theory will eventually supersede the old theory.

# Measurement

by David M. Lane

## *Prerequisites*

- Values of Pearson's Correlation
- Variance Sum Law
- Chapter 3: Measures of Variability

## *Learning Objectives*

1. Define reliability
2. Describe reliability in terms of true scores and error
3. Compute reliability from the true score and error variance
4. Define the standard error of measurement and state why it is valuable
5. State the effect of test length on reliability
6. Distinguish between reliability and validity
7. Define three types of validity
8. State the how reliability determines the upper limit to validity

The measurement of psychological attributes such as self-esteem can be complex. A good measurement scale should be both reliable and valid. These concepts will be discussed in turn.

## **Reliability**

The notion of reliability revolves around whether you would get at least approximately the same result if you measure something twice with the same measurement instrument. A common way to define reliability is the correlation between parallel forms of a test. Letting “test” represent a parallel form of the test, the symbol  $r_{\text{test,test}}$  is used to denote the reliability of the test.

## **True Scores and Error**

Assume you wish to measure a person's mean response time to the onset of a stimulus. For simplicity, assume that there is no learning over tests which, of

course, is not really true. The person is given 1,000 trials on the task and you obtain the response time on each trial.

The mean response time over the 1,000 trials can be thought of as the person's "true" score, or at least a very good approximation of it. Theoretically, the true score is the mean that would be approached as the number of trials increases indefinitely.

An individual response time can be thought of as being composed of two parts: the true score and the error of measurement. Thus if the person's true score were 345 and their response on one of the trials were 358, then the error of measurement would be 13. Similarly, if the response time were 340, the error of measurement would be -5.

Now consider the more realistic example of a class of students taking a 100-point true/false exam. Let's assume that each student knows the answer to some of the questions and has no idea about the other questions. For the sake of simplicity, we are assuming there is no partial knowledge of any of the answers and for a given question a student either knows the answer or guesses. Finally, assume the test is scored such that a student receives one point for a correct answer and loses a point for an incorrect answer. In this example, a student's true score is the number of questions they know the answer to and their error score is their score on the questions they guessed on. For example, assume a student knew 90 of the answers and guessed correctly on 7 of the remaining 10 (and therefore incorrectly on 3). Their true score would be 90 since that is the number of answers they knew. Their error score would be  $7 - 3 = 4$  and therefore their actual test score would be  $90 + 4$ .

Every test score can be thought of as the sum of two independent components, the true score and the error score. This can be written as:

$$y_{test} = y_{true} + y_{error}$$

The following expression follows directly from the Variance Sum Law:

$$\sigma_{Test}^2 = \sigma_{True}^2 + \sigma_{error}^2$$

### **Reliability in Terms of True Scores and Error**

It can be shown that the reliability of a test,  $r_{test,test}$ , is the ratio of true-score variance to test-score variance. This can be written as:

$$r_{test,test} = \frac{\sigma_{True}^2}{\sigma_{Test}^2} = \frac{\sigma_{True}^2}{\sigma_{True}^2 + \sigma_{error}^2}$$

It is important to understand the implications of the role the variance of true scores plays in the definition of reliability: If a test were given in two populations for which the variance of the true scores differed, the reliability of the test would be higher in the population with the higher true-score variance. Therefore, reliability is not a property of a test per se but the reliability of a test in a given population.

## Assessing Error of Measurement

The reliability of a test does not show directly how close the test scores are to the true scores. That is, it does not reveal how much a person's test score would vary across parallel forms of the test. By definition, the mean over a large number of parallel tests would be the true score. The standard deviation of a person's test scores would indicate how much the test scores vary from the true score. This standard deviation is called the standard error of measurement. In practice, it is not practical to give a test over and over to the same person and/or assume that there are no practice effects. Instead, the following formula is used to estimate the standard error of measurement.

$$S_{measurement} = S_{test} \sqrt{1 - r_{test,test}}$$

where  $s_{measurement}$  is the standard error of measurement,  $s_{test}$  is the standard deviation of the test scores, and  $r_{test,test}$  is the reliability of the test. Taking the extremes, if the reliability is 0, then the standard error of measurement is equal to the standard deviation of the test; if the reliability is perfect (1.0) then the standard error of measurement is 0.

## Increasing Reliability

It is important to make measures as reliable as is practically possible. Suppose an investigator is studying the relationship between spatial ability and a set of other variables. The higher the reliability of the test of spatial ability, the higher the correlations will be. Similarly, if an experimenter seeks to determine whether a particular exercise regimen decreases blood pressure, the higher the reliability of

the measure of blood pressure, the more sensitive the experiment. More precisely, the higher the reliability the higher the power of the experiment. Power is covered in detail in Chapter 13. Finally, if a test is being used to select students for college admission or employees for jobs, the higher the reliability of the test the stronger will be the relationship to the criterion.

Two basic ways of increasing reliability are (1) to improve the quality of the items and (2) to increase the number of items. Items that are either too easy so that almost everyone gets them correct or too difficult so that almost no one gets them correct are not good items: they provide very little information. In most contexts, items which about half the people get correct are the best (other things being equal).

Items that do not correlate with other items can usually be improved. Sometimes the item is confusing or ambiguous.

Increasing the number of items increases reliability in the manner shown by the following formula:

$$r_{new,new} = \frac{kr_{test,test}}{1 + (k - 1)r_{test,test}}$$

where  $k$  is the factor by which the test length is increased,  $r_{new,new}$  is the reliability of the new longer test, and  $r_{test,test}$  is the current reliability. For example, if a test with 50 items has a reliability of .70 then the reliability of a test that is 1.5 times longer (75 items) would be calculated as follows

$$r_{new,new} = \frac{(1.5)(0.70)}{1 + (1.5 - 1)(0.70)}$$

which equals 0.78. Thus increasing the number of items from 50 to 75 would increase the reliability from 0.70 to 0.78.

It is important to note that this formula assumes the new items have the same characteristics as the old items. Obviously adding poor items would not increase the reliability as expected and might even decrease the reliability.

## **Validity**

The validity of a test refers to whether the test measures what it is supposed to measure. The three most common types of validity are face validity, empirical validity, and construct validity. We consider these types of validity below.

### **Face Validity**

A test's face validity refers to whether the test appears to measure what it is supposed to measure. That is, does the test "on its face" appear to measure what it is supposed to be measuring. An Asian history test consisting of a series of questions about Asian history would have high face validity. If the test included primarily questions about American history then it would have little or no face validity as a test of Asian history.

### **Predictive Validity**

Predictive validity (sometimes called empirical validity) refers to a test's ability to predict a relevant behavior. For example, the main way in which SAT tests are validated is by their ability to predict college grades. Thus, to the extent these tests are successful at predicting college grades they are said to possess predictive validity.

### **Construct Validity**

Construct validity is more difficult to define. In general, a test has construct validity if its pattern of correlations with other measures is in line with the construct it is purporting to measure. Construct validity can be established by showing a test has both convergent and divergent validity. A test has convergent validity if it correlates with other tests that are also measures of the construct in question. Divergent validity is established by showing the test does not correlate highly with tests of other constructs. Of course, some constructs may overlap so the establishment of convergent and divergent validity can be complex.

To take an example, suppose one wished to establish the construct validity of a new test of spatial ability. Convergent and divergent validity could be established by showing the test correlates relatively highly with other measures of spatial ability but less highly with tests of verbal ability or social intelligence.

## **Reliability and Predictive Validity**

The reliability of a test limits the size of the correlation between the test and other measures. In general, the correlation of a test with another measure will be lower than the test's reliability. After all, how could a test correlate with something else as high as it correlates with a parallel form of itself? Theoretically it is possible for a test to correlate as high as the square root of the reliability with another measure. For example, if a test has a reliability of 0.81 then it could correlate as high as 0.90 with another measure. This could happen if the other measure were a perfectly reliable test of the same construct as the test in question. In practice, this is very unlikely.

A correlation above the upper limit set by reliabilities can act as a red flag. For example, Vul, Harris, Winkielman, and Paschler (2009) found that in many studies the correlations between various fMRI activation patterns and personality measures were higher than their reliabilities would allow. A careful examination of these studies revealed serious flaws in the way the data were analyzed.

# Basics of Data Collection

by Heidi Zeimer

## *Prerequisites*

- None

## *Learning Objectives*

1. Describe how a variable such as height should be recorded
2. Choose a good response scale for a questionnaire

Most statistical analyses require that your data be in numerical rather than verbal form (you can't punch letters into your calculator). Therefore, data collected in verbal form must be coded so that it is represented by numbers. To illustrate, consider the data in Table 1.

Table 1. Example Data

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5'4"	Lots
Amber	Blonde	Female	Social Science	5'7"	Very little
Paul	Blonde	Male	History	6'1"	Moderate
Christopher	Black	Male	Biology	5'10"	Lots
Sonya	Brown	Female	Psychology	5'4"	Little

Can you conduct statistical analyses on the above data or must you re-code it in some way? For example, how would you go about computing the average height of the 5 students. You cannot enter students' heights in their current form into a statistical program -- the computer would probably give you an error message because it does not understand notation such as 5'4". One solution is to change all the numbers to inches. So, 5'4" becomes  $(5 \times 12) + 4 = 64$ , and 6'1" becomes  $(6 \times 12) + 1 = 73$ , and so forth. In this way, you are converting height in feet and inches to simply height in inches. From there, it is very easy to ask a statistical program to calculate the mean height in inches for the 5 students.

You may ask, “Why not simply ask subjects to write their height in inches in the first place?” Well, the number one rule of data collection is to ask for information in such a way as it will be most accurately reported. Most people know their height in feet and inches and cannot quickly and accurately convert it into inches “on the fly.” So, in order to preserve data accuracy, it is best for researchers to make the necessary conversions.

Let’s take another example. Suppose you wanted to calculate the mean amount of computer experience for the five students shown in Table 1. One way would be to convert the verbal descriptions to numbers as shown in Table 2. Thus, “Very Little” would be converted to “1” and “Little” would be converted to “2.”

Table 2. Conversion of verbal descriptions to numbers

1	2	3	4	5
Very Little	Little	Moderate	Lots	Very Lots

## Measurement Examples

### Example #1: How much information should I record?

Say you are volunteering at a track meet at your college, and your job is to record each runner’s time as they pass the finish line for each race. Their times are shown in large red numbers on a digital clock with eight digits to the right of the decimal point, and you are told to record the entire number in your tablet. Thinking eight decimal places is a bit excessive, you only record runners’ times to one decimal place. The track meet begins, and runner number one finishes with a time of 22.93219780 seconds. You dutifully record her time in your tablet, but only to one decimal place, that is 22.9. Race number two finishes and you record 32.7 for the winning runner. The fastest time in Race number three is 25.6. Race number four winning time is 22.9, Race number five is.... But wait! You suddenly realize your mistake; you now have a tie between runner one and runner four for the title of Fastest Overall Runner! You should have recorded more information from the digital clock -- that information is now lost, and you cannot go back in time and record running times to more decimal places.

The point is that you should think very carefully about the scales and specificity of information needed in your research before you begin collecting data. If you believe you might need additional information later but are not sure,

measure it; you can always decide to not use some of the data, or “collapse” your data down to lower scales if you wish, but you cannot expand your data set to include more information after the fact. In this example, you probably would not need to record eight digits to the right of the decimal point. But recording only one decimal digit is clearly too few.

## Example #2

Pretend for a moment that you are teaching five children in middle school (yikes!), and you are trying to convince them that they must study more in order to earn better grades. To prove your point, you decide to collect actual data from their recent math exams, and, toward this end, you develop a questionnaire to measure their study time and subsequent grades. You might develop a questionnaire which looks like the following:

1. Please write your name: \_\_\_\_\_
2. Please indicate how much you studied for this math exam:  
a lot.....moderate.....little
3. Please circle the grade you received on the math exam:  
A B C D F

Given the above questionnaire, your obtained data might look like the following:

Name	Amount Studied	Grade
John	Little	C
Sally	Moderate	B
Alexander	Lots	A
Linda	Moderate	A
Thomas	Little	B

Eyeballing the data, it seems as if the children who studied more received better grades, but it’s difficult to tell. “Little,” “lots,” and “B,” are imprecise, qualitative terms. You could get more precise information by asking specifically how many hours they studied and their exact score on the exam. The data then might look as follows:

Name	Hours studied	% Correct

John	5	71
Sally	9	83
Alexander	13	97
Linda	12	91
Thomas	7	85

Of course, this assumes the students would know how many hours they studied. Rather than trust the students' memories, you might ask them to keep a log of their study time as they study.

# Sampling Bias

by David M. Lane

## *Prerequisites*

- Inferential Statistics (including sampling)

## *Learning Objectives*

1. Recognize sampling bias
2. Distinguish among self-selection bias, undercoverage bias, and survivorship bias

Descriptions of various types of sampling such as *simple random sampling* and *stratified random sampling* are covered in the inferential statistics section of Chapter 1. This section discusses various types of sampling biases including self-selection bias and survivorship bias. Examples of other sampling biases that are not easily categorized will also be given.

It is important to keep in mind that sampling bias refers to the method of sampling, not the sample itself. There is no guarantee that random sampling will result in a sample representative of the population just as not every sample obtained using a biased sampling method will be greatly non-representative of the population.

## **Self-Selection Bias**

Imagine that a university newspaper ran an ad asking for students to volunteer for a study in which intimate details of their sex lives would be discussed. Clearly the sample of students who would volunteer for such a study would not be representative of the students at the university. Similarly, an online survey about computer use is likely to attract people more interested in technology than is typical. In both of these examples, people who “self-select” themselves for the experiment are likely to differ in important ways from the population the experimenter wishes to draw conclusions about. Many of the admittedly “non-scientific” polls taken on television or web sites suffer greatly from self-selection bias.

A self-selection bias can result when the non-random component occurs after the potential subject has enlisted in the experiment. Considering again the hypothetical experiment in which subjects are to be asked intimate details of their sex lives, assume that the subjects did not know what the experiment was going to be about until they showed up. Many of the subjects would then likely leave the experiment resulting in a biased sample.

## **Undercoverage Bias**

A common type of sampling bias is to sample too few observations from a segment of the population. A commonly-cited example of undercoverage is the poll taken by the Literary Digest in 1936 that indicated that Landon would win an election against Roosevelt by a large margin when, in fact, it was Roosevelt who won by a large margin. A common explanation is that poorer people were undercovered because they were less likely to have telephones and that this group was more likely to support Roosevelt.

A detailed analysis by Squire (1988) showed that it was not just an undercoverage bias that resulted in the faulty prediction of the election results. He concluded that, in addition to the undercoverage described above, there was a nonresponse bias (a form of self-selection bias) such that those favoring Landon were more likely to return their survey than were those favoring Roosevelt.

## **Survivorship Bias**

Survivorship bias occurs when the observations recorded at the end of the investigation are a non-random set of those present at the beginning of the investigation. Gains in stock funds is an area in which survivorship bias often plays a role. The basic problem is that poorly-performing funds are often either eliminated or merged into other funds. Suppose one considers a sample of stock funds that exist in the present and then calculates the mean 10-year appreciation of those funds. Can these results be validly generalized to other stock funds of the same type? The problem is that the poorly-performing stock funds that are not still in existence (did not survive for 10 years) are not included. Therefore, there is a bias toward selecting better-performing funds. There is good evidence that this survivorship bias is substantial (Malkiel, 1995).

In World War II, the statistician Abraham Wald analyzed the distribution of hits from anti-aircraft fire on aircraft returning from missions. The idea was that this information would be useful for deciding where to place extra armor. A naive

approach would be to put armor at locations that were frequently hit to reduce the damage there. However, this would ignore the survivorship bias occurring because only a subset of aircraft return. Wald's approach was the opposite: if there were few hits in a certain location on returning planes, then hits in that location were likely to bring a plane down. Therefore, he recommended that locations without hits on the returning planes should be given extra armor. A detailed and mathematical description of Wald's work can be found in Mangel and Samaniego (1984.)

# Experimental Designs

by David M. Lane

## *Prerequisites*

- Chapter 1: Variables

## *Learning Objectives*

1. Distinguish between between-subject and within-subject designs
2. State the advantages of within-subject designs
3. Define “multi-factor design” and “factorial design”
4. Identify the levels of a variable in an experimental design
5. Describe when counterbalancing is used

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one *independent variable* or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

## **Between-Subjects Designs**

In a *between-subjects* design, the various experimental treatments are given to different groups of subjects. For example, in the “*Teacher Ratings*” case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor’s lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the *independent variable* is “Condition” and has two levels (charismatic teacher and punitive teacher). It is a *between-subjects* variable because different subjects were used for the two levels of the independent variable: subjects were in either the “charismatic teacher” or the “punitive teacher” condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. The inferential statistics applicable to testing the difference between the means of the two conditions covered in Chapter 12. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using *inferential statistics*, it can be calculated that the probability of finding a difference as large or larger than the one obtained if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the “Smiles and Leniency” case study, the independent variable is “type of smile” and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

## **Multi-Factor Between-Subject Designs**

In the “*Bias Against Associates of the Obese*” experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as an Associate's Weight (2) x Associate's Relationship (2) *factorial design*. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an *interaction* between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

## **Within-Subjects Designs**

A within-subjects design differs from a between-subjects design in that the same subjects perform at all levels of the *independent variable*. For example consider the “ADHD Treatment” case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under each of the four levels of the independent variable “dose,” the design is a *within-subjects design* and dose is a *within-subjects variable*. Within-subjects designs are sometimes called *repeated-measures designs*.

## **Counterbalancing**

In a within-subject design it is important not to *confound* the order in which a task is performed with the experimental treatment. For example, consider the problem that would have occurred if, in the ADHD study, every subject had received the doses in the same order starting with the lowest and continuing to the highest. It is not unlikely that experience with the delay of gratification task would have an effect. If practice on this task leads to better performance, then it would appear that higher doses caused the better performance when, in fact, it was the practice that caused the better performance.

One way to address this problem is to *counterbalance* the order of presentations. In other words, subjects would be given the doses in different orders in such a way that each dose was given in each sequential position an equal number of times. An example of counterbalancing is shown in Table 1.

Table 1. Counterbalanced order for four subjects.

Subject	0 mg/kg	.15 mg/kg	.30 mg/kg	.60 mg/kg
1	First	Second	Third	Fourth
2	Second	Third	Fourth	First
3	Third	Fourth	First	Second
4	Fourth	First	Second	Third

It should be kept in mind that counterbalancing is not a satisfactory solution if there are complex dependencies between which treatment precedes which and the dependent variable. In these cases, it is usually better to use a between-subjects design than a within-subjects design.

### Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more *power* than between-subjects designs. That is, this makes within-subjects designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called “repeated-measures” designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

### Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the “*Weapons and Aggression*” case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

# Causation

by David M. Lane

## *Prerequisites*

- Chapter 1: What are Statistics
- Chapter 3: Measures of Variability
- Chapter 4: Pearson's Correlation
- Chapter 6: Experimental Designs

## *Learning Objectives*

1. Explain how experimentation allows causal inferences
2. Explain the role of unmeasured variables
3. Explain the “third-variable” problem
4. Explain how causation can be inferred in non-experimental designs

The concept of causation is a complex one in the philosophy of science. Since a full coverage of this topic is well beyond the scope of this text, we focus on two specific topics: (1) the establishment of causation in experiments and (2) the establishment of causation in non-experimental designs.

## **Establishing Causation in Experiments**

Consider a simple experiment in which subjects are *sampled randomly* from a *population* and then *assigned randomly* to either the experimental group or the control group. Assume the condition means on the *dependent variable* differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random

assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in “minutes slept” could have been due to a chance difference between the control group and the experimental group rather than due to the drug's effect.

This problem seems intractable since, by definition, it is impossible to measure an “unmeasured variable” just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the difference obtained in the experiment. If that probability is low, then it is inferred (that's why they call it *inferential statistics*) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

## **Causation in Non-Experimental Designs**

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the “*third-variable problem*” and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975) to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better though, admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010) found a strong correlation between public debt and GDP growth. Although some have argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 6: Causation

A low level of HDL have long been known to be a risk factor for heart disease. Taking niacin has been shown to increase HDL levels and has been recommended for patients with low levels of HDL. The assumption of this recommendation is that niacin causes HDL to increase thus causing a lower risk for heart disease.

## **What do you think?**

What experimental design involving niacin would test whether the relationship between HDL and heart disease is causal?

You could randomly assign patients with low levels of HDL to a condition in which they received niacin or to one in which they did not. A finding that niacin increased HDL without decreasing heart disease would cast doubt on the causal relationship. This is exactly what was found in a study conducted by the NIH. See the description of the results [here](#).

## References

- Harlow, H. (1958) The nature of love. *American Psychologist*, 13, 673-685.
- Li, C. (1975) *Path analysis: A primer*. Boxwood Press, Pacific Grove. CA .
- Malkiel, B. G. (1995) Returns from investing in equity mutual funds 1971 to 1991. *The Journal of Finance*, 50, 549-572.
- Mangel, M. & Samaniego, F. J. (1984) Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association*, 79, 259-267.
- Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, <http://www.nber.org/papers/w15639>
- Squire, P. (1988) Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52, 125-133.
- Vul, E., Harris, C., Winkielman, P., & Paschler, H. (2009) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, 4, 274-290.

## Exercises

1. To be a scientific theory, the theory must be potentially \_\_\_\_\_.
2. What is the difference between a faith-based explanation and a scientific explanation?
3. What does it mean for a theory to be parsimonious?
4. Define reliability in terms of parallel forms.
5. Define true score.
6. What is the reliability if the true score variance is 80 and the test score variance is 100?
7. What statistic relates to how close a score on one test will be to a score on a parallel form?
8. What is the effect of test length on the reliability of a test?
9. Distinguish between predictive validity and construct validity.
10. What is the theoretical maximum correlation of a test with a criterion if the test has a reliability of .81?
11. An experiment solicits subjects to participate in a highly stressful experiment. What type of sampling bias is likely to occur?
12. Give an example of survivorship bias not presented in this text.
13. Distinguish “between-subject” variables from “within-subjects” variables.
14. Of the variables “gender” and “trials,” which is likely to be a between-subjects variable and which a within-subjects variable?
15. Define interaction.
16. What is counterbalancing used for?
17. How does randomization deal with the problem of pre-existing differences between groups?
18. Give an example of the “third-variable problem” other than those in this text.

# 7. Normal Distributions

- A. Introduction
- B. History
- C. Areas of Normal Distributions
- D. Standard Normal
- E. Exercises

Most of the statistical analyses presented in this book are based on the bell-shaped or normal distribution. The introductory section defines what it means for a distribution to be normal and presents some important properties of normal distributions. The interesting history of the discovery of the normal distribution is described in the second section. Methods for calculating probabilities based on the normal distribution are described in Areas of Normal Distributions. A frequently used normal distribution is called the Standard Normal distribution and is described in the section with that name. The binomial distribution can be approximated by a normal distribution. The section Normal Approximation to the Binomial shows this approximation.

# Introduction to Normal Distributions

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Central Tendency
- Chapter 3: Variability

## *Learning Objectives*

1. Describe the shape of normal distributions
2. State 7 features of normal distributions

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the “bell curve,” although the tonal qualities of such a bell would be less than pleasing. It is also called the “Gaussian curve” after the mathematician Karl Friedrich Gauss. As you will see in the section on the history of the normal distribution, although Gauss played an important role in its history, de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about “the normal distribution” since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure 1 shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

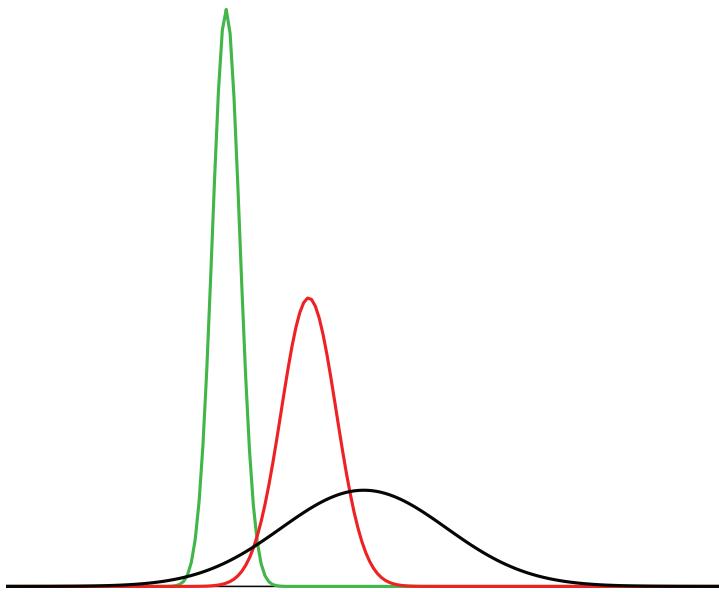


Figure 1. Normal distributions differing in mean and standard deviation.

The density of the normal distribution (the height for a given value on the x-axis) is shown below. The parameters  $\mu$  and  $\sigma$  are the mean and standard deviation, respectively, and define the normal distribution. The symbol  $e$  is the base of the natural logarithm and  $\pi$  is the constant pi.

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will not be referring back to it in later sections.

Seven features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean ( $\mu$ ) and the standard deviation ( $\sigma$ ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

# History of the Normal Distribution

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Central Tendency
- Chapter 3: Variability
- Chapter 5: Binomial Distribution

## *Learning Objectives*

1. Name the person who discovered the normal distribution and state the problem he applied it to
2. State the relationship between the normal and binomial distributions
3. State who related the normal distribution to errors
4. Describe briefly the central limit theorem
5. State who was the first to prove the central limit theorem

In the chapter on probability, we saw that the binomial distribution could be used to solve problems such as “If a fair coin is flipped 100 times, what is the probability of getting 60 or more heads?” The probability of exactly  $x$  heads out of  $N$  flips is computed using the formula:

$$P(x) = \frac{N!}{x!(N-x)!} \pi^x (1-\pi)^{N-x}$$

where  $x$  is the number of heads (60),  $N$  is the number of flips (100), and  $\pi$  is the probability of a head (0.5). Therefore, to solve this problem, you compute the probability of 60 heads, then the probability of 61 heads, 62 heads, etc., and add up all these probabilities. Imagine how long it must have taken to compute binomial probabilities before the advent of calculators and computers.

Abraham de Moivre, an 18th century statistician and consultant to gamblers, was often called upon to make these lengthy computations. de Moivre noted that when the number of events (coin flips) increased, the shape of the binomial

distribution approached a very smooth curve. Binomial distributions for 2, 4, 12, and 24 flips are shown in Figure 1.

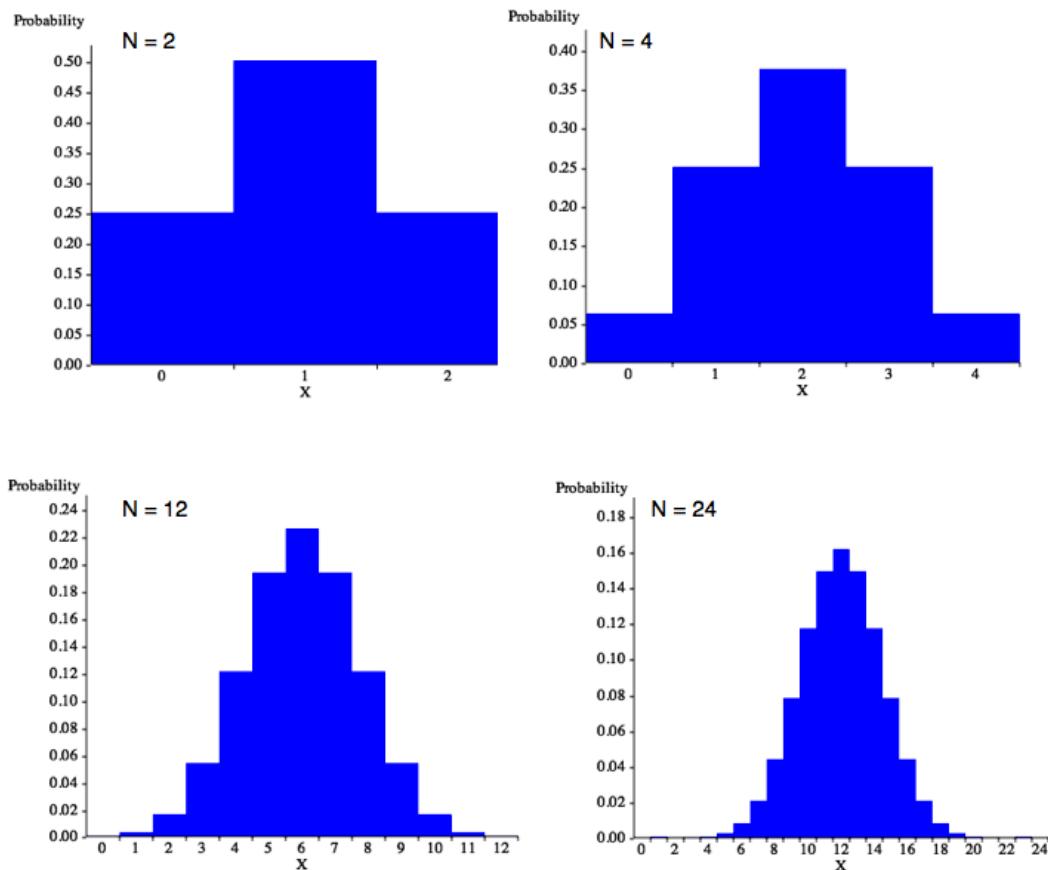


Figure 1. Examples of binomial distributions. The heights of the blue bars represent the probabilities.

de Moivre reasoned that if he could find a mathematical expression for this curve, he would be able to solve problems such as finding the probability of 60 or more heads out of 100 coin flips much more easily. This is exactly what he did, and the curve he discovered is now called the “normal curve.”

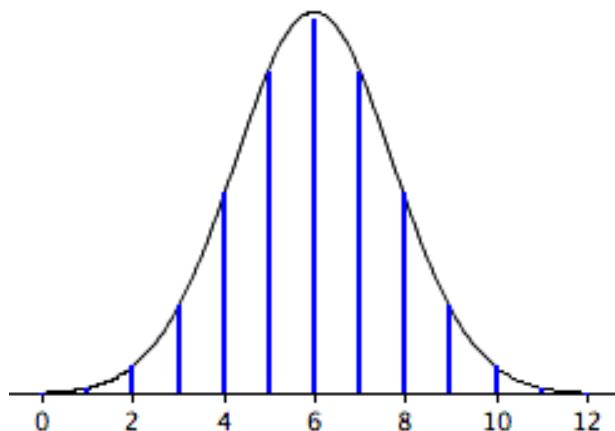


Figure 2. The normal approximation to the binomial distribution for 12 coin flips. The smooth curve is the normal distribution. Note how well it approximates the binomial probabilities represented by the heights of the blue lines.

The importance of the normal curve stems primarily from the fact that the distributions of many natural phenomena are at least approximately normally distributed. One of the first applications of the normal distribution was to the analysis of errors of measurement made in astronomical observations, errors that occurred because of imperfect instruments and imperfect observers. Galileo in the 17th century noted that these errors were symmetric and that small errors occurred more frequently than large errors. This led to several hypothesized distributions of errors, but it was not until the early 19th century that it was discovered that these errors followed a normal distribution. Independently, the mathematicians Adrain in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors were fit well by this distribution.

This same distribution had been discovered by Laplace in 1778 when he derived the extremely important central limit theorem, the topic of a later section of this chapter. Laplace showed that even if a distribution is not normally distributed, the means of repeated samples from the distribution would be very nearly normally distributed, and that the larger the sample size, the closer the distribution of means would be to a normal distribution.

Most statistical procedures for testing differences between means assume normal distributions. Because the distribution of means is very close to normal, these tests work well even if the original distribution is only roughly normal.

Quetelet was the first to apply the normal distribution to human characteristics. He noted that characteristics such as height, weight, and strength were normally distributed.

# Areas Under Normal Distributions

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 3: Central Tendency
- Chapter 3: Variability
- Chapter 7: Introduction to Normal Distributions

## *Learning Objectives*

1. State the proportion of a normal distribution within 1 standard deviation of the mean
2. State the proportion of a normal distribution that is more than 1.96 standard deviations from the mean
3. Use the normal calculator to calculate an area for a given  $X$
4. Use the normal calculator to calculate  $X$  for a given area

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. Figure 1 shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.

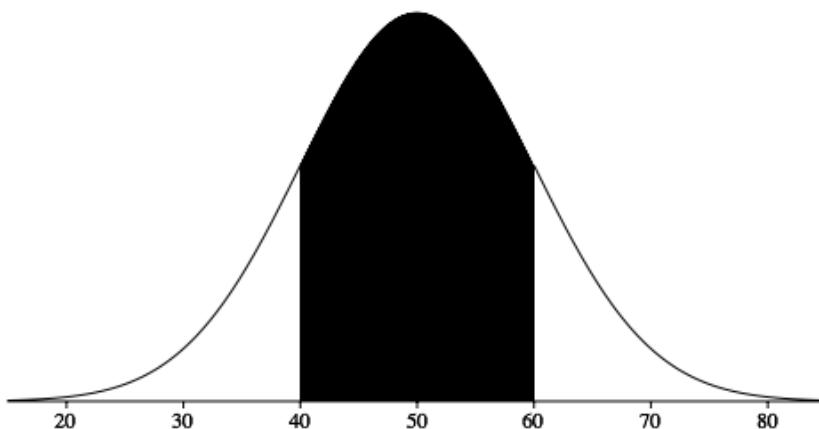


Figure 1. Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

Figure 2 shows a normal distribution with a mean of 100 and a standard deviation of 20. As in Figure 1, 68% of the distribution is within one standard deviation of the mean.

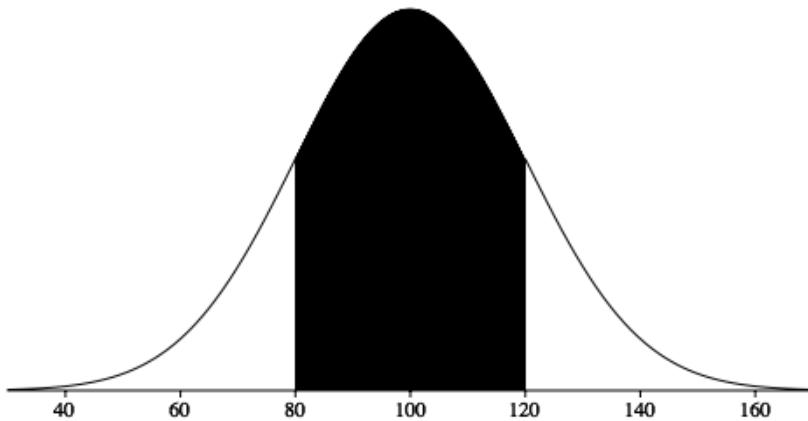


Figure 2. Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within one standard deviation (20) of the mean (100).

The normal distributions shown in Figures 1 and 2 are specific examples of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

Figure 3 shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.



Figure 3. A normal distribution with a mean of 75 and a standard deviation of 10. 95% of the area is within 1.96 standard deviations of the mean.

Areas under the normal distribution can be calculated with this [online calculator](#).

# Standard Normal Distribution

by David M. Lane

## *Prerequisites*

- Chapter 3: Effects of Linear Transformations
- Chapter 7: Introduction to Normal Distributions

## *Learning Objectives*

1. State the mean and standard deviation of the standard normal distribution
2. Use a Z table
3. Use the normal calculator
4. Transform raw data to Z scores

As discussed in the introductory section, normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution.

Areas of the normal distribution are often represented by tables of the standard normal distribution. A portion of a table of the standard normal distribution is shown in Table 1.

Table 1. A portion of a table of the standard normal distribution.

<b>Z</b>	<b>Area below</b>
-2.5	0.0062
-2.49	0.0064
-2.48	0.0066
-2.47	0.0068
-2.46	0.0069
-2.45	0.0071
-2.44	0.0073
-2.43	0.0075
-2.42	0.0078
-2.41	0.008
-2.4	0.0082

-2.39	0.0084
-2.38	0.0087
-2.37	0.0089
-2.36	0.0091
-2.35	0.0094
-2.34	0.0096
-2.33	0.0099
-2.32	0.0102

The first column titled “Z” contains values of the standard normal distribution; the second column contains the area below Z. Since the distribution has a mean of 0 and a standard deviation of 1, the Z column is equal to the number of standard deviations below (or above) the mean. For example, a Z of -2.5 represents a value 2.5 standard deviations below the mean. The area below Z is 0.0062.

The same information can be obtained using the following Java applet. Figure 1 shows how it can be used to compute the area below a value of -2.5 on the standard normal distribution. Note that the mean is set to 0 and the standard deviation is set to 1.

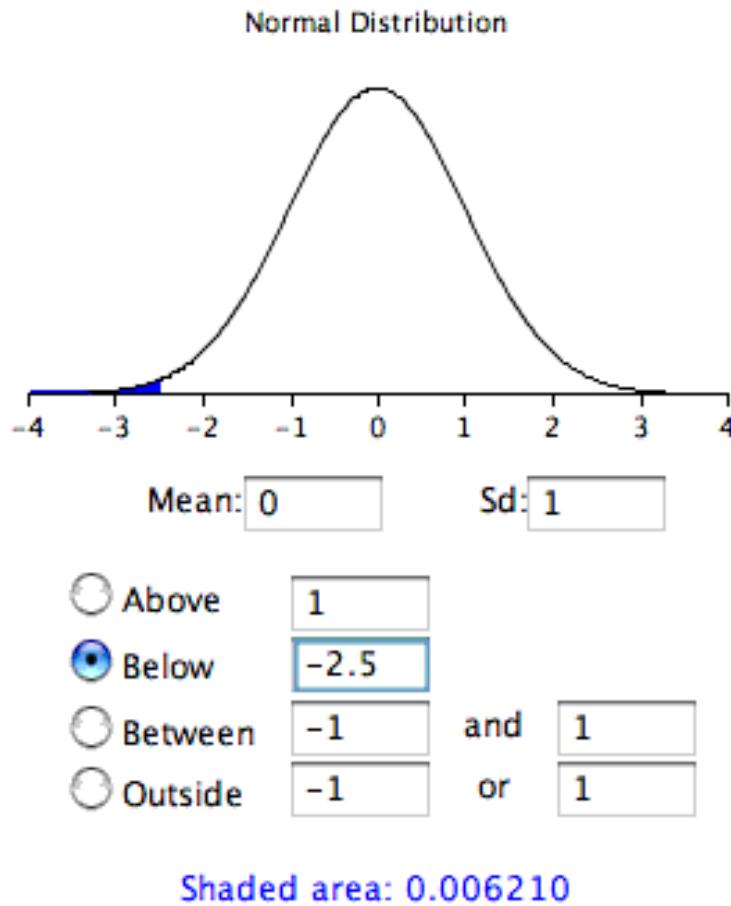


Figure 1. An example from the applet.

A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

$$Z = (X - \mu) / \sigma$$

where  $Z$  is the value on the standard normal distribution,  $X$  is the value on the original distribution,  $\mu$  is the mean of the original distribution, and  $\sigma$  is the standard deviation of the original distribution.

As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26? Applying the formula, we obtain

$$Z = (26 - 50) / 10 = -2.4.$$

From Table 1, we can see that 0.0082 of the distribution is below -2.4. There is no need to transform to  $Z$  if you use the applet as shown in Figure 2.

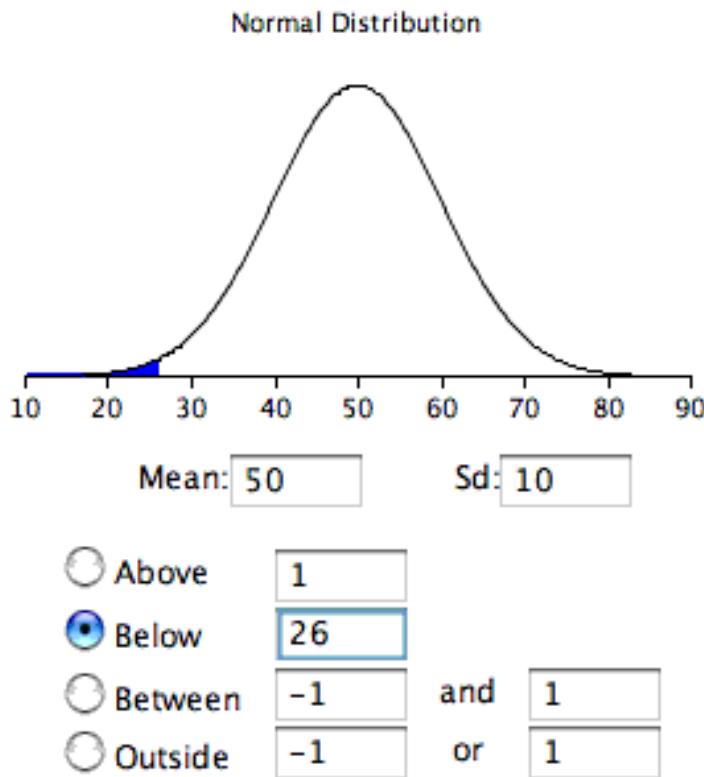


Figure 2. Area below 26 in a normal distribution with a mean of 50 and a standard deviation of 10.

If all the values in a distribution are transformed to Z scores, then the distribution will have a mean of 0 and a standard deviation of 1. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called standardizing the distribution.

# Normal Approximation to the Binomial

by David M. Lane

## *Prerequisites*

- Chapter 5: Binomial Distribution
- Chapter 7: History of the Normal Distribution
- Chapter 7: Areas of Normal Distributions

## *Learning Objectives*

1. State the relationship between the normal distribution and the binomial distribution
2. Use the normal distribution to approximate the binomial distribution
3. State when the approximation is adequate

In the section on the history of the normal distribution, we saw that the normal distribution can be used to approximate the binomial distribution. This section shows how to compute these approximations.

Let's begin with an example. Assume you have a fair coin and wish to know the probability that you would get 8 heads out of 10 flips. The binomial distribution has a mean of  $\mu = N\pi = (10)(0.5) = 5$  and a variance of  $\sigma^2 = N\pi(1-\pi) = (10)(0.5)(0.5) = 2.5$ . The standard deviation is therefore 1.5811. A total of 8 heads is  $(8 - 5)/1.5811 = 1.897$  standard deviations above the mean of the distribution. The question then is, "What is the probability of getting a value exactly 1.897 standard deviations above the mean?" You may be surprised to learn that the answer is 0: The probability of any one specific point is 0. The problem is that the binomial distribution is a discrete probability distribution, whereas the normal distribution is a continuous distribution.

The solution is to round off and consider any value from 7.5 to 8.5 to represent an outcome of 8 heads. Using this approach, we figure out the area under a normal curve from 7.5 to 8.5. The area in green in Figure 1 is an approximation of the probability of obtaining 8 heads.

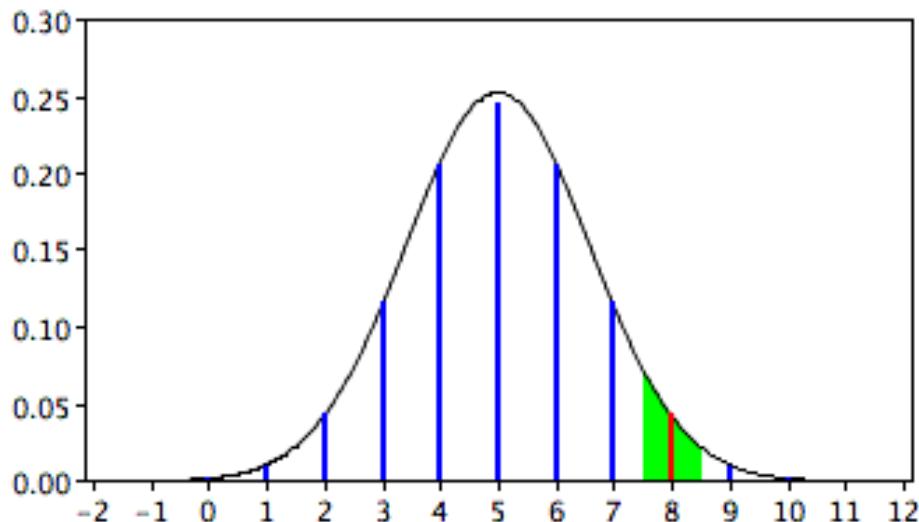


Figure 1. Approximating the probability of 8 heads with the normal distribution.

The solution is therefore to compute this area. First we compute the area below 8.5, and then subtract the area below 7.5.

The results of using the normal area calculator to find the area below 8.5 are shown in Figure 2. The results for 7.5 are shown in Figure 3.

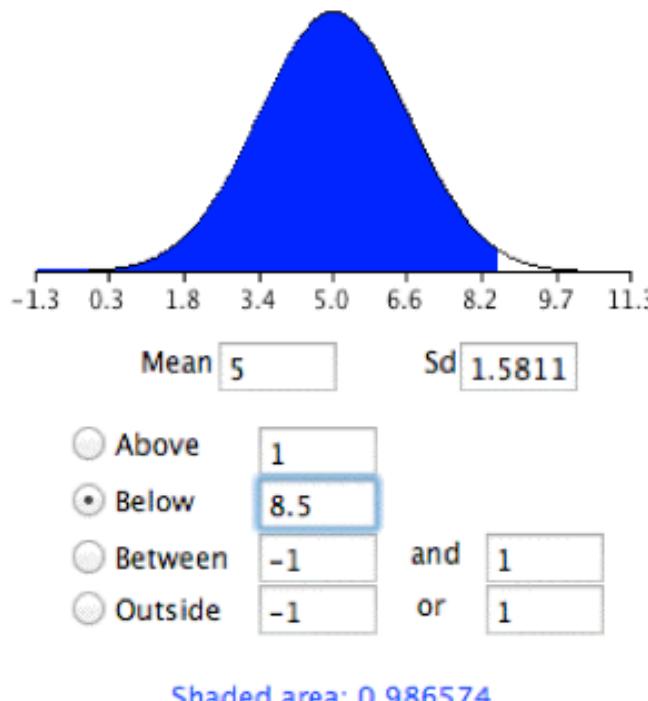


Figure 2. Area below 8.5

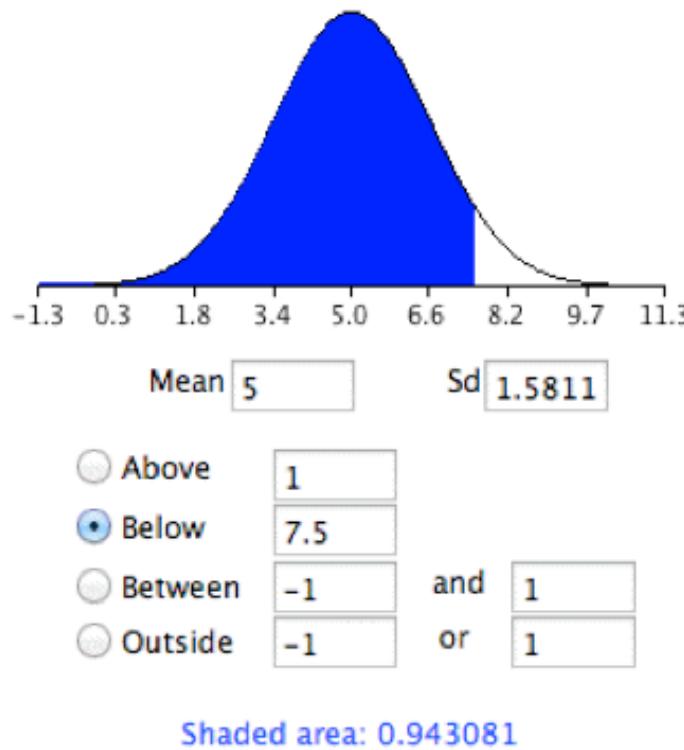


Figure 3. Area below 7.5.

The difference between the areas is 0.044, which is the approximation of the binomial probability. For these parameters, the approximation is very accurate. The demonstration in the next section allows you to explore its accuracy with different parameters.

If you did not have the normal area calculator, you could find the solution using a table of the standard normal distribution (a Z table) as follows:

1. Find a Z score for 8.5 using the formula  $Z = (8.5 - 5)/1.5811 = 2.21$ .
2. Find the area below a Z of 2.21 = 0.987.
3. Find a Z score for 7.5 using the formula  $Z = (7.5 - 5)/1.5811 = 1.58$ .
4. Find the area below a Z of 1.58 = 0.943.
5. Subtract the value in step 4 from the value in step 2 to get 0.044.

The same logic applies when calculating the probability of a range of outcomes. For example, to calculate the probability of 8 to 10 flips, calculate the area from 7.5 to 10.5.

The accuracy of the approximation depends on the values of  $N$  and  $\pi$ . A rule of thumb is that the approximation is good if both  $N\pi$  and  $N(1-\pi)$  are both greater than 10.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 7: Areas Under the Normal Distribution
- Chapter 7: Shapes of Distributions

Risk analyses often are based on the assumption of normal distributions. Critics have said that extreme events in reality are more frequent than would be expected assuming normality. The assumption has even been called a "Great Intellectual Fraud."

A [recent article](#) discussing how to protect investments against extreme events defined "tail risk" as "A tail risk, or extreme shock to financial markets, is technically defined as an investment that moves more than three standard deviations from the mean of a normal distribution of investment returns."

## **What do you think?**

Tail risk can be evaluated by assuming a normal distribution and computing the probability of such an event. Is that how "tail risk" should be evaluated?

Events more than three standard deviations from the mean are very rare for normal distributions. However, they are not as rare for other distributions such as highly-skewed distributions. If the normal distribution is used to assess the probability of tail events defined this way, then the "tail risk" will be underestimated.

## Exercises

### *Prerequisites*

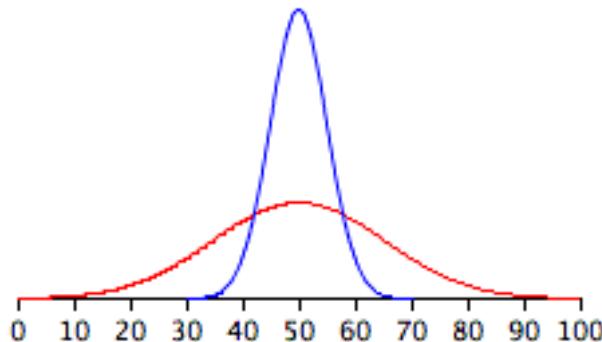
- All material presented in the Normal Distributions chapter

1. If scores are normally distributed with a mean of 35 and a standard deviation of 10, what percent of the scores is:
  - a. greater than 34?
  - b. smaller than 42?
  - c. between 28 and 34?
2. What are the mean and standard deviation of the standard normal distribution?  
(b) What would be the mean and standard deviation of a distribution created by multiplying the standard normal distribution by 8 and then adding 75?
3. The normal distribution is defined by two parameters. What are they?
4. What proportion of a normal distribution is within one standard deviation of the mean? (b) What proportion is more than 2.0 standard deviations from the mean?  
(c) What proportion is between 1.25 and 2.1 standard deviations above the mean?
5. A test is normally distributed with a mean of 70 and a standard deviation of 8.  
(a) What score would be needed to be in the 85th percentile? (b) What score would be needed to be in the 22nd percentile?
6. Assume a normal distribution with a mean of 70 and a standard deviation of 12. What limits would include the middle 65% of the cases?
7. A normal distribution has a mean of 20 and a standard deviation of 4. Find the Z scores for the following numbers: (a) 28 (b) 18 (c) 10 (d) 23
8. Assume the speed of vehicles along a stretch of I-10 has an approximately normal distribution with a mean of 71 mph and a standard deviation of 8 mph.
  - a. The current speed limit is 65 mph. What is the proportion of vehicles less than or equal to the speed limit?
  - b. What proportion of the vehicles would be going less than 50 mph?

- c. A new speed limit will be initiated such that approximately 10% of vehicles will be over the speed limit. What is the new speed limit based on this criterion?
- d. In what way do you think the actual distribution of speeds differs from a normal distribution?
9. A variable is normally distributed with a mean of 120 and a standard deviation of 5. One score is randomly sampled. What is the probability it is above 127?
10. You want to use the normal distribution to approximate the binomial distribution. Explain what you need to do to find the probability of obtaining exactly 7 heads out of 12 flips.
11. A group of students at a school takes a history test. The distribution is normal with a mean of 25, and a standard deviation of 4. (a) Everyone who scores in the top 30% of the distribution gets a certificate. What is the lowest score someone can get and still earn a certificate? (b) The top 5% of the scores get to compete in a statewide history contest. What is the lowest score someone can get and still go onto compete with the rest of the state?
12. Use the normal distribution to approximate the binomial distribution and find the probability of getting 15 to 18 heads out of 25 flips. Compare this to what you get when you calculate the probability using the binomial distribution. Write your answers out to four decimal places.
13. True/false: For any normal distribution, the mean, median, and mode will be equal.
14. True/false: In a normal distribution, 11.5% of scores are greater than  $Z = 1.2$ .
15. True/false: The percentile rank for the mean is 50% for any normal distribution.
16. True/false: The larger the  $n$ , the better the normal distribution approximates the binomial distribution.
17. True/false: A Z-score represents the number of standard deviations above or below the mean.

18. True/false: Abraham de Moivre, a consultant to gamblers, discovered the normal distribution when trying to approximate the binomial distribution to make his computations easier.

Answer questions 19 - 21 based on the graph below:



19. True/false: The standard deviation of the blue distribution shown below is about 10.
20. True/false: The red distribution has a larger standard deviation than the blue distribution.
21. True/false: The red distribution has more area underneath the curve than the blue distribution does.

### *Questions from Case Studies*

#### Angry Moods (AM) case study

22. For this problem, use the Anger Expression (AE) scores.
- Compute the mean and standard deviation.
  - Then, compute what the 25th, 50th and 75th percentiles would be if the distribution were normal.
  - Compare the estimates to the actual 25th, 50th, and 75th percentiles.

#### Physicians' Reactions (PR) case study

23. (PR) For this problem, use the time spent with the overweight patients. (a) Compute the mean and standard deviation of this distribution. (b) What is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person? (c) Now assume this distribution is normal (and has the same mean and standard deviation). Now what is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person?

The following questions are from ARTIST (reproduced with permission)



24. A set of test scores are normally distributed. Their mean is 100 and standard deviation is 20. These scores are converted to standard normal z scores. What would be the mean and median of this distribution?

- a. 0
- b. 1
- c. 50
- d. 100

25. Suppose that weights of bags of potato chips coming from a factory follow a normal distribution with mean 12.8 ounces and standard deviation .6 ounces. If the manufacturer wants to keep the mean at 12.8 ounces but adjust the standard deviation so that only 1% of the bags weigh less than 12 ounces, how small does he/she need to make that standard deviation?

26. A student received a standardized (z) score on a test that was  $-.57$ . What does this score tell about how this student scored in relation to the rest of the class? Sketch a graph of the normal curve and shade in the appropriate area.

27. Suppose you take 50 measurements on the speed of cars on Interstate 5, and that these measurements follow roughly a Normal distribution. Do you expect the standard deviation of these 50 measurements to be about 1 mph, 5 mph, 10 mph, or 20 mph? Explain.
28. Suppose that combined verbal and math SAT scores follow a normal distribution with mean 896 and standard deviation 174. Suppose further that Peter finds out that he scored in the top 3% of SAT scores. Determine how high Peter's score must have been.
29. Heights of adult women in the United States are normally distributed with a population mean of  $\mu = 63.5$  inches and a population standard deviation of  $\sigma = 2.5$ . A medical researcher is planning to select a large random sample of adult women to participate in a future study. What is the standard value, or z-value, for an adult woman who has a height of 68.5 inches?
30. An automobile manufacturer introduces a new model that averages 27 miles per gallon in the city. A person who plans to purchase one of these new cars wrote the manufacturer for the details of the tests, and found out that the standard deviation is 3 miles per gallon. Assume that in-city mileage is approximately normally distributed.
- What is the probability that the person will purchase a car that averages less than 20 miles per gallon for in-city driving?
  - What is the probability that the person will purchase a car that averages between 25 and 29 miles per gallon for in-city driving?

# 8. Advanced Graphs

- A. Q-Q Plots
- B. Contour Plots
- C. 3D Plots

# Quantile-Quantile (q-q) Plots

by David Scott

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 1: Percentiles
- Chapter 2: Histograms
- Chapter 4: Introduction to Bivariate Data
- Chapter 7: Introduction to Normal Distributions

## *Learning Objectives*

1. State what q-q plots are used for.
2. Describe the shape of a q-q plot when the distributional assumption is met.
3. Be able to create a normal q-q plot.

## Introduction

The quantile-quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line.

Before delving into the details of q-q plots, we first describe two related graphical methods for assessing distributional assumptions: the histogram and the cumulative distribution function (CDF). As will be seen, q-q plots are more general than these alternatives.

## Assessing Distributional Assumptions

As an example, consider data measured from a physical device such as the spinner depicted in Figure 1. The red arrow is spun around the center, and when the arrow stops spinning, the number between 0 and 1 is recorded. Can we determine if the spinner is fair?

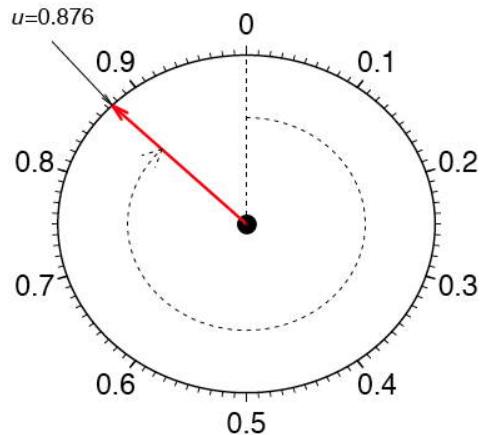


Figure 1. A physical device that gives samples from a uniform distribution.

If the spinner is fair, then these numbers should follow a uniform distribution. To investigate whether the spinner is fair, spin the arrow  $n$  times, and record the measurements by  $\{\mu_1, \mu_2, \dots, \mu_n\}$ . In this example, we collect  $n = 100$  samples. The histogram provides a useful visualization of these data. In Figure 2, we display three different histograms on a probability scale. The histogram should be flat for a uniform sample, but the visual perception varies depending on whether the histogram has 10, 5, or 3 bins. The last histogram looks flat, but the other two histograms are not obviously flat. It is not clear which histogram we should base our conclusion on.

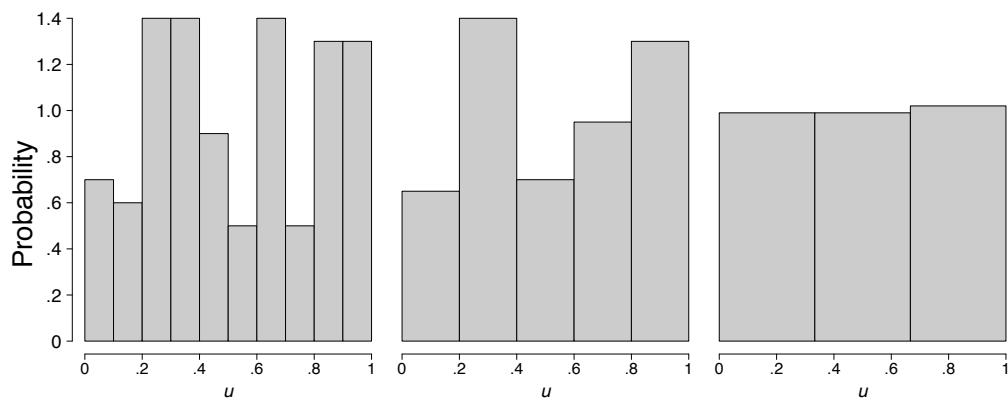


Figure 2. Three histograms of a sample of 100 uniform points.

Alternatively, we might use the cumulative distribution function (CDF), which is denoted by  $F(\mu)$ . The CDF gives the probability that the spinner gives a value less than or equal to  $\mu$ , that is, the probability that the red arrow lands in the interval  $[0, \mu]$ . By simple arithmetic,  $F(\mu) = \mu$ , which is the diagonal straight line  $y = x$ . The CDF based upon the sample data is called the empirical CDF (ECDF), is denoted by

$$\hat{F}_n(\mu)$$

and is defined to be the fraction of the data less than or equal to  $\mu$ ; that is,

$$\hat{F}_n(u) = \frac{\# u_i \leq u}{n}.$$

In general, the ECDF takes on a ragged staircase appearance.

For the spinner sample analyzed in Figure 2, we computed the ECDF and CDF, which are displayed in Figure 3. In the left frame, the ECDF appears close to the line  $y = x$ , shown in the middle frame. In the right frame, we overlay these two curves and verify that they are indeed quite close to each other. Observe that we do not need to specify the number of bins as with the histogram.

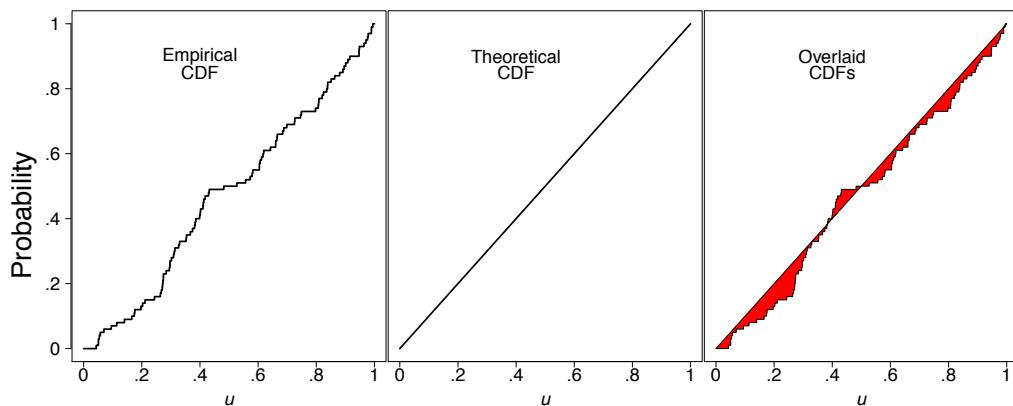


Figure 3. The empirical and theoretical cumulative distribution functions of a sample of 100 uniform points.

## q-q plot for uniform data

The q-q plot for uniform data is very similar to the empirical CDF graphic, except with the axes reversed. The q-q plot provides a visual comparison of the sample quantiles to the corresponding theoretical quantiles. In general, if the points in a q-q plot depart from a straight line, then the assumed distribution is called into question.

Here we define the  $q^{\text{th}}$  quantile of a batch of  $n$  numbers as a number  $\xi_q$  such that a fraction  $q \times n$  of the sample is less than  $\xi_q$ , while a fraction  $(1 - q) \times n$  of the sample is greater than  $\xi_q$ . The best known quantile is the median,  $\xi_{0.5}$ , which is located in the middle of the sample.

Consider a small sample of 5 numbers from the spinner

$$\mu_1 = 0.41, \mu_2 = 0.24, \mu_3 = 0.59, \mu_4 = 0.03, \mu_5 = 0.67.$$

Based upon our description of the spinner, we expect a uniform distribution to model these data. If the sample data were “perfect,” then on average there would be an observation in the middle of each of the 5 intervals: 0 to .2, .2 to .4, .4 to .6, and so on. Table 1 shows the 5 data points (sorted in ascending order) and the theoretically expected value of each based on the assumption that the distribution is uniform (the middle of the interval).

Table 1. Computing the Expected Quantile Values.

Data ( $\mu$ )	Rank ( $i$ )	Middle of the $i^{\text{th}}$ Interval
0.03	1	Middle of the 1 <sup>st</sup>
0.24	2	0.3
0.24	2 3	0.5
0.59	4	0.7
0.67	5	0.9

The theoretical and empirical CDFs are shown in Figure 4 and the q-q plot is shown in the left frame of Figure 5.

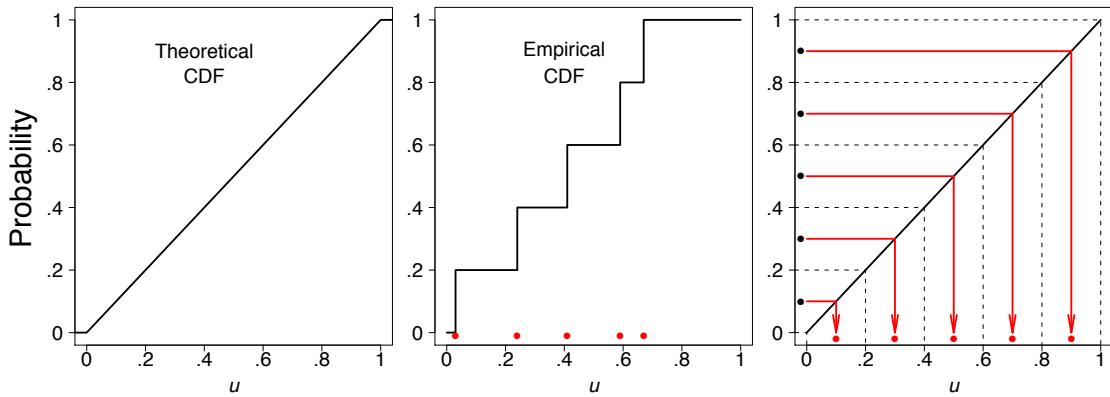


Figure 4. The theoretical and empirical CDFs of a small sample of 5 uniform points, together with the expected values of the 5 points (red dots in the right frame).

In general, we consider the full set of sample quantiles to be the sorted data values

$$\mu_{(1)} < \mu_{(2)} < \mu_{(3)} < \dots < \mu_{(n-1)} < \mu_{(n)},$$

where the parentheses in the subscript indicate the data have been ordered. Roughly speaking, we expect the first ordered value to be in the middle of the interval  $(0, 1/n)$ , the second to be in the middle of the interval  $(1/n, 2/n)$ , and the last to be in the middle of the interval  $((n - 1)/n, 1)$ . Thus, we take as the theoretical quantile the value

$$\xi_q = q \approx \frac{i - 0.5}{n},$$

where  $q$  corresponds to the  $i^{\text{th}}$  ordered sample value. We subtract the quantity 0.5 so that we are exactly in the middle of the interval  $((i - 1)/n, i/n)$ . These ideas are depicted in the right frame of Figure 4 for our small sample of size  $n = 5$ .

We are now prepared to define the q-q plot precisely. First, we compute the  $n$  expected values of the data, which we pair with the  $n$  data points sorted in ascending order. For the uniform density, the q-q plot is composed of the  $n$  ordered pairs

$$\left( \frac{i - 0.5}{n}, u_{(i)} \right), \quad \text{for } i = 1, 2, \dots, n.$$

This definition is slightly different from the ECDF, which includes the points  $(u_{(i)}, i/n)$ . In the left frame of Figure 5, we display the q-q plot of the 5 points in Table 1. In the right two frames of Figure 5, we display the q-q plot of the same batch of numbers used in Figure 2. In the final frame, we add the diagonal line  $y = x$  as a point of reference.

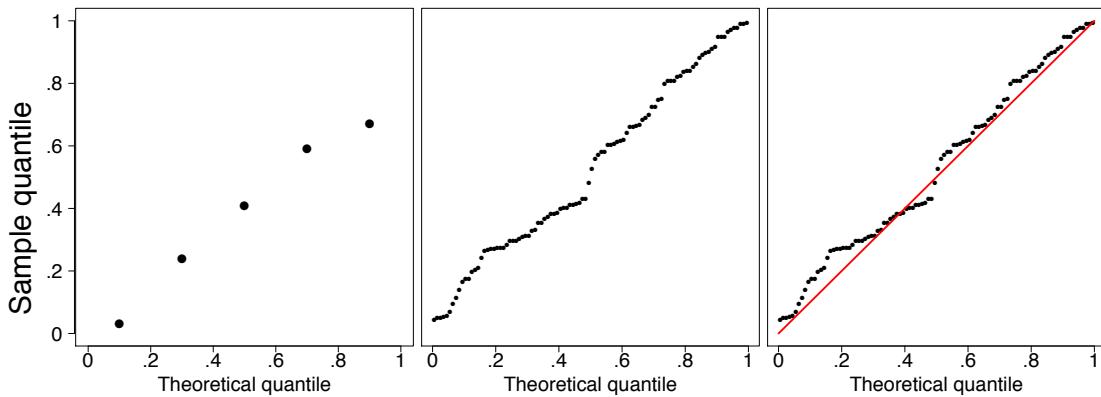


Figure 5. (Left) q-q plot of the 5 uniform points. (Right) q-q plot of a sample of 100 uniform points.

The sample size should be taken into account when judging how close the q-q plot is to the straight line. We show two other uniform samples of size  $n = 10$  and  $n = 1000$  in Figure 6. Observe that the q-q plot when  $n = 1000$  is almost identical to the line  $y = x$ , while such is not the case when the sample size is only  $n = 10$ .

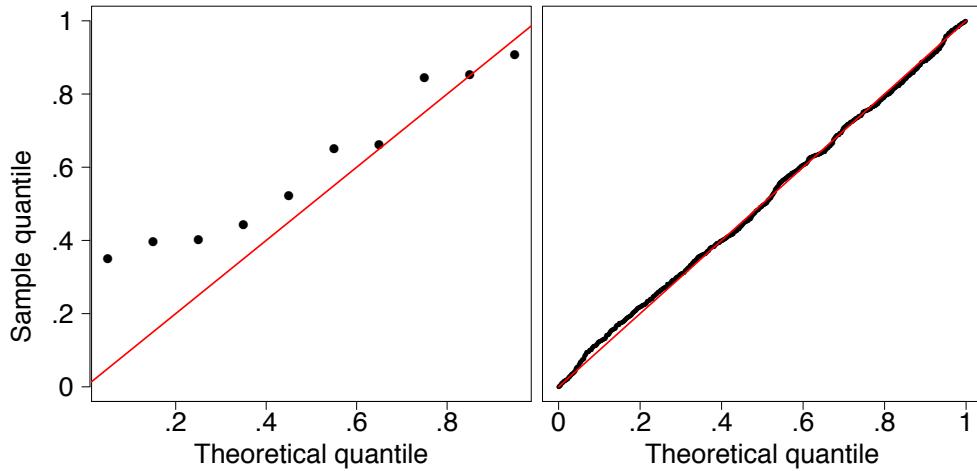


Figure 6. q-q plots of a sample of 10 and 1000 uniform points.

In Figure 7, we show the q-q plots of two random samples that are not uniform. In both examples, the sample quantiles match the theoretical quantiles only at the median and at the extremes. Both samples seem to be symmetric around the median. But the data in the left frame are closer to the median than would be expected if the data were uniform. The data in the right frame are further from the median than would be expected if the data were uniform.

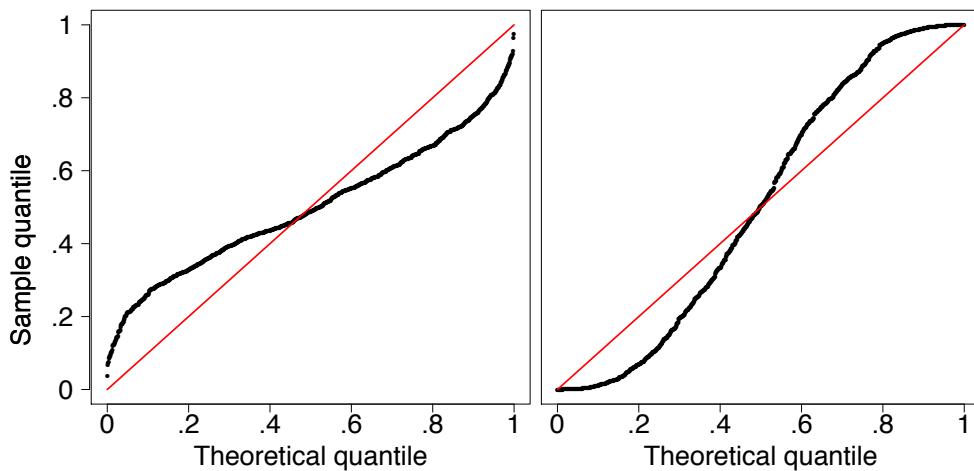


Figure 7. q-q plots of two samples of size 1000 that are not uniform.

In fact, the data were generated in the R language from beta distributions with parameters  $a = b = 3$  on the left and  $a = b = 0.4$  on the right. In Figure 8 we display histograms of these two data sets, which serve to clarify the true shapes of the densities. These are clearly non-uniform.

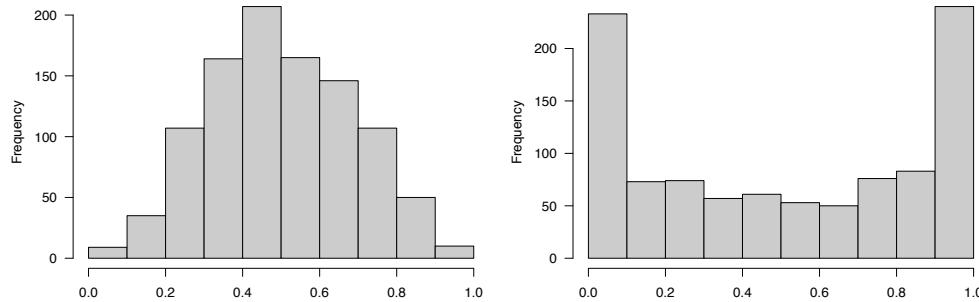


Figure 8. Histograms of the two non-uniform data sets.

### q-q plot for normal data

The definition of the q-q plot may be extended to any continuous density. The q-q plot will be close to a straight line if the assumed density is correct. Because the cumulative distribution function of the uniform density was a straight line, the q-q plot was very easy to construct. For data that are not uniform, the theoretical quantiles must be computed in a different manner.

Let  $\{z_1, z_2, \dots, z_n\}$  denote a random sample from a normal distribution with mean  $\mu = 0$  and standard deviation  $\sigma = 1$ . Let the ordered values be denoted by

$$z_{(1)} < z_{(2)} < z_{(3)} < \dots < z_{(n-1)} < z_{(n)}.$$

These  $n$  ordered values will play the role of the sample quantiles.

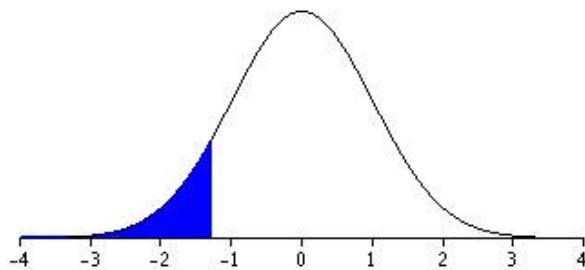
Let us consider a sample of 5 values from a distribution to see how they compare with what would be expected for a normal distribution. The 5 values in ascending order are shown in the first column of Table 2.

Table 2. Computing the Expected Quantile Values for Normal Data.  
of the Two Non-Uniform Data Sets.

Data (z)	Rank (i)	Middle of the i <sup>th</sup> Interval	z
-1.96	1	0.1	-1.28
-0.78	2	0.3	-0.52
0.31	3	0.5	0
1.15	4	0.7	0.52
1.62	5	0.9	1.28

Just as in the case of the uniform distribution, we have 5 intervals. However, with a normal distribution the theoretical quantile is not the middle of the interval but rather the inverse of the normal distribution for the middle of the interval. Taking the first interval as an example, we want to know the z value such that 0.1 of the area in the normal distribution is below z. This can be computed using the Inverse Normal Calculator as shown in Figure 9. Simply set the “Shaded Area” field to the middle of the interval (0.1) and click on the “Below” button. The result is -1.28. Therefore, 10% of the distribution is below a z value of -1.28.

Normal Distribution



Mean:

Sd:

Shaded Area:

Above

Below: -1.2816

Between

Outside

Figure 9. Example of the Inverse Normal Calculator for finding a value of the expected quantile from a normal distribution.

The q-q plot for the data in Table 2 is shown in the left frame of Figure 11.

In general, what should we take as the corresponding theoretical quantiles? Let the cumulative distribution function of the normal density be denoted by  $\Phi(z)$ . In the previous example,  $\Phi(-1.28) = 0.10$  and  $\Phi(0.00) = 0.50$ . Using the quantile notation, if  $\xi_q$  is the  $q^{\text{th}}$  quantile of a normal distribution, then

$$\Phi(\xi_q) = q.$$

That is, the probability a normal sample is less than  $\xi_q$  is in fact just  $q$ .

Consider the first ordered value,  $z_{(1)}$ . What might we expect the value of  $\Phi(z_{(1)})$  to be? Intuitively, we expect this probability to take on a value in the interval  $(0, 1/n)$ . Likewise, we expect  $\Phi(z_{(2)})$  to take on a value in the interval  $(1/n, 2/n)$ . Continuing, we expect  $\Phi(z_{(n)})$  to fall in the interval  $((n-1)/n, 1/n)$ . Thus, the theoretical quantile we desire is defined by the inverse (not reciprocal) of the normal CDF. In particular, the theoretical quantile corresponding to the empirical quantile  $z_{(i)}$  should be

$$\Phi^{-1} \left( \frac{i - 0.5}{n} \right) \quad \text{for } i = 1, 2, \dots, n.$$

The empirical CDF and theoretical quantile construction for the small sample given in Table 2 are displayed in Figure 10. For the larger sample of size 100, the first few expected quantiles are -2.576, -2.170, and -1.960.

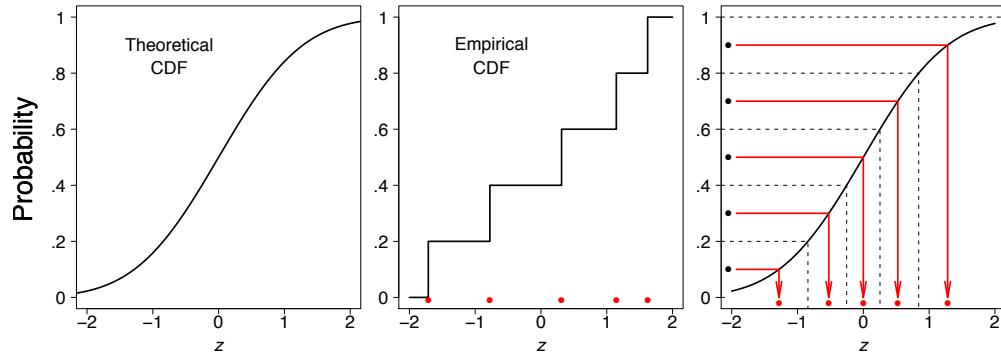


Figure 10. The empirical CDF of a small sample of 5 normal points, together with the expected values of the 5 points (red dots in the right frame).

In the left frame of Figure 11, we display the q-q plot of the small normal sample given in Table 2. The remaining frames in Figure 11 display the q-q plots of normal random samples of size  $n = 100$  and  $n = 1000$ . As the sample size increases, the points in the q-q plots lie closer to the line  $y = x$ .

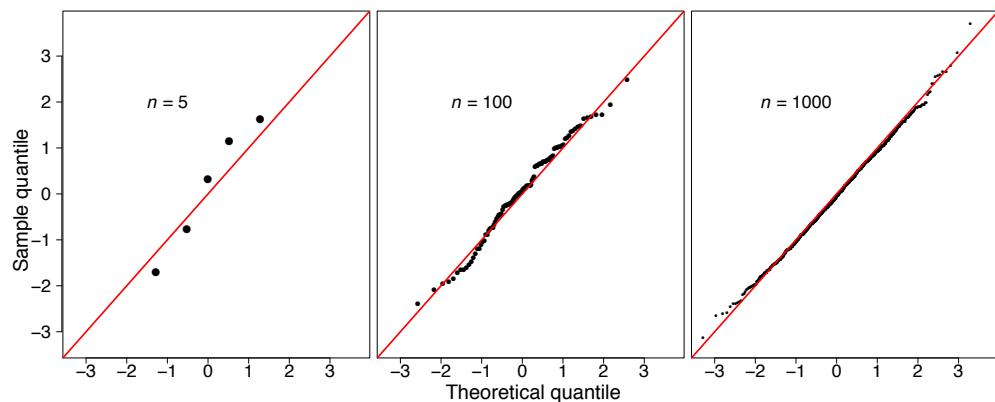


Figure 11. q-q plots of normal data.

As before, a normal q-q plot can indicate departures from normality. The two most common examples are skewed data and data with heavy tails (large kurtosis). In

Figure 12 we show normal q-q plots for a chi-squared (skewed) data set and a Student's-t (kurtotic) data set, both of size  $n = 1000$ . The data were first standardized. The red line is again  $y = x$ . Notice, in particular, that the data from the t distribution follow the normal curve fairly closely until the last dozen or so points on each extreme.

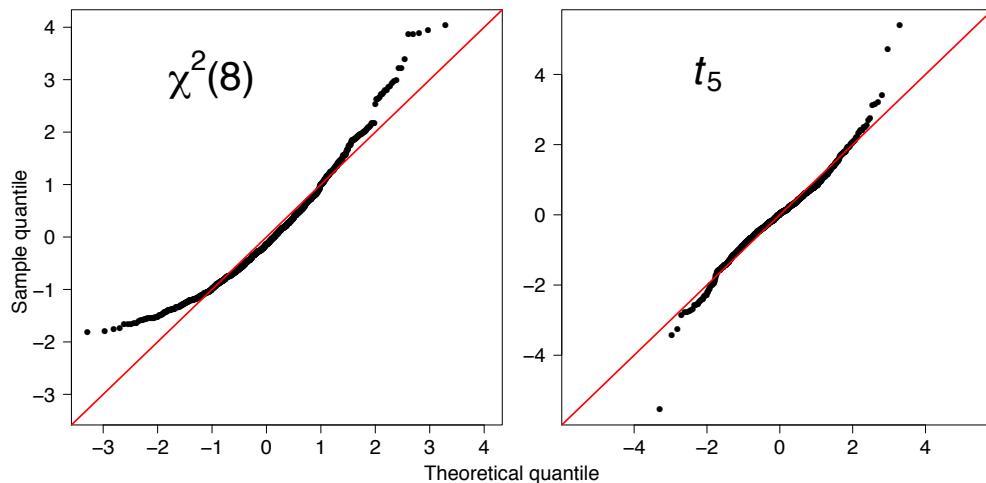


Figure 12. q-q plots for standardized non-normal data ( $n = 1000$ ).

### q-q plots for normal data with general mean and scale

Our previous discussion of q-q plots for normal data all assumed that our data were standardized. One approach to constructing q-q plots is to first standardize the data and then proceed as described previously. An alternative is to construct the plot directly from raw data.

In this section we present a general approach for data that are not standardized. Why did we standardize the data in Figure 12? The q-q plot is comprised of the  $n$  points

$$\left( \Phi^{-1} \left( \frac{i - 0.5}{n} \right), z_{(i)} \right) \quad \text{for } i = 1, 2, \dots, n.$$

If the original data  $\{z_i\}$  are normal, but have an arbitrary mean  $\mu$  and standard deviation  $\sigma$ , then the line  $y = x$  will not match the expected theoretical quantile. Clearly, the linear transformation

$$\mu + \sigma \xi q$$

would provide the  $q$ th theoretical quantile on the transformed scale. In practice, with a new data set

$$\{x_1, x_2, \dots, x_n\},$$

the normal q-q plot would consist of the  $n$  points

Instead of plotting the line  $y = x$  as a reference line, the line

$$y = M + s \cdot x$$

should be composed, where  $M$  and  $s$  are the sample moments (mean and standard deviation) corresponding to the theoretical moments  $\mu$  and  $\sigma$ . Alternatively, if the data are standardized, then the line  $y = x$  would be appropriate, since now the sample mean would be 0 and the sample standard deviation would be 1.

### **Example: SAT Case Study**

The SAT case study followed the academic achievements of 105 college students majoring in computer science. The first variable is their verbal SAT score and the second is their grade point average (GPA) at the university level. Before we compute inferential statistics using these variables, we should check if their distributions are normal. In Figure 13, we display the q-q plots of the verbal SAT and university GPA variables.

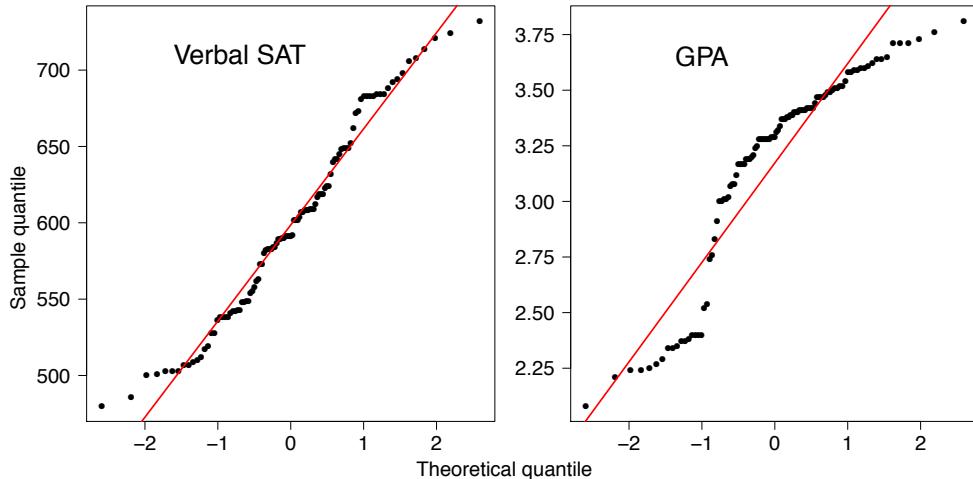


Figure 13. q-q plots for the student data ( $n = 105$ ).

The verbal SAT seems to follow a normal distribution reasonably well, except in the extreme tails. However, the university GPA variable is highly non-normal. Compare the GPA q-q plot to the simulation in the right frame of Figure 7. These figures are very similar, except for the region where  $x \approx -1$ . To follow these ideas, we computed histograms of the variables and their scatter diagram in Figure 14. These figures tell quite a different story. The university GPA is bimodal, with about 20% of the students falling into a separate cluster with a grade of C. The scatter diagram is quite unusual. While the students in this cluster all have below average verbal SAT scores, there are as many students with low SAT scores whose GPAs were quite respectable. We might speculate as to the cause(s): different distractions, different study habits, but it would only be speculation. But observe that the raw correlation between verbal SAT and GPA is a rather high 0.65, but when we exclude the cluster, the correlation for the remaining 86 students falls a little to 0.59.

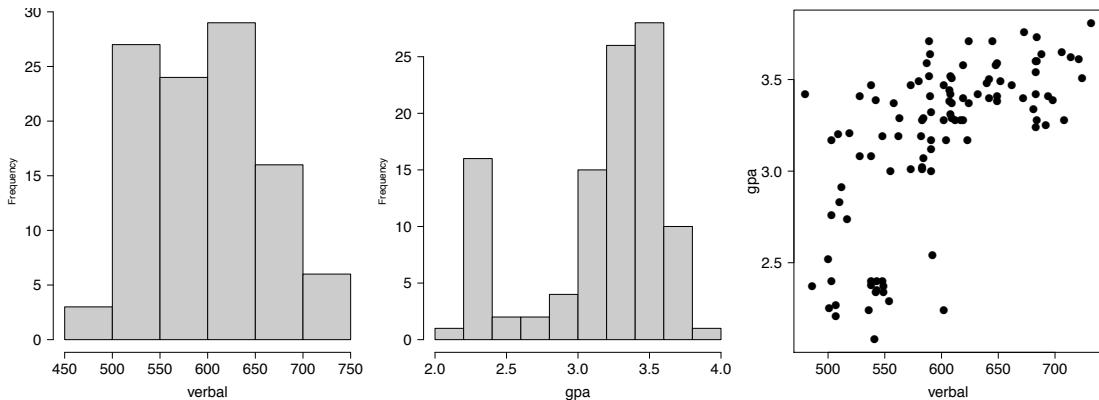


Figure 14. Histograms and scatter diagram of the verbal SAT and GPA variables for the 105 students.

## Discussion

Parametric modeling usually involves making assumptions about the shape of data, or the shape of residuals from a regression fit. Verifying such assumptions can take many forms, but an exploration of the shape using histograms and q-q plots is very effective. The q-q plot does not have any design parameters such as the number of bins for a histogram.

In an advanced treatment, the q-q plot can be used to formally test the null hypothesis that the data are normal. This is done by computing the correlation coefficient of the  $n$  points in the q-q plot. Depending upon  $n$ , the null hypothesis is rejected if the correlation coefficient is less than a threshold. The threshold is already quite close to 0.95 for modest sample sizes.

We have seen that the q-q plot for uniform data is very closely related to the empirical cumulative distribution function. For general density functions, the so-called probability integral transform takes a random variable  $X$  and maps it to the interval  $(0, 1)$  through the CDF of  $X$  itself, that is,

$$Y = F_X(X)$$

which has been shown to be a uniform density. This explains why the q-q plot on standardized data is always close to the line  $y = x$  when the model is correct. Finally, scientists have used special graph paper for years to make relationships linear (straight lines). The most common example used to be semi-log paper, on which points following the formula  $y = ae^{bx}$  appear linear. This follows of course since  $\log(y) = \log(a) + bx$ , which is the equation for a straight line. The q-q plots

may be thought of as being “probability graph paper” that makes a plot of the ordered data values into a straight line. Every density has its own special probability graph paper.

# Contour Plots

by David Lane

## *Prerequisites*

- none

## *Learning Objectives*

1. Describe a contour plot.
2. Interpret a contour plot

Contour plots portray data for three variables in two dimensions. The plot contains a number of contour lines. Each contour line is shown in an X-Y plot and has a constant value on a third variable. Consider the Figure 1 that contains data on the fat, non-sugar carbohydrates, and calories present in a variety of breakfast cereals. Each line shows the carbohydrate and fat levels for cereals with the same number of calories. Note that the number of calories is not determined exactly by the fat and non-sugar carbohydrates since cereals also differ in sugar and protein.

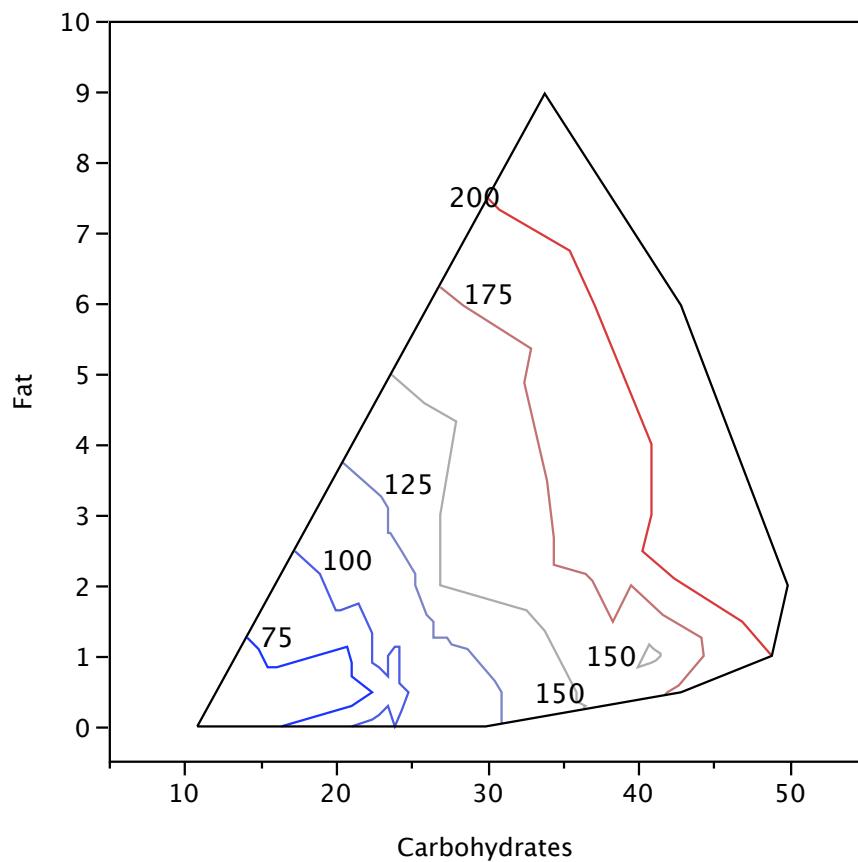


Figure 1. A contour plot showing calories as a function of fat and carbohydrates.

An alternative way to draw the plot is shown in Figure 2. The areas with the same number of calories are shaded.

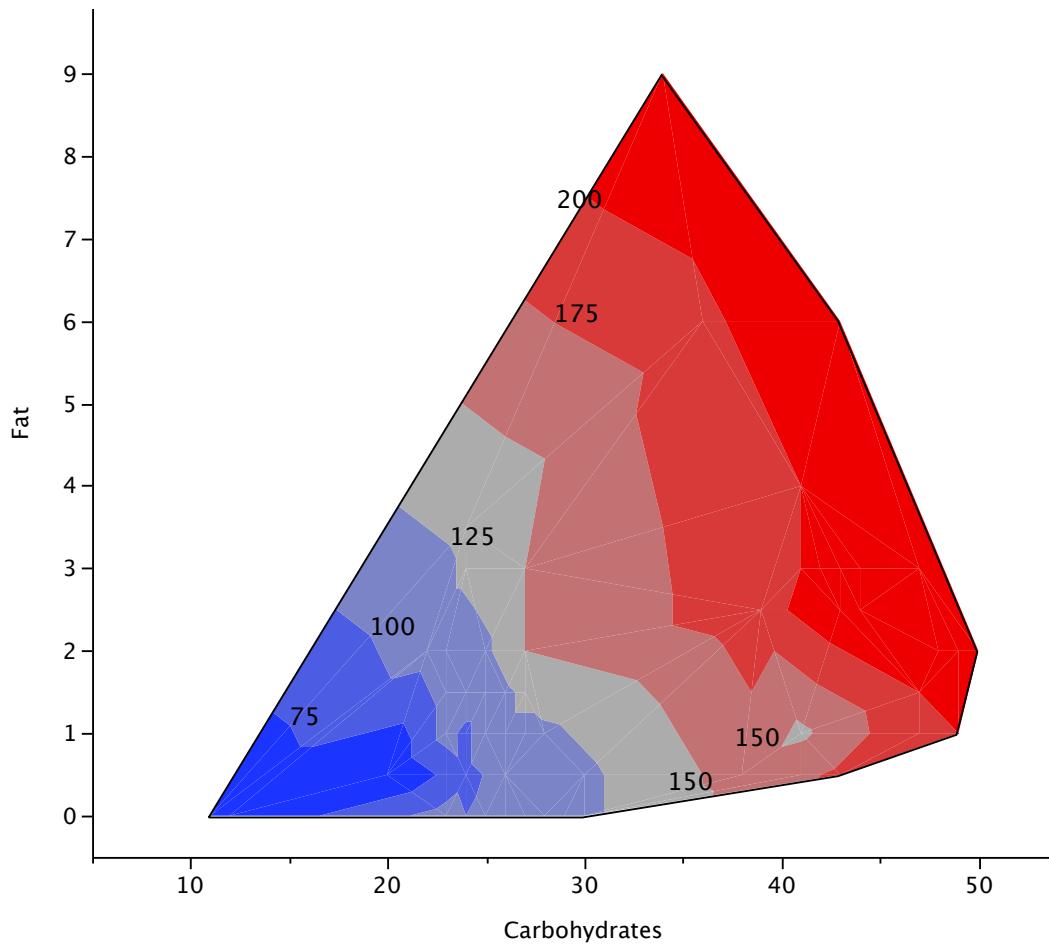


Figure 2. A contour plot showing calories as a function of fat and carbohydrates with areas shaded. An area represents values less than or equal to the label to the right of the area.

# 3D Plots

by David Lane

## *Prerequisites*

- Chapter 4: Introduction to Bivariate Data

## *Learning Objectives*

1. Describe a 3D Plot.
2. Give an example of the value of a 3D plot.

Just as two-dimensional scatter plots show the data in two dimensions, 3D plots show data in three dimensions. Figure 1 shows a 3D scatter plot of the fat, non-sugar carbohydrates, and calories from a variety of cereal types.

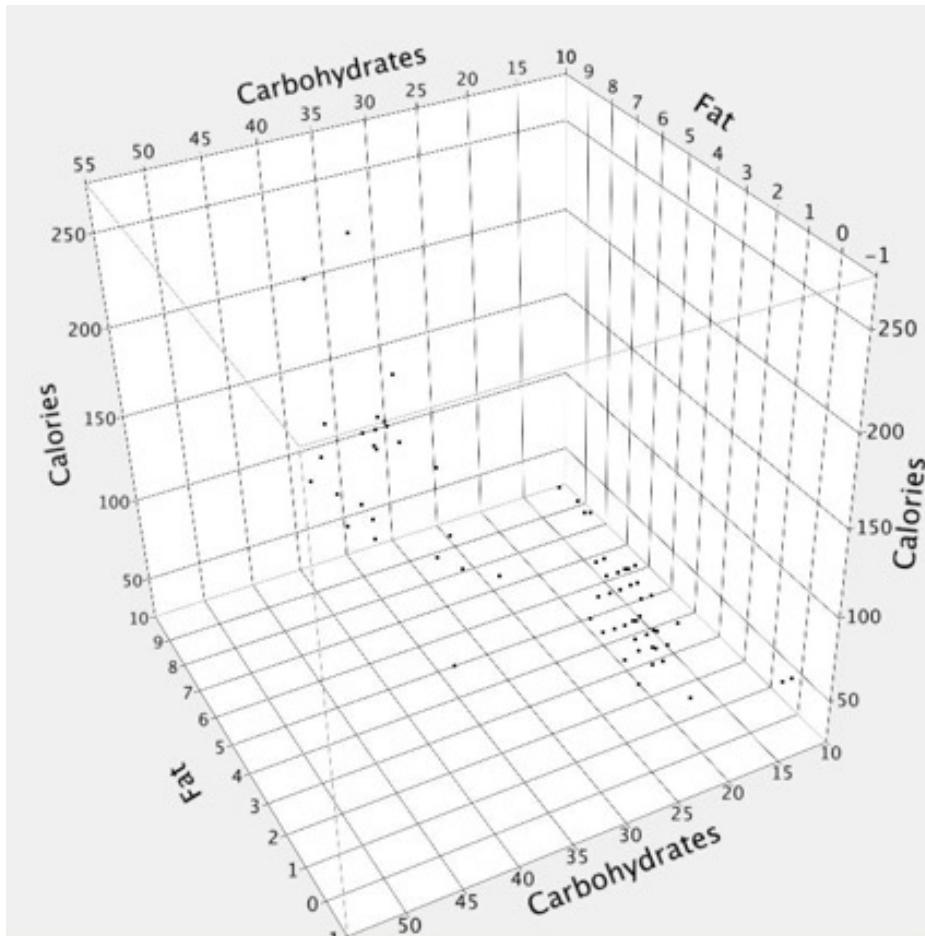


Figure 1. A 3D scatter plot showing fat, non-sugar carbohydrates, and calories from a variety of cereal types.

Many statistical packages allow you to rotate the axes interactively to view the data from a different vantage point. Figure 2 is an example.

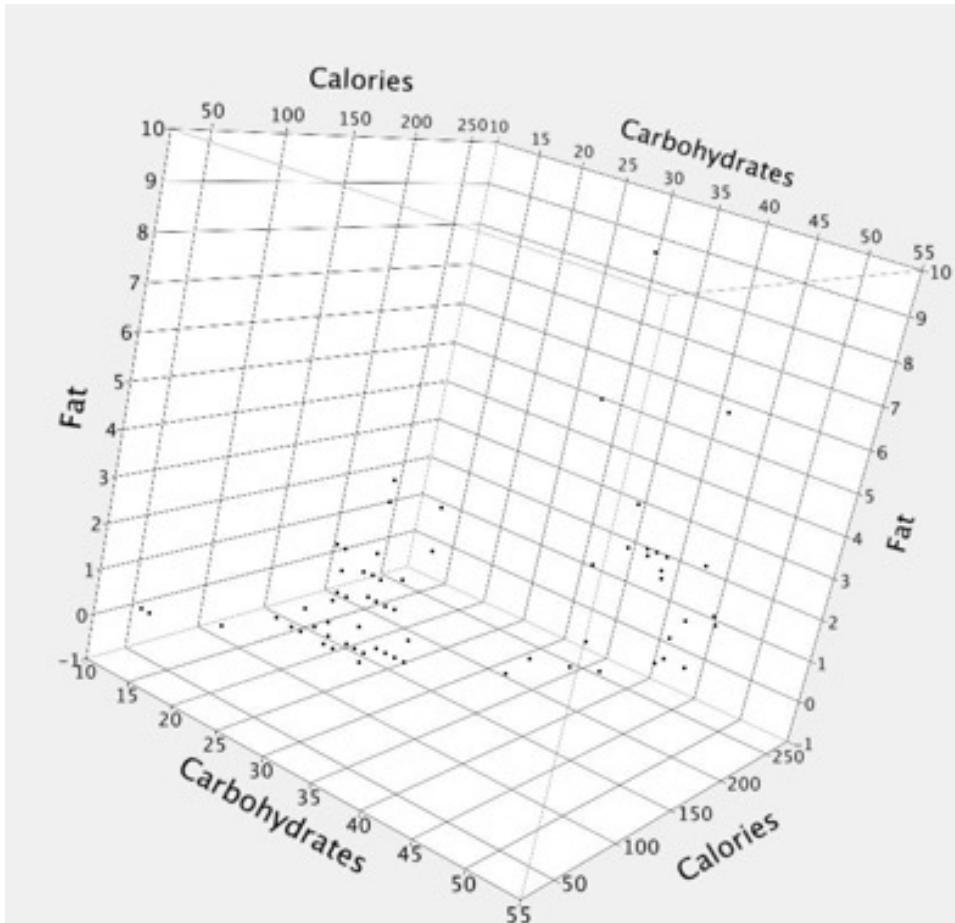


Figure 2. An alternative 3D scatter plot showing fat, non-sugar carbohydrates, and calories.

A fourth dimension can be represented as long as it is represented as a nominal variable. Figure 3 represents the different manufacturers by using different colors.

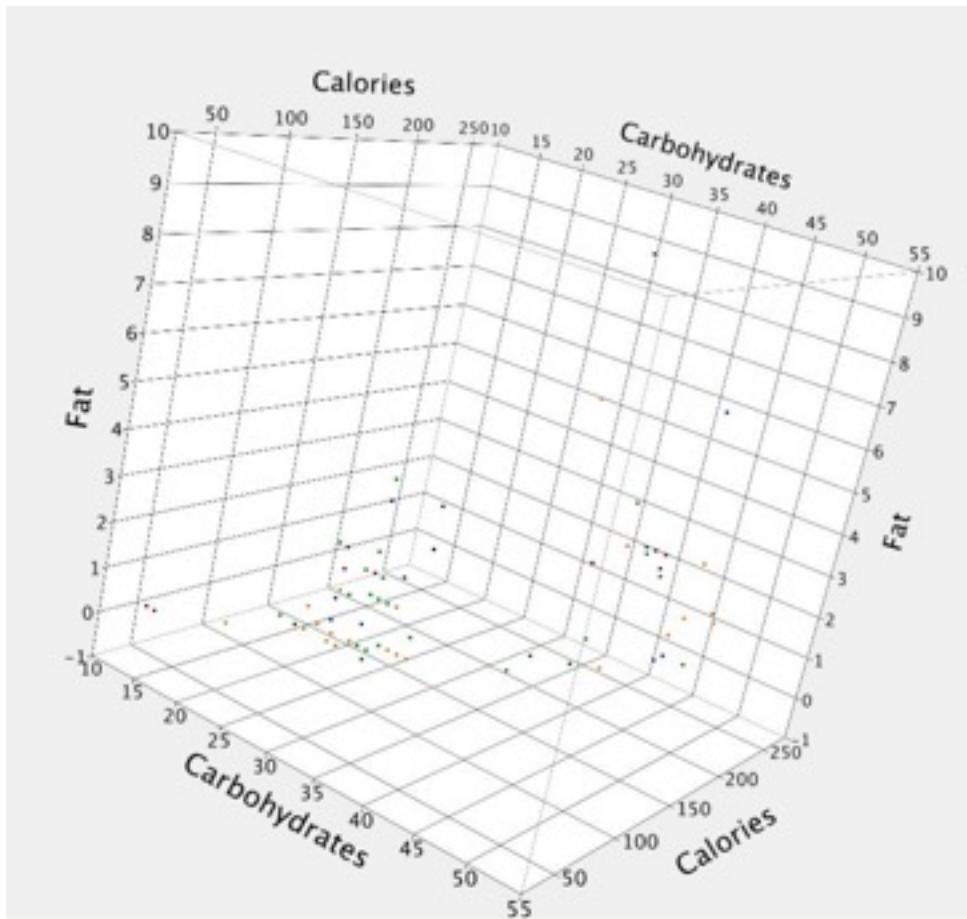


Figure 3. The different manufacturers are color coded.

Interactively rotating 3D plots can sometimes reveal aspects of the data not otherwise apparent. Figure 4 shows data from a pseudo random number generator. Figure 4 does not show anything systematic and the random number generator appears to generate data with properties similar to those of true random numbers.

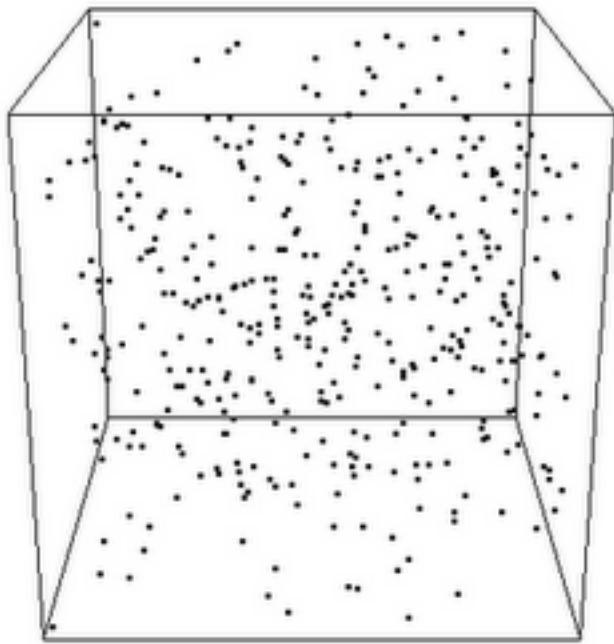


Figure 4. A 3D scatter plot showing 400 values of X, Y, and Z from a pseudo random number generator.

Figure 5 shows a different perspective on these data. Clearly they were not generated by a random process.

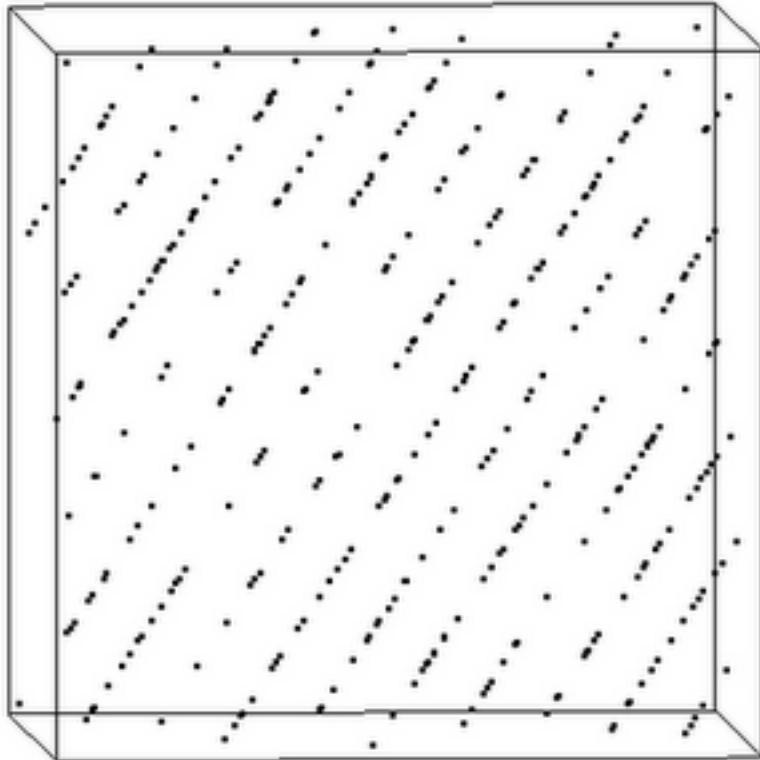


Figure 5. A different perspective on the 3D scatter plot showing 400 values of X, Y, and Z from a pseudo random number generator.

Figures 4 and 5 are reproduced with permission from [R snippets](#) by Bogumil Kaminski.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 8: Contour Plots

This [web page](#) portrays altitudes in the United States.

## **What do you think?**

What part of the state of Texas (North, South, East, or West) contains the highest elevation?

West Texas

## Exercises

1. What are Q-Q plots useful for?
2. For the following data, plot the theoretically expected z score as a function of the actual z score (a Q-Q plot).

0	0.5	0.8	1.3	2.1
0	0.6	0.9	1.4	2.1
0	0.6	1	1.4	2.1
0	0.6	1	1.5	2.1
0	0.6	1.1	1.6	2.1
0	0.6	1.1	1.7	2.1
0.1	0.6	1.2	1.7	2.3
0.1	0.6	1.2	1.7	2.5
0.1	0.6	1.2	1.8	2.7
0.1	0.6	1.2	1.8	3
0.1	0.7	1.2	1.9	4.2
0.2	0.7	1.2	1.9	5
0.2	0.8	1.3	2	5.7
0.3	0.8	1.3	2	12.4
0.3	0.8	1.3	2	15.2
0.4	0.8	1.3	2.1	

3. For the data in problem 2, describe how the data differ from a normal distribution.
4. For the “SAT and College GPA” case study data, create a contour plot looking at College GPA as a function of Math SAT and High School GPA. Naturally, you should use a computer to do this.
5. For the “SAT and College GPA” case study data, create a 3D plot using the variables College GPA, Math SAT, and High School GPA. Naturally, you should use a computer to do this.

# 9. Sampling Distributions

## *Prerequisites*

- none

- A. Introduction
- B. Sampling Distribution of the Mean
- C. Sampling Distribution of Difference Between Means
- D. Sampling Distribution of Pearson's r
- E. Sampling Distribution of a Proportion
- F. Exercises

The concept of a sampling distribution is perhaps the most basic concept in inferential statistics. It is also a difficult concept because a sampling distribution is a theoretical distribution rather than an empirical distribution.

The introductory section defines the concept and gives an example for both a discrete and a continuous distribution. It also discusses how sampling distributions are used in inferential statistics.

The remaining sections of the chapter concern the sampling distributions of important statistics: the Sampling Distribution of the Mean, the Sampling Distribution of the Difference Between Means, the Sampling Distribution of r, and the Sampling Distribution of a Proportion.

# Introduction to Sampling Distributions

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 1: Inferential Statistics

## *Learning Objectives*

1. Define inferential statistics
2. Graph a probability distribution for the mean of a discrete variable
3. Describe a sampling distribution in terms of “all possible outcomes”
4. Describe a sampling distribution in terms of repeated sampling
5. Describe the role of sampling distributions in inferential statistics
6. Define the standard error of the mean

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall that inferential statistics concern generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter. (In this example, the sample statistics are the sample means and the population parameter is the population mean.) As the later portions of this chapter show, these determinations are based on sampling distributions.

## **Discrete Distributions**

We will illustrate the concept of sampling distributions with a simple example. Figure 1 shows three pool balls, each with a number on it. Suppose two of the balls are selected randomly (with replacement) and the average of their numbers is computed. All possible outcomes are shown below in Table 1.



Figure 1. The pool balls.

Table 1. All possible outcomes when two balls are sampled with replacement.

Outcome	Ball 1	Ball 2	Mean
1	1	1	1
2	1	2	1.5
3	1	3	2
4	2	1	1.5
5	2	2	2
6	2	3	2.5
7	3	1	2
8	3	2	2.5
9	3	3	3

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in Table 2. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Table 2. Frequencies of means for  $N = 2$ .

Mean	Frequency	Relative Frequency
1	1	0.111
1.5	2	0.222
2	3	0.333
2.5	2	0.222
3	1	0.111

Figure 2 shows a relative frequency distribution of the means based on Table 2. This distribution is also a probability distribution since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.

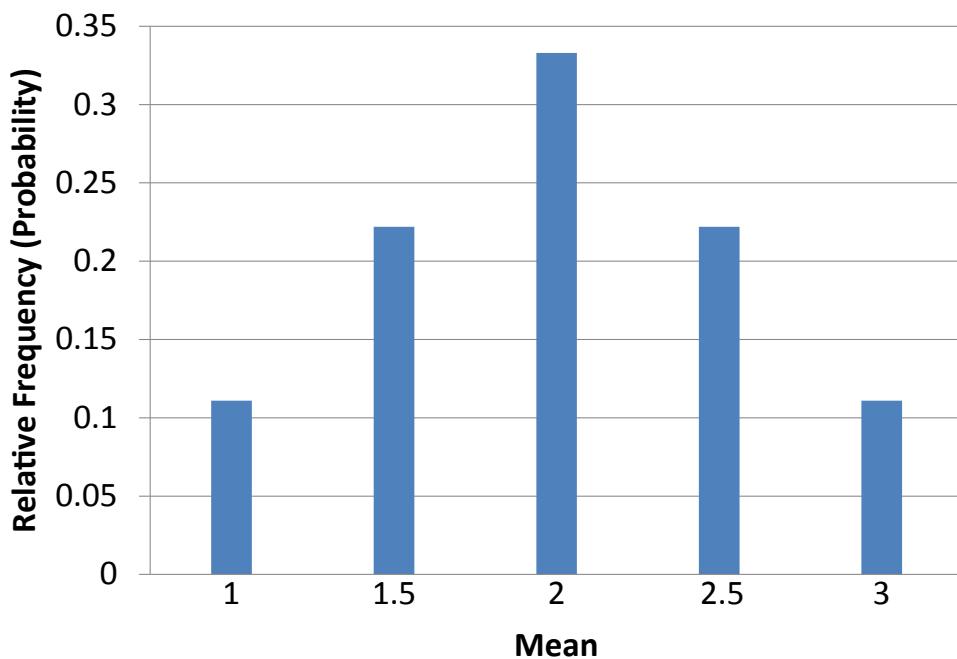


Figure 2. Distribution of means for  $N = 2$ .

The distribution shown in Figure 2 is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 ( $N = 2$ ). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions. The pool balls have only the values 1, 2, and 3, and a sample mean can have one of only five values shown in Table 2.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement) and the mean of the two balls is computed and recorded. Then this process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution shown in Figure 2. As the number of samples approaches infinity, the relative frequency distribution will approach the sampling distribution. This means that you

can conceive of a sampling distribution as being a relative frequency distribution based on a very large number of samples. To be strictly correct, the relative frequency distribution approaches the sampling distribution as the number of samples approaches infinity.

It is important to keep in mind that every statistic, not just the mean, has a sampling distribution. For example, Table 3 shows all possible outcomes for the range of two numbers (larger number minus the smaller number). Table 4 shows the frequencies for each of the possible ranges and Figure 3 shows the sampling distribution of the range.

Table 3. All possible outcomes when two balls are sampled with replacement.

Outcome	Ball 1	Ball 2	Range
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

Table 4. Frequencies of ranges for  $N = 2$ .

Range	Frequency	Relative Frequency
0	3	0.333
1	4	0.444
2	2	0.222

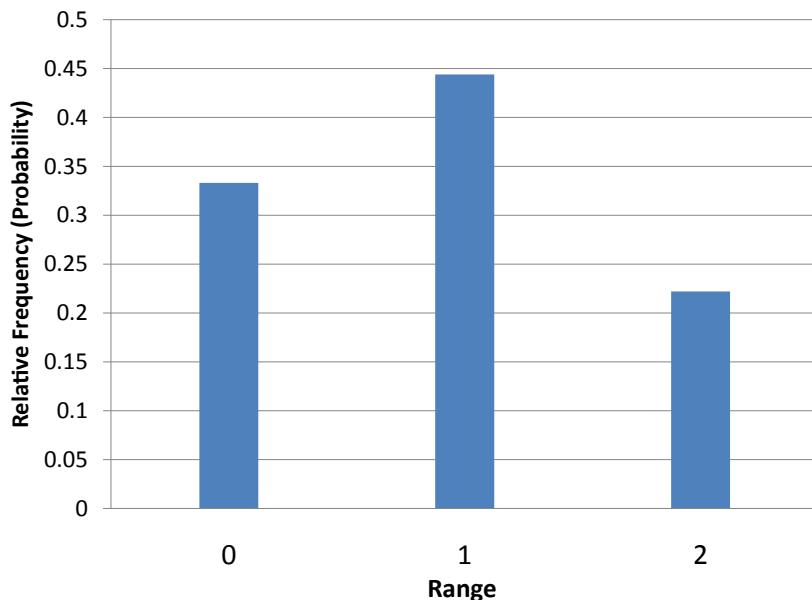


Figure 3. Distribution of ranges for  $N = 2$ .

It is also important to keep in mind that there is a sampling distribution for various sample sizes. For simplicity, we have been using  $N = 2$ . The sampling distribution of the range for  $N = 3$  is shown in Figure 4.

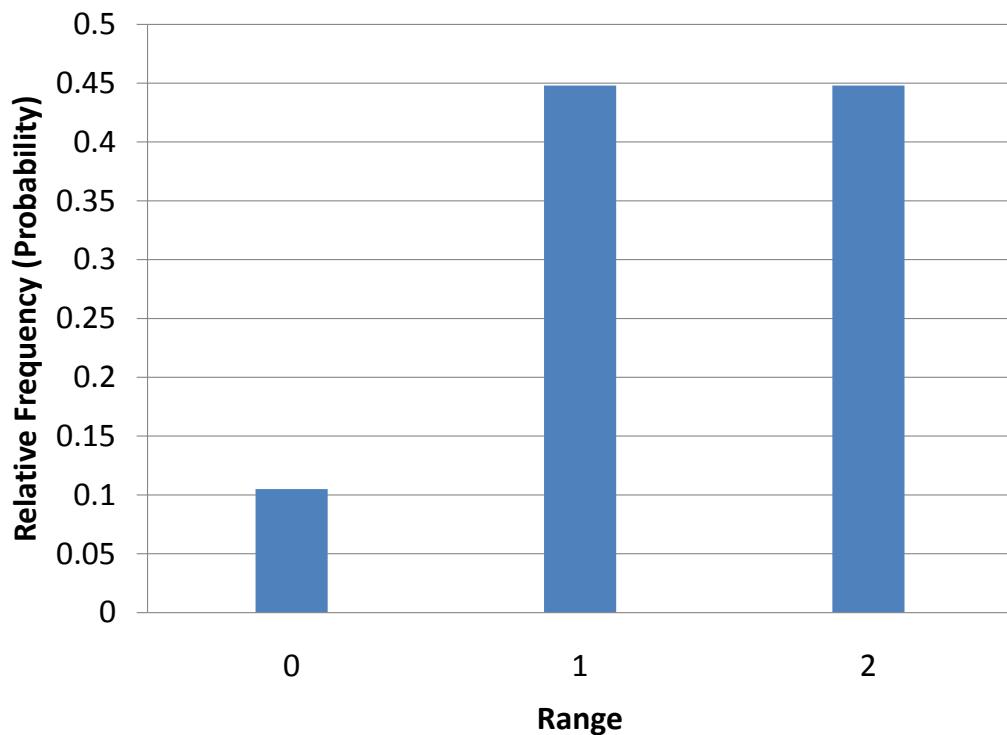


Figure 4. Distribution of ranges for  $N = 3$ .

## Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? (Although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two balls. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this example since there are 1,000,000 possible outcomes (1,000 for the first ball  $\times$  1,000 for the second). Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions. Specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous

distributions, the probability of obtaining any single value is zero. Therefore, as discussed in the section “Distributions” in Chapter 1, these values are called probability densities rather than probabilities.

## **Sampling Distributions and Inferential Statistics**

As we stated in the beginning of this chapter, sampling distributions are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: you collect sample data, and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Keep in mind that all statistics have sampling distributions, not just the mean. In later sections we will be discussing the sampling distribution of the variance, the sampling distribution of the difference between means, and the sampling distribution of Pearson's correlation, among others.

# Sampling Distribution of the Mean

by David M. Lane

## *Prerequisites*

- Chapter 3: Variance Sum Law I
- Chapter 9: Introduction to Sampling Distributions

## *Learning Objectives*

1. State the mean and variance of the sampling distribution of the mean
2. Compute the standard error of the mean
3. State the central limit theorem

The sampling distribution of the mean was defined in the section introducing sampling distributions. This section reviews some important properties of the sampling distribution of the mean.

## **Mean**

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean  $\mu$ , then the mean of the sampling distribution of the mean is also  $\mu$ . The symbol  $\mu_M$  is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_M = \mu$$

## **Variance**

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_m^2 = \frac{\sigma^2}{N}$$

That is, the variance of the sampling distribution of the mean is the population variance divided by  $N$ , the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

(optional paragraph) This expression can be derived very easily from the variance sum law. Let's begin by computing the variance of the sampling distribution of the

sum of three numbers sampled from a population with variance  $\sigma^2$ . The variance of the sum would be  $\sigma^2 + \sigma^2 + \sigma^2$ . For  $N$  numbers, the variance would be  $N\sigma^2$ . Since the mean is  $1/N$  times the sum, the variance of the sampling distribution of the mean would be  $1/N^2$  times the variance of the sum, which equals  $\sigma^2/N$ .

The standard error of the mean is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as:

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

The standard error is represented by a  $\sigma$  because it is a standard deviation. The subscript (M) indicates that the standard error in question is the standard error of the mean.

## Central Limit Theorem

The central limit theorem states that:

*Given a population with a finite mean  $\mu$  and a finite non-zero variance  $\sigma^2$ , the sampling distribution of the mean approaches a normal distribution with a mean of  $\mu$  and a variance of  $\sigma^2/N$  as  $N$ , the sample size, increases.*

The expressions for the mean and variance of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as  $N$  increases. If you have used the “Central Limit Theorem Demo,” ([external link](#); requires Java) you have already seen this for yourself. As a reminder, Figure 1 shows the results of the simulation for  $N = 2$  and  $N = 10$ . The parent population was a uniform distribution. You can see that the distribution for  $N = 2$  is far from a normal distribution. Nonetheless, it does show that the scores are denser in the middle than in the tails. For  $N = 10$  the distribution is quite close to a normal distribution. Notice that the means of the two distributions are the same, but that the spread of the distribution for  $N = 10$  is smaller.

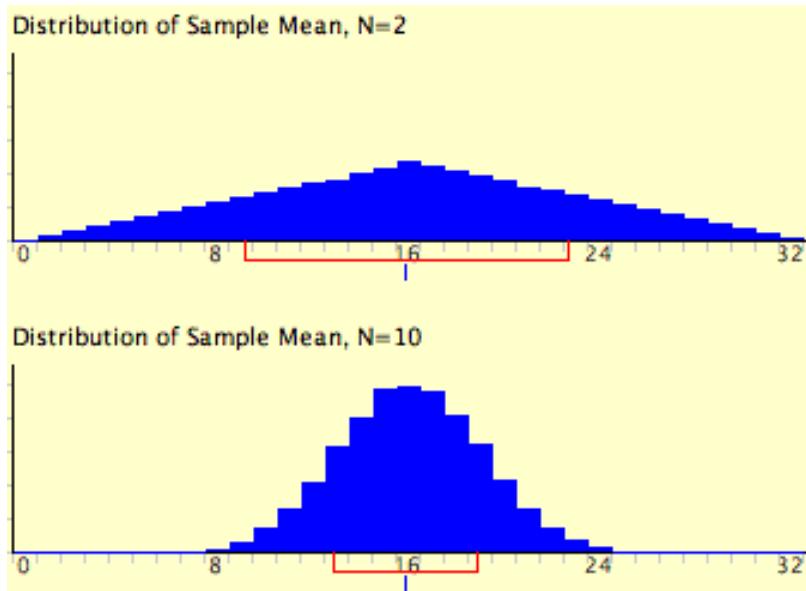


Figure 1. A simulation of a sampling distribution. The parent population is uniform. The blue line under “16” indicates that 16 is the mean. The red line extends from the mean plus and minus one standard deviation.

Figure 2 shows how closely the sampling distribution of the mean approximates a normal distribution even when the parent population is very non-normal. If you look closely you can see that the sampling distributions do have a slight positive skew. The larger the sample size, the closer the sampling distribution of the mean would be to a normal distribution.

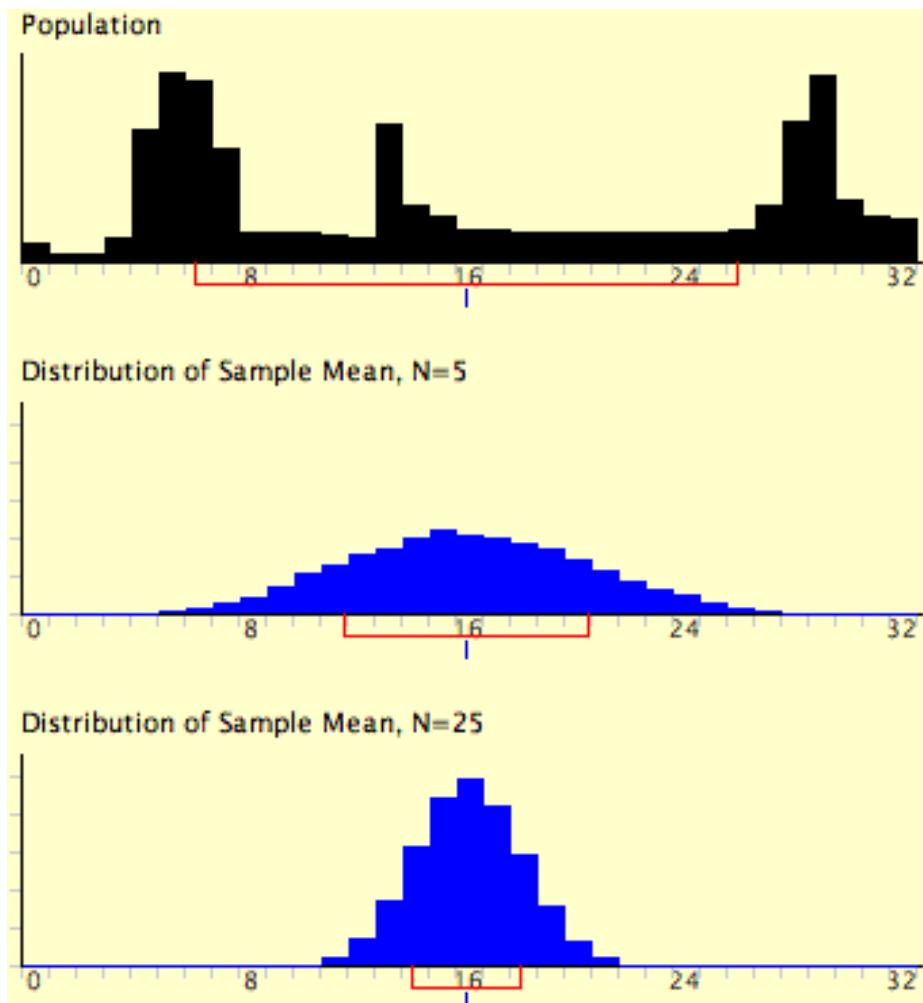


Figure 2. A simulation of a sampling distribution. The parent population is very non-normal.

# Sampling Distribution of Difference Between Means

by David M. Lane

## *Prerequisites*

- Chapter 3: Variance Sum Law I
- Chapter 9: Sampling Distributions
- Chapter 9: Sampling Distribution of the Mean

## *Learning Objectives*

1. State the mean and variance of the sampling distribution of the difference between means
2. Compute the standard error of the difference between means
3. Compute the probability of a difference between means being above a specified value

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again: (1) sample  $n_1$  scores from Population 1 and  $n_2$  scores from Population 2, (2) compute the means of the two samples ( $M_1$  and  $M_2$ ), and (3) compute the difference between means,  $M_1 - M_2$ . The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the difference between means is:

$$\mu_{M_1 - M_2} = \mu_1 - \mu_2$$

which says that the mean of the distribution of differences between sample means is equal to the difference between population means. For example, say that the mean test score of all 12-year-olds in a population is 34 and the mean of 10-year-olds is 25. If numerous samples were taken from each age group and the mean

difference computed each time, the mean of these numerous differences between sample means would be  $34 - 25 = 9$ .

From the variance sum law, we know that:

$$\sigma_{M_1 - M_2}^2 = \sigma_{M_1}^2 + \sigma_{M_2}^2$$

which says that the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2. Recall the formula for the variance of the sampling distribution of the mean:

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

Since we have two populations and two samples sizes, we need to distinguish between the two variances and sample sizes. We do this by using the subscripts 1 and 2. Using this convention, we can write the formula for the variance of the sampling distribution of the difference between means as:

$$\sigma_{M_1 - M_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Since the standard error of a sampling distribution is the standard deviation of the sampling distribution, the standard error of the difference between means is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Just to review the notation, the symbol on the left contains a sigma ( $\sigma$ ), which means it is a standard deviation. The subscripts  $M_1 - M_2$  indicate that it is the standard deviation of the sampling distribution of  $M_1 - M_2$ .

Now let's look at an application of this formula. Assume there are two species of green beings on Mars. The mean height of Species 1 is 32 while the mean height of Species 2 is 22. The variances of the two species are 60 and 70,

respectively, and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2. What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more? Without doing any calculations, you probably know that the probability is pretty high since the difference in population means is 10. But what exactly is the probability?

First, let's determine the sampling distribution of the difference between means. Using the formulas above, the mean is

$$\mu_{M_1 - M_2} = 32 - 22 = 10$$

The standard error is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$

The sampling distribution is shown in Figure 1. Notice that it is normally distributed with a mean of 10 and a standard deviation of 3.317. The area above 5 is shaded blue.

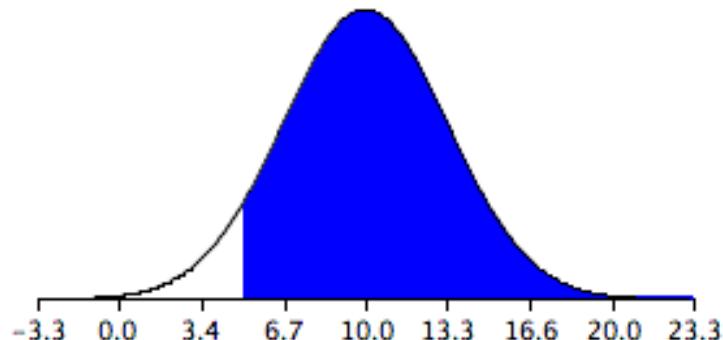


Figure 1. The sampling distribution of the difference between means.

The last step is to determine the area that is shaded blue. Using either a Z table or the normal calculator, the area can be determined to be 0.934. Thus the probability that the mean of the sample from Species 1 will exceed the mean of the sample from Species 2 by 5 or more is 0.934.

As shown below, the formula for the standard error of the difference between means is much simpler if the sample sizes and the population variances

are equal. When the variances and samples sizes are the same, there is no need to use the subscripts 1 and 2 to differentiate these terms.

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

This simplified version of the formula can be used for the following problem: The mean height of 15-year-old boys (in cm) is 175 and the variance is 64. For girls, the mean is 165 and the variance is 64. If eight boys and eight girls were sampled, what is the probability that the mean height of the sample of girls would be higher than the mean height of the sample of boys? In other words, what is the probability that the mean height of girls minus the mean height of boys is greater than 0?

As before, the problem can be solved in terms of the sampling distribution of the difference between means (girls - boys). The mean of the distribution is  $165 - 175 = -10$ . The standard deviation of the distribution is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{(2)(64)}{8}} = 4$$

A graph of the distribution is shown in Figure 2. It is clear that it is unlikely that the mean height for girls would be higher than the mean height for boys since in the population boys are quite a bit taller. Nonetheless it is not inconceivable that the girls' mean could be higher than the boys' mean.

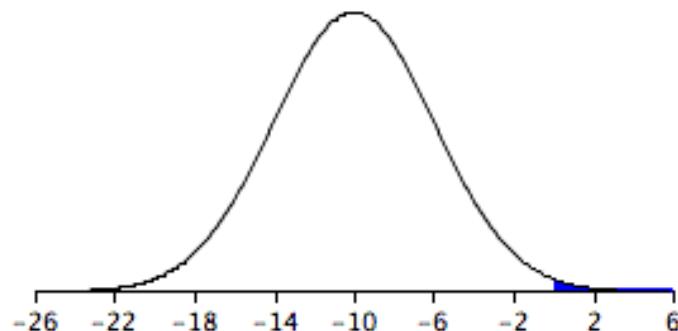


Figure 2. Sampling distribution of the difference between mean heights.

A difference between means of 0 or higher is a difference of  $10/4 = 2.5$  standard deviations above the mean of -10. The probability of a score 2.5 or more standard deviations above the mean is 0.0062.

# Sampling Distribution of Pearson's $r$

by David M. Lane

## *Prerequisites*

- Chapter 4: Values of the Pearson Correlation
- Chapter 9: Introduction to Sampling Distributions

## *Learning Objectives*

1. State how the shape of the sampling distribution of  $r$  deviates from normality
2. Transform  $r$  to  $z'$
3. Compute the standard error of  $z'$
4. Calculate the probability of obtaining an  $r$  above a specified value

Assume that the correlation between quantitative and verbal SAT scores in a given population is 0.60. In other words,  $\rho = 0.60$ . If 12 students were sampled randomly, the sample correlation,  $r$ , would not be exactly equal to 0.60. Naturally different samples of 12 students would yield different values of  $r$ . The distribution of values of  $r$  after repeated samples of 12 students is the sampling distribution of  $r$ .

The shape of the sampling distribution of  $r$  for the above example is shown in Figure 1. You can see that the sampling distribution is not symmetric: it is negatively skewed. The reason for the skew is that  $r$  cannot take on values greater than 1.0 and therefore the distribution cannot extend as far in the positive direction as it can in the negative direction. The greater the value of  $\rho$ , the more pronounced the skew.

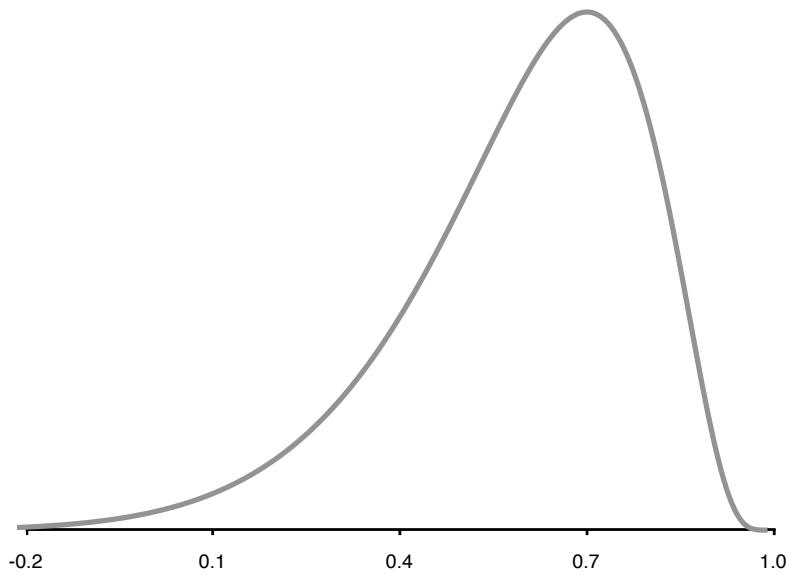
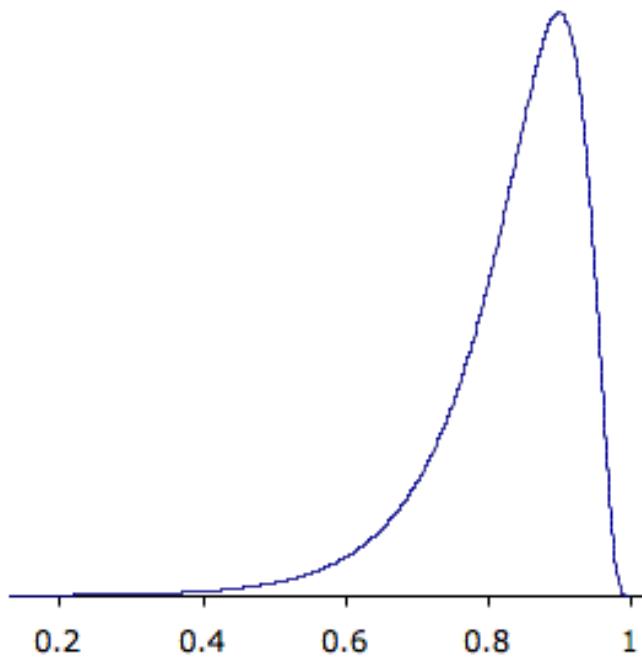


Figure 1. The sampling distribution of  $r$  for  $N = 12$  and  $Q = 0.60$ .

Figure 2 shows the sampling distribution for  $Q = 0.90$ . This distribution has a very short positive tail and a long negative tail.



## Figure 2. The sampling distribution of $r$ for $N = 12$ and $\rho = 0.90$ .

Referring back to the SAT example, suppose you wanted to know the probability that in a sample of 12 students, the sample value of  $r$  would be 0.75 or higher. You might think that all you would need to know to compute this probability is the mean and standard error of the sampling distribution of  $r$ . However, since the sampling distribution is not normal, you would still not be able to solve the problem. Fortunately, the statistician Fisher developed a way to transform  $r$  to a variable that is normally distributed with a known standard error. The variable is called  $z'$  and the formula for the transformation is given below.

$$z' = 0.5 \ln [(1+r) / (1-r)]$$

The details of the formula are not important here since normally you will use either a table or calculator ([external link](#)) to do the transformation. What is important is that  $z'$  is normally distributed and has a standard error of

$$\frac{1}{\sqrt{N - 3}}$$

where  $N$  is the number of pairs of scores.

Let's return to the question of determining the probability of getting a sample correlation of 0.75 or above in a sample of 12 from a population with a correlation of 0.60. The first step is to convert both 0.60 and 0.75 to their  $z'$  values, which are 0.693 and 0.973, respectively. The standard error of  $z'$  for  $N = 12$  is 0.333. Therefore, the question is reduced to the following: given a normal distribution with a mean of 0.693 and a standard deviation of 0.333, what is the probability of obtaining a value of 0.973 or higher? The answer can be found directly from the normal calculator ([external link](#)) to be 0.20. Alternatively, you could use the formula:

$$z = (X - \mu) / \sigma = (0.973 - 0.693) / 0.333 = 0.841$$

and use a table to find that the area above 0.841 is 0.20.

# **Sampling Distribution of p**

by David M. Lane

## *Prerequisites*

- Chapter 5: Binomial Distribution
- Chapter 7: Normal Approximation to the Binomial
- Chapter 9: Introduction to Sampling Distributions

## *Learning Objectives*

1. Compute the mean and standard deviation of the sampling distribution of p
2. State the relationship between the sampling distribution of p and the normal distribution

Assume that in an election race between Candidate A and Candidate B, 0.60 of the voters prefer Candidate A. If a random sample of 10 voters were polled, it is unlikely that exactly 60% of them (6) would prefer Candidate A. By chance the proportion in the sample preferring Candidate A could easily be a little lower than 0.60 or a little higher than 0.60. The sampling distribution of p is the distribution that would result if you repeatedly sampled 10 voters and determined the proportion (p) that favored Candidate A.

The sampling distribution of p is a special case of the sampling distribution of the mean. Table 1 shows a hypothetical random sample of 10 voters. Those who prefer Candidate A are given scores of 1 and those who prefer Candidate B are given scores of 0. Note that seven of the voters prefer candidate A so the sample proportion (p) is

$$p = 7/10 = 0.70$$

As you can see, p is the mean of the 10 preference scores.

Table 1. Sample of voters.

Voter	Preference
1	1
2	0
3	1
4	1
5	1
6	0
7	1
8	0
9	1
10	1

The distribution of  $p$  is closely related to the binomial distribution. The binomial distribution is the distribution of the total number of successes (favoring Candidate A, for example), whereas the distribution of  $p$  is the distribution of the mean number of successes. The mean, of course, is the total divided by the sample size,  $N$ . Therefore, the sampling distribution of  $p$  and the binomial distribution differ in that  $p$  is the mean of the scores (0.70) and the binomial distribution is dealing with the total number of successes (7).

The binomial distribution has a mean of

$$\mu = N\pi$$

Dividing by  $N$  to adjust for the fact that the sampling distribution of  $p$  is dealing with means instead of totals, we find that the mean of the sampling distribution of  $p$  is:

$$\mu_p = \pi$$

The standard deviation of the binomial distribution is:

$$\sqrt{N\pi(1 - \pi)}$$

Dividing by N because p is a mean not a total, we find the standard error of p:

$$\sigma_p = \frac{\sqrt{N\pi(1 - \pi)}}{N} = \sqrt{\frac{\pi(1 - \pi)}{N}}$$

Returning to the voter example,  $\pi = 0.60$  (Don't confuse  $\pi = 0.60$ , the population proportion, with  $p = 0.70$ , the sample proportion) and  $N = 10$ . Therefore, the mean of the sampling distribution of p is 0.60. The standard error is

$$\sigma_p = \sqrt{\frac{0.60(1 - .60)}{10}} = 0.155$$

The sampling distribution of p is a discrete rather than a continuous distribution. For example, with an N of 10, it is possible to have a p of 0.50 or a p of 0.60, but not a p of 0.55.

The sampling distribution of p is approximately normally distributed if N is fairly large and  $\pi$  is not close to 0 or 1. A rule of thumb is that the approximation is good if both  $N\pi$  and  $N(1 - \pi)$  are greater than 10. The sampling distribution for the voter example is shown in Figure 1. Note that even though  $N(1 - \pi)$  is only 4, the approximation is quite good.

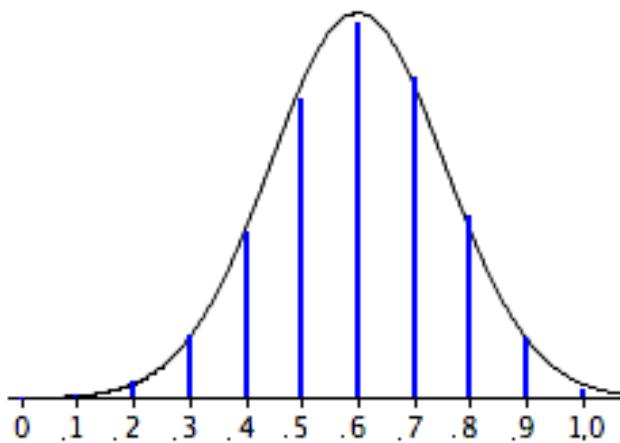


Figure 1. The sampling distribution of p. Vertical bars are the probabilities; the smooth curve is the normal approximation.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 9: Introduction
- Chapter 9: Sampling Distribution of the Mean

The monthly jobs report always gets a lot of attention. Presidential candidates refer to the report when it favors their position. Referring to the August 2012 report in which only 96,000 jobs were created, Republican presidential challenger Mitt Romney stated "the weak jobs report is devastating news for American workers and American families ... a harsh indictment of the president's handling of the economy." When the September 2012 report was released showing 114,000 jobs were created (and the previous report was revised upwards), some supporters of Romney claimed the data were tampered with for political reasons. The most famous statement, "Unbelievable jobs numbers...these Chicago guys will do anything..can't debate so change numbers," was made by former Chairman and CEO of General Electric.

## **What do you think?**

The standard error of the monthly estimate is 100,000. Given that, what do you think of the difference between the two job reports?

The difference between the two reports is very small given that the standard error is 100,000. It is not sensible to take any single jobs report too seriously.

## Exercises

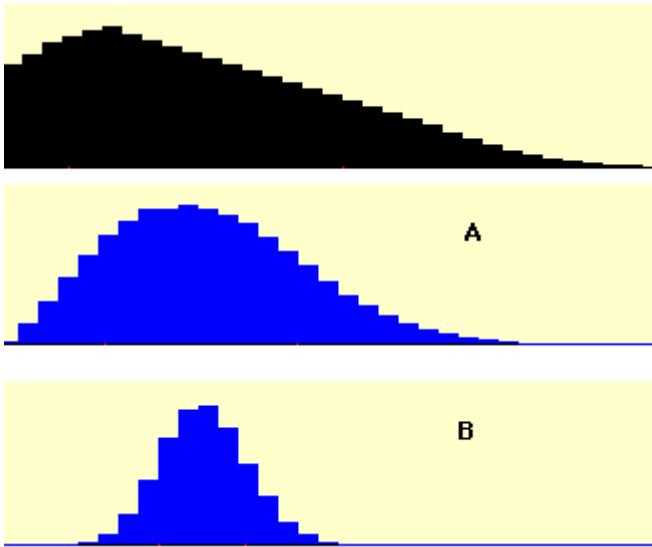
### *Prerequisites*

All material presented in the Sampling Distributions chapter

1. A population has a mean of 50 and a standard deviation of 6. (a) What are the mean and standard deviation of the sampling distribution of the mean for  $N = 16$ ? (b) What are the mean and standard deviation of the sampling distribution of the mean for  $N = 20$ ?
2. Given a test that is normally distributed with a mean of 100 and a standard deviation of 12, find:
  - a. the probability that a single score drawn at random will be greater than 110
  - b. the probability that a sample of 25 scores will have a mean greater than 105
  - c. the probability that a sample of 64 scores will have a mean greater than 105
  - d. the probability that the mean of a sample of 16 scores will be either less than 95 or greater than 105
3. What term refers to the standard deviation of a sampling distribution?
4. (a) If the standard error of the mean is 10 for  $N = 12$ , what is the standard error of the mean for  $N = 22$ ? (b) If the standard error of the mean is 50 for  $N = 25$ , what is it for  $N = 64$ ?
5. A questionnaire is developed to assess women's and men's attitudes toward using animals in research. One question asks whether animal research is wrong and is answered on a 7-point scale. Assume that in the population, the mean for women is 5, the mean for men is 4, and the standard deviation for both groups is 1.5. Assume the scores are normally distributed. If 12 women and 12 men are selected randomly, what is the probability that the mean of the women will be more than 2 points higher than the mean of the men?
6. If the correlation between reading achievement and math achievement in the population of fifth graders were 0.60, what would be the probability that in a sample of 28 students, the sample correlation coefficient would be greater than 0.65?

7. If numerous samples of  $N = 15$  are taken from a uniform distribution and a relative frequency distribution of the means is drawn, what would be the shape of the frequency distribution?
8. A normal distribution has a mean of 20 and a standard deviation of 10. Two scores are sampled randomly from the distribution and the second score is subtracted from the first. What is the probability that the difference score will be greater than 5? Hint: Read the Variance Sum Law section of Chapter 3.
9. What is the shape of the sampling distribution of  $r$ ? In what way does the shape depend on the size of the population correlation?
10. If you sample one number from a standard normal distribution, what is the probability it will be 0.5?
11. A variable is normally distributed with a mean of 120 and a standard deviation of 5. Four scores are randomly sampled. What is the probability that the mean of the four scores is above 127?
12. The correlation between self-esteem and extraversion is .30. A sample of 84 is taken. a. What is the probability that the correlation will be less than 0.10? b. What is the probability that the correlation will be greater than 0.25?
13. The mean GPA for students in School A is 3.0; the mean GPA for students in School B is 2.8. The standard deviation in both schools is 0.25. The GPAs of both schools are normally distributed. If 9 students are randomly sampled from each school, what is the probability that:  
a. the sample mean for School A will exceed that of School B by 0.5 or more?  
b. the sample mean for School B will be greater than the sample mean for School A?
14. In a city, 70% of the people prefer Candidate A. Suppose 30 people from this city were sampled.  
a. What is the mean of the sampling distribution of  $p$ ?  
b. What is the standard error of  $p$ ?

- c. What is the probability that 80% or more of this sample will prefer Candidate A?
15. When solving problems where you need the sampling distribution of  $r$ , what is the reason for converting from  $r$  to  $z'$ ?
16. In the population, the mean SAT score is 1000. Would you be more likely (or equally likely) to get a sample mean of 1200 if you randomly sampled 10 students or if you randomly sampled 30 students? Explain.
17. True/false: The standard error of the mean is smaller when  $N = 20$  than when  $N = 10$ .
18. True/false: The sampling distribution of  $r = .8$  becomes normal as  $N$  increases.
19. True/false: You choose 20 students from the population and calculate the mean of their test scores. You repeat this process 100 times and plot the distribution of the means. In this case, the sample size is 100.
20. True/false: In your school, 40% of students watch TV at night. You randomly ask 5 students every day if they watch TV at night. Every day, you would find that 2 of the 5 do watch TV at night.
21. True/false: The median has a sampling distribution.
22. True/false: Refer to the figure below. The population distribution is shown in black, and its corresponding sampling distribution of the mean for  $N = 10$  is labeled “A.”



*Questions from Case Studies*

Angry Moods (AM) case study

23. (AM)

- a. How many men were sampled?
- b. How many women were sampled?

24. (AM) What is the mean difference between men and women on the Anger-Out scores?

25. (AM) Suppose in the population, the Anger-Out score for men is two points higher than it is for women. The population variances for men and women are both 20. Assume the Anger-Out scores for both genders are normally distributed. Given this information about the population parameters:

- (a) What is the mean of the sampling distribution of the difference between means?
- (b) What is the standard error of the difference between means?
- (c) What is the probability that you would have gotten this mean difference (see #24) or less in your sample?

Animal Research (AR) case study

26. (AR) How many people were sampled to give their opinions on animal research?
27. (AR) What is the correlation in this sample between the belief that animal research is wrong and belief that animal research is necessary?
28. (AR) Suppose the correlation between the belief that animal research is wrong and the belief that animal research is necessary is  $-.68$  in the population.
- (a) Convert  $-.68$  to  $z'$ .
  - (b) Find the standard error of this sampling distribution.
  - (c) Assuming the data used in this study was randomly sampled, what is the probability that you would get this correlation or stronger (closer to  $-1$ )?

# 10. Estimation

- A. Introduction
- B. Degrees of Freedom
- C. Characteristics of Estimators
- D. Confidence Intervals
  - 1. Introduction
  - 2. Confidence Interval for the Mean
  - 3. t distribution
  - 4. Confidence Interval for the Difference Between Means
  - 5. Confidence Interval for Pearson's Correlation
  - 6. Confidence Interval for a Proportion

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

The concept of degrees of freedom and its relationship to estimation is discussed in Section B. “Characteristics of Estimators” discusses two important concepts: bias and precision.

Point estimates are usually supplemented by interval estimates called confidence intervals. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval:  $0.46 < \pi < 0.60$ . The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%. The sections on confidence intervals show how to compute confidence intervals for a variety of parameters.

# Introduction to Estimation

by David M. Lane

## *Prerequisites*

- Chapter 3 Measures of Central Tendency
- Chapter 3: Variability

## *Learning Objectives*

1. Define statistic
2. Define parameter
3. Define point estimate
4. Define interval estimate
5. Define margin of error

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

Point estimates are usually supplemented by interval estimates called confidence intervals. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval:  $0.46 < \pi < 0.60$ . The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

In an experiment on memory for chess positions, the mean recall for tournament players was 63.8 and the mean for non-players was 33.1. Therefore a point estimate of the difference between population means is 30.7. The 95% confidence interval on the difference between means extends from 19.05 to 42.35. You will see how to compute this kind of interval in another section.

# Degrees of Freedom

by David M. Lane

## *Prerequisites*

- Chapter 3: Measures of Variability
- Chapter 10: Introduction to Estimation

## *Learning Objectives*

1. Define degrees of freedom
2. Estimate the variance from a sample of 1 if the population mean is known
3. State why deviations from the sample mean are not independent
4. State the general formula for degrees of freedom in terms of the number of values and the number of estimated parameters
5. Calculate  $s^2$

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean  $(8-6)^2 = 4$  is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df. If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance,  $(5-6)^2 = 1$ . We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on two independently and randomly selected Martians. The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

As you are probably thinking, it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean ( $\mu$ ) with the sample mean (M). The process of estimating the mean affects our degrees of freedom as shown below.

Returning to our problem of estimating the variance in Martian heights, let's assume we do not know the population mean and therefore we have to estimate it from the sample. We have sampled two Martians and found that their heights are 8 and 5. Therefore M, our estimate of the population mean, is

$$M = (8+5)/2 = 6.5.$$

We can now compute two estimates of variance:

$$\text{Estimate 1} = (8-6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5-6.5)^2 = 2.25$$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of M. Since the first Martian's height of 8 influenced M, it also influenced Estimate 2. If the first height had been, for example, 10, then M would have been 7.5 and Estimate 2 would have been  $(5-7.5)^2 = 6.25$  instead of 2.25. The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be  $13-5 = 8$ .

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter ( $\mu$ ) on the way to estimating the parameter of interest ( $\sigma^2$ ). Therefore, the estimate of variance has  $2 - 1 = 1$  degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to  $N - 1$  where N is the number of observations.

Recall from the section on variability that the formula for estimating the variance in a sample is:

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

The denominator of this formula is the degrees of freedom.

# Characteristics of Estimators

by David M. Lane

## *Prerequisites*

- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability
- Chapter 9: Introduction to Sampling Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 10: Introduction to Estimation
- Chapter 10: Degrees of Freedom

## *Learning Objectives*

1. Define bias
2. Define sampling variability
3. Define expected value
4. Define relative efficiency

This section discusses two important characteristics of statistics used as point estimates of parameters: bias and sampling variability. Bias refers to whether an estimator tends to either over or underestimate the parameter. Sampling variability refers to how much the estimate varies from sample to sample.

Have you ever noticed that some bathroom scales give you very different weights each time you weigh yourself? With this in mind, let's compare two scales. Scale 1 is a very high-tech digital scale and gives essentially the same weight each time you weigh yourself; it varies by at most 0.02 pounds from weighing to weighing. Although this scale has the potential to be very accurate, it is calibrated incorrectly and, on average, overstates your weight by one pound. Scale 2 is a cheap scale and gives very different results from weighing to weighing. However, it is just as likely to underestimate as overestimate your weight. Sometimes it vastly overestimates it and sometimes it vastly underestimates it. However, the average of a large number of measurements would be your actual weight. Scale 1 is biased since, on average, its measurements are one pound higher than your actual weight. Scale 2, by contrast, gives unbiased estimates of your weight. However, Scale 2 is highly variable and its measurements are often very far from

your true weight. Scale 1, in spite of being biased, is fairly accurate. Its measurements are never more than 1.02 pounds from your actual weight.

We now turn to more formal definitions of variability and precision. However, the basic ideas are the same as in the bathroom scale example.

## Bias

A statistic is biased if the long-term average value of the statistic is not the parameter it is estimating. More formally, a statistic is biased if the mean of the sampling distribution of the statistic is not equal to the parameter. The mean of the sampling distribution of a statistic is sometimes referred to as the expected value of the statistic.

As we saw in the section on the sampling distribution of the mean, the mean of the sampling distribution of the (sample) mean is the population mean ( $\mu$ ). Therefore the sample mean is an unbiased estimate of  $\mu$ . Any given sample mean may underestimate or overestimate  $\mu$ , but there is no systematic tendency for sample means to either under or overestimate  $\mu$ .

In the section on variability, we saw that the formula for the variance in a population is

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

whereas the formula to estimate the variance from a sample is

$$s^2 = \frac{\sum(X - M)^2}{N - 1}$$

Notice that the denominators of the formulas are different:  $N$  for the population and  $N-1$  for the sample. If  $N$  is used in the formula for  $s^2$ , then the estimates tend to be too low and therefore biased. The formula with  $N-1$  in the denominator gives an unbiased estimate of the population variance. Note that  $N-1$  is the degrees of freedom.

## Sampling Variability

The sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its standard error ; the smaller the standard error, the less the sampling variability. For example, the standard error of

the mean is a measure of the sampling variability of the mean. Recall that the formula for the variance of the sampling distribution of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

The larger the sample size (N), the smaller the standard error of the mean and therefore the lower the sampling variability.

Statistics differ in their sampling variability even with the same sample size. For example, for normal distributions, the standard error of the median is larger than the standard error of the mean. The smaller the standard error of a statistic, the more efficient the statistic. The relative efficiency of two statistics is typically defined as the ratio of their standard errors. However, it is sometimes defined as the ratio of their squared standard errors.

# **Confidence Intervals**

by David M. Lane

- A. Introduction
- B. Confidence Interval for the Mean
- C. t distribution
- D. Confidence Interval for the Difference Between Means
- E. Confidence Interval for Pearson's Correlation
- F. Confidence Interval for a Proportion

These sections show how to compute confidence intervals for a variety of parameters.

# Introduction to Confidence Intervals

by David M. Lane

## *Prerequisites*

- Chapter 5: Introduction to Probability
- Chapter 10: Introduction to Estimation
- Chapter 10: Characteristics of Estimators

## *Learning Objectives*

1. Define confidence interval
2. State why a confidence interval is not the probability the interval contains the parameter

Say you were interested in the mean weight of 10-year-old girls living in the United States. Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds. This sample mean of 90 is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean. For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure (presented in the next section) that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

Naturally, 5% of the intervals would not contain the population mean.

It is natural to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple. One problem is that the computation of a confidence interval does not take into account any other information you might have about the value of the population mean. For example, if numerous prior studies had all found sample means above 110, it would not make sense to conclude that there is a 0.95 probability that the population mean is between 72.85 and 107.15. What about situations in which there is no prior information about the value of the population mean? Even here the interpretation is complex. The problem is that there can be more than one procedure that produces intervals that contain the population parameter 95% of the time. Which procedure produces the “true” 95% confidence interval? Although the various methods are equal from a purely mathematical point of view, the standard method of computing confidence intervals has two desirable properties: each interval is symmetric about the point estimate and each interval is contiguous. Recall from the introductory section in the chapter on probability that, for some purposes, probability is best thought of as subjective. It is reasonable, although not required by the laws of probability, that one adopt a subjective probability of 0.95 that a 95% confidence interval, as typically computed, contains the parameter in question.

Confidence intervals can be computed for various parameters, not just the mean. For example, later in this chapter you will see how to compute a confidence interval for  $\rho$ , the population value of Pearson's  $r$ , based on sample data.

# t Distribution

by David M. Lane

## *Prerequisites*

- Chapter 7: Normal Distribution,
- Chapter 7: Areas Under Normal Distributions
- Chapter 10: Degrees of Freedom

## *Learning Objectives*

1. State the difference between the shape of the t distribution and the normal distribution
2. State how the difference between the shape of the t distribution and normal distribution is affected by the degrees of freedom
3. Use a t table to find the value of t to use in a confidence interval
4. Use the t calculator to find the value of t to use in a confidence interval

In the introduction to normal distributions it was shown that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean. Therefore, if you randomly sampled a value from a normal distribution with a mean of 100, the probability it would be within  $1.96\sigma$  of 100 is 0.95. Similarly, if you sample N values from the population, the probability that the sample mean (M) will be within  $1.96 \sigma_M$  of 100 is 0.95.

Now consider the case in which you have a normal distribution but you do not know the standard deviation. You sample N values and compute the sample mean (M) and estimate the standard error of the mean ( $\sigma_M$ ) with  $s_M$ . What is the probability that M will be within 1.96  $s_M$  of the population mean ( $\mu$ )? This is a difficult problem because there are two ways in which M could be more than 1.96  $s_M$  from  $\mu$ : (1) M could, by chance, be either very high or very low and (2)  $s_M$  could, by chance, be very low. Intuitively, it makes sense that the probability of being within 1.96 standard errors of the mean should be smaller than in the case when the standard deviation is known (and cannot be underestimated). But exactly how much smaller? Fortunately, the way to work out this type of problem was solved in the early 20th century by W. S. Gosset who determined the distribution of a mean divided by its estimate of the standard error. This distribution is called the Student's t distribution or sometimes just the t distribution. Gosset worked out the t

distribution and associated statistical tests while working for a brewery in Ireland. Because of a contractual agreement with the brewery, he published the article under the pseudonym “Student.” That is why the t test is called the “Student's t test.”

The t distribution is very similar to the normal distribution when the estimate of variance is based on many degrees of freedom, but has relatively more scores in its tails when there are fewer degrees of freedom. Figure 1 shows t distributions with 2, 4, and 10 degrees of freedom and the standard normal distribution. Notice that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t distribution is therefore leptokurtic. The t distribution approaches the normal distribution as the degrees of freedom increase.

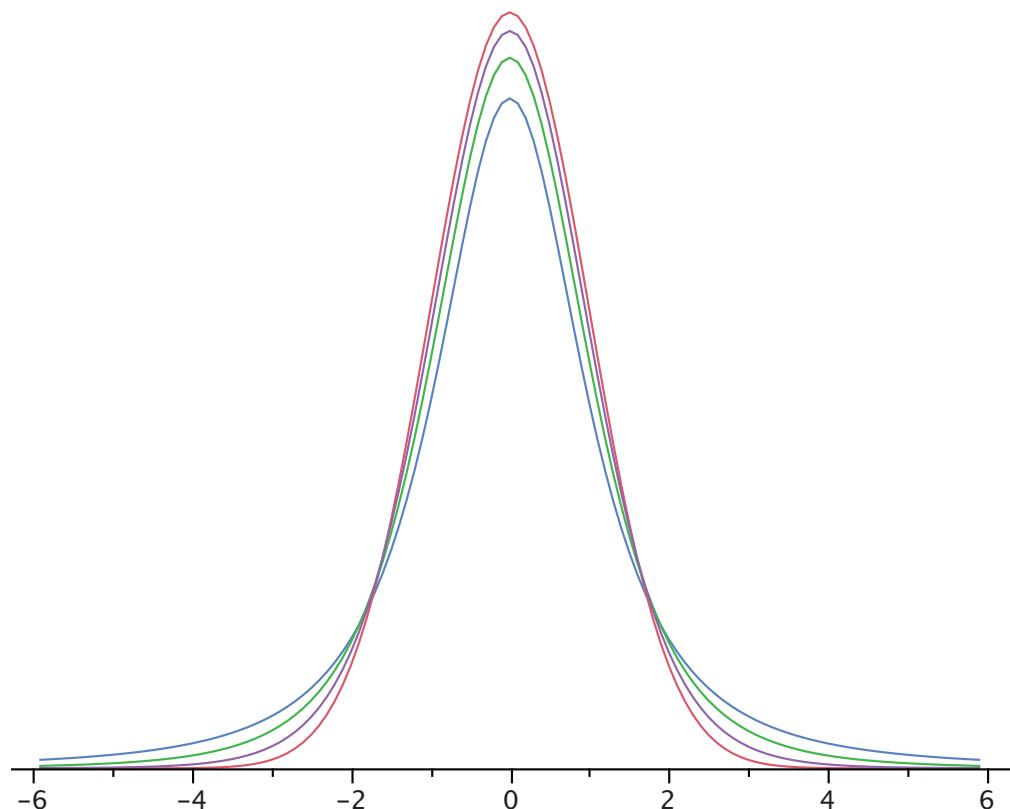


Figure 1. A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the lowest is the standard normal distribution.

Since the t distribution is leptokurtic, the percentage of the distribution within 1.96 standard deviations of the mean is less than the 95% for the normal distribution. Table 1 shows the number of standard deviations from the mean required to contain 95% and 99% of the area of the t distribution for various degrees of freedom. These are the values of t that you use in a confidence interval. The corresponding values for the normal distribution are 1.96 and 2.58 respectively. Notice that with few degrees of freedom, the values of t are much higher than the corresponding values for a normal distribution and that the difference decreases as the degrees of freedom increase. The values in Table 1 can be obtained from the “Find t for a confidence interval” calculator.

Table 1. Abbreviated t table.

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Returning to the problem posed at the beginning of this section, suppose you sampled 9 values from a normal population and estimated the standard error of the mean ( $\sigma_M$ ) with  $s_M$ . What is the probability that  $M$  would be within 1.96 $s_M$  of  $\mu$ ? Since the sample size is 9, there are  $N - 1 = 8$  df. From Table 1 you can see that with 8 df the probability is 0.95 that the mean will be within 2.306  $s_M$  of  $\mu$ . The probability that it will be within 1.96  $s_M$  of  $\mu$  is therefore lower than 0.95.

A “t distribution” calculator can be used to find that 0.086 of the area of a t distribution is more than 1.96 standard deviations from the mean, so the probability that  $M$  would be less than 1.96 $s_M$  from  $\mu$  is  $1 - 0.086 = 0.914$ .

As expected, this probability is less than 0.95 that would have been obtained if  $\sigma_M$  had been known instead of estimated.

# Confidence Interval for the Mean

by David M. Lane

## *Prerequisites*

- Chapter 7: Areas Under Normal Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 10: Introduction to Estimation
- Chapter 10: Introduction to Confidence Intervals
- Chapter 10: t distribution

## *Learning Objectives*

1. Use the inverse normal distribution calculator to find the value of  $z$  to use for a confidence interval
2. Compute a confidence interval on the mean when  $\sigma$  is known
3. Determine whether to use a  $t$  distribution or a normal distribution
4. Compute a confidence interval on the mean when  $\sigma$  is estimated

When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean, there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from the section on the sampling distribution of the mean that the mean of the sampling distribution is  $\mu$  and the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of  $36/3 = 12$ . Note that the standard deviation of a sampling distribution is its standard error. Figure 1 shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48

to 113.52. These limits were computed by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

$$\begin{aligned} 90 - (1.96)(12) &= 66.48 \\ 90 + (1.96)(12) &= 113.52 \end{aligned}$$

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean; 12 is the standard error of the mean.

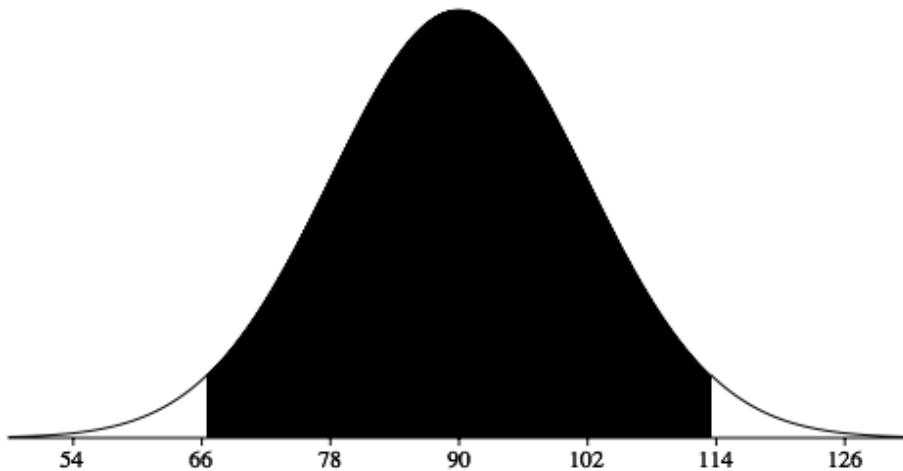


Figure 1. The sampling distribution of the mean for  $N=9$ . The middle 95% of the distribution is shaded.

Figure 1 shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the mean ( $M$ ) from a sample, and create an interval ranging from  $M - 23.52$  to  $M + 23.52$ , this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

$$\begin{aligned} \text{Lower limit} &= M - Z \cdot .95\sigma_m \\ \text{Upper limit} &= M + Z \cdot .95\sigma_m \end{aligned}$$

where  $Z_{.95}$  is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area and  $\sigma_M$  is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the standard deviation ( $\sigma$ ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when  $\sigma$  is known is easier than when  $\sigma$  has to be estimated, and serves a pedagogical purpose. Later in this section we will show how to compute a confidence interval for the mean when  $\sigma$  has to be estimated.

Suppose the following five numbers were sampled from a normal distribution with a standard deviation of 2.5: 2, 3, 5, 6, and 9. To compute the 95% confidence interval, start by computing the mean and standard error:

$$M = (2 + 3 + 5 + 6 + 9) / 5 = 5.$$

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{2.5}{\sqrt{5}} = 1.118$$

$Z_{.95}$  can be found using the normal distribution calculator and specifying that the area is 0.95 and indicating that you want the area to be between the cutoff points. The value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been 2.58.

The confidence interval can then be computed as follows:

$$\begin{aligned} \text{Lower limit} &= 5 - (1.96)(1.118) = 2.81 \\ \text{Upper limit} &= 5 + (1.96)(1.118) = 7.19 \end{aligned}$$

You should use the t distribution rather than the normal distribution when the variance is not known and has to be estimated from sample data. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution is leptokurtic, which means it has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78

standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of  $t$  to be used in a confidence interval can be looked up in a table of the  $t$  distribution. A small version of such a table is shown in Table 1. The first column,  $df$ , stands for degrees of freedom, and for confidence intervals on the mean,  $df$  is equal to  $N - 1$ , where  $N$  is the sample size.

Table 1. Abbreviated  $t$  table.

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the “inverse  $t$  distribution” calculator to find the  $t$  values to use in confidence intervals.

Assume that the following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9 and that the standard deviation is not known. The first steps are to compute the sample mean and variance:

$$\begin{aligned} M &= 5 \\ s^2 &= 7.5 \end{aligned}$$

The next step is to estimate the standard error of the mean. If we knew the population variance, we could use the following formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Instead we compute an estimate of the standard error ( $s_M$ ):

$$s_m = \frac{s}{\sqrt{N}} = 1.225$$

The next step is to find the value of  $t$ . As you can see from Table 1, the value for the 95% confidence interval for  $df = N - 1 = 4$  is 2.776. The confidence interval is then computed just as it is with  $\sigma_M$ . The only differences are that  $s_M$  and  $t$  rather than  $\sigma_M$  and  $Z$  are used.

$$\begin{aligned} \text{Lower limit} &= 5 - (2.776)(1.225) = 1.60 \\ \text{Upper limit} &= 5 + (2.776)(1.225) = 8.40 \end{aligned}$$

More generally, the formula for the 95% confidence interval on the mean is:

$$\begin{aligned} \text{Lower limit} &= M - (t_{CL})(s_M) \\ \text{Upper limit} &= M + (t_{CL})(s_M) \end{aligned}$$

where  $M$  is the sample mean,  $t_{CL}$  is the  $t$  for the confidence level desired (0.95 in the above example), and  $s_M$  is the estimated standard error of the mean.

We will finish with an analysis of the Stroop Data. Specifically, we will compute a confidence interval on the mean difference score. Recall that 47 subjects named the color of ink that words were written in. The names conflicted so that, for example, they would name the ink color of the word “blue” written in red ink. The correct response is to say “red” and ignore the fact that the word is “blue.” In a second condition, subjects named the ink color of colored rectangles.

Table 2. Response times in seconds for 10 subjects.

Naming Colored Rectangle	Interference	Difference
17	38	21
15	58	43
18	35	17
20	39	19
18	33	15
20	32	12
20	45	25
19	52	33
17	31	14
21	29	8

Table 2 shows the time difference between the interference and color-naming conditions for 10 of the 47 subjects. The mean time difference for all 47 subjects is 16.362 seconds and the standard deviation is 7.470 seconds. The standard error of the mean is 1.090. A t table shows the critical value of t for  $47 - 1 = 46$  degrees of freedom is 2.013 (for a 95% confidence interval). Therefore the confidence interval is computed as follows:

$$\begin{aligned} \text{Lower limit} &= 16.362 - (2.013)(1.090) = 14.17 \\ \text{Upper limit} &= 16.362 + (2.013)(1.090) = 18.56 \end{aligned}$$

Therefore, the interference effect (difference) for the whole population is likely to be between 14.17 and 18.56 seconds.

# Difference between Means

by David M. Lane

## *Prerequisites*

- Chapter 9: Sampling Distribution of Difference between Means
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Mean

## *Learning Objectives*

1. State the assumptions for computing a confidence interval on the difference between means
2. Compute a confidence interval on the difference between means
3. Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. We take as an example the data from the “Animal Research” case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 1.

Table 1. Means and Variances in Animal Research study.

Condition	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is the difference in the population. The difference in sample means is used to estimate the difference in population means. The accuracy of the estimate is revealed by a confidence interval.

In order to construct a confidence interval, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.

3. Each value is sampled independently from each other value.

The consequences of violating these assumptions are discussed in Chapter 12. For now, suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference.

A confidence interval on the difference between means is computed using the following formula:

$$\text{Lower Limit} = M_1 - M_2 - (t_{CL})(S_{M_1 - M_2})$$

$$\text{Upper Limit} = M_1 - M_2 + (t_{CL})(S_{M_1 - M_2})$$

where  $M_1 - M_2$  is the difference between sample means,  $t_{CL}$  is the  $t$  for the desired level of confidence, and  $(S_{M_1 - M_2})$  is the estimated standard error of the difference between sample means. The meanings of these terms will be made clearer as the calculations are demonstrated.

We continue to use the data from the “Animal Research” case study and will compute a confidence interval on the difference between the mean score of the females and the mean score of the males. For this calculation, we will assume that the variances in each of the two populations are equal.

The first step is to compute the estimate of the standard error of the difference between means  $(S_{M_1 - M_2})$ . Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference in means in the population is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate  $\sigma^2$  and use that estimate in place of  $\sigma^2$ . Since we are assuming the population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where  $MSE$  is our estimate of  $\sigma^2$ . In this example,

$$MSE = (2.743 + 2.985)/2 = 2.864.$$

Note that  $MSE$  stands for “mean square error” and is the mean squared deviation of each score from its group’s mean.

Since  $n$  (the number of scores in each condition) is 17,

$$s_{M_1-M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805.$$

The next step is to find the  $t$  to use for the confidence interval ( $t_{CL}$ ). To calculate  $t_{CL}$ , we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which  $MSE$  is based. This is equal to  $(n_1 - 1) + (n_2 - 1)$  where  $n_1$  is the sample size of the first group and  $n_2$  is the sample size of the second group. For this example,  $n_1 = n_2 = 17$ . When  $n_1 = n_2$ , it is conventional to use “ $n$ ” to refer to the sample size of each group. Therefore, the degrees of freedom is  $16 + 16 = 32$ .

From either the above calculator or a  $t$  table, you can find that the  $t$  for a 95% confidence interval for 32 df is 2.037.

We now have all the components needed to compute the confidence interval. First, we know the difference between means:

$$M_1 - M_2 = 5.353 - 3.882 = 1.471$$

We know the standard error of the difference between means is

$$s_{M_1-M_2} = 0.5805$$

and that the  $t$  for the 95% confidence interval with 32 df is

$$t_{CL} = 2.037$$

Therefore, the 95% confidence interval is

$$\text{Lower Limit} = 1.471 - (2.037)(0.5805) = 0.29$$

$$\text{Upper Limit} = 1.471 + (2.037)(0.5805) = 2.65$$

We can write the confidence interval as:

$$0.29 \leq \mu_f - \mu_m \leq 2.65$$

where  $\mu_f$  is the population mean for females and  $\mu_m$  is the population mean for males. This analysis provides evidence that the mean for females is higher than the mean for males, and that the difference between means in the population is likely to be between 0.29 and 2.65.

## Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 2.

Table 2. Example Data

Group 1	Group 2
3	5
4	6
5	7

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. For the data in Table 2, the reformatted data look as follows:

Table 3. Reformatted Data

G	Y
1	3
1	4
1	5
2	5
2	6
2	7

### Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the sample with the larger sample size more than the sample with the smaller sample size. Computationally this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2$$

where  $M_1$  is the mean for group 1 and  $M_2$  is the mean for group 2. Consider the following small example:

Table 4. Example Data

Group 1	Group 2
3	2
4	4
5	

$$M_1 = 4 \text{ and } M_2 = 3.$$

$$SSE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4$$

Then, MSE is computed by:

$$\text{MSE} = \text{SSE}/\text{df}$$

where the degrees of freedom (df) is computed as before:

$$\begin{aligned}\text{df} &= (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3. \\ \text{MSE} &= \text{SSE}/\text{df} = 4/3 = 1.333.\end{aligned}$$

The formula

$$s_{M_1-M_2} = \sqrt{\frac{2\text{MSE}}{n}}$$

is replaced by

$$s_{M_1-M_2} = \sqrt{\frac{2\text{MSE}}{n_h}}$$

where  $n_h$  is the harmonic mean of the sample sizes and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4$$

and

$$s_{M_1-M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054.$$

$t_{\text{CL}}$  for 3 df and the 0.05 level = 3.182.

Therefore the 95% confidence interval is

$$\text{Lower Limit} = 1 - (3.182)(1.054) = -2.35$$

$$\text{Upper Limit} = 1 + (3.182)(1.054) = 4.35$$

We can write the confidence interval as:

$$-2.35 \leq \mu_1 - \mu_2 \leq 4.35$$

# Correlation

by David M. Lane

## *Prerequisites*

- Chapter 4: Values of the Pearson Correlation
- Chapter 9: Sampling Distribution of Pearson's  $r$
- Chapter 10: Confidence Intervals

## *Learning Objectives*

1. State why the  $z'$  transformation is necessary
2. Compute the standard error of  $z'$
3. Compute a confidence interval on  $\rho$

The computation of a confidence interval on the population value of Pearson's correlation ( $\rho$ ) is complicated by the fact that the sampling distribution of  $r$  is not normally distributed. The solution lies with Fisher's  $z'$  transformation described in the section on the sampling distribution of Pearson's  $r$ . The steps in computing a confidence interval for  $\rho$  are:

1. Convert  $r$  to  $z'$
2. Compute a confidence interval in terms of  $z'$
3. Convert the confidence interval back to  $r$ .

Let's take the data from the case study Animal Research as an example. In this study, students were asked to rate the degree to which they thought animal research is wrong and the degree to which they thought it is necessary. As you might have expected, there was a negative relationship between these two variables: the more that students thought animal research is wrong, the less they thought it is necessary. The correlation based on 34 observations is -0.654. The problem is to compute a 95% confidence interval on  $\rho$  based on this  $r$  of -0.654.

The conversion of  $r$  to  $z'$  can be done using a [calculator](#). This calculator shows that the  $z'$  associated with an  $r$  of -0.654 is -0.78.

The sampling distribution of  $z'$  is approximately normally distributed and has a standard error of

$$\frac{1}{\sqrt{N - 3}}$$

For this example,  $N = 34$  and therefore the standard error is 0.180. The  $Z$  for a 95% confidence interval ( $Z_{.95}$ ) is 1.96, as can be found using the normal distribution calculator (setting the shaded area to .95 and clicking on the “Between” button). The confidence interval is therefore computed as:

$$\begin{aligned}\text{Lower limit} &= -0.78 - (1.96)(0.18) = -1.13 \\ \text{Upper limit} &= -0.78 + (1.96)(0.18) = -0.43\end{aligned}$$

The final step is to convert the endpoints of the interval back to  $r$  using a table or the calculator. The  $r$  associated with a  $z'$  of -1.13 is -0.81 and the  $r$  associated with a  $z'$  of -0.43 is -0.40. Therefore, the population correlation ( $\rho$ ) is likely to be between -0.81 and -0.40. The 95% confidence interval is:

$$-0.81 \leq \rho \leq -0.40$$

To calculate the 99% confidence interval, you use the  $Z$  for a 99% confidence interval of 2.58 as follows:

$$\begin{aligned}\text{Lower limit} &= -0.775 - (2.58)(0.18) = -1.24 \\ \text{Upper limit} &= -0.775 + (2.58)(0.18) = -0.32\end{aligned}$$

Converting back to  $r$ , the confidence interval is:

$$-0.84 \leq \rho \leq -0.31$$

Naturally, the 99% confidence interval is wider than the 95% confidence interval.

# Proportion

by David M. Lane

## *Prerequisites*

- Chapter 7: Introduction to the Normal Distribution
- Chapter 7: Normal Approximation to the Binomial
- Chapter 9: Sampling Distribution of the Mean
- Chapter 9: Sampling Distribution of a Proportion
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Mean

## *Learning Objectives*

1. Estimate the population proportion from sample proportions
2. Apply the correction for continuity
3. Compute a confidence interval

A candidate in a two-person election commissions a poll to determine who is ahead. The pollster randomly chooses 500 registered voters and determines that 260 out of the 500 favor the candidate. In other words, 0.52 of the sample favors the candidate. Although this point estimate of the proportion is informative, it is important to also compute a confidence interval. The confidence interval is computed based on the mean and standard deviation of the sampling distribution of a proportion. The formulas for these two parameters are shown below:

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{N}}$$

Since we do not know the population parameter  $\pi$ , we use the sample proportion  $p$  as an estimate. The estimated standard error of  $p$  is therefore

$$s_p = \sqrt{\frac{p(1 - p)}{N}}$$

We start by taking our statistic ( $p$ ) and creating an interval that ranges  $(Z_{.95})(s_p)$  in both directions where  $Z_{.95}$  is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area. (See the section on the confidence interval for the mean). The value of  $Z_{.95}$  is computed with the normal calculator and is equal to 1.96. We then make a slight adjustment to correct for the fact that the distribution is discrete rather than continuous.

$s_p$  is calculated as shown below:

$$s_p = \sqrt{\frac{.52(1 - .52)}{500}} = 0.0223$$

To correct for the fact that we are approximating a discrete distribution with a continuous distribution (the normal distribution), we subtract  $0.5/N$  from the lower limit and add  $0.5/N$  to the upper limit of the interval. Therefore the confidence interval is

$$p \pm Z_{.95} \sqrt{\frac{p(1 - p)}{N}} \pm \frac{0.5}{N}$$

$$\text{Lower: } 0.52 - (1.96)(0.0223) - 0.001 = 0.475$$

$$\text{Upper: } 0.52 + (1.96)(0.0223) + 0.001 = 0.565$$

$$.475 \leq \pi \leq .565$$

Since the interval extends 0.045 in both directions, the margin of error is 0.045. In terms of percent, between 47.5% and 56.5% of the voters favor the candidate and the margin of error is 4.5%. Keep in mind that the margin of error of 4.5% is the margin of error for the percent favoring the candidate and not the margin of error for the difference between the percent favoring the candidate and the percent favoring the opponent. The margin of error for the difference is 9%, twice the margin of error for the individual percent. Keep this in mind when you hear reports in the media; the media often get this wrong.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 10: Proportions

In July of 2011, Gene Munster of Piper Jaffray reported the results of a survey in a note to clients. This research was reported throughout the media. Perhaps the fullest description was presented on the CNNMoney website (A service of CNN, Fortune, and Money) in an article entitled "Survey: iPhone retention 94% vs. Android 47%." The data were collected by asking people in food courts and baseball stadiums what their current phone was and what phone they planned to buy next. The data were collected in the summer of 2011. Below is a portion of the data:

Phone	Keep	Change	Proportion
iPhone	58	4	0.94
Android	17	19	0.47

## What do you think?

The article contains the strong caution: "It's only a tiny sample, so large conclusions must not be drawn." This caution appears to be a welcome change from the overstating of findings typically found in the media. But has this report understated the importance of the study? Perhaps it is valid to draw some "large conclusions."?

The confidence interval on the proportion extends from 0.87 to 1.0 (some methods give the interval from 0.85 to 0.97). Even the lower bound indicates the vast majority of iPhone owners plan to buy another iPhone. A strong conclusion can be made even with this sample size.



## Exercises

### *Prerequisites*

- All material presented in the Estimation Chapter

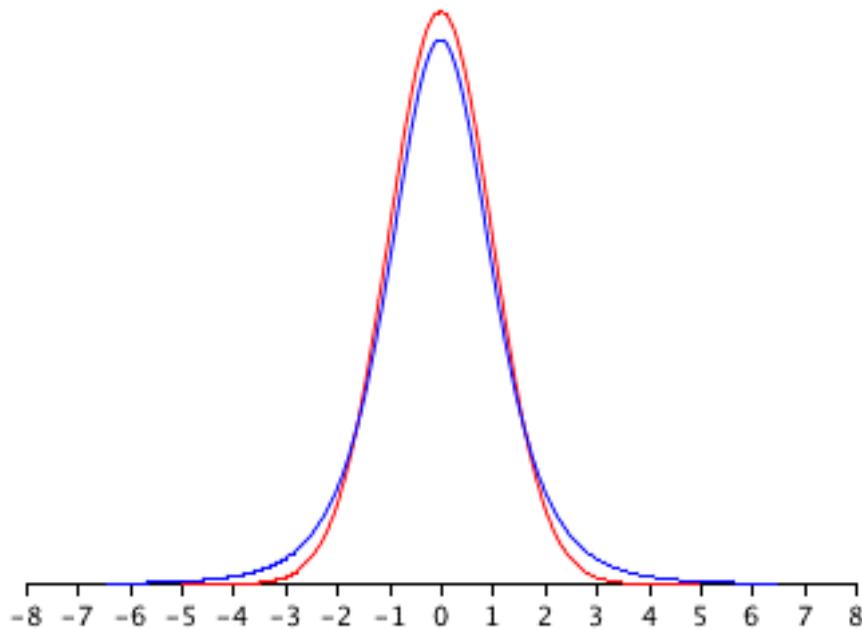
1. When would the mean grade in a class on a final exam be considered a statistic?  
When would it be considered a parameter?
2. Define bias in terms of expected value.
3. Is it possible for a statistic to be unbiased yet very imprecise? How about being very accurate but biased?
4. Why is a 99% confidence interval wider than a 95% confidence interval?
5. When you construct a 95% confidence interval, what are you 95% confident about?
6. What is the difference in the computation of a confidence interval between cases in which you know the population standard deviation and cases in which you have to estimate it?
7. Assume a researcher found that the correlation between a test he or she developed and job performance was 0.55 in a study of 28 employees. If correlations under .35 are considered unacceptable, would you have any reservations about using this test to screen job applicants?
8. What is the effect of sample size on the width of a confidence interval?
9. How does the *t* distribution compare with the normal distribution? How does this difference affect the size of confidence intervals constructed using *z* relative to those constructed using *t*? Does sample size make a difference?
10. The effectiveness of a blood-pressure drug is being investigated. How might an experimenter demonstrate that, on average, the reduction in systolic blood pressure is 20 or more?

11. A population is known to be normally distributed with a standard deviation of 2.8. (a) Compute the 95% confidence interval on the mean based on the following sample of nine: 8, 9, 10, 13, 14, 16, 17, 20, 21. (b) Now compute the 99% confidence interval using the same data.
12. A person claims to be able to predict the outcome of flipping a coin. This person is correct 16/25 times. Compute the 95% confidence interval on the proportion of times this person can predict coin flips correctly. What conclusion can you draw about this test of his ability to predict the future?
13. What does it mean that the variance (computed by dividing by N) is a biased statistic?
14. A confidence interval for the population mean computed from an N of 16 ranges from 12 to 28. A new sample of 36 observations is going to be taken. You can't know in advance exactly what the confidence interval will be because it depends on the random sample. Even so, you should have some idea of what it will be. Give your best estimation.
15. You take a sample of 22 from a population of test scores, and the mean of your sample is 60. (a) You know the standard deviation of the population is 10. What is the 99% confidence interval on the population mean. (b) Now assume that you do not know the population standard deviation, but the standard deviation in your sample is 10. What is the 99% confidence interval on the mean now?
16. You read about a survey in a newspaper and find that 70% of the 250 people sampled prefer Candidate A. You are surprised by this survey because you thought that more like 50% of the population preferred this candidate. Based on this sample, is 50% a possible population proportion? Compute the 95% confidence interval to be sure.
17. Heights for teenage boys and girls were calculated. The mean height for the sample of 12 boys was 174 cm and the variance was 62. For the sample of 12 girls, the mean was 166 cm and the variance was 65. Assuming equal variances and normal distributions in the population, (a) What is the 95% confidence interval on the difference between population means? (b) What is the 99%

confidence interval on the difference between population means? (c) Do you think it is very unlikely that the mean difference in the population is about 5? Why or why not?

18. You were interested in how long the average psychology major at your college studies per night, so you asked 10 psychology majors to tell you the amount they study. They told you the following times: 2, 1.5, 3, 2, 3.5, 1, 0.5, 3, 2, 4.  
(a) Find the 95% confidence interval on the population mean. (b) Find the 90% confidence interval on the population mean.
19. True/false: As the sample size gets larger, the probability that the confidence interval will contain the population mean gets higher.
20. True/false: You have a sample of 9 men and a sample of 8 women. The degrees of freedom for the t value in your confidence interval on the difference between means is 16.
21. True/false: Greek letters are used for statistics as opposed to parameters.
22. True/false: In order to construct a confidence interval on the difference between means, you need to assume that the populations have the same variance and are both normally distributed.

23. True/false: The red distribution represents the t distribution and the blue distribution represents the normal distribution.



### *Questions from Case Studies*

#### Angry Moods (AM) case study

24. (AM) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the “Anger-Out” scores, compute a 99% confidence interval on the difference between gender means.
25. (AM) Calculate the 95% confidence interval for the difference between the mean Anger-In score for the athletes and non-athletes. What can you conclude?
26. (AM) Find the 95% confidence interval on the population correlation between the Anger- Out and Control-Out scores.

#### Flatulence (F) case study

27. (F) Compare men and women on the variable “perday.” Compute the 95% confidence interval on the difference between means.

28. (F) What is the 95% confidence interval of the mean time people wait before farting in front of a romantic partner.

#### Animal Research (AR) case study

29. (AR) What percentage of the women studied in this sample strongly agreed (gave a rating of 7) that using animals for research is wrong?

30. (AR) Use the proportion you computed in #29. Compute the 95% confidence interval on the population proportion of women who strongly agree that animal research is wrong.

31. (AR) Compute a 95% confidence interval on the difference between the gender means with respect to their beliefs that animal research is wrong.

#### ADHD Treatment (AT) case study

32. (AT) What is the correlation between the participants' correct number of responses after taking the placebo and their correct number of responses after taking 0.60 mg/kg of MPH? Compute the 95% confidence interval on the population correlation.

#### Weapons and Aggression (WA) case study

33. (WA) Recall that the hypothesis is that a person can name an aggressive word more quickly if it is preceded by a weapon word prime than if it is preceded by a neutral word prime. The first step in testing this hypothesis is to compute the difference between (a) the naming time of aggressive words when preceded by a neutral word prime and (b) the naming time of aggressive words when preceded by a weapon word prime separately for each of the 32 participants. That is, compute an  $- aw$  for each participant.

a. (WA) Would the hypothesis of this study be supported if the difference were positive or if it were negative?

b. What is the mean of this difference score?

c. What is the standard deviation of this difference score?

d. What is the 95% confidence interval of the mean difference score?

e. What does the confidence interval computed in (d) say about the hypothesis.

### Diet and Health (DH) case study

34. (DH) Compute a 95% confidence interval on the proportion of people who are healthy on the AHA diet.

	Cancers	Deaths	Nonfatal illness	Healthy	Total
<b>AHA</b>	15	24	25	239	303
<b>Mediterranean</b>	7	14	8	273	302
<b>Total</b>	22	38	33	512	605

The following questions are from ARTIST (reproduced with permission)



35. Suppose that you take a random sample of 10,000 Americans and find that 1,111 are left-handed. You perform a test of significance to assess whether the sample data provide evidence that more than 10% of all Americans are left-handed, and you calculate a test statistic of 3.70 and a p-value of .0001. Furthermore, you calculate a 99% confidence interval for the proportion of left-handers in America to be (.103,.119). Consider the following statements: The sample provides strong evidence that more than 10% of all Americans are left-handed. The sample provides evidence that the proportion of left-handers in America is much larger than 10%. Which of these two statements is the more appropriate conclusion to draw? Explain your answer based on the results of the significance test and confidence interval.

36. A student wanted to study the ages of couples applying for marriage licenses in his county. He studied a sample of 94 marriage licenses and found that in 67 cases the husband was older than the wife. Do the sample data provide strong evidence that the husband is usually older than the wife among couples

applying for marriage licenses in that county? Explain briefly and justify your answer.

37. Imagine that there are 100 different researchers each studying the sleeping habits of college freshmen. Each researcher takes a random sample of size 50 from the same population of freshmen. Each researcher is trying to estimate the mean hours of sleep that freshmen get at night, and each one constructs a 95% confidence interval for the mean. Approximately how many of these 100 confidence intervals will NOT capture the true mean?
- a. None
  - b. 1 or 2
  - c. 3 to 7
  - d. about half
  - e. 95 to 100
  - f. other

# 11. Logic of Hypothesis Testing

- A. Introduction
- B. Significance Testing
- C. Type I and Type II Errors
- D. One- and Two-Tailed Tests
- E. Interpreting Significant Results
- F. Interpreting Non-Significant Results
- G. Steps in Hypothesis Testing
- H. Significance Testing and Confidence Intervals
- I. Misconceptions
- J. Exercises

When interpreting an experimental finding, a natural question arises as to whether the finding could have occurred by chance. Hypothesis testing is a statistical procedure for testing whether chance is a plausible explanation of an experimental finding. Misconceptions about hypothesis testing are common among practitioners as well as students. To help prevent these misconceptions, this chapter goes into more detail about the logic of hypothesis testing than is typical for an introductory-level text.

# Introduction

by David M. Lane

## *Prerequisites*

- Chapter 5: Binomial Distribution

## *Learning Objectives*

1. Describe the logic by which it can be concluded that someone can distinguish between two things
2. State whether random assignment ensures that all uncontrolled sources of variation will be equal
3. Define precisely what the probability is that is computed to reach the conclusion that a difference is not due to chance
4. Distinguish between the probability of an event and the probability of a state of the world
5. Define “null hypothesis”
6. Be able to determine the null hypothesis from a description of an experiment
7. Define “alternative hypothesis”

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from the binomial distribution and the binomial distribution calculator shows it to be 0.0106. This is a pretty low probability, and therefore

someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for normal-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference ( $31.4 - 24.7 = 6.7$  minutes) if the difference were, in fact, due solely to chance. Using methods presented in Chapter 12, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

## The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial distribution, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he

cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and therefore are difficult to apply in some contexts.

## The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{\text{obese}} = \mu_{\text{average}}$$

or as

$$\mu_{\text{obese}} - \mu_{\text{average}} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where  $\rho$  is the population correlation (not to be confused with  $r$ , the correlation in the sample).

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that  $\pi = 0.5$ .

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers

hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The alternative hypothesis is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{\text{obese}} = \mu_{\text{average}}$$

is rejected, then there are two alternatives:

$$\begin{aligned}\mu_{\text{obese}} &< \mu_{\text{average}} \\ \mu_{\text{obese}} &> \mu_{\text{average}}.\end{aligned}$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two populations means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960) showed how it is justified to draw a conclusion about the direction of the difference.

# Significance Testing

by David M. Lane

## *Prerequisites*

- Chapter 5: Binomial Distribution
- Chapter 11: Introduction to Hypothesis Testing

## *Learning Objectives*

1. Describe how a probability value is used to cast doubt on the null hypothesis
2. Define “statistically significant”
3. Distinguish between statistical significance and practical significance
4. Distinguish between two approaches to significance testing

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the  $\alpha$  level or simply  $\alpha$ . It is also called the *significance level*.

When the null hypothesis is rejected, the effect is said to be *statistically significant*. For example, in the Physicians Reactions case study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what “significant” usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

**Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.**

Why does the word “significant” in the phrase “statistically significant” mean something so different from other uses of the word? Interestingly, this is because the meaning of “significant” in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was “significant” if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of “significant” changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher) a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis. If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an  $\alpha$  level before analyzing the data. If the data analysis results in a probability value below the  $\alpha$  level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

# Type I and II Errors

by David M. Lane

## *Prerequisites*

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing

## *Learning Objectives*

1. Define Type I and Type II errors
2. Interpret significant and non-significant differences
3. Explain why the null hypothesis should not be accepted when the effect is not significant

In the Physicians' Reactions case study, the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05 then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the  $\alpha$  level or simply  $\alpha$ . It is also called the significance level. As discussed in the introduction to hypothesis testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the  $\alpha$  level: the lower the  $\alpha$  level the lower the Type I error rate. It might seem that  $\alpha$  is the probability of a Type I error. However, this is not correct. Instead,  $\alpha$  is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called  $\beta$  (beta). The probability of correctly rejecting a false null hypothesis equals  $1 - \beta$  and is called power. Power is covered in detail in Chapter 13.

# One- and Two-Tailed Tests

by David M. Lane

## *Prerequisites*

- Chapter 6: Binomial Distribution
- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance

## *Learning Objectives*

1. Define one- and two-tailed tests
2. State the difference between one- and two-tailed hypotheses
3. State when it is valid to use a one-tailed test

In the James Bond case study, Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure 1 shows a graph of the binomial. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a “one-tailed probability.”

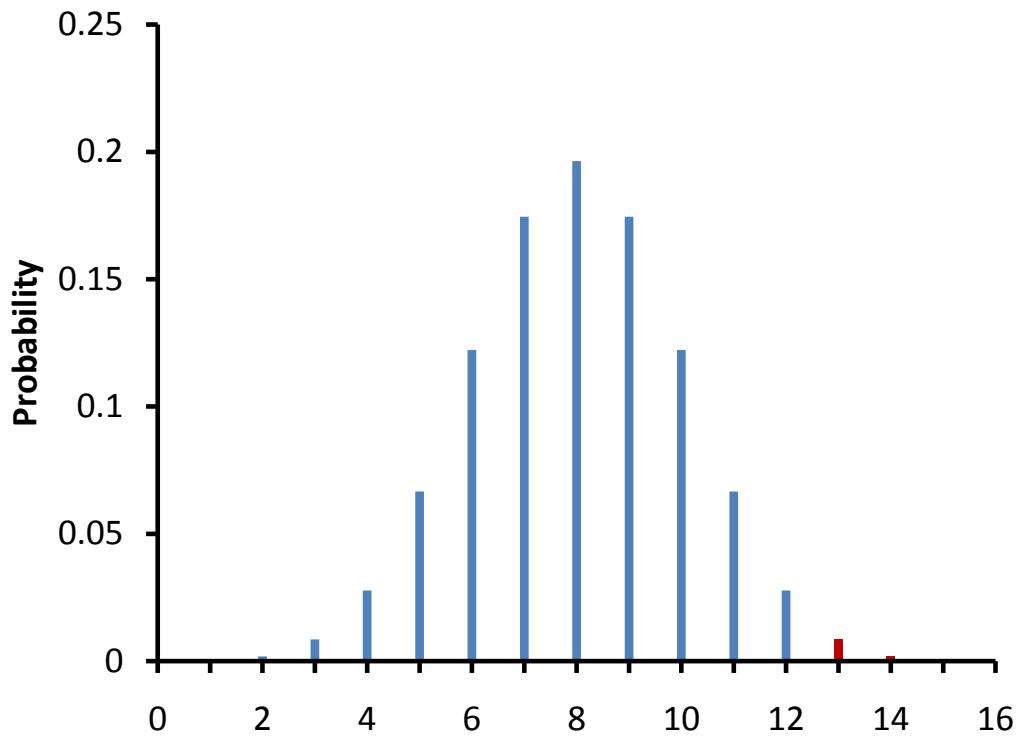


Figure 1. The binomial distribution. The upper (right-hand) tail is red.

A slightly different question can be asked of the data: “What is the probability of getting a result as extreme or more extreme than the one observed”? Since the chance expectation is 8/16, a result of 3/13 is equally as extreme as 13/16. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when  $\pi = 0.5$ , this probability is exactly double the probability of 0.0106 computed previously. Therefore,  $p = 0.0212$ . A probability calculated in both tails of a distribution is called a two-tailed probability (see Figure 2).

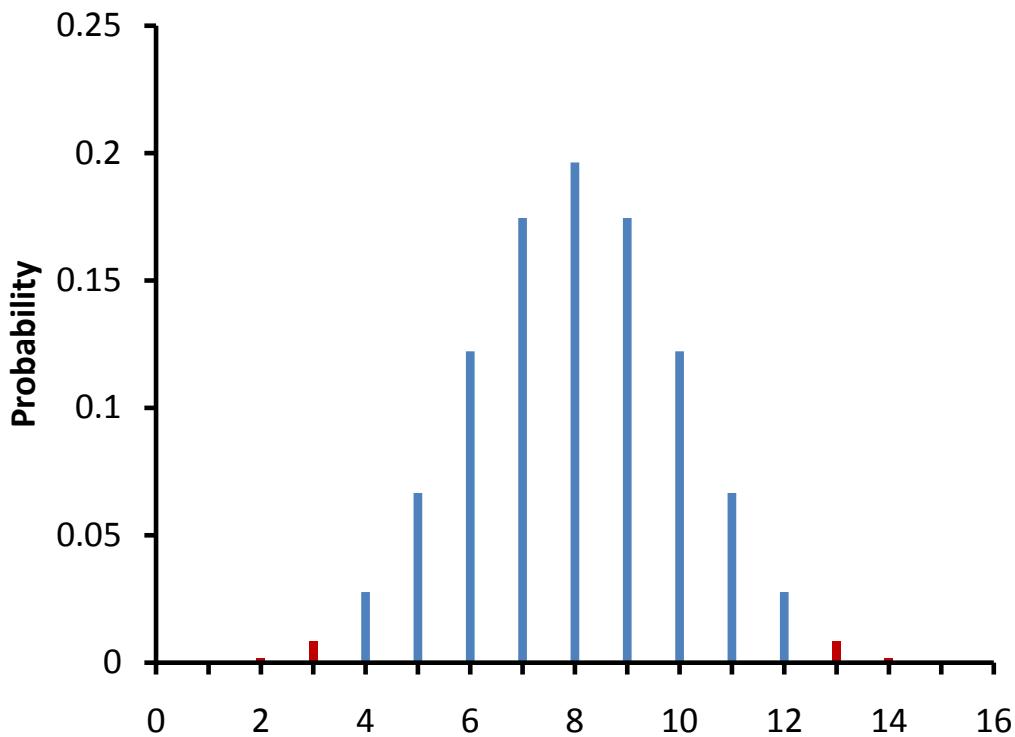


Figure 2. The binomial distribution. Both tails are red.

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond's performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is  $\pi = 0.5$ . By contrast, the null hypothesis for the one-tailed test is  $\pi \leq 0.5$ . Accordingly, we reject the two-tailed hypothesis if the sample proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected only if the sample proportion is much greater

than 0.5. The alternative hypothesis in the two-tailed test is  $\pi \neq 0.5$ . In the one-tailed test it is  $\pi > 0.5$ .

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

# Interpreting Significant Results

by David M. Lane

## *Prerequisites*

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance, Type I and II Errors
- Chapter 11: One and Two-Tailed Tests

## *Learning Objectives*

1. Discuss whether rejection of the null hypothesis should be an all-or-none proposition
2. State the usefulness of a significance test when it is extremely likely that the null hypothesis of no difference is false even before doing the experiment

When a probability value is below the  $\alpha$  level, the effect is *statistically significant* and the null hypothesis is rejected. However, not all statistically significant effects should be treated the same way. For example, you should have less confidence that the null hypothesis is false if  $p = 0.049$  than  $p = 0.003$ . Thus, rejecting the null hypothesis is not an all-or-none proposition.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. Consider the one-tailed test in the James Bond case study: Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred and the question is whether he is better than chance on this task. The null hypothesis for this one-tailed test is that  $\pi \leq 0.5$  where  $\pi$  is the probability of being correct on any given trial. If this null hypothesis is rejected, then the alternative hypothesis that  $\pi > 0.5$  is accepted. If  $\pi$  is greater than 0.5, then Mr. Bond is better than chance on this task.

Now consider the two-tailed test used in the Physicians' Reactions case study. The null hypothesis is:

$$\mu_{\text{obese}} = \mu_{\text{average}}.$$

If this null hypothesis is rejected, then there are two alternatives:

$$\begin{aligned}\mu_{\text{obese}} &< \mu_{\text{average}} \\ \mu_{\text{obese}} &> \mu_{\text{average}}.\end{aligned}$$

Naturally, the direction of the sample means determines which alternative is adopted. If the sample mean for the obese patients is significantly lower than the sample mean for the average-weight patients, then one should conclude that the population mean for the obese patients is lower than the sample mean for the average-weight patients.

There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established. This point is also made in the section on the relationship between confidence intervals and significance tests.

*Optional*

Some textbooks have incorrectly stated that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Instead, they say that all one can conclude is that the population means differ. The validity of concluding the direction of the effect is clear if you note that a two-tailed test at the 0.05 level is equivalent to two separate one-tailed tests each at the 0.025 level. The two null hypotheses are then

$$\begin{aligned}\mu_{\text{obese}} &\geq \mu_{\text{average}} \\ \mu_{\text{obese}} &\leq \mu_{\text{average}}.\end{aligned}$$

If the former of these is rejected, then the conclusion is that the population mean for obese patients is lower than that for average-weight patients. If the latter is rejected, then the conclusion is that the population mean for obese patients is higher than that for average-weight patients. See Kaiser (1960).

# Interpreting Non-Significant Results

by David M. Lane

## *Prerequisites*

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing
- Chapter 11: Type I and II Errors

## *Learning Objectives*

1. State what it means to accept the null hypothesis
2. Explain why the null hypothesis should not be accepted
3. Describe how a non-significant result can increase confidence that the null hypothesis is false
4. Discuss the problems of affirming a negative conclusion

When a significance test results in a high probability value, it means that the data provide little or no evidence that the null hypothesis is false. However, the high probability value is not evidence that the null hypothesis is true. The problem is that it is impossible to distinguish a null effect from a very small effect. For example, in the James Bond Case Study, suppose Mr. Bond is, in fact, just barely better than chance at judging whether a martini was shaken or stirred. Assume he has a 0.51 probability of being correct on a given trial ( $\pi = 0.51$ ). Let's say Experimenter Jones (who did not know  $\pi = 0.51$ ) tested Mr. Bond and found he was correct 49 times out of 100 tries. How would the significance test come out? The experimenter's significance test would be based on the assumption that Mr. Bond has a 0.50 probability of being correct on each trial ( $\pi = 0.50$ ). Given this assumption, the probability of his being correct 49 or more times out of 100 is 0.62. This means that the probability value is 0.62, a value very much higher than the conventional significance level of 0.05. This result, therefore, does not give even a hint that the null hypothesis is false. However, we know (but Experimenter Jones does not) that  $\pi = 0.51$  and not 0.50 and therefore that the null hypothesis is false. So, if Experimenter Jones had concluded that the null hypothesis was true based on the statistical analysis, he or she would have been mistaken. Concluding that the null hypothesis is true is called *accepting the null hypothesis*. To do so is a serious error.

**Do not accept the null hypothesis when you do not reject it.**

So how should the non-significant result be interpreted? The experimenter should report that there is no credible evidence Mr. Bond can tell whether a martini was shaken or stirred, but that there is no proof that he cannot. It is generally impossible to prove a negative. What if I claimed to have been Socrates in an earlier life? Since I have no evidence for this claim, I would have great difficulty convincing anyone that it is true. However, no one would be able to prove definitively that I was not.

Often a non-significant finding increases one's confidence that the null hypothesis is false. Consider the following hypothetical example. A researcher develops a treatment for anxiety that he or she believes is better than the traditional treatment. A study is conducted to test the relative effectiveness of the two treatments: 20 subjects are randomly divided into two groups of 10. One group receives the new treatment and the other receives the traditional treatment. The mean anxiety level is lower for those receiving the new treatment than for those receiving the traditional treatment. However, the difference is not significant. The statistical analysis shows that a difference as large or larger than the one obtained in the experiment would occur 11% of the time even if there were no true difference between the treatments. In other words, the probability value is 0.11. A naive researcher would interpret this finding as evidence that the new treatment is no more effective than the traditional treatment. However, the sophisticated researcher, although disappointed that the effect was not significant, would be encouraged that the new treatment led to less anxiety than the traditional treatment. The data support the thesis that the new treatment is better than the traditional one even though the effect is not statistically significant. This researcher should have more confidence that the new treatment is better than he or she had before the experiment was conducted. However, the support is weak and the data are inconclusive. What should the researcher do? A reasonable course of action would be to do the experiment again. Let's say the researcher repeated the experiment and again found the new treatment was better than the traditional treatment. However, once again the effect was not significant and this time the probability value was 0.07. The naive researcher would think that two out of two experiments failed to find significance and therefore the new treatment is unlikely to be better than the traditional treatment. The sophisticated researcher would note that two out of two times the new treatment was better than the traditional treatment. Moreover, two experiments each providing weak support that the new treatment is better, when taken together, can provide strong support. Using a method for combining

probabilities, it can be determined that combining the probability values of 0.11 and 0.07 results in a probability value of 0.045. Therefore, these two non-significant findings taken together result in a significant finding.

Although there is never a statistical basis for concluding that an effect is exactly zero, a statistical analysis can demonstrate that an effect is most likely small. This is done by computing a confidence interval. If all effect sizes in the interval are small, then it can be concluded that the effect is small. For example, suppose an experiment tested the effectiveness of a treatment for insomnia. Assume that the mean time to fall asleep was 2 minutes shorter for those receiving the treatment than for those in the control group and that this difference was not significant. If the 95% confidence interval ranged from -4 to 8 minutes, then the researcher would be justified in concluding that the benefit is eight minutes or less. However, the researcher would not be justified in concluding the null hypothesis is true, or even that it was supported.

# Steps in Hypothesis Testing

by David M. Lane

## *Prerequisites*

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

## *Learning Objectives*

1. Be able to state the null hypothesis for both one-tailed and two-tailed tests
  2. Differentiate between a significance level and a probability level
  3. State the four steps involved in significance testing
- 
1. The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is  $\mu_1 - \mu_2 = 0$  which is equivalent to  $\mu_1 = \mu_2$ . For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that  $\mu_1$  is larger than  $\mu_2$ , then the null hypothesis (the reverse of the prediction) is  $\mu_2 - \mu_1 \geq 0$ . This is equivalent to  $\mu_1 \leq \mu_2$ .
  2. The second step is to specify the  $\alpha$  level which is also known as the significance level. Typical values are 0.05 and 0.01.
  3. The third step is to compute the probability value (also known as the  $p$  value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
  4. Finally, compare the probability value with the  $\alpha$  level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional  $\alpha$  level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

# Significance Testing and Confidence Intervals

by David M. Lane

## *Prerequisites*

- Chapter 10: Confidence Intervals Introduction
- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing

## *Learning Objectives*

1. Determine from a confidence interval whether a test is significant
2. Explain why a confidence interval makes clear that one should not accept the null hypothesis

There is a close relationship between confidence intervals and significance tests. Specifically, if a statistic is significantly different from 0 at the 0.05 level then the 95% confidence interval will not contain 0. All values in the confidence interval are plausible values for the parameter whereas values outside the interval are rejected as plausible values for the parameter. In the Physicians' Reactions case study, the 95% confidence interval for the difference between means extends from 2.00 to 11.26. Therefore, any value lower than 2.00 or higher than 11.26 is rejected as a plausible value for the population difference between means. Since zero is lower than 2.00, it is rejected as a plausible value and a test of the null hypothesis that there is no difference between means is significant. It turns out that the p value is 0.0057. There is a similar relationship between the 99% confidence interval and significance at the 0.01 level.

Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative). Therefore, a significant finding allows the researcher to specify the direction of the effect. There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established because the values in the confidence interval are either all positive or all negative.

If the 95% confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then the effect will not be significant at the 0.05 level. Looking at non-significant effects in terms of confidence intervals makes clear why the null hypothesis should not be accepted when it is not rejected: Every value in the confidence interval is a plausible value of the parameter. Since zero is in the interval, it cannot be rejected. However, there is an infinite number of other values in the interval (assuming continuous measurement), and none of them can be rejected either.

# Misconceptions

by David M. Lane

## *Prerequisites*

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

## *Learning Objectives*

1. State why the probability value is not the probability the null hypothesis is false
2. Explain why a low probability value does not necessarily mean there is a large effect
3. Explain why a non-significant outcome does not mean the null hypothesis is probably true

Misconceptions about significance testing are common. This section lists three important ones.

1. **Misconception:** The probability value is the probability that the null hypothesis is false.

Proper interpretation: The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.

2. **Misconception:** A low probability value indicates a large effect.

Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.

3. **Misconception:** A non-significant outcome means that the null hypothesis is probably true.

Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 11: Interpreting Non-Significant Results

Research in March, 2012 reported here found evidence for the existence of the Higgs Boson particle. However, the evidence for the existence of the particle was not statistically significant.

## **What do you think?**

Did the researchers conclude that their investigation had been a failure or did they conclude they have evidence of the particle, just not strong enough evidence to draw a confident conclusion?

One of the investigators stated, "We see some tantalizing evidence but not significant enough to make a stronger statement." Therefore, they were encouraged by the result. In a subsequent study, the evidence was significant.

## **References**

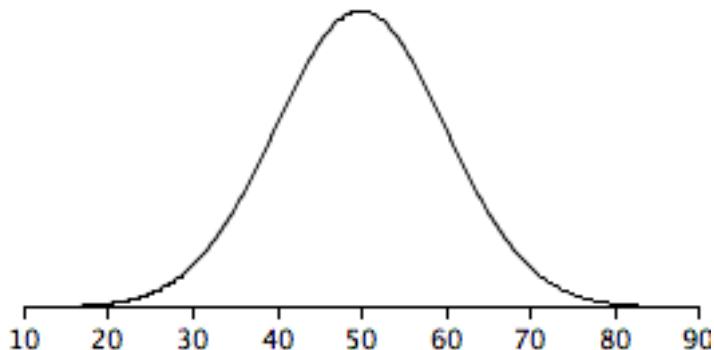
- Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, 67, 160-167

## Exercises

### *Prerequisites*

- All material presented in the Logic of Hypothesis Testing chapter
1. An experiment is conducted to test the claim that James Bond can taste the difference between a Martini that is shaken and one that is stirred. What is the null hypothesis?
  2. The following explanation is incorrect. What three words should be added to make it correct?

The probability value is the probability of obtaining a statistic as different (add three words here) from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming that the null hypothesis is true.
  3. Why do experimenters test hypotheses they think are false?
  4. State the null hypothesis for:
    - a. An experiment testing whether echinacea decreases the length of colds.
    - b. A correlational study on the relationship between brain size and intelligence.
    - c. An investigation of whether a self-proclaimed psychic can predict the outcome of a coin flip.
    - d. A study comparing a drug with a placebo on the amount of pain relief. (A one-tailed test was used.)
  5. Assume the null hypothesis is that  $\mu = 50$  and that the graph shown below is the sampling distribution of the mean (M). Would a sample value of  $M = 60$  be significant in a two-tailed test at the .05 level? Roughly what value of M would be needed to be significant?



6. A researcher develops a new theory that predicts that vegetarians will have more of a particular vitamin in their blood than non-vegetarians. An experiment is conducted and vegetarians do have more of the vitamin, but the difference is not significant. The probability value is 0.13. Should the experimenter's confidence in the theory increase, decrease, or stay the same?
7. A researcher hypothesizes that the lowering in cholesterol associated with weight loss is really due to exercise. To test this, the researcher carefully controls for exercise while comparing the cholesterol levels of a group of subjects who lose weight by dieting with a control group that does not diet. The difference between groups in cholesterol is not significant. Can the researcher claim that weight loss has no effect?
8. A significance test is performed and  $p = .20$ . Why can't the experimenter claim that the probability that the null hypothesis is true is .20?
9. For a drug to be approved by the FDA, the drug must be shown to be safe and effective. If the drug is significantly more effective than a placebo, then the drug is deemed effective. What do you know about the effectiveness of a drug once it has been approved by the FDA (assuming that there has not been a Type I error)?
10. When is it valid to use a one-tailed test? What is the advantage of a one-tailed test? Give an example of a null hypothesis that would be tested by a one-tailed test.
11. Distinguish between probability value and significance level.

12. Suppose a study was conducted on the effectiveness of a class on “How to take tests.” The SAT scores of an experimental group and a control group were compared. (There were 100 subjects in each group.) The mean score of the experimental group was 503 and the mean score of the control group was 499. The difference between means was found to be significant,  $p = .037$ . What do you conclude about the effectiveness of the class?
13. Is it more conservative to use an alpha level of .01 or an alpha level of .05? Would beta be higher for an alpha of .05 or for an alpha of .01?
14. Why is “ $H_0: M_1 = M_2$ ” not a proper null hypothesis?
15. An experimenter expects an effect to come out in a certain direction. Is this sufficient basis for using a one-tailed test? Why or why not?
16. How do the Type I and Type II error rates of one-tailed and two-tailed tests differ?
17. A two-tailed probability is .03. What is the one-tailed probability if the effect were in the specified direction? What would it be if the effect were in the other direction?
18. You choose an alpha level of .01 and then analyze your data.
- What is the probability that you will make a Type I error given that the null hypothesis is true?
  - What is the probability that you will make a Type I error given that the null hypothesis is false?
19. Why doesn’t it make sense to test the hypothesis that the sample mean is 42?
20. True/false: It is easier to reject the null hypothesis if the researcher uses a smaller alpha ( $\alpha$ ) level.
21. True/false: You are more likely to make a Type I error when using a small sample than when using a large sample.

22. True/false: You accept the alternative hypothesis when you reject the null hypothesis.
23. True/false: You do not accept the null hypothesis when you fail to reject it.
24. True/false: A researcher risks making a Type I error any time the null hypothesis is rejected.

# 12. Testing Means

- A. Single Mean
- B. Difference between Two Means (Independent Groups)
- C. All Pairwise Comparisons Among Means
- D. Specific Comparisons
- E. Difference between Two Means (Correlated Pairs)
- F. Specific Comparisons (Correlated Observations)
- G. Pairwise Comparisons (Correlated Observations)
- H. Exercises

Many, if not most experiments are designed to compare means. The experiment may involve only one sample mean that is to be compared to a specific value. Or the experiment could be testing differences among many different experimental conditions, and the experimenter could be interested in comparing each mean with each of the other means. This chapter covers methods of comparing means in many different experimental situations.

The topics covered here in sections C, D, F, and G are typically covered in other texts in a chapter on Analysis of Variance. We prefer to cover them here since they bear no necessary relationship to analysis of variance. As discussed by Wilkinson (1999), it is not logical to consider the procedures in this chapter as tests to be performed subsequent to an analysis of variance. Nor is it logical to call them post-hoc tests as some computer programs do.

# Testing a Single Mean

by David M. Lane

## *Prerequisites*

- Chapter 7: Normal Distributions
- Chapter 7: Areas Under Normal Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 9: Introduction to Sampling Distributions
- Chapter 10: t Distribution
- Chapter 11: Logic of Hypothesis Testing

## *Learning Objectives*

1. Compute the probability of a sample mean being at least as high as a specified value when  $\sigma$  is known
2. Compute a two-tailed probability
3. Compute the probability of a sample mean being at least as high as a specified value when  $\sigma$  is estimated
4. State the assumptions required for item 3 above

This section shows how to test the null hypothesis that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50. In other words, the null hypothesis is that the population mean ( $\mu$ ) is 50. The (hypothetical) data are shown in Table 1. The data in Table 1 have a sample mean ( $M$ ) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table 1. Distribution of scores.

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from  $\mu$  by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As shown in Chapter 9, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_M = \mu$$

and

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

respectively. It is clear that  $\mu_M = 50$ . In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation ( $\sigma$ ).

The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know  $\sigma$  and therefore you would use  $s$ , the sample estimate of  $\sigma$ . However, it is

instructive to see how the probability is computed if  $\sigma$  is known before proceeding to see how it is calculated when  $\sigma$  is estimated.

For the current example, if the null hypothesis is true, then based on the binomial distribution, one can compute that variance of the number correct is

$$\begin{aligned}\sigma^2 &= N\pi(1-\pi) \\ &= 100(0.5)(1-0.5) \\ &= 25.\end{aligned}$$

Therefore,  $\sigma = 5$ . For a  $\sigma$  of 5 and an  $N$  of 9, the standard deviation of the sampling distribution of the mean is  $5/3 = 1.667$ . Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or more when the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute this probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use the normal distribution calculator ([external link](#)) as shown in Figure 1.

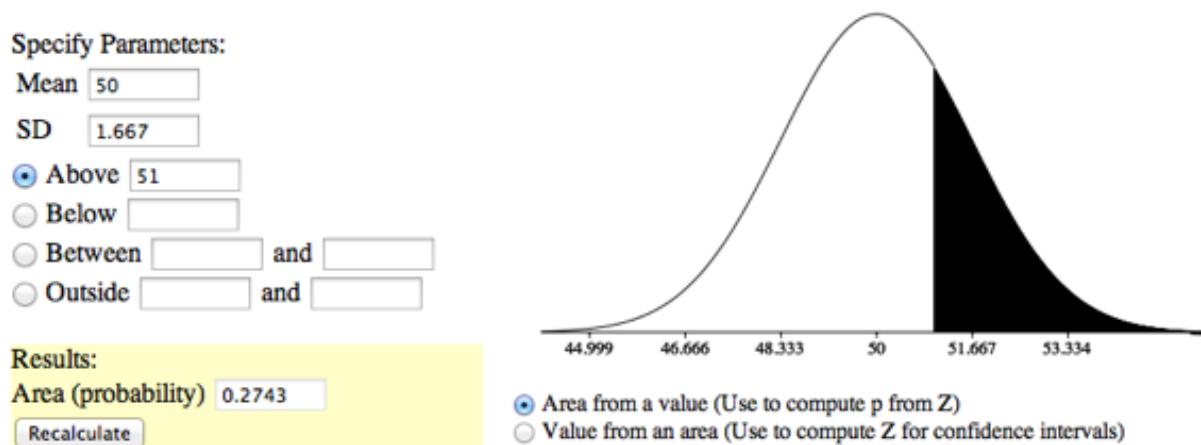


Figure 1. Probability of a sample mean being 51 or greater.

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area **above** 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results of the normal distribution calculator are shown in Figure 2.

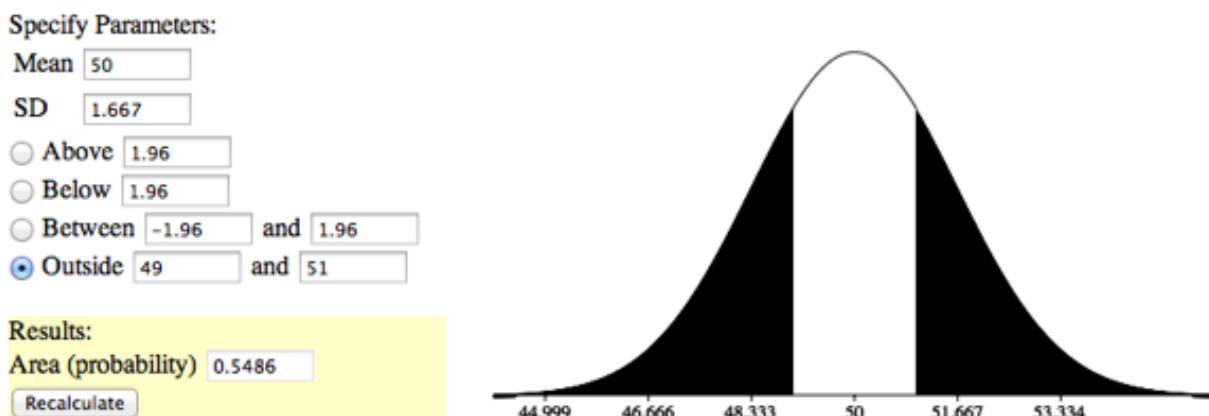


Figure 2. Probability of a sample mean being less than or equal to 49 or greater than or equal to 51.

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure 1.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by computing Z based on the formula

$$Z = \frac{M - \mu}{\sigma_M}$$

where  $Z$  is the value on the standard normal distribution,  $M$  is the sample mean,  $\mu$  is the hypothesized value of the mean, and  $\sigma_M$  is the standard error of the mean. For this example,  $Z = (51-50)/1.667 = 0.60$ . The normal calculator with a mean of 0 and a standard deviation of 1 is shown in Figure 3.

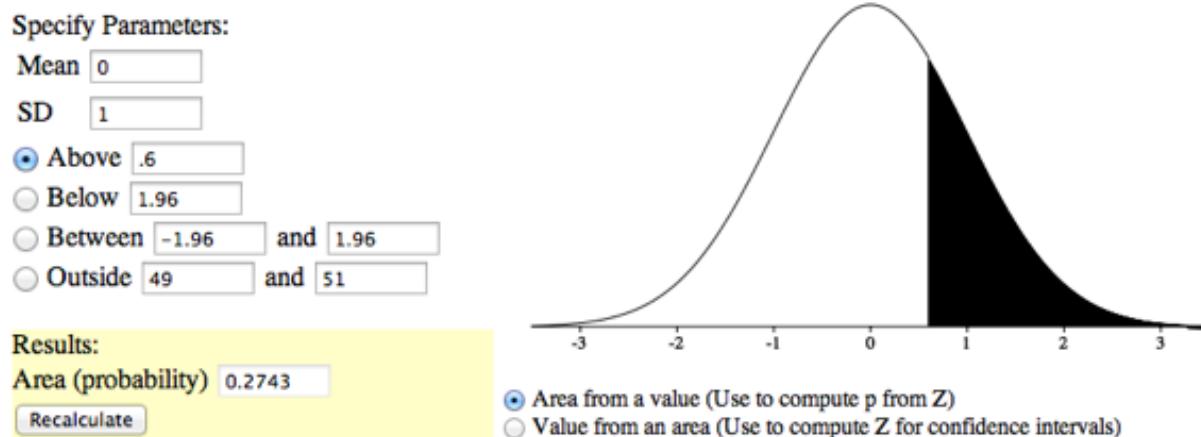


Figure 3. Calculation using the standardized normal distribution.

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analyses it is very rare that you would know  $\sigma$  and wish to estimate  $\mu$ . Typically  $\sigma$  is not known and is estimated in a sample by  $s$ , and  $\sigma_M$  is estimated by  $s_M$ . For our next example, we will consider the data in the “ADHD Treatment” case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table 2 shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled “Diff” that shows the difference in performance between the 0.6 mg (D60) and the 0 mg (D0) conditions. These difference scores are positive for children who performed better in the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

Table 2. DOG scores as a function of dosage.

D0	D60	Diff
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

To test this null hypothesis, we compute  $t$  using a special case of the following formula:

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

The special case of this formula applicable to testing a single mean is

$$t = \frac{M - \mu}{S_M}$$

where  $t$  is the value we compute for the significance test,  $M$  is the sample mean,  $\mu$  is the hypothesized value of the population mean, and  $S_M$  is the estimated standard error of the mean. Notice the similarity of this formula to the formula for  $Z$ .

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean ( $M$ ) of the  $N = 24$  difference scores is 4.958, the hypothesized value of  $\mu$  is 0, and the standard deviation ( $s$ ) is 7.538. The estimate of the standard error of the mean is computed as:

$$S_m = \frac{s}{\sqrt{N}} = \frac{7.5382}{\sqrt{24}} = 1.54$$

Therefore,  $t = 4.96/1.54 = 3.22$ . The probability value for  $t$  depends on the degrees of freedom. The number of degrees of freedom is equal to  $N - 1 = 23$ . A  $t$  distribution calculator shows that a  $t$  less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.

## Review of Assumptions

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

# Differences between Two Means (Independent Groups)

by David M. Lane

## *Prerequisites*

- Chapter 9: Sampling Distribution of Difference between Means
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Difference between Means
- Chapter 11: Logic of Hypothesis Testing
- Chapter 12: Testing a Single Mean

## *Learning Objectives*

1. State the assumptions for testing the difference between two means
2. Estimate the population variance assuming homogeneity of variance
3. Compute the standard error of the difference between means
4. Compute t and p for the difference between means
5. Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. This section covers how to test for differences between means from two separate groups of subjects. A later section describes how to test for differences between the means of two conditions in designs where only one group of subjects is used and each subject is tested in each condition.

We take as an example the data from the “Animal Research” case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 1.

Table 1. Means and Variances in Animal Research study.

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very

important. What is important is whether there is a difference in the population means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent. The analysis of data with two scores per subject is shown in the section on the correlated t test later in this chapter.

Small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

We saw the following general formula for significance testing in the section on testing a single mean:

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0. The hypothesized value is the null hypothesis that the difference between population means is 0.

We continue to use the data from the “Animal Research” case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males. For this calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$$M_1 - M_2 = 5.3529 - 3.8824 = 1.4705.$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means so the estimated standard error of the statistic is ( $S_{M_1-M_2}$ ). Recall from the relevant section in the chapter on

sampling distributions that the formula for the standard error of the difference between means is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate  $\sigma^2$  and use that estimate in place of  $\sigma^2$ . Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of  $\sigma^2$ . In this example,

$$MSE = (2.743 + 2.985)/2 = 2.864.$$

Since  $n$  (the number of scores **in each group**) is 17,

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

The next step is to compute  $t$  by plugging these values into the formula:

$$t = 1.4705/.5805 = 2.533.$$

Finally, we compute the probability of getting a  $t$  as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to  $(n_1 - 1) + (n_2 - 1)$ , where  $n_1$  is the sample size of the first group and  $n_2$  is the sample size of the second group. For this example,  $n_1 = n_2 = 17$ . When  $n_1 = n_2$ , it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees for freedom is  $16 + 16 = 32$ .

Once we have the degrees of freedom, we can use a t distribution calculator to find that the probability value for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. This is the probability of a  $t < -2.533$  or a  $t > 2.533$ . A one-tailed test would result in a probability of 0.0082, which is half the two-tailed probability.

## Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 2.

Table 2. Example Data.

Group 1	Group 2
3	2
4	6
5	8

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 2 is shown in Table 3.

Table 3. Reformatted Data

G	Y
1	3
1	4
1	5
2	2
2	6
2	8

## Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the group

with the larger sample size more than the group with the smaller sample size. Computationally, this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2$$

where  $M_1$  is the mean for group 1 and  $M_2$  is the mean for group 2. Consider the following small example:

Table 4. Unequal n

Group 1	Group 2
3	2
4	4
5	

$$M_1 = 4 \text{ and } M_2 = 3.$$

$$\begin{aligned} SSE &= (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 \\ &= 4 \end{aligned}$$

Then, MSE is computed by:

$$MSE = \frac{SSE}{df}$$

The formula

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}}$$

is replaced by

$$S_{M_1-M_2} = \sqrt{\frac{2MSE}{n_h}}$$

where  $n_h$  is the harmonic mean of the sample sizes and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4$$

and

$$S_{M_1-M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054$$

Therefore,

$$t = (4-3)/1.054 = 0.949$$

and the two-tailed  $p = 0.413$ .

# All Pairwise Comparisons Among Means

by David M. Lane

## *Prerequisites*

- Chapter 12: Difference Between Two Means (Independent Groups)

## *Learning Objectives*

1. Define pairwise comparison
2. Describe the problem with doing t tests among all pairs of means
3. Calculate the Tukey HSD test
4. Explain why Tukey test should not necessarily be considered a follow-up test

Many experiments are designed to compare more than two conditions. We will take as an example the case study “Smiles and Leniency.” In this study, the effect of different types of smiles on the leniency showed to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table 1.

The problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the Type I error rate is inflated. Figure 1 shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

Table 1. Six Comparisons among Means.

false vs felt			felt vs miserable		
false vs miserable			felt vs neutral		
false vs neutral			miserable vs neutral		

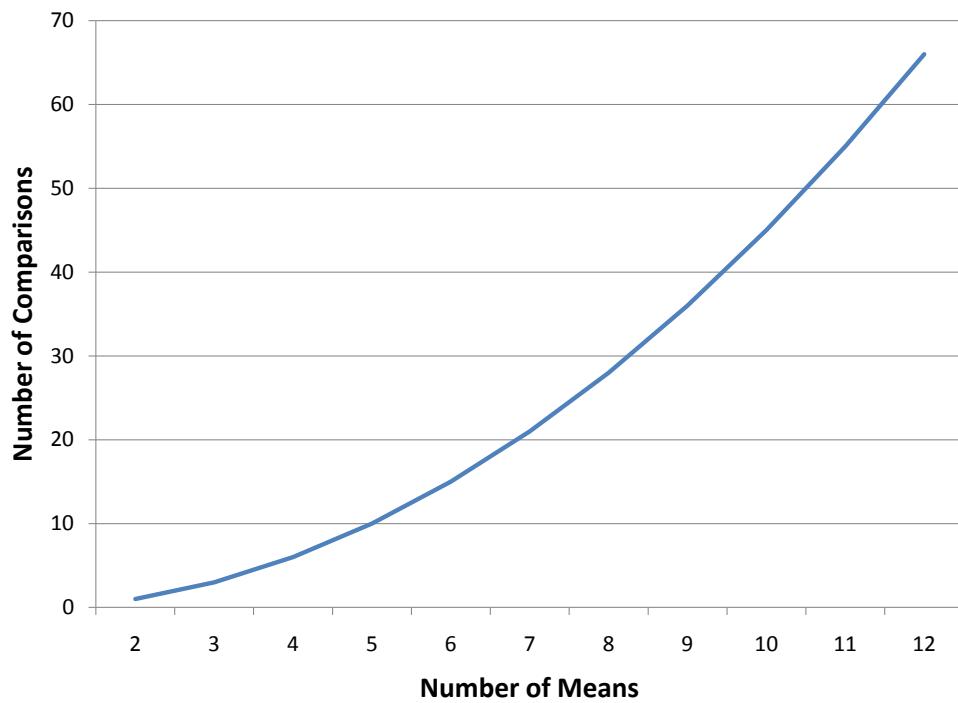


Figure 1. Number of pairwise comparisons as a function of the number of means.

Figure 2 shows the probability of a Type I error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

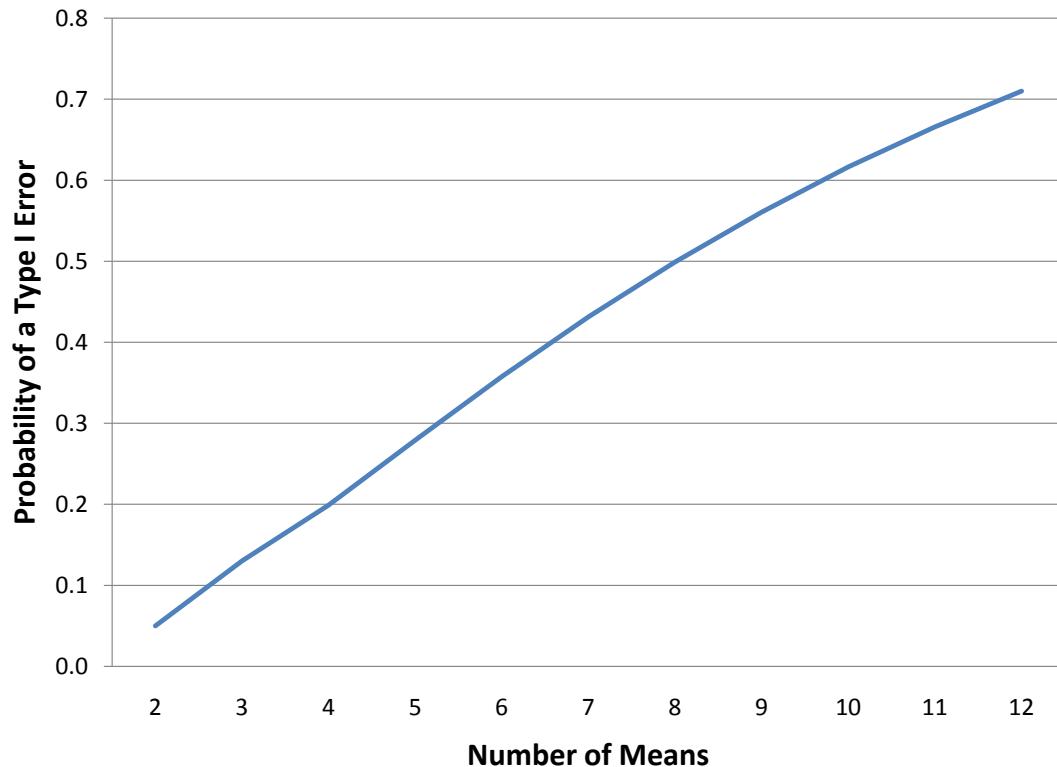


Figure 2. Probability of a Type I Error as a Function of the Number of Means.

The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or Tukey HSD for short. The Tukey HSD is based on a variation of the *t distribution* that takes into account the number of means being compared. This distribution is called the *studentized range distribution*.

Let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups *t* test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

Condition	Mean	Variance
FALSE	5.37	3.34
Felt	4.91	2.83

Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.

3. Compute

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

for each pair of means, where  $M_i$  is one mean,  $M_j$  is the other mean, and  $n$  is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute  $p$  for each comparison using the Studentized Range Calculator ([external link](#); requires Java). The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment,  $df = 136 - 4 = 132$ .

The tests for these data are shown in Table 2. The only significant comparison is between the false smile and the neutral smile.

Table 2. Six Pairwise Comparisons.

Comparison	$M_i - M_j$	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.01
Felt - Miserable	0	0	1
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject

it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in-between.

## Assumptions

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

## Computer Analysis

For most computer programs, you should format your data the same way you do for independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

Although full-featured statistics programs such as SAS, SPSS, R, and others can compute Tukey's test, smaller programs (including Analysis Lab) may not. However, these programs are generally able to compute a procedure known as Analysis of Variance (ANOVA). This procedure will be described in detail in a later chapter. Its relevance here is that an ANOVA computes the MSE that is used in the calculation of Tukey's test. For example, the following shows the ANOVA summary table for the “Smiles and Leniency” data.

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.4650	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The column labeled MS stands for “Mean Square” and therefore the value 2.6489 in the “Error” row and the MS column is the “Mean Squared Error” or MSE. Recall that this is the same value computed here (2.65) when rounded off.

## **Tukey's Test Need Not Be A Follow-Up to ANOVA**

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Wilkinson and the Task Force on Statistical Inference (1999).

## **Computations for Unequal Sample Sizes (optional)**

The calculation of MSE for unequal sample sizes is similar to its calculation in an independent-groups t test. Here are the steps:

1. Compute a Sum of Squares Error (SSE) using the following formula

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2 + \dots + \sum (X - M_k)^2$$

where  $M_i$  is the mean of the  $i^{\text{th}}$  group and  $k$  is the number of groups.

2. Compute the degrees of freedom error ( $df_e$ ) by subtracting the number of groups ( $k$ ) from the total number of observations ( $N$ ). Therefore,

$$df_e = N - k.$$

Compute MSE by dividing SSE by  $df_e$ :

$$MSE = SSE / df_e.$$

For each comparison of means, use the harmonic mean of the  $n$ 's for the two means ( $n_h$ ).

All other aspects of the calculations are the same as when you have equal sample sizes.

# Specific Comparisons (Independent Groups)

by David M. Lane

## *Prerequisites*

- Chapter 12: Difference Between Two Means (Independent Groups)

## *Learning Objectives*

1. Define linear combination
2. Specify a linear combination in terms of coefficients
3. Do a significance test for a specific comparison

There are many situations in which the comparisons among means are more complicated than simply comparing one mean with another. This section shows how to test these more complex comparisons. The methods in this section assume that the comparison among means was decided on before looking at the data.

Therefore these comparisons are called *planned comparisons*. A different procedure is necessary for *unplanned comparisons*.

Let's begin with the made-up data from a hypothetical experiment shown in Table 1. Twelve subjects were selected from a population of high-self-esteem subjects (esteem = 1) and an additional 12 subjects were selected from a population of low-self-esteem subjects (esteem = 2). Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they succeeded (outcome = 1) and the other half were told they failed (outcome = 2). Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects all together.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table 1. Data from Hypothetical Experiment.

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7
2	2	6

The means of the four conditions are shown in Table 2.

Table 2. Mean ratings of self-attributions of success or failure.

Outcome	Esteem	Mean
Success	High Self-Esteem	7.333
	Low Self-Esteem	5.5
Failure	High Self-Esteem	4.833
	Low Self-Esteem	7.833

There are several questions we can ask about the data. We begin by asking whether, on average, subjects who were told they succeeded differed significantly from subjects who were told they failed. The means for subjects in the success condition are 7.333 for the high-self-esteem subjects and 5.500 for the low-self-esteem subjects. Therefore, the mean for all subjects in the success condition is  $(7.333 + 5.500)/2 = 6.4167$ . Similarly, the mean for all subjects in the failure condition is  $(4.833 + 7.833)/2 = 6.333$ . The question is: How do we do a significance test for this difference of  $6.4167 - 6.333 = 0.083$ ?

The first step is to express this difference in terms of a linear combination using a set of coefficients and the means. This may sound complex, but it is really pretty easy. We can compute the mean of the success conditions by multiplying each success mean by 0.5 and then adding the result. In other words, we compute

$$\begin{aligned}
 (.5)(7.333) + (.5)(5.500) \\
 = 3.67 + 2.75 \\
 = 6.42
 \end{aligned}$$

Similarly we can compute the mean of the failure conditions by multiplying each failure mean by 0.5 and then adding the result:

$$\begin{aligned}
 (.5)(4.833) + (.5)(7.833) \\
 = 2.417 + 3.917 \\
 = 6.33
 \end{aligned}$$

The difference between the two means can be expressed as

$$\begin{aligned}
 .5 \times 7.333 + .5 \times 5.500 - (.5 \times 4.833 + .5 \times 7.833) = \\
 .5 \times 7.333 + .5 \times 5.500 - .5 \times 4.833 - .5 \times 7.833
 \end{aligned}$$

We therefore can compute the difference between the “success” mean and the “failure” mean by multiplying each “success” mean by 0.5, each “failure” mean by -0.5 and adding the results. In Table 3, the coefficient column is the multiplier and the product column in the result of the multiplication. If we add up the four values in the product column we get:

$$L = 3.667 + 2.750 - 2.417 - 3.917 = 0.083$$

This is the same value we got when we computed the difference between means previously (within rounding error). We call the value “L” for “linear combination.”

Table 3. Coefficients for comparing low and high self-esteem.

Outcome	Esteem	Mean	Coeff	Product
Success	High Self-Esteem	7.333	0.5	3.667
	Low Self-Esteem	5.5	0.5	2.75
Failure	High Self-Esteem	4.833	-0.5	-2.417
	Low Self-Esteem	7.833	-0.5	-3.917

Now, the question is whether our value of L is significantly different from 0. The general formula for L is

$$L = \sum c_i M_i$$

where  $c_i$  is the  $i^{\text{th}}$  coefficient and  $M_i$  is the  $i^{\text{th}}$  mean. As shown above,  $L = 0.083$ . The formula for testing L for significance is shown below:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}}$$

In this example,

$$\sum c_i^2 = .5^2 + .5^2 + (-.5)^2 + (-.5)^2 = 1$$

MSE is the mean of the variances. The four variances are shown in Table 4. Their mean is 1.625. Therefore  $MSE = 1.625$ .

Table 4. Variances of attributions of success or failure to oneself.

Outcome	Esteem	Variance
Success	High Self-Esteem	1.867
	Low Self-Esteem	1.1
Failure	High Self-Esteem	2.167
	Low Self-Esteem	1.367

The value of  $n$  is the number of subjects in each group. Here  $n = 6$ .

Putting it all together,

$$t = \frac{0.083}{\sqrt{\frac{(1)(1.625)}{6}}} = 0.16.$$

We need to know the degrees for freedom in order to compute the probability value. The degrees of freedom is

$$df = N - k$$

where  $N$  is the total number of subjects (24) and  $k$  is the number of groups (4). Therefore,  $df = 20$ . Using the Online Calculator, we find that the two-tailed probability value is 0.874. Therefore, the difference between the “success” condition and the “failure” condition is not significant.

A more interesting question about the results is whether the effect of outcome (success or failure) differs depending on the self-esteem of the subject. For example, success may make high-self-esteem subjects **more** likely to attribute the outcome to themselves, whereas success may make low-self-esteem subjects **less** likely to attribute the outcome to themselves.

To test this, we have to test a difference between differences. Specifically, is the difference between success and failure outcomes for the high-self-esteem subjects different from the difference between success and failure outcomes for the low-self-esteem subjects? The means in Table 5 suggest that this is the case. For the high-self-esteem subjects, the difference between the success and failure

attribution scores is  $7.333 - 4.833 = 2.500$ . For low-self-esteem subjects, the difference is  $5.500 - 7.833 = -2.333$ . The difference between differences is  $2.500 - (-2.333) = 4.833$ .

The coefficients to test this difference between differences are shown in Table 5.

*Table 5. Coefficients for testing differences between differences.*

Self-Esteem	Outcome	Mean	Coefficient	Product
High	Success	7.333	1	7.333
	Failure	4.833	-1	-4.833
Low	Success	5.5	-1	-5.5
	Failure	7.833	1	7.833

If it is hard to see where these coefficients came from, consider that our difference between differences was computed this way:

$$\begin{aligned}
 & (7.33 - 4.83) - (5.5 - 7.83) \\
 &= 7.3 - 4.83 - 5.5 + 7.83 \\
 &= (1)7.3 + (-1)4.83 + (-1)5.5 + (1)7.83
 \end{aligned}$$

The values in parentheses are the coefficients.

To continue the calculations,

$$L = 4.83$$

$$\sum c_i^2 = 1^2 + (-1)^2 + (-1)^2 + (1)^2 = 4$$

$$t = \frac{4.83}{\sqrt{\frac{(4)(1.625)}{6}}} = 4.64$$

The two-tailed p value is 0.0002. Therefore, the difference between differences is highly significant.

In a later chapter on Analysis of Variance, you will see that comparisons such as this are testing what is called an *interaction*. In general, there is an interaction when the effect of one variable differs as a function of the level of another variable. . In this example, the effect of the outcome variable is different depending on the subject's self-esteem. For the high-self-esteem subjects, success led to more self-attribution than did failure; for the low-self-esteem subjects, success led to less self-attribution than did failure.

## Multiple Comparisons

The more comparisons you make, the greater your chance of a Type I error. It is useful to distinguish between two error rates: (1) the *per-comparison error rate* and (2) the *familywise error rate*. The per-comparison error rate is the probability of a Type I error for a particular comparison. The *familywise error rate* is the probability of making one or more Type I errors in a family or set of comparisons. In the attribution experiment discussed previously, we computed two comparisons. If we use the 0.05 level for each comparison, then the per-comparison rate is simply 0.05. The familywise rate can be complex. Fortunately, there is a simple approximation that is fairly accurate when the number of comparisons is small. Defining  $\alpha$  as the per-comparison error rate and  $c$  as the number of comparisons, the following inequality always holds true for the familywise error rate (FW):

$$FW \leq c\alpha$$

This inequality is called the *Bonferroni inequality*. In practice, FW can be approximated by  $c\alpha$ . This is a conservative approximation since FW can never be greater than  $c\alpha$  and is generally less than  $c\alpha$ .

The Bonferroni inequality can be used to control the familywise error rate as follows: If you want the familywise error rate to be  $\alpha$ , you use  $\alpha/c$  as the per-comparison error rate. This correction, called the *Bonferroni correction*, will generally result in a familywise error rate less than  $\alpha$ . Alternatively, you could multiply the by  $c$  and use the original  $\alpha$  level.

Should the familywise error rate be controlled? Unfortunately, there is no clear-cut answer to this question. The disadvantage of controlling the familywise error rate is that it makes it more difficult to obtain a significant result for any given comparison: The more comparisons you do, the lower the per-comparison rate must be and therefore the harder it is to reach significance. That is, the power

is lower when you control the familywise error rate. The advantage is that you have a lower chance of making a Type I error.

One consideration is the definition of a family of comparisons. Let's say you conducted a study in which you were interested in whether there was a difference between male and female babies in the age at which they started crawling. After you finished analyzing the data, a colleague of yours had a totally different research question: Do babies who are born in the winter differ from those born in the summer in the age they start crawling? Should the familywise rate be controlled or should it be allowed to be greater than 0.05? Our view is that there is no reason you should be penalized (by lower power) just because your colleague used the same data to address a different research question. Therefore, the familywise error rate need not be controlled. Consider the two comparisons done on the attribution example at the beginning of this section: These comparisons are testing completely different hypotheses. Therefore, controlling the familywise rate is not necessary.

Now consider a study designed to investigate the relationship between various variables and the ability of subjects to predict the outcome of a coin flip. One comparison is between males and females; a second comparison is between those over 40 and those under 40; a third is between vegetarians and non-vegetarians; and a fourth is between firstborns and others. The question of whether these four comparisons are testing different hypotheses depends on your point of view. On the one hand, there is nothing about whether age makes a difference that is related to whether diet makes a difference. In that sense, the comparisons are addressing different hypotheses. On the other hand, the whole series of comparisons could be seen as addressing the general question of whether anything affects the ability to predict the outcome of a coin flip. If nothing does, then allowing the familywise rate to be high means that there is a high probability of reaching the wrong conclusion.

## **Orthogonal Comparisons**

In the preceding sections, we talked about comparisons being independent. Independent comparisons are often called orthogonal comparisons. There is a simple test to determine whether two comparisons are orthogonal: If the sum of the products of the coefficients is 0, then the comparisons are orthogonal. Consider again the experiment on the attribution of success or failure. Table 6 shows the coefficients previously presented in Table 3 and in Table 5. The column "C1"

contains the coefficients from the comparison shown in Table 3; the column “C2” contains the coefficients from the comparison shown in Table 5. The column labeled “Product” is the product of these two columns. Note that the sum of the numbers in this column is 0. Therefore, the two comparisons are orthogonal.

Table 6. Coefficients for two orthogonal comparisons.

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	1	0.5
	Low Self-Esteem	0.5	-1	-0.5
Failure	High Self-Esteem	-0.5	-1	0.5
	Low Self-Esteem	-0.5	1	-0.5

Table 7 shows two comparisons that are not orthogonal. The first compares the high-self-esteem subjects to the low-self-esteem subjects; the second considers only those in the success group and compares high-self-esteem subjects to low-self-esteem subjects. The failure group is ignored by using 0's as coefficients. Clearly the comparison of high-self-esteem subjects to low-self-esteem subjects for the whole sample is not independent of the comparison for the success group only. You can see that the sum of the products of the coefficients is 0.5 and not 0.

Table 7. Coefficients for two non-orthogonal comparisons.

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	0.5	0.25
	Low Self-Esteem	-0.5	-0.5	0.25
Failure	High Self-Esteem	0.5	0	0
	Low Self-Esteem	-0.5	0	0

# Difference Between Two Means (Correlated Pairs)

by David M. Lane

## *Prerequisites*

- Chapter 4: Values of the Pearson Correlation
- Chapter 10: t Distribution
- Chapter 11: Hypothesis Testing
- Chapter 12: Testing a Single Mean
- Chapter 12: Difference Between Two Means (Independent Groups)

## *Learning Objectives*

1. Determine whether you have correlated pairs or independent groups
2. Compute a t test for correlated pairs

Let's consider how to analyze the data from the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. In this section, we will be concerned only with testing the difference between the mean of the placebo (D0) condition and the mean of the highest dosage condition (D60). The first question is why the difference between means should not be tested using the procedure described in the section Difference Between Two Means (Independent Groups). The answer lies in the fact that in this experiment we do not have independent groups. The scores in the D0 condition are from the same subjects as the scores in the D60 condition. There is only one group of subjects, each subject being tested in both the D0 and D60 conditions.

Figure 1 shows a scatter plot of the 60-mg scores (D60) as a function of the 0-mg scores (D0). It is clear that children who get more correct in the D0 condition tend to get more correct in the D60 condition. The correlation between the two conditions is high:  $r = 0.80$ . Clearly these two variables are not independent.

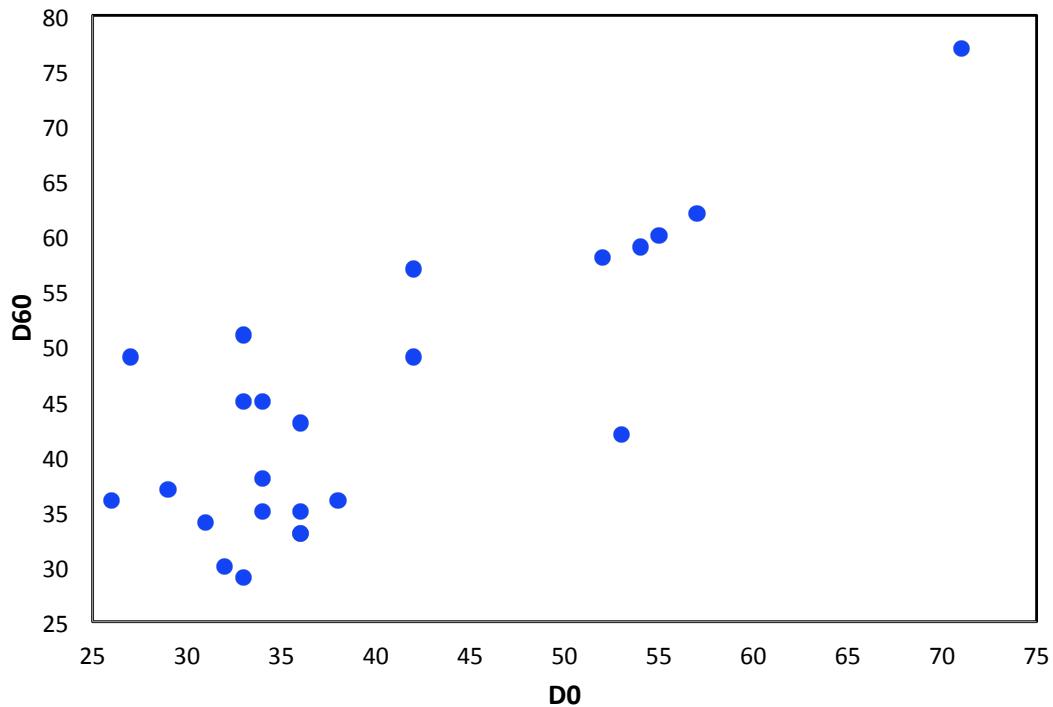


Figure 1. Number of correct responses made in the 60-mg condition as a function of the number of correct responses in the 0-mg condition.

## Computations

You may recall that the method to test the difference between these means was presented in the section on “Testing a Single Mean.” The computational procedure is to compute the difference between the D60 and the D0 conditions for each child and test whether the mean difference is significantly different from 0. The difference scores are shown in Table 1. As shown in the section on testing a single mean, the mean difference score is 4.96 which is significantly different from 0:  $t = 3.22$ ,  $df = 23$ ,  $p = 0.0038$ . This  $t$  test has various names including “*correlated t test*” and “*related-pairs t test*.”

In general, the correlated  $t$  test is computed by first computing the differences between the two scores for each subject. Then, a test of a single mean is computed on the mean of these difference scores.

Table 1. DOG scores as a function of dosage.

D0	D60	D60-D0
57	62	5
27	49	22

32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

If you had mistakenly used the method for an independent-groups t test with these data, you would have found that  $t = 1.42$ ,  $df = 46$ , and  $p = 0.15$ . That is, the difference between means would not have been found to be statistically significant. This is a typical result: correlated t tests almost always have greater power than independent-groups t tests. This is because in correlated t tests, each difference score is a comparison of performance in one condition with the performance of that same subject in another condition. This makes each subject “their own control” and

keeps differences between subjects from entering into the analysis. The result is that the standard error of the difference between means is smaller in the correlated t test and, since this term is in the denominator of the formula for t, results in a larger t.

### **Details about the Standard Error of the Difference between Means (Optional)**

To see why the standard error of the difference between means is smaller in a correlated t test, consider the variance of difference scores. As shown in the section on the Variance Sum Law, the variance of the sum or difference of the two variables X and Y is:

$$s_{X \pm Y}^2 = s_X^2 + s_Y^2 \pm 2rs_Xs_Y$$

Therefore, the variance of difference scores is the variance in the first condition (X) plus the variance in the second condition (Y) minus twice the product of (1) the correlation, (2) the standard deviation of X, and (3) the standard deviation of Y. For the current example,  $r = 0.80$  and the variances and standard deviations are shown in Table 2.

Table 2. Variances and Standard Deviations

	<b>D0</b>	<b>D60</b>	<b>D60 - D0</b>
Variance	128.02	151.78	56.82
Sd	11.31	12.32	7.54

The variance of the difference scores of 56.82 can be computed as:

$$128.02 + 151.78 - (2)(0.80)(11.31)(12.32)$$

which is equal to 56.82 except for rounding error. Notice that the higher the correlation, the lower the standard error of the mean.

# Specific Comparisons (Correlated Observations)

by David M. Lane

## *Prerequisites*

- Chapter 10: t Distribution
- Chapter 12: Hypothesis Testing, Testing a Single Mean
- Chapter 12: Specific Comparisons
- Chapter 12: Difference Between Two Means (Correlated Pairs)

## *Learning Objectives*

1. Determine whether to use the formula for correlated comparisons or independent-groups comparisons
2. Compute t for a comparison for repeated-measures data

In the "Weapons and Aggression" case study, subjects were asked to read words presented on a computer screen as quickly as they could. Some of the words were aggressive words such as injure or shatter. Others were control words such as relocate or consider. These two types of words were preceded by words that were either the names of weapons, such as shotgun or grenade, or non-weapon words, such as rabbit or fish. For each subject, the mean reading time across words was computed for these four conditions. The four conditions are labeled as shown in Table 1. Table 2 shows the data from five subjects.

Table 1. Description of Conditions.

Variable	Description
aw	The time in milliseconds (msec) to name an aggressive word following a weapon word prime.
an	The time in milliseconds (msec) to name an aggressive word following a non-weapon word prime.
cw	The time in milliseconds (msec) to name a control word following a weapon word prime.
cn	The time in milliseconds (msec) to name a control word following a non-weapon word prime.

Table 2. Data from Five Subjects

Subject	aw	an	cw	cn
1	447	440	432	452
2	427	437	469	451
3	417	418	445	434
4	348	371	353	344
5	471	443	462	463

One question was whether reading times would be shorter when the preceding word was a weapon word (aw and cw conditions) than when it was a non-weapon word (an and cn conditions). In other words, is

$$L_1 = (an + cn) - (aw + cw)$$

greater than 0? This is tested for significance by computing  $L_1$  for each subject and then testing whether the mean value of  $L_1$  is significantly different from 0. Table 3 shows  $L_1$  for the first five subjects.  $L_1$  for Subject 1 was computed by

$$L_1 = (440 + 452) - (447 + 432) = 892 - 879 = 13$$

Table 3.  $L_1$  for Five Subjects

Subject	aw	an	cw	cn	$L_1$
1	447	440	432	452	13
2	427	437	469	451	-8
3	417	418	445	434	-10
4	348	371	353	344	14
5	471	443	462	463	-27

Once  $L_1$  is computed for each subject, the significance test described in the section “Testing a Single Mean” can be used. First we compute the mean and the standard error of the mean for  $L_1$ . There were 32 subjects in the experiment. Computing  $L_1$  for the 32 subjects, we find that the mean and standard error of the mean are 5.875 and 4.2646, respectively. We then compute

$$t = \frac{M - \mu}{s_M}$$

where  $M$  is the sample mean,  $\mu$  is the hypothesized value of the population mean (0 in this case), and  $s_M$  is the estimated standard error of the mean. The calculations show that  $t = 1.378$ . Since there were 32 subjects, the degrees of freedom is  $32 - 1 = 31$ . The  $t$  distribution calculator shows that the two-tailed probability is 0.178.

A more interesting question is whether the priming effect (the difference between words preceded by a non-weapon word and words preceded by a weapon word) is different for aggressive words than it is for non-aggressive words. That is, do weapon words prime aggressive words more than they prime non-aggressive words? The priming of aggressive words is  $(an - aw)$ . The priming of non-aggressive words is  $(cn - cw)$ . The comparison is the difference:

$$L_2 = (an - aw) - (cn - cw).$$

Table 4 shows  $L_2$  for five of the 32 subjects.

Table 4.  $L_2$  for Five Subjects

Subject	aw	an	cw	cn	$L_2$
1	447	440	432	452	-27
2	427	437	469	451	28
3	417	418	445	434	12
4	348	371	353	344	32
5	471	443	462	463	-29

The mean and standard error of the mean for all 32 subjects are 8.4375 and 3.9128, respectively. Therefore,  $t = 2.156$  and  $p = 0.039$ .

## Multiple Comparisons

Issues associated with doing multiple comparisons are the same for related observations as they are for multiple comparisons among independent groups.

## Orthogonal Comparisons

The most straightforward way to assess the degree of dependence between two comparisons is to correlate them directly. For the weapons and aggression data, the comparisons  $L_1$  and  $L_2$  are correlated 0.24. Of course, this is a sample correlation and only estimates what the correlation would be if  $L_1$  and  $L_2$  were correlated in the population. Although mathematically possible, orthogonal comparisons with correlated observations are very rare.

# Pairwise Comparisons (Correlated Observations)

by David M. Lane

## *Prerequisites*

- Chapter 12: Difference between Two Means (Independent Groups)
- Chapter 12: All Pairwise Comparisons Among Means
- Chapter 12: Difference Between Two Means
- Chapter 12: Difference Between Two Means ( Correlated Pairs)
- Chapter 12: Specific Comparisons (Independent Groups)
- Chapter 12: Specific Comparisons (Correlated Observations)

## *Learning Objectives*

1. Compute the Bonferroni correction
2. Calculate pairwise comparisons using the Bonferroni correction

In the section on all pairwise comparisons among independent groups, the *Tukey HSD test* was the recommended procedure. However, when you have one group with several scores from the same subjects, the Tukey test makes an assumption that is unlikely to hold: The variance of difference scores is the same for all pairwise differences between means.

The standard practice for pairwise comparisons with *correlated observations* is to compare each pair of means using the method outlined in the section “Difference Between Two Means (Correlated Pairs)” with the addition of the *Bonferroni correction* described in the section “Specific Comparisons.” For example, suppose you were going to do all pairwise comparisons among four means and hold the familywise error rate at 0.05. Since there are six possible pairwise comparisons among four means, you would use  $0.05/6 = 0.0083$  for the per-comparison error rate.

As an example, consider the case study “Stroop Interference.” There were three tasks each performed by 47 subjects. In the “words” task, subjects read the names of 60 color words written in black ink; in the “color” task, subjects named the colors of 60 rectangles; in the “interference” task, subjects named the ink color of 60 conflicting color words. The times to read the stimuli were recorded. In order to compute all pairwise comparisons, the difference in times for each pair of conditions for each subject is calculated. Table 1 shows these scores for five of the 47 subjects.

Table 1. Pairwise Differences

W-C	W-I	C-I
-3	-24	-21
2	-41	-43
-1	-18	-17
-4	-23	-19
-2	-17	-15

The means, standard deviations (Sd), and *standard error of the mean* (Sem), t, and p for all 47 subjects are shown in Table 2. The t's are computed by dividing the means by the standard errors of the mean. Since there are 47 subjects, the degrees of freedom is 46. Notice how different the standard deviations are. For the Tukey test to be valid, all population values of the standard deviation would have to be the same.

Table 2. Pairwise Comparisons.

Comparison	Mean	Sd	Sem	t	p
W-C	-4.15	2.99	0.44	-9.53	<0.001
W-I	-20.51	7.84	1.14	-17.93	<0.001
C-I	-16.36	7.47	1.09	-15.02	<0.001

Using the Bonferroni correction for three comparisons, the p value has to be below  $0.05/3 = 0.0167$  for an effect to be significant at the 0.05 level. For these data, all p values are far below that, and therefore all pairwise differences are significant.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 12: Single Mean

Research on the effectiveness of surgery for weight loss [reported here](#) found that "The surgery was associated with significantly greater weight loss [than the control group who dieted] through 2 years (61.3 versus 11.2 pounds,  $p < 0.001$ )."

## What do you think?

What test could have been used and how would it have been computed?

For each subject a difference score between their initial weight and final weight could be computed. A t test of whether the mean difference score differs significantly from 0 could then be computed. The mean difference score will equal the difference between the mean weight losses of the two groups ( $61.3 - 11.2 = 50.1$ ).

## References

- Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

## Exercises

### Prerequisites

- All material presented in the Testing Means chapter

1. The scores of a random sample of 8 students on a physics test are as follows: 60, 62, 67, 69, 70, 72, 75, and 78.
  - a. Test to see if the sample mean is significantly different from 65 at the .05 level. Report the t and p values.
  - b. The researcher realizes that she accidentally recorded the score that should have been 76 as 67. Are these corrected scores significantly different from 65 at the .05 level?
2. A (hypothetical) experiment is conducted on the effect of alcohol on perceptual motor ability. Ten subjects are each tested twice, once after having two drinks and once after having two glasses of water. The two tests were on two different days to give the alcohol a chance to wear off. Half of the subjects were given alcohol first and half were given water first. The scores of the 10 subjects are shown below. The first number for each subject is their performance in the “water” condition. Higher scores reflect better performance. Test to see if alcohol had a significant effect. Report the t and p values.

water	alcohol
16	13
15	13
11	10
20	18
19	17
14	11
13	10
15	15
14	11
16	16

3. The scores on a (hypothetical) vocabulary test of a group of 20 year olds and a group of 60 year olds are shown below.

20 yr olds	60 yr olds
27	26

26	29
21	29
24	29
15	27
18	16
17	20
12	27
13	

- a. Test the mean difference for significance using the .05 level.
- b. List the assumptions made in computing your answer.
4. The sampling distribution of a statistic is normally distributed with an estimated standard error of 12 (df = 20). (a) What is the probability that you would have gotten a mean of 107 (or more extreme) if the population parameter were 100? Is this probability significant at the .05 level (two-tailed)? (b) What is the probability that you would have gotten a mean of 95 or less (one-tailed)? Is this probability significant at the .05 level? You may want to use the t Distribution calculator for this problem.
5. How do you decide whether to use an independent groups t test or a correlated t test (test of dependent means)?
6. An experiment compared the ability of three groups of subjects to remember briefly-presented chess positions. The data are shown below.

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

- a. Using the Tukey HSD procedure, determine which groups are significantly different from each other at the .05 level.

- b. Now compare each pair of groups using t-tests. Make sure to control for the familywise error rate (at 0.05) by using the Bonferroni correction. Specify the alpha level you used.
7. Below are data showing the results of six subjects on a memory test. The three scores per subject are their scores on three trials (a, b, and c) of a memory task. Are the subjects getting better each trial? Test the linear effect of trial for the data.

a	b	c
4	6	7
3	7	8
2	8	5
1	4	7
4	6	9
2	4	2

- a. Compute L for each subject using the contrast weights -1, 0, and 1. That is, compute  $(-1)(a) + (0)(b) + (1)(c)$  for each subject.
- b. Compute a one-sample t-test on this column (with the L values for each subject) you created.
8. Participants threw darts at a target. In one condition, they used their preferred hand; in the other condition, they used their other hand. All subjects performed in both conditions (the order of conditions was counterbalanced). Their scores are shown below.

Preferred	Non-preferred
12	7
7	9
11	8
13	10
10	9

- a. Which kind of t-test should be used?
- b. Calculate the two-tailed t and p values using this t test.
- c. Calculate the one-tailed t and p values using this t test.
9. Assume the data in the previous problem were collected using two different groups of subjects: One group used their preferred hand and the other group used

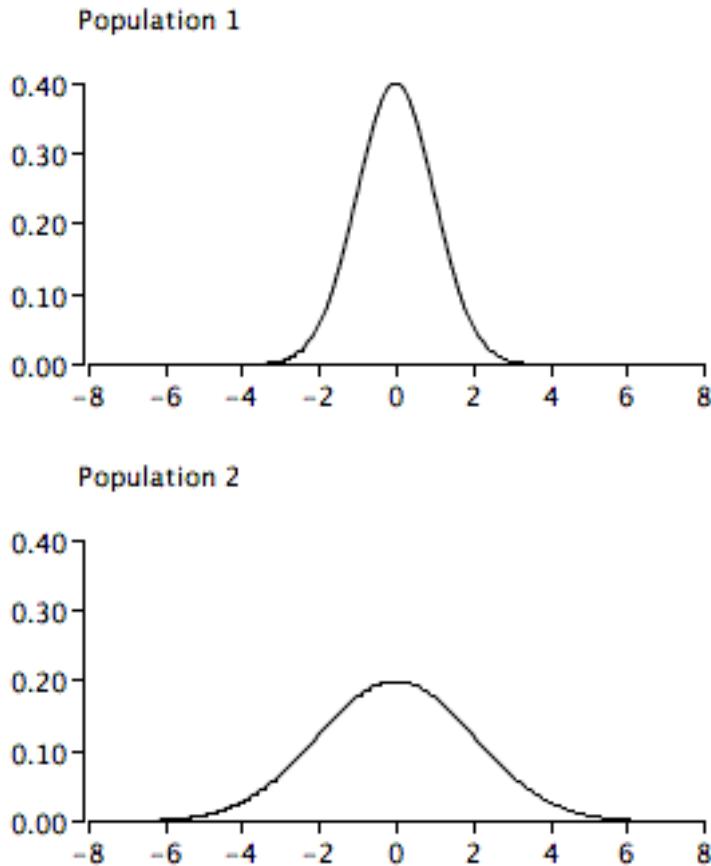
their non-preferred hand. Analyze the data and compare the results to those for the previous problem.

10. You have 4 means, and you want to compare each mean to every other mean.  
(a) How many tests total are you going to compute? (b) What would be the chance of making at least one Type I error if the Type I error for each test was .05 and the tests were independent? (c) Are the tests independent and how does independence/non-independence affect the probability in (b).
11. In an experiment, participants were divided into 4 groups. There were 20 participants in each group, so the degrees of freedom (error) for this study was  $80 - 4 = 76$ . Tukey's HSD test was performed on the data. (a) Calculate the p value for each pair based on the Q value given below. You will want to use the Studentized Range Calculator. (b) Which differences are significant at the .05 level?

Comparison of Groups	Q
A - B	3.4
A - C	3.8
A - D	4.3
B - C	1.7
B - D	3.9
C - D	3.7

12. If you have 5 groups in your study, why shouldn't you just compute a t test of each group mean with each other group mean?
13. You are conducting a study to see if students do better when they study all at once or in intervals. One group of 12 participants took a test after studying for one hour continuously. The other group of 12 participants took a test after studying for three twenty minute sessions. The first group had a mean score of 75 and a variance of 120. The second group had a mean score of 86 and a variance of 100.
  - a. What is the calculated t value? Are the mean test scores of these two groups significantly different at the .05 level?
  - b. What would the t value be if there were only 6 participants in each group? Would the scores be significant at the .05 level?

14. A new test was designed to have a mean of 80 and a standard deviation of 10. A random sample of 20 students at your school take the test, and the mean score turns out to be 85. Does this score differ significantly from 80?
15. You perform a one-sample t test and calculate a t statistic of 3.0. The mean of your sample was 1.3 and the standard deviation was 2.6. How many participants were used in this study?
16. True/false: The contrasts  $(-3, 1 1 1)$  and  $(0, 0, -1, 1)$  are orthogonal.
17. True/false: If you are making 4 comparisons between means, then based on the Bonferroni correction, you should use an alpha level of .01 for each test.
18. True/false: Correlated t tests almost always have greater power than independent t tests.
19. True/false: The graph below represents a violation of the homogeneity of variance assumption.



20. True/false: When you are conducting a one-sample t test and you know the population standard deviation, you look up the critical t value in the table based on the degrees of freedom.

*Questions from Case Studies*

Angry Moods (AM) case study

21. (AM) Do athletes or non-athletes calm down more when angry? Conduct a t test to see if the difference between groups in Control-In scores is statistically significant.
22. (AM) Do people in general have a higher Anger-Out or Anger-In score? Conduct a t test on the difference between means of these two scores. Are these two means independent or dependent?

Smiles and Leniency (SL) case study

23. (SL) Compare each mean to the neutral mean. Be sure to control for the familywise error rate.
24. (SL) Does a “felt smile” lead to more leniency than other types of smiles? (a) Calculate L (the linear combination) using the following contrast weights false: -1, felt: 2, miserable: -1, neutral: 0. (b) Perform a significance test on this value of L.

#### Animal Research (AR) case study

25. (AR) Conduct an independent samples t test comparing males to females on the belief that animal research is necessary.
26. (AR) Based on the t test you conducted in the previous problem, are you able to reject the null hypothesis if alpha = 0.05? What about if alpha = 0.1?
27. (AR) Is there any evidence that the t test assumption of homogeneity of variance is violated in the t test you computed in #25?

#### ADHD Treatment (AT) case study

28. (AT) Compare each dosage with the dosage below it (compare d0 and d15, d15 and d30, and d30 and d60). Remember that the patients completed the task after every dosage. (a) If the familywise error rate is .05, what is the alpha level you will use for each comparison when doing the Bonferroni correction? (b) Which differences are significant at this level?
29. (AT) Does performance increase linearly with dosage?
- Plot a line graph of this data.
  - Compute L for each patient. To do this, create a new variable where you multiply the following coefficients by their corresponding dosages and then sum up the total:  $(-3)d0 + (-1)d15 + (1)d30 + (3)d60$  (see #7). What is the mean of L?
  - Perform a significance test on L. Compute the 95% confidence interval for L.

# 13. Power

- A. Introduction
- B. Example Calculations
- C. Factors Affecting Power
- D. Exercises

# Introduction to Power

by David M. Lane

## *Prerequisites*

- Chapter 11: Significance Testing
- Chapter 11: Type I and Type II Errors
- Chapter 11: Misconceptions

## *Learning Objectives*

1. Define power
2. Identify situations in which it is important to estimate power

Suppose you work for a foundation whose mission is to support researchers in mathematics education and your role is to evaluate grant proposals and decide which ones to fund. You receive a proposal to evaluate a new method of teaching high-school algebra. The research plan is to compare the achievement of students taught by the new method with the achievement of students taught by the traditional method. The proposal contains good theoretical arguments why the new method should be superior and the proposed methodology is sound. In addition to these positive elements, there is one important question still to be answered: Does the experiment have a high probability of providing strong evidence that the new method is better than the standard method if, in fact, the new method is actually better? It is possible, for example, that the proposed sample size is so small that even a fairly large population difference would be difficult to detect. That is, if the sample size is small, then even a fairly large difference in sample means might not be significant. If the difference is not significant, then no strong conclusions can be drawn about the population means. It is not justified to conclude that the null hypothesis that the population means are equal is true just because the difference is not significant. Of course, it is not justified to conclude that this null hypothesis is false. Therefore, when an effect is not significant, the result is inconclusive. You may prefer that your foundation's money be used to fund a project that has a higher probability of being able to make a strong conclusion.

Power is defined as the probability of correctly rejecting a false null hypothesis. In terms of our example, it is the probability that given there is a difference between the population means of the new method and the standard method, the sample means will be significantly different. The probability of failing

to reject a false null hypothesis is often referred to as  $\beta$  (the Greek letter beta). Therefore power can be defined as:

$$\text{power} = 1 - \beta.$$

It is very important to consider power while designing an experiment. You should avoid spending a lot of time and/or money on an experiment that has little chance of finding a *significant* effect.

# Example Calculations

by David M. Lane

## Prerequisites

- Chapter 5: Binomial Distribution
- Chapter 12: Testing a Single Mean
- Chapter 13: Introduction to Power

## Learning Objectives

1. Compute power using the binomial distribution
2. Compute power using the normal distribution
3. Use a power calculator to compute power for the t distribution

In the “Shaking and Stirring Martinis” case study, the question was whether Mr. Bond could tell the difference between martinis that were stirred and martinis that were shaken. For the sake of this example, assume he can tell the difference and is able to correctly state whether a martini had been shaken or stirred 0.75 of the time. Now, suppose an experiment is being conducted to investigate whether Mr. Bond can tell the difference. Specifically, is Mr. Bond correct more than 0.50 of the time? We know that he is (that's an assumption of the example). However, the experimenter does not know and asks Mr. Bond to judge 16 martinis. The experimenter will do a *significance* test based on the binomial distribution. Specifically, if a *one tailed* test is significant at the 0.05 level, then he or she will conclude that Mr. Bond can tell the difference. The probability value is computed assuming the *null hypothesis* is true ( $\pi = 0.50$ ). Therefore, the experimenter will determine how many times Mr. Bond is correct, and compute the probability of being correct that many or more times given that the null hypothesis is true. The question is: what is the probability the experimenter will correctly reject the null hypothesis that  $\pi = 0.50$ ? In other words, what is the power of this experiment?

The binomial distribution for  $N = 16$  and  $\pi = 0.50$  is shown in Figure 1. The probability of being correct on 11 or more trials is 0.105 and the probability of being correct on 12 or more trials is 0.038. Therefore, the probability of being correct on 12 or more trials is less than 0.05. This means that the null hypothesis will be rejected if Mr. Bond is correct on 12 or more trials and will not be rejected otherwise.

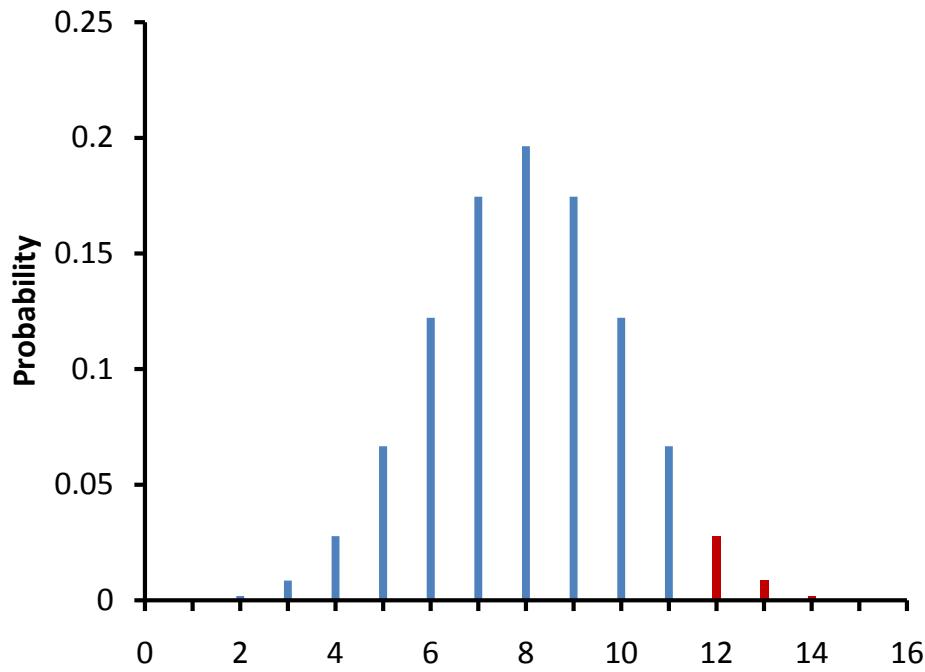


Figure 1. The binomial distribution for  $N = 16$  and  $\pi = 0.50$ .

We know that Mr. Bond is correct 0.75 of the time. (Obviously the experimenter does not know this or there would be no need for an experiment.) The binomial distribution with  $N = 16$  and  $\pi = 0.75$  is shown in Figure 2.

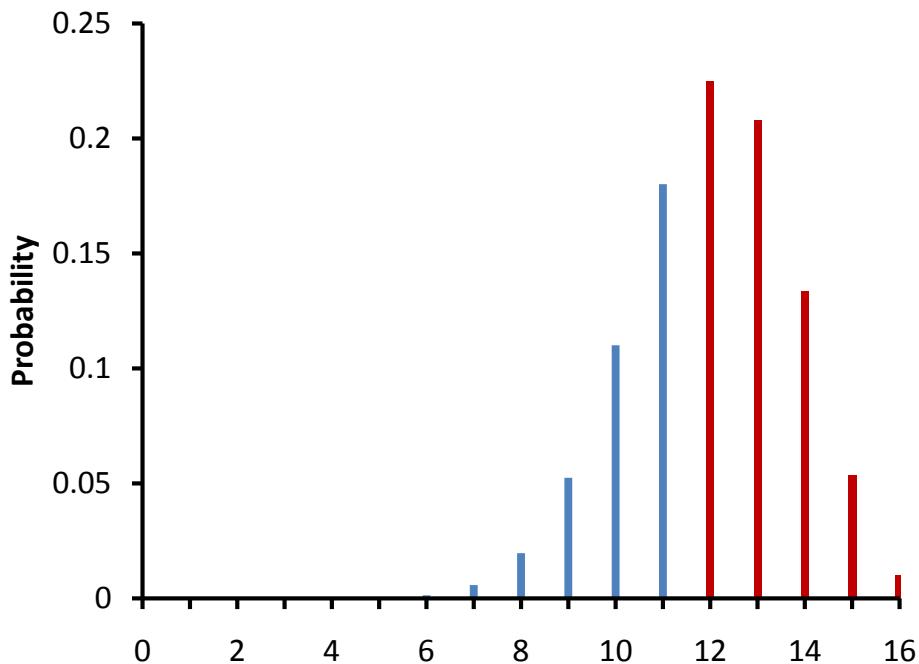


Figure 2. The binomial distribution for  $N = 16$  and  $\pi = 0.75$ .

The probability of being correct on 12 or more trials is 0.63. Therefore, the power of the experiment is 0.63.

To sum up, the probability of being correct on 12 or more trials given that the null hypothesis is true is less than 0.05. Therefore, if Mr. Bond is correct on 12 or more trials, the null hypothesis will be rejected. Given Mr. Bond's true ability to be correct on 0.75 of the trials, the probability he will be correct on 12 or more trials is 0.63. Therefore power is 0.63.

In the section on testing a single mean for significance in Chapter 12, the first example was based on the assumption that the experimenter knew the population variance. Although this is rarely true in practice, the example is very useful for pedagogical purposes. For the same reason, the following example assumes the experimenter knows the population variance. Power calculators are available for situations in which the experimenter does not know the population variance.

Suppose a math achievement test were known to have a mean of 75 and a standard deviation of 10. A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean for the new method is 80. The researcher plans to sample 25 subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. What is the probability that the researcher will correctly reject the false null hypothesis that the population mean for the new method is 75 or lower? The following shows how this probability is computed.

The researcher assumes that the population standard deviation with the new method is the same as with the old method (10) and that the distribution is normal. Since the population standard deviation is assumed to be known, the researcher can use the *normal distribution* rather than the t distribution to compute the p value. Recall that the standard error of the mean ( $\sigma_M$ ) is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

which is equal to  $10/5 = 2$  in this example. As can be seen in Figure 3, if the null hypothesis that the population mean equals 75 is true, then the probability of a sample mean being greater than or equal to 78.29 is 0.05. Therefore, the experimenter will reject the null hypothesis if the sample mean,  $M$ , is 78.29 or larger.

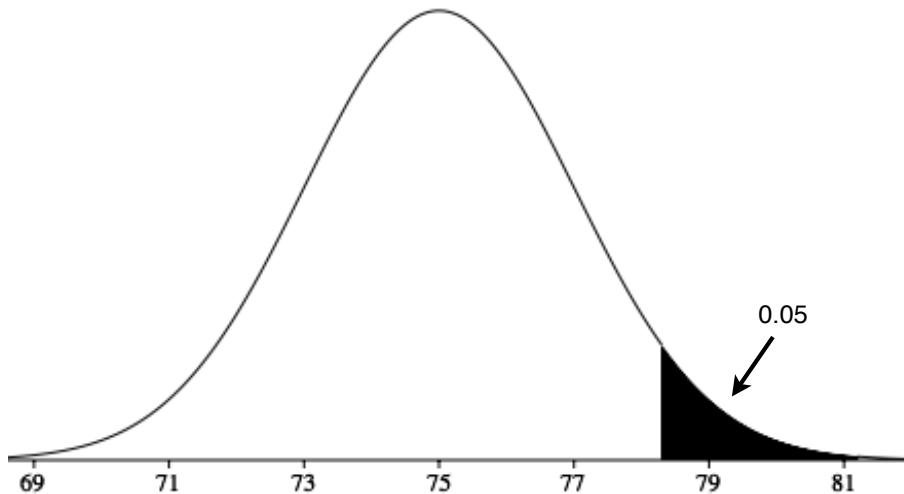


Figure 3. The sampling distribution of the mean if the null hypothesis is true.

The question, then, is what is the probability the experimenter gets a sample mean greater than 78.29 given that the population mean is 80? Figure 4 shows that this probability is 0.80.

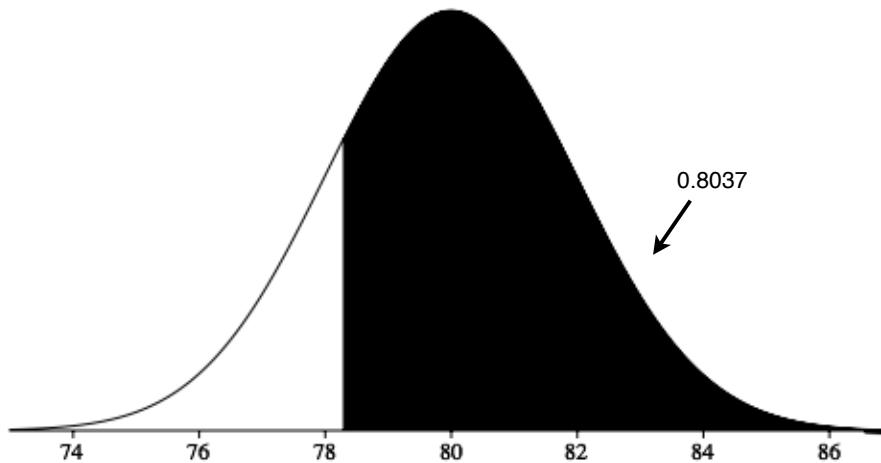


Figure 4. The sampling distribution of the mean if the population mean is 80.

The test is significant if the sample mean is 78.29 or higher.

Therefore, the probability that the experimenter will reject the null hypothesis that the population mean for the new method is 75 or lower is 0.80. In other words, power = 0.80.

Calculation of power is more complex for t tests and for Analysis of Variance. There are many programs that compute power.

# Factors Affecting Power

by David M. Lane

## *Prerequisites*

- Chapter 11: Significance Testing
- Chapter 11: Type I and Type II Errors
- Chapter 11: One- and Two-Tailed Tests
- Chapter 13: Introduction to Power
- Chapter 13: Example Calculations

## *Learning Objectives*

1. State five factors affecting power
2. State what the effect of each of the factors is

Several factors affect the power of a statistical test. Some of the factors are under the control of the experimenter, whereas others are not. The following example will be used to illustrate the various factors.

Suppose a math achievement test were known to be normally distributed with a mean of 75 and a *standard deviation* of  $\sigma$ . A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean  $\mu$  for the new method is larger than 75. The researcher plans to sample N subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. In this section, we consider factors that affect the probability that the researcher will correctly reject the false *null hypothesis* that the population mean is 75. In other words, factors that affect power.

## **Sample Size**

Figure 1 shows that the larger the sample size, the higher the power. Since sample size is typically under an experimenter's control, increasing sample size is one way to increase power. However, it is sometimes difficult and/or expensive to use a large sample size.

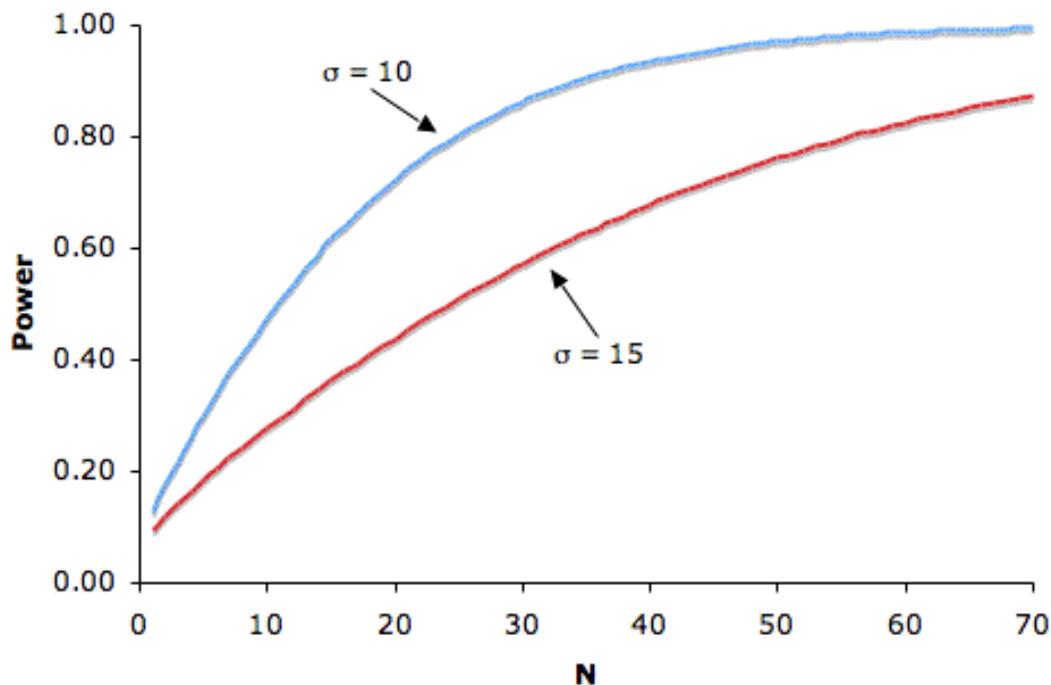


Figure 1. The relationship between sample size and power for  $H_0: \mu = 75$ , real  $\mu = 80$ , one-tailed  $\alpha = 0.05$ , for  $\sigma$ 's of 10 and 15.

## Standard Deviation

Figure 1 also shows that power is higher when the standard deviation is small than when it is large. For all values of N, power is higher for the standard deviation of 10 than for the standard deviation of 15 (except, of course, when N = 0).

Experimenters can sometimes control the standard deviation by sampling from a homogeneous population of subjects, by reducing random measurement error, and/or by making sure the experimental procedures are applied very consistently.

## Difference between Hypothesized and True Mean

Naturally, the larger the effect size, the more likely it is that an experiment would find a significant effect. Figure 2 shows the effect of increasing the difference between the mean specified by the null hypothesis (75) and the population mean  $\mu$  for standard deviations of 10 and 15.

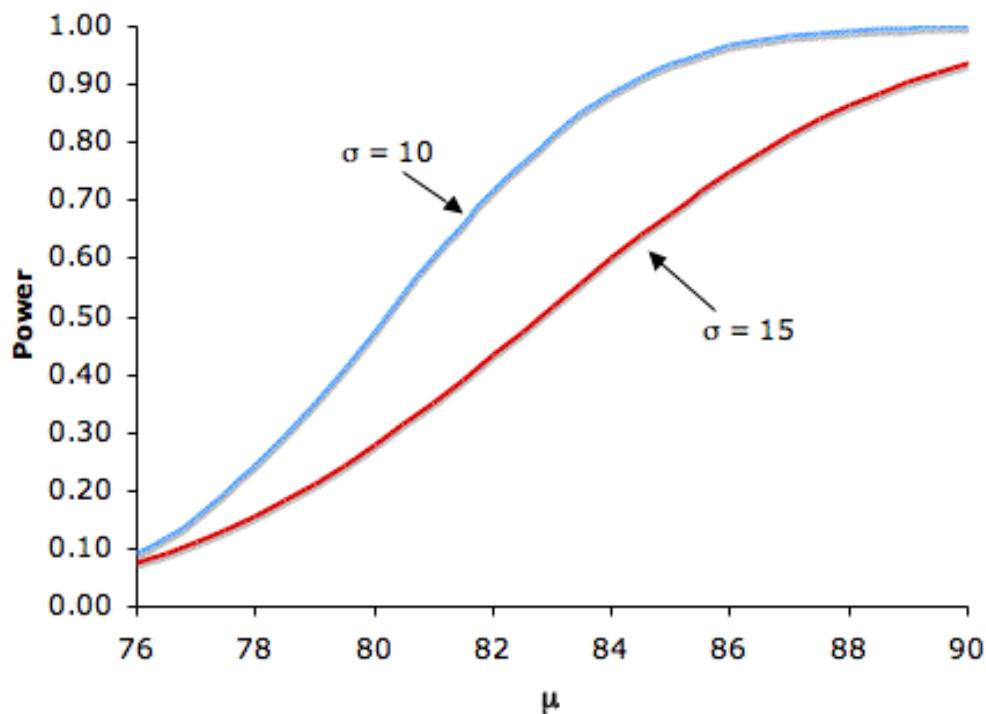


Figure 2. The relationship between  $\mu$  and power for  $H_0: \mu = 75$ , one-tailed  $\alpha = 0.05$ , for  $\sigma$ 's of 10 and 15.

### Significance Level

There is a trade-off between the *significance level* and power: the more stringent (lower) the significance level, the lower the power. Figure 3 shows that power is lower for the 0.01 level than it is for the 0.05 level. Naturally, the stronger the evidence needed to reject the null hypothesis, the lower the chance that the null hypothesis will be rejected.

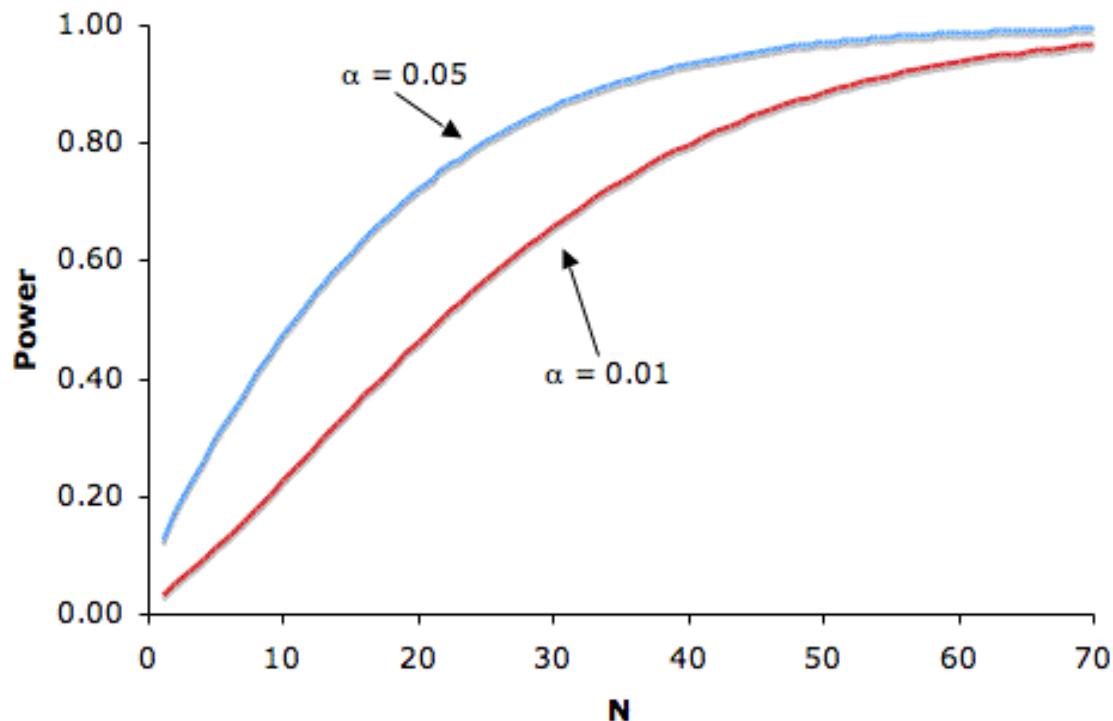


Figure 3. The relationship between significance level and power with one-tailed tests:  $\mu = 75$ , real  $\mu = 80$ , and  $\sigma = 10$ .

### One- versus Two-Tailed Tests

Power is higher with a *one-tailed* test than with a *two-tailed* test as long as the hypothesized direction is correct. A one-tailed test at the 0.05 level has the same power as a two-tailed test at the 0.10 level. A one-tailed test, in effect, raises the significance level.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 13:

A research design to compare three drugs for the treatment of Alzheimer's disease is [described here](#). For the first two years of the study, researchers will follow the subjects with scans and memory tests.

## What do you think?

The data could be analyzed as a between-subjects design or as a within-subjects design. What type of analysis would be done for each type of design and how would the choice of designs affect power?

For a between-subjects design, the subjects in the different conditions would be compared after two years. For a within-subjects design, the change in subjects' scores in the different conditions would be compared. The latter would be more powerful.

## Exercises

### *Prerequisites*

1. Define power in your own words.
2. List 3 measures one can take to increase the power of an experiment. Explain why your measures result in greater power.

3. Population 1 mean = 36

Population 2 mean = 45

Both population variances are 10.

What is the probability that a t test will find a significant difference between means at the 0.05 level? Give results for both one- and two-tailed tests. Hint: the power of a one-tailed test at 0.05 level is the power of a two-tailed test at 0.10.

4. Rank order the following in terms of power.

	<b>Population 1 Mean</b>	<b>n</b>	<b>Population 2 Mean</b>	<b>Standard Deviation</b>
a	29	20	43	12
b	34	15	40	6
c	105	24	50	27
d	170	2	120	10

5. Alan, while snooping around his grandmother's basement stumbled upon a shiny object protruding from under a stack of boxes. When he reached for the object a genie miraculously materialized and stated: "You have found my magic coin. If you flip this coin an infinite number of times you will notice that heads will show 60% of the time." Soon after the genie's declaration he vanished, never to be seen again. Alan, excited about his new magical discovery, approached his friend Ken and told him about what he had found. Ken was skeptical of his friend's story, however, he told Alan to flip the coin 100 times and to record how many flips resulted with heads.

(a) What is the probability that Alan will be able convince Ken that his coin has special powers by finding a p value below 0.05 (one tailed).

Use the Binomial Calculator (and some trial and error)

(b) If Ken told Alan to flip the coin only 20 times, what is the probability that Alan will not be able to convince Ken (by failing to reject the null hypothesis at the 0.05 level)?

# 14. Regression

- A. Introduction to Simple Linear Regression
- B. Partitioning Sums of Squares
- C. Standard Error of the Estimate
- D. Inferential Statistics for  $b$  and  $r$
- E. Influential Observations
- F. Regression Toward the Mean
- G. Introduction to Multiple Regression
- H. Exercises

This chapter is about prediction. Statisticians are often called upon to develop methods to predict one variable from other variables. For example, one might want to predict college grade point average from high school grade point average. Or, one might want to predict income from the number of years of education.

# Introduction to Linear Regression

by David M. Lane

## *Prerequisites*

- Chapter 3: Measures of Variability
- Chapter 4: Describing Bivariate Data

## *Learning Objectives*

1. Define linear regression
2. Identify errors of prediction in a scatter plot with a regression line

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

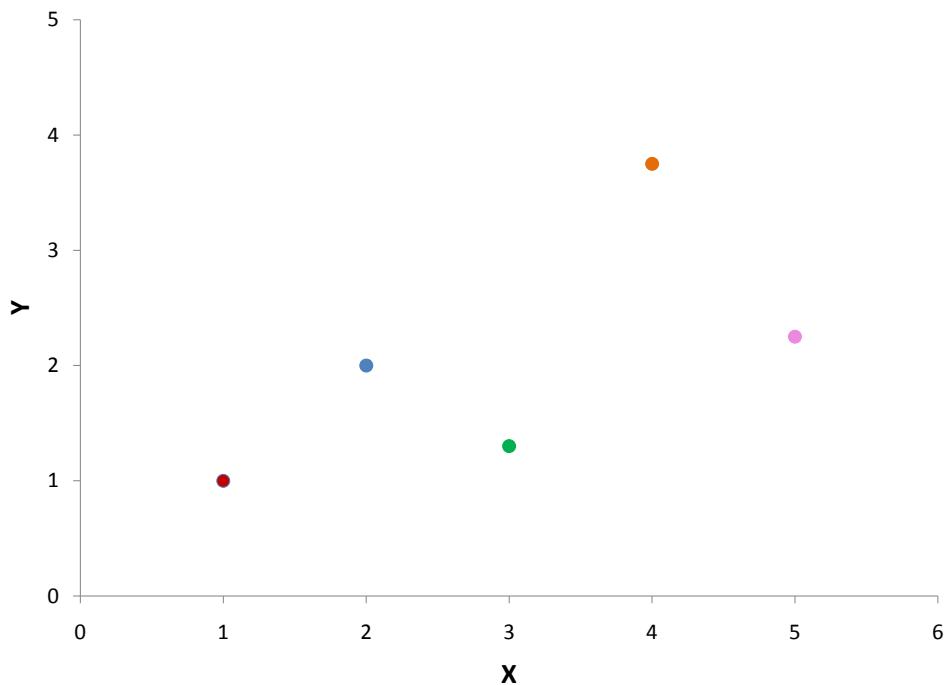


Figure 1. A scatter plot of the example data.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

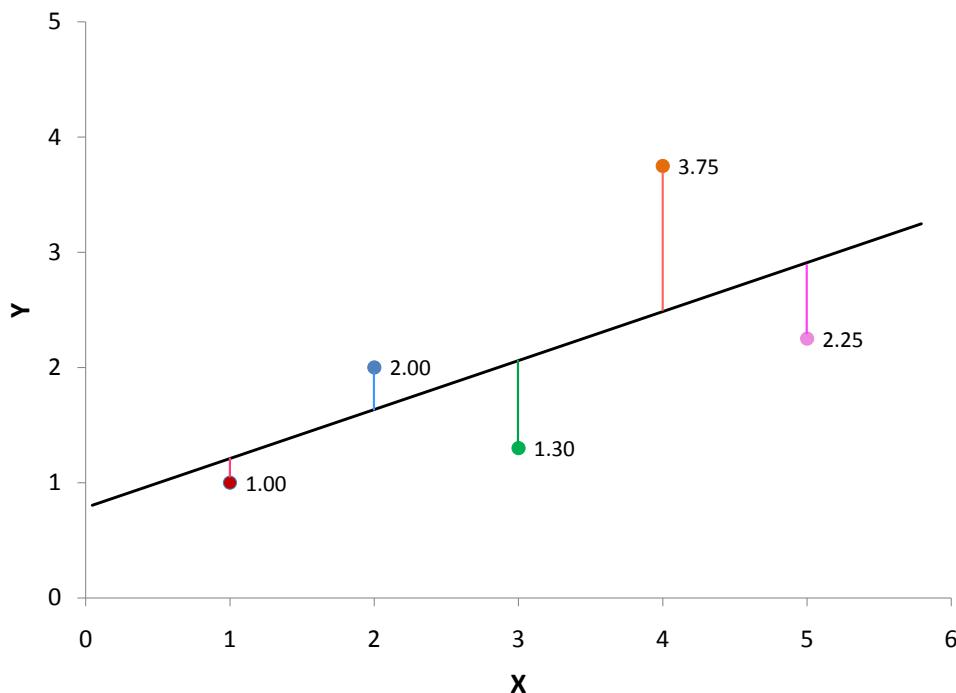


Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values ( $Y'$ ) and the errors of prediction ( $Y - Y'$ ). For example, the first point has a  $Y$  of 1.00 and a predicted  $Y$  of 1.21. Therefore, its error of prediction is -0.21.

Table 2. Example data.

X	Y	Y'	Y-Y'	(Y-Y') <sup>2</sup>
1	1	1.21	-0.21	0.044
2	2	1.635	0.365	0.133
3	1.3	2.06	-0.76	0.578
4	3.75	2.485	1.265	1.6
5	2.25	2.91	-0.66	0.436

You may have noticed that we did not specify what is meant by “best-fitting line.” By far the most commonly used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was

used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A$$

where  $Y'$  is the predicted score,  $b$  is the slope of the line, and  $A$  is the  $Y$  intercept. The equation for the line in Figure 2 is

$$Y' = 0.425X + 0.785$$

For  $X = 1$ ,

$$Y' = (0.425)(1) + 0.785 = 1.21.$$

For  $X = 2$ ,

$$Y' = (0.425)(2) + 0.785 = 1.64.$$

## Computing the Regression Line

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy are given here for anyone who is interested. The calculations are based on the statistics shown in Table 3.  $M_x$  is the mean of  $X$ ,  $M_y$  is the mean of  $Y$ ,  $s_x$  is the standard deviation of  $X$ ,  $s_y$  is the standard deviation of  $Y$ , and  $r$  is the correlation between  $X$  and  $Y$ .

Table 3. Statistics for computing the regression line

$M_x$	$M_y$	$s_x$	$s_y$	$r$
3	2.06	1.581	1.072	0.627

The slope ( $b$ ) can be calculated as follows:

$$b = r \frac{s_y}{s_x}$$

and the intercept ( $A$ ) can be calculated as

$$A = M_Y - bM_X.$$

For these data,

$$b = (0.627) \frac{1.072}{1.581} = 0.425$$

$$A = 2.06 - (0.425)(3) = 0.785$$

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

### Standardized Variables

The regression equation is simpler if variables are *standardized* so that their means are equal to 0 and standard deviations are equal to 1, for then  $b = r$  and  $A = 0$ . This makes the regression line:

$$Z_Y' = (r)(Z_X)$$

where  $Z_Y'$  is the predicted standard score for  $Y$ ,  $r$  is the correlation, and  $Z_X$  is the standardized score for  $X$ . Note that the slope of the regression equation for standardized variables is  $r$ .

Figure 3 shows a scatterplot with the regression line predicting the standardized Verbal SAT from the standardized Math SAT.

### A Real Example

The case study, “SAT and College GPA” contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure 3 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

$$\text{Univ GPA}' = (0.675)(\text{High School GPA}) + 1.097$$

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

$$\text{University GPA}' = (0.675)(3) + 1.097 = 3.12.$$

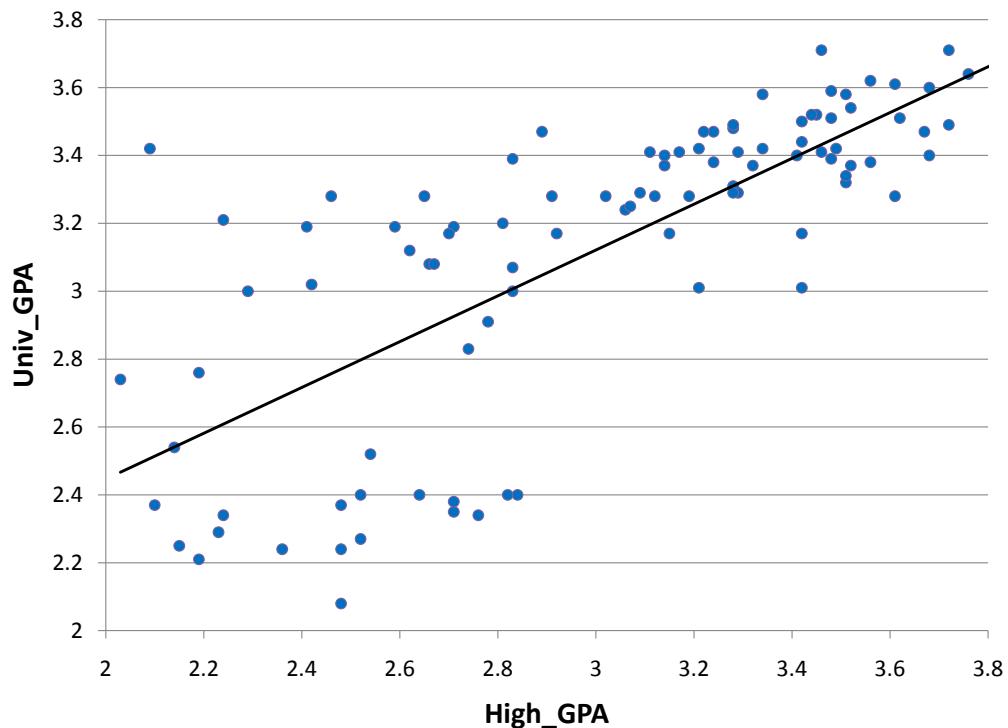


Figure 3. University GPA as a function of High School GPA.

## Assumptions

It may surprise you, but the calculations shown in this section are assumption free. Of course, if the relationship between X and Y is not linear, a different shaped function could fit the data better. *Inferential statistics* in regression are based on several assumptions, and these assumptions are presented in a later section of this chapter.

# Partitioning the Sums of Squares

by David M. Lane

## *Prerequisites*

- Chapter 14: Introduction to Linear Regression

## *Learning Objectives*

1. Compute the sum of squares  $Y$
2. Convert raw scores to deviation scores
3. Compute predicted scores from a regression equation
4. Partition sum of squares  $Y$  into sum of squares predicted and sum of squares error
5. Define  $r^2$  in terms of sum of squares explained and sum of squares  $Y$

One useful aspect of regression is that it can divide the variation in  $Y$  into two parts: the variation of the predicted scores and the variation in the errors of prediction. The variation of  $Y$  is called the sum of squares  $Y$  and is defined as the sum of the squared deviations of  $Y$  from the mean of  $Y$ . In the population, the formula is

$$SSY = \sum (Y - \mu_Y)^2$$

where  $SSY$  is the sum of squares  $Y$ ,  $Y$  is an individual value of  $Y$ , and  $\mu_Y$  is the mean of  $Y$ . A simple example is given in Table 1. The mean of  $Y$  is 2.06 and  $SSY$  is the sum of the values in the third column and is equal to 4.597.

Table 1. Example of  $SSY$ .

$Y$	$Y - \mu_Y$	$(Y - \mu_Y)^2$
1	-1.06	1.1236
2	-0.06	0.0036
1.3	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

When computed in a sample, you should use the sample mean,  $M$ , in place of the population mean:

$$SSY = \sum (Y - M_Y)^2$$

It is sometimes convenient to use formulas that use *deviation scores* rather than raw scores. Deviation scores are simply deviations from the mean. By convention, small letters rather than capitals are used for deviation scores. Therefore, the score,  $y$  indicates the difference between  $Y$  and the mean of  $Y$ . Table 2 shows the use of this notation. The numbers are the same as in Table 1.

Table 2. Example of SSY using Deviation Scores.

Y	y	$y^2$
1	-1.06	1.1236
2	-0.06	0.0036
1.3	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361
10.3	0	4.597

The data in Table 3 are reproduced from the introductory section. The column X has the values of the *predictor variable* and the column Y has the *criterion variable*. The third column,  $y$ , contains the the differences between the column Y and the mean of Y.

Table 3. Example data. The last row contains column sums.

X	Y	y	y <sup>2</sup>	Y'	y'	y' <sup>2</sup>	Y-Y'	(Y-Y') <sup>2</sup>
1	1	-1.06	1.1236	1.21	-0.85	0.7225	-0.21	0.044
2	2	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3	1.3	-0.76	0.5776	2.06	0	0	-0.76	0.578
4	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.6
5	2.25	0.19	0.0361	2.91	0.85	0.7225	-0.66	0.436
15	10.3	0	4.597	10.3	0	1.806	0	2.791

The fourth column,  $y^2$ , is simply the square of the y column. The column  $Y'$  contains the predicted values of Y. In the introductory section, it was shown that the equation for the regression line for these data is

$$Y' = 0.425X + 0.785.$$

The values of  $Y'$  were computed according to this equation. The column  $y'$  contains deviations of  $Y'$  from the mean of  $Y'$  and  $y'^2$  is the square of this column. The next-to-last column,  $Y-Y'$ , contains the actual scores (Y) minus the predicted scores ( $Y'$ ). The last column contains the squares of these errors of prediction.

We are now in a position to see how the SSY is partitioned. Recall that SSY is the sum of the squared deviations from the mean. It is therefore the sum of the  $y^2$  column and is equal to 4.597. SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE). The sum of squares predicted is the sum of the squared deviations of the predicted scores from the mean predicted score. In other words, it is the sum of the  $y'^2$  column and is equal to 1.806. The sum of squares error is the sum of the squared errors of prediction. It is therefore the sum of the  $(Y-Y')^2$  column and is equal to 2.791. This can be summed up as:

$$\begin{aligned} SSY &= SSY' + SSE \\ 4.597 &= 1.806 + 2.791 \end{aligned}$$

There are several other notable features about Table 3. First, notice that the sum of  $y$  and the sum of  $y'$  are both zero. This will always be the case because these variables were created by subtracting their respective means from each value. Also, notice that the mean of  $Y - Y'$  is 0. This indicates that although some  $Y$  values are higher than their respective predicted  $Y$  values and some are lower, the average difference is zero.

The  $SSY$  is the total variation, the  $SSY'$  is the variation explained, and the  $SSE$  is the variation unexplained. Therefore, the proportion of variation explained can be computed as:

$$\text{Proportion explained} = \frac{SSY'}{SSY}$$

Similarly, the proportion not explained is:

$$\text{Proportion not explained} = \frac{SSE}{SSY}$$

There is an important relationship between the proportion of variation explained and Pearson's correlation:  $r^2$  is the proportion of variation explained. Therefore, if  $r = 1$ , then, naturally, the proportion of variation explained is 1; if  $r = 0$ , then the proportion explained is 0. One last example: for  $r = 0.4$ , the proportion of variation explained is 0.16.

Since the variance is computed by dividing the variation by  $N$  (for a population) or  $N-1$  (for a sample), the relationships spelled out above in terms of variation also hold for variance. For example,

$$\sigma_{total}^2 = \sigma_{Y'}^2 + \sigma_e^2$$

where the first term is the variance total, the second term is the variance of  $Y'$ , and the last term is the variance of the errors of prediction ( $Y - Y'$ ). Similarly,  $r^2$  is the proportion of variance explained as well as the proportion of variation explained.

## Summary Table

It is often convenient to summarize the partitioning of the data in a table such as Table 4. The *degrees of freedom* column (df) shows the degrees of freedom for

each source of variation. The degrees of freedom for the sum of squares explained is equal to the number of predictor variables. This will always be 1 in simple regression. The error degrees of freedom is equal to the total number of observations minus 2. In this example, it is  $5 - 2 = 3$ . The total degrees of freedom is the total number of observations minus 1.

Table 4. Summary Table for Example Data

Source	Sum of Squares	df	Mean Square
Explained	1.806	1	1.806
Error	2.791	3	0.93
Total	4.597	4	

# Standard Error of the Estimate

by David M. Lane

## Prerequisites

- Chapter 3: Measures of Variability
- Chapter 14: Introduction to Linear Regression
- Chapter 14: Partitioning Sums of Squares

## Learning Objectives

1. Make judgments about the size of the standard error of the estimate from a scatter plot
2. Compute the standard error of the estimate based on errors of prediction
3. Compute the standard error using Pearson's correlation
4. Estimate the standard error of the estimate based on a sample

Figure 1 shows two regression examples. You can see that in Graph A, the points are closer to the line than they are in Graph B. Therefore, the predictions in Graph A are more accurate than in Graph B.

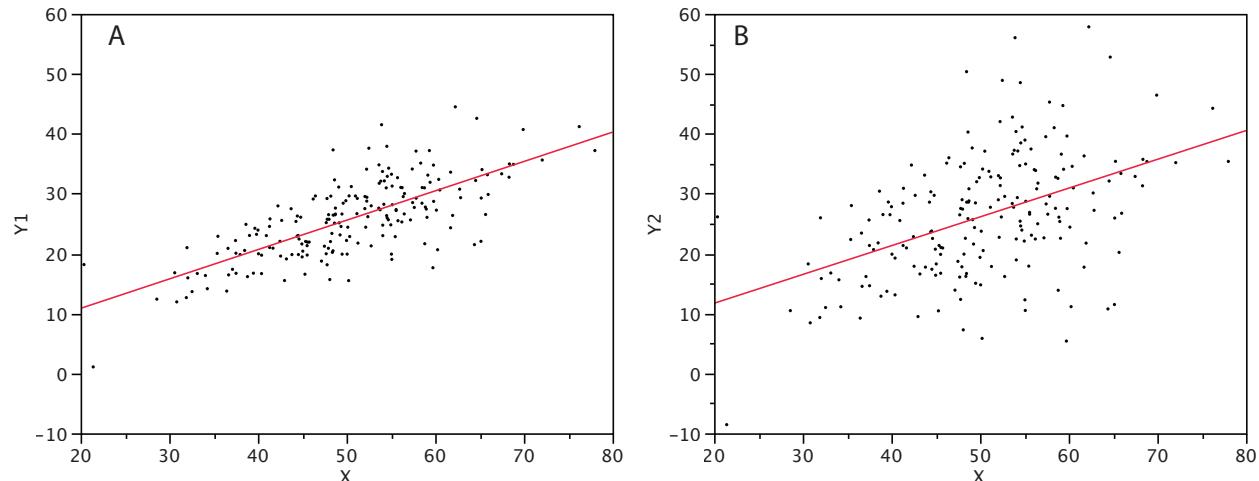


Figure 1. Regressions differing in accuracy of prediction.

The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

where  $\sigma_{est}$  is the standard error of the estimate,  $Y$  is an actual score,  $Y'$  is a predicted score, and  $N$  is the number of pairs of scores. The numerator is the sum of squared differences between the actual scores and the predicted scores.

Note the similarity of the formula for  $\sigma_{est}$  to the formula for  $\sigma$ :

$$\sigma = \sqrt{\frac{\sum (Y - \mu)^2}{N}}$$

In fact,  $\sigma_{est}$  is the standard deviation of the errors of prediction (each  $Y - Y'$  is an error of prediction).

Assume the data in Table 1 are the data from a population of five  $X, Y$  pairs.

Table 1. Example data.

	$X$	$Y$	$Y'$	$Y - Y'$	$(Y - Y')^2$
	1	1	1.21	-0.21	0.044
	2	2	1.635	0.365	0.133
	3	1.3	2.06	-0.76	0.578
	4	3.75	2.485	1.265	1.6
	5	2.25	2.91	-0.66	0.436
Sum	15	10.3	10.3	0	2.791

The last column shows that the sum of the squared errors of prediction is 2.791. Therefore, the standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747$$

There is a version of the formula for the standard error in terms of Pearson's correlation:

$$\sigma_{est} = \sqrt{\frac{(1 - \rho^2)SSY}{N}}$$

where  $\rho$  is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2$$

For the data in Table 1,  $m_y = 10.30$ ,  $SSY = 4.597$  and  $r = 0.6268$ . Therefore,

$$\sigma_{est} = \sqrt{\frac{(1 - 0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747$$

which is the same value computed previously.

Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is  $N-2$  rather than  $N$ . The reason  $N-2$  is used rather than  $N-1$  is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares. Formulas for a sample comparable to the ones for a population are shown below:

$$s_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N - 2}}$$

$$s_{est} = \sqrt{\frac{2.791}{3}} = 0.964$$

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

# **Inferential Statistics for $b$ and $r$**

by David M. Lane

## *Prerequisites*

- Chapter 9: Sampling Distribution of  $r$
- Chapter 9: Confidence Interval for  $r$

## *Learning Objectives*

1. State the assumptions that inferential statistics in regression are based upon
2. Identify heteroscedasticity in a scatter plot
3. Compute the standard error of a slope
4. Test a slope for significance
5. Construct a confidence interval on a slope
6. Test a correlation for significance
7. Construct a confidence interval on a correlation

This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation. As you will see, if the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

## **Assumptions**

Although no assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics. Naturally, these assumptions refer to the population, not the sample.

1. Linearity: The relationship between the two variables is linear.
2. Homoscedasticity: The variance around the regression line is the same for all values of  $X$ . A clear violation of this assumption is shown in Figure 1. Notice that the predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs are not very good. In other words, the points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students are not.

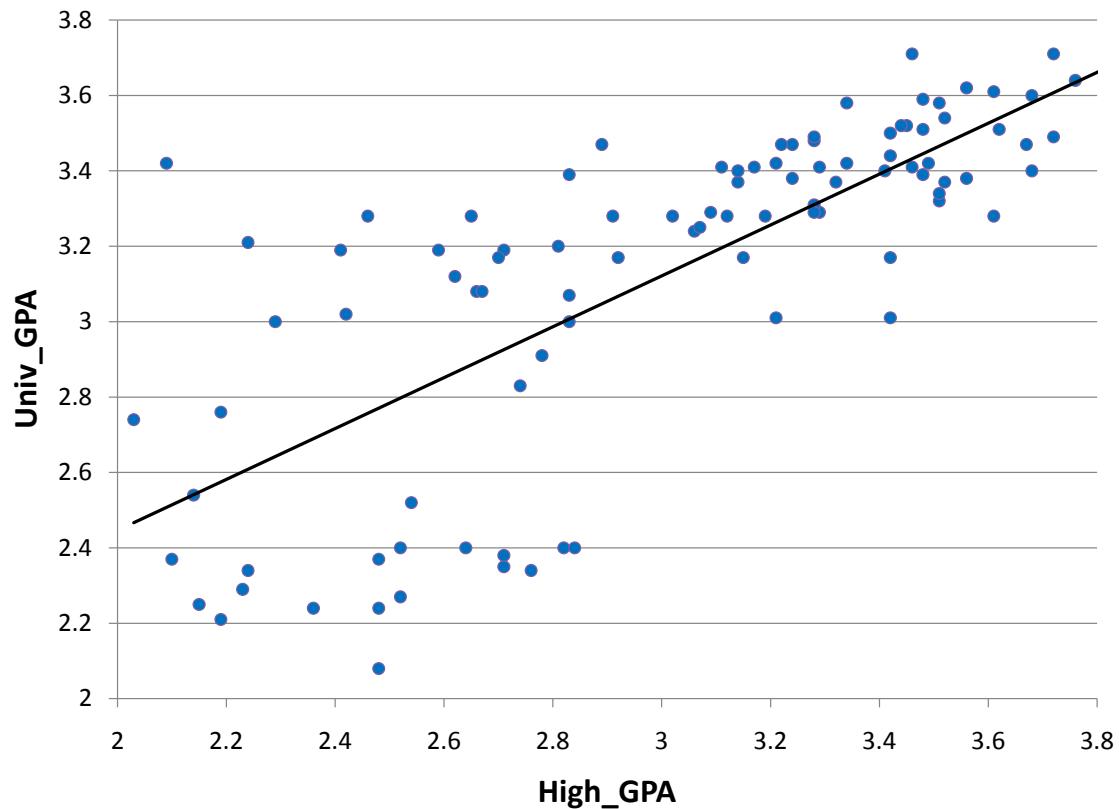


Figure 1. University GPA as a function of High School GPA.

3. The errors of prediction are distributed normally. This means that the distributions of deviations from the regression line are normally distributed. It does not mean that X or Y is normally distributed.

### Significance Test for the Slope (b)

Recall the general formula for a t test:

$$t = \frac{\text{statistics} - \text{hypothesized value}}{\text{estimated standard error of the statistic}}$$

As applied here, the statistic is the sample value of the slope (b) and the hypothesized value is 0. The degrees of freedom for this test are:

$$df = N - 2$$

where N is the number of pairs of scores.

The estimated standard error of  $b$  is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}}$$

where  $s_b$  is the estimated standard error of  $b$ ,  $s_{est}$  is the standard error of the estimate, and  $SSX$  is the sum of squared deviations of  $X$  from the mean of  $X$ .  $SSX$  is calculated as

$$SSX = \sum (X - M_x)^2$$

where  $M_x$  is the mean of  $X$ . As shown previously, the standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

These formulas are illustrated with the data shown in Table 1. These data are reproduced from the introductory section. The column  $X$  has the values of the *predictor variable* and the column  $Y$  has the values of the *criterion variable*. The third column,  $x$ , contains the differences between the values of column  $X$  and the mean of  $X$ . The fourth column,  $x^2$ , is the square of the  $x$  column. The fifth column,  $y$ , contains the differences between the values of column  $Y$  and the mean of  $Y$ . The last column,  $y^2$ , is simply the square of the  $y$  column.

Table 1. Example data.

	<b>X</b>	<b>Y</b>	<b>x</b>	<b>x<sup>2</sup></b>	<b>y</b>	<b>y<sup>2</sup></b>
	1	1	-2	4	-1.06	1.1236
	2	2	-1	1	-0.06	0.0036
	3	1.3	0	0	-0.76	0.5776
	4	3.75	1	1	1.69	2.8561
	5	2.25	2	4	0.19	0.0361
<b>Sum</b>	15	10.3	0	10	0	4.597

The computation of the standard error of the estimate ( $s_{\text{est}}$ ) for these data is shown in the section on the standard error of the estimate. It is equal to 0.964.

$$s_{\text{est}} = 0.964$$

SSX is the sum of squared deviations from the mean of X. It is, therefore, equal to the sum of the  $x^2$  column and is equal to 10.

$$\text{SSX} = 10.00$$

We now have all the information to compute the standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

As shown previously, the slope (b) is 0.425. Therefore,

$$t = \frac{0.425}{0.305} = 1.39$$

$$\text{df} = N-2 = 5-2 = 3.$$

The p value for a two-tailed t test is 0.26. Therefore, the slope is not significantly different from 0.

## Confidence Interval for the Slope

The method for computing a confidence interval for the population slope is very similar to methods for computing other confidence intervals. For the 95% confidence interval, the formula is:

$$\begin{aligned}\text{lower limit: } b - (t_{.95}) (s_b) \\ \text{upper limit: } b + (t_{.95}) (s_b)\end{aligned}$$

where  $t_{.95}$  is the value of  $t$  to use for the 95% confidence interval.

The values of  $t$  to be used in a confidence interval can be looked up in a table of the  $t$  distribution. A small version of such a table is shown in Table 2. The first column,  $df$ , stands for degrees of freedom.

Table 2. Abbreviated  $t$  table.

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the “inverse  $t$  distribution” calculator ([external link](#); requires Java) to find the  $t$  values to use in a confidence interval.

Applying these formulas to the example data,

$$\begin{aligned}\text{lower limit: } 0.425 - (3.182) (0.305) &= -0.55 \\ \text{upper limit: } 0.425 + (3.182) (0.305) &= 1.40\end{aligned}$$

## Significance Test for the Correlation

The formula for a significance test of Pearson's correlation is shown below:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

where N is the number of pairs of scores. For the example data,

$$t = \frac{0.627\sqrt{5-2}}{\sqrt{1-0.627^2}} = 1.39$$

Notice that this is the same t value obtained in the t test of b. As in that test, the degrees of freedom is  $N-2 = 5-2 = 3$ .

# Influential Observations

by David M. Lane

## *Prerequisites*

- Chapter 14: Introduction to Linear Regression

## *Learning Objectives*

1. Define “influence”
2. Describe what makes a point influential
3. Define “leverage”
4. Define “distance”

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of influential observations and to take them into consideration when interpreting the results.

## **Influence**

The influence of an observation can be thought of in terms of how much the predicted scores for other observations would differ if the observation in question were not included. Cook's D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

An observation's influence is a function of two factors: (1) how much the observation's value on the predictor variable differs from the mean of the predictor variable and (2) the difference between the predicted score for the observation and its actual score. The former factor is called the observation's leverage. The latter factor is called the observation's *distance*.

## Calculation of Cook's D (Optional)

The first step in calculating the value of Cook's D for an observation is to predict all the scores in the data once using a regression equation based on all the observations and once using all the observations except the observation in question. The second step is to compute the sum of the squared differences between these two sets of predictions. The final step is to divide this result by 2 times the MSE (see the section on partitioning the variance).

## Leverage

The *leverage* of an observation is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation. For example, an observation with the mean on the predictor variable has no influence on the slope of the regression line regardless of its value on the criterion variable. On the other hand, an observation that is extreme on the predictor variable has, depending on its distance, the potential to affect the slope greatly.

## Calculation of Leverage (h)

The first step is to standardize the predictor variable so that it has a mean of 0 and a standard deviation of 1. Then, the leverage (h) is computed by squaring the observation's value on the standardized predictor variable, adding 1, and dividing by the number of observations.

## Distance

The distance of an observation is based on the error of prediction for the observation: The greater the error of prediction, the greater the distance. The most commonly used measure of distance is the *studentized residual*. The studentized residual for an observation is closely related to the error of prediction for that observation divided by the standard deviation of the errors of prediction. However, the predicted score is derived from a regression equation in which the observation in question is not counted. The details of the computation of a studentized residual are a bit complex and are beyond the scope of this work.

An observation with a large distance will not have that much influence if its leverage is low. It is the combination of an observation's leverage and distance that determines its influence.

## Example

Table 1 shows the leverage, studentized residual, and influence for each of the five observations in a small dataset.

Table 1. Example Data.

ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.4
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.2	1.22	0.19
E	8	7	0.73	-1.68	8.86

h is the leverage, R is the studentized residual, and D is Cook's measure of influence.

Observation A has fairly high leverage, a relatively high residual, and moderately high influence.

Observation B has small leverage and a relatively small residual. It has very little influence.

Observation C has small leverage and a relatively high residual. The influence is relatively low.

Observation D has the lowest leverage and the second highest residual. Although its residual is much higher than Observation A, its influence is much less because of its low leverage.

Observation E has by far the largest leverage and the largest residual. This combination of high leverage and high residual makes this observation extremely influential.

Figure 1 shows the regression line for the whole dataset (blue) and the regression line if the observation in question is not included (red) for all observations. The observation in question is circled. Naturally, the regression line for the whole dataset is the same in all panels. The residual is calculated relative to the line for which the observation in question is not included in the analysis. This can be seen most clearly for Observation E which lies very close to the regression line

computed when it is included but very far from the regression line when it is excluded from the calculation of the line.

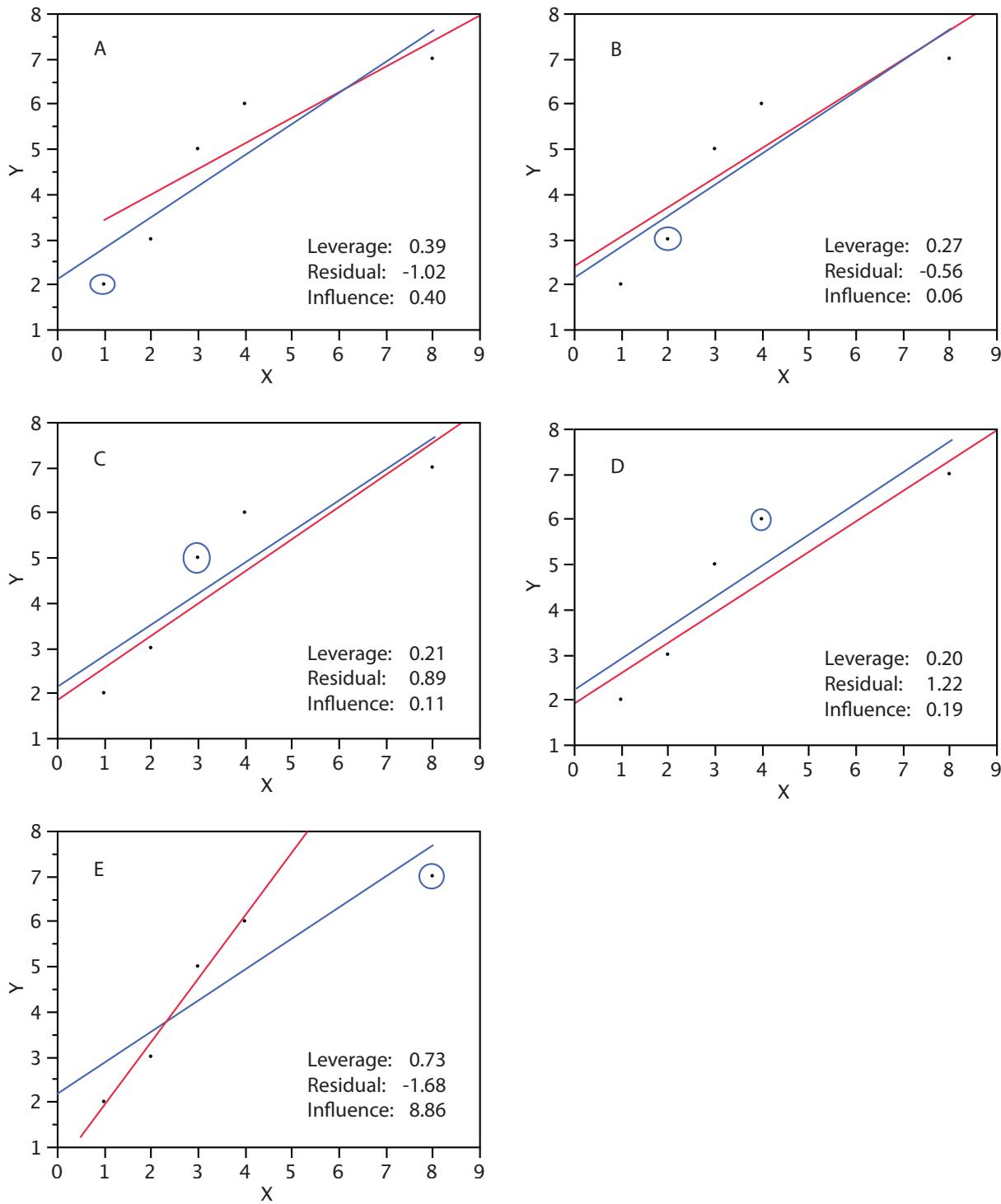


Figure 1. Illustration of leverage, residual, and influence.

The circled points are not included in the calculation of the red regression line. All points are included in the calculation of the blue regression line.

# Regression Toward the Mean

by David M. Lane

## *Prerequisites*

- Chapter 14: Regression Introduction

## *Learning Objectives*

1. Explain what regression towards the mean is
2. State the conditions under which regression toward the mean occurs
3. Identify situations in which neglect of regression toward the mean leads to incorrect conclusions
4. Explain how regression toward the mean relates to a regression equation.

Regression toward the mean involves outcomes that are at least partly due to chance. We begin with an example of a task that is entirely chance: Imagine an experiment in which a group of 25 people each predicted the outcomes of flips of a fair coin. For each subject in the experiment, a coin is flipped 12 times and the subject predicts the outcome of each flip. Figure 1 shows the results of a simulation of this “experiment.” Although most subjects were correct from 5 to 8 times out of 12, one simulated subject was correct 10 times. Clearly, this subject was very lucky and probably would not do as well if he or she performed the task a second time. In fact, the best prediction of the number of times this subject would be correct on the retest is 6 since the probability of being correct on a given trial is 0.5 and there are 12 trials.

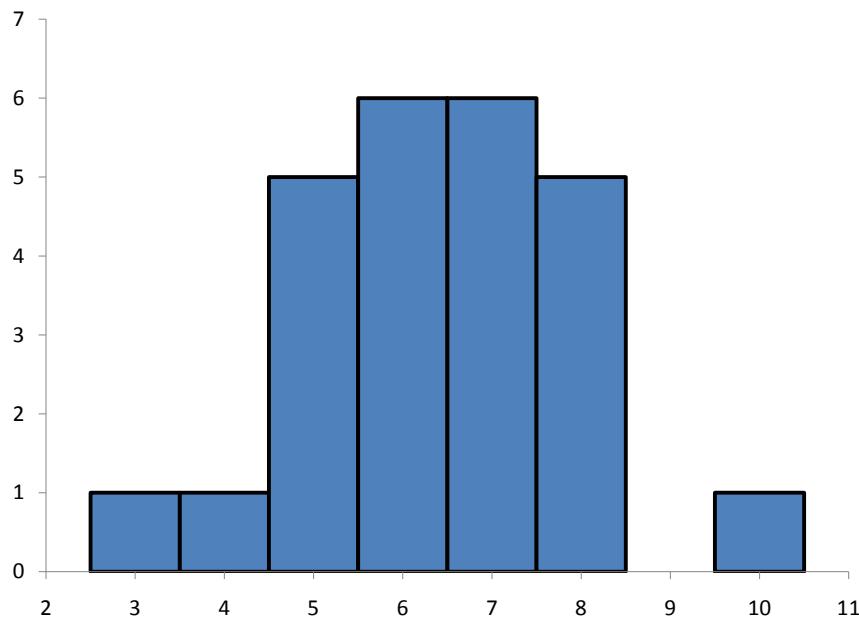


Figure 1. Histogram of results of a simulated experiment.

More technically, the best prediction for the subject's result on the retest is the mean of the binomial distribution with  $N = 12$  and  $p = 0.50$ . This distribution is shown in Figure 2 and has a mean of 6.

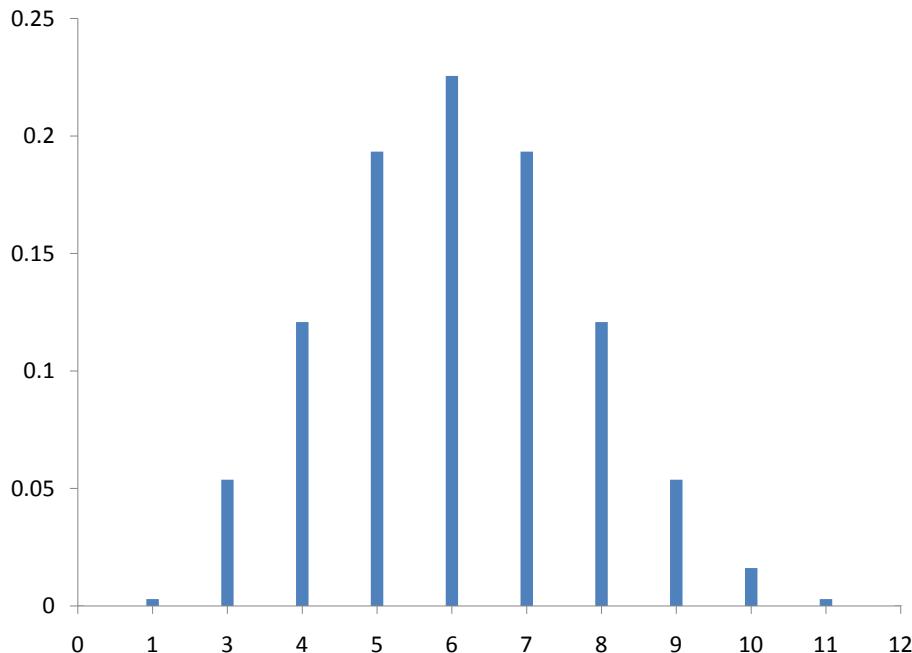


Figure 2. Binomial Distribution for  $N = 12$  and  $p = .50$ .

The point here is that no matter how many coin flips a subject predicted correctly, the best prediction of their score on a retest is 6.

Now we consider a test we will call “Test A” that is partly chance and partly skill: Instead of predicting the outcomes of 12 coin flips, each subject predicts the outcomes of 6 coin flips and answers 6 true/false questions about world history. Assume that the mean score on the 6 history questions is 4. A subject's score on Test A has a large chance component but also depends on history knowledge. If a subject scored very high on this test (such as a score of 10/12), it is likely that they did well on both the history questions and the coin flips. For example, if they only got four of the history questions correct, they would have had to have gotten all six of the coin predictions correct, and this would have required exceptionally good luck. If given a second test (Test B) that also included coin predictions and history questions, their knowledge of history would be helpful and they would again be expected to score above the mean. However, since their high performance on the coin portion of Test A would not be predictive of their coin performance on Test B, they would not be expected to fare as well on Test B as on Test A. Therefore, the best prediction of their score on Test B would be somewhere between their score on Test A and the mean of Test B. This tendency of subjects with high values on a measure that includes chance and skill to score closer to the mean on a retest is called “*regression toward the mean*.”

The essence of the regression-toward-the-mean phenomenon is that people with high scores tend to be above average in skill and in luck, and that only the skill portion is relevant to future performance. Similarly, people with low scores tend to be below average in skill and luck and their bad luck is not relevant to future performance. This does not mean that all people who score high have above average luck. However, on average they do.

Almost every measure of behavior has a chance and a skill component to it. Take a student's grade on a final exam as an example. Certainly, the student's knowledge of the subject will be a major determinant of his or her grade. However, there are aspects of performance that are due to chance. The exam cannot cover everything in the course and therefore must represent a subset of the material. Maybe the student was lucky in that the one aspect of the course the student did not understand well was not well represented on the test. Or, maybe, the student was not sure which of two approaches to a problem would be better but, more or less by chance, chose the right one. Other chance elements come into play as well. Perhaps the student was awakened early in the morning by a random phone call, resulting in fatigue and lower performance. And, of course, guessing on multiple choice questions is another source of randomness in test scores.

There will be regression toward the mean in a test-retest situation whenever there is less than a perfect ( $r = 1$ ) relationship between the test and the retest. This follows from the formula for a regression line with standardized variables shown below.

$$Z_Y = (r) (Z_X)$$

From this equation it is clear that if the absolute value of  $r$  is less than 1, then the predicted value of  $Z_Y$  will be closer to 0, the mean for standardized scores, than is  $Z_X$ . Also, note that if the correlation between  $X$  and  $Y$  is 0, as it would be for a task that is all luck, the predicted standard score for  $Y$  is its mean, 0, regardless of the score on  $X$ .

Figure 3 shows a scatter plot with the regression line predicting the standardized Verbal SAT from the standardized Math SAT. Note that the slope of the line is equal to the correlation of 0.835 between these variables.

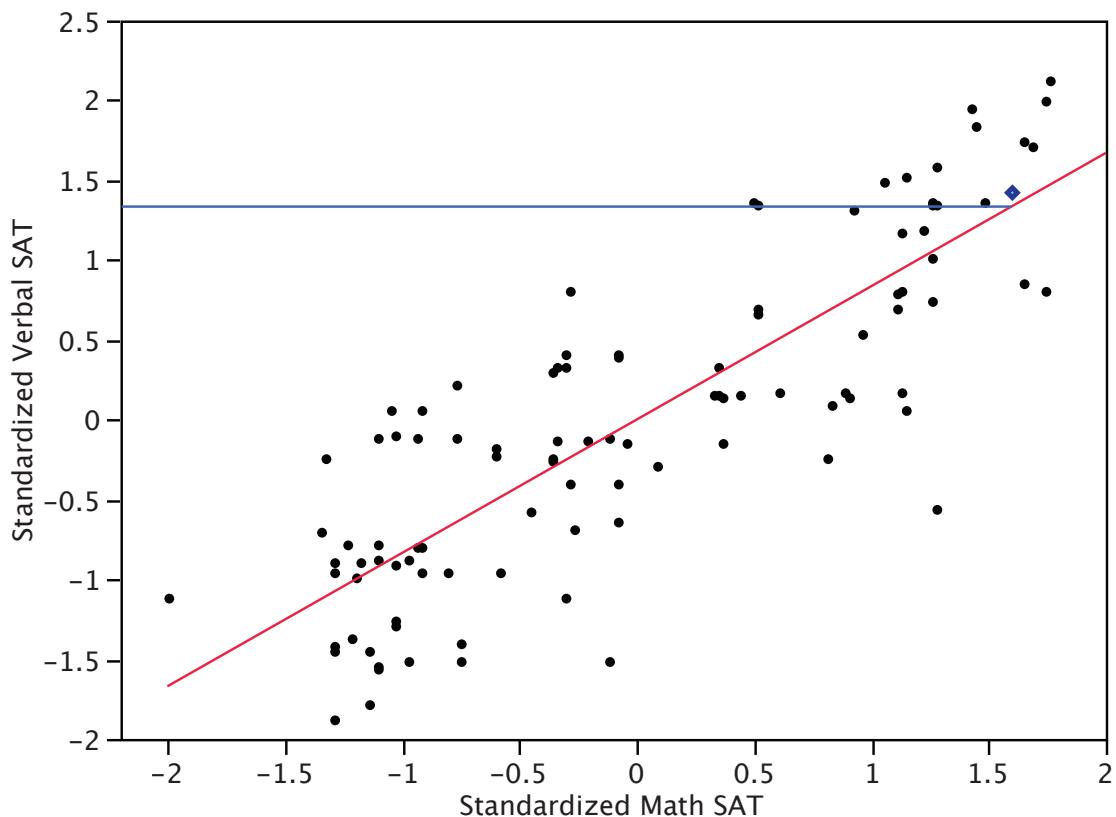


Figure 3. Prediction of Standardized Verbal SAT from Standardized Math SAT.

The point represented by a blue diamond has a value of 1.6 on the standardized Math SAT. This means that this student scored 1.6 standard deviations above the mean on Math SAT. The predicted score is  $(r)(1.6) = (0.835)(1.6) = 1.34$ . The horizontal line on the graph shows the value of the predicted score. The key point is that although this student scored 1.6 standard deviations above the mean on Math SAT, he or she is only predicted to score 1.34 standard deviations above the mean on Verbal SAT. Thus, the prediction is that the Verbal SAT score will be closer to the mean of 0 than is the Math SAT score. Similarly, a student scoring far below the mean on Math SAT will be predicted to score higher on Verbal SAT.

Regression toward the mean occurs in any situation in which observations are selected on the basis of performance on a task that has a random component. If you choose people on the basis of their performance on such a task, you will be choosing people partly on the basis of their skill and partly on the basis of their luck on the task. Since their luck cannot be expected to be maintained from trial to trial, the best prediction of a person's performance on a second trial will be somewhere between their performance on the first trial and the mean performance on the first trial. The degree to which the score is expected to "regress toward the mean" in this manner depends on the relative contributions of chance and skill to the task: the greater the role of chance, the more the regression toward the mean.

## **Errors Resulting From Failure to Understand Regression Toward the Mean**

Failure to appreciate regression toward the mean is common and often leads to incorrect interpretations and conclusions. One of the best examples is provided by Nobel Laureate Daniel Kahneman in his autobiography ([external link](#)). Dr. Kahneman was attempting to teach flight instructors that praise is more effective than punishment. He was challenged by one of the instructors who relayed that in his experience praising a cadet for executing a clean maneuver is typically followed by a lesser performance, whereas screaming at a cadet for bad execution is typically followed by improved performance. This, of course, is exactly what would be expected based on regression toward the mean. A pilot's performance, although based on considerable skill, will vary randomly from maneuver to maneuver. When a pilot executes an extremely clean maneuver, it is likely that he or she had a bit of luck in their favor in addition to their considerable skill. After the praise but not because of it, the luck component will probably disappear and

the performance will be lower. Similarly, a poor performance is likely to be partly due to bad luck. After the criticism but not because of it, the next performance will likely be better. To drive this point home, Kahneman had each instructor perform a task in which a coin was tossed at a target twice. He demonstrated that the performance of those who had done the best the first time deteriorated, whereas the performance of those who had done the worst improved.

Regression toward the mean is frequently present in sports performance. A good example is provided by Schall and Smith (2000), who analyzed many aspects of baseball statistics including the batting averages of players in 1998. They chose the 10 players with the highest batting averages (BAs) in 1998 and checked to see how well they did in 1999. According to what would be expected based on regression toward the mean, these players should, on average, have lower batting averages in 1999 than they did in 1998. As can be seen in Table 1, 7/10 of the players had lower batting averages in 1999 than they did in 1998. Moreover, those who had higher averages in 1999 were only slightly higher, whereas those who were lower were much lower. The average decrease from 1998 to 1999 was 33 points. Even so, most of these players had excellent batting averages in 1999 indicating that skill was an important component of their 1998 averages.

Table 1. How the Ten Players with the Highest BAs in 1998 did in 1999.

1998	1999	Difference
363	379	16
354	298	-56
339	342	3
337	281	-56
336	249	-87
331	298	-33
328	297	-31
328	303	-25
327	257	-70
327	332	5

Figure 4 shows the batting averages of the two years. The decline from 1998 to 1999 is clear. Note that although the mean decreased from 1998, some players increased their batting averages. This illustrates that regression toward the mean

does not occur for every individual. Although the predicted scores for every individual will be lower, some of the predictions will be wrong.

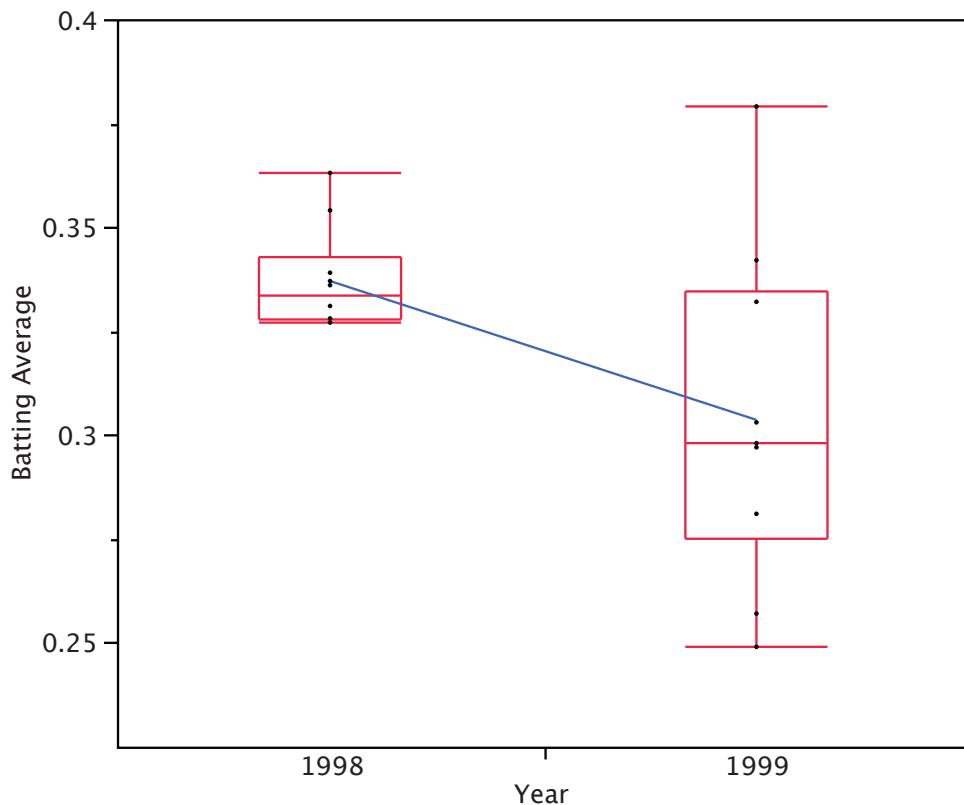


Figure 4. Quantile plots of the batting averages. The line connects the means of the plots.

Regression toward the mean plays a role in the so-called “Sophomore Slump,” a good example of which is that a player who wins “rookie of the year” typically does less well in his second season. A related phenomenon is called the Sports Illustrated Cover Jinx.

An experiment without a control group can *confound* regression effects with real effects. For example, consider a hypothetical experiment to evaluate a reading-improvement program. All first graders in a school district were given a reading achievement test and the 50 lowest-scoring readers were enrolled in the program. The students were retested following the program and the mean improvement was large. Does this necessarily mean the program was effective? No, it could be that the initial poor performance of the students was due, in part, to bad luck. Their luck

would be expected to improve in the retest, which would increase their scores with or without the treatment program.

For a real example, consider an experiment that sought to determine whether the drug propranolol would increase the SAT scores of students thought to have test anxiety ([external link](#)). Propranolol was given to 25 high-school students chosen because IQ tests and other academic performance indicated that they had not done as well as expected on the SAT. On a retest taken after receiving propranolol, students improved their SAT scores an average of 120 points. This was a *significantly* greater increase than the 38 points expected simply on the basis of having taken the test before. The problem with the study is that the method of selecting students likely resulted in a disproportionate number of students who had bad luck when they first took the SAT. Consequently, these students would likely have increased their scores on a retest with or without the propranolol. This is not to say that propranolol had no effect. However, since possible propranolol effects and regression effects were confounded, no firm conclusions should be drawn.

Randomly assigning students to either the propranolol group or a control group would have improved the experimental design. Since the regression effects would then not have been systematically different for the two groups, a significant difference would have provided good evidence for a propranolol effect.

# Introduction to Multiple Regression

by David M. Lane

## *Prerequisites*

- Chapter 14: Simple Linear Regression
- Chapter 14: Partitioning Sums of Squares
- Chapter 14: Standard Error of the Estimate
- Chapter 14: Inferential Statistics for  $b$  and  $r$

## *Learning Objectives*

1. State the regression equation
2. Define “regression coefficient”
3. Define “beta weight”
4. Explain what  $R$  is and how it is related to  $r$
5. Explain why a regression weight is called a “partial slope”
6. Explain why the sum of squares explained in a multiple regression model is usually less than the sum of the sums of squares in simple regression
7. Define  $R^2$  in terms of proportion explained
8. Test  $R^2$  for significance
9. Test the difference between a complete and reduced model for significance
10. State the assumptions of multiple regression and specify which aspects of the analysis require assumptions

In simple linear regression, a criterion variable is predicted from one predictor variable. In multiple regression, the criterion is predicted by two or more variables. For example, in the SAT case study, you might want to predict a student's university grade point average on the basis of their High-School GPA (HSGPA) and their total SAT score (verbal + math). The basic idea is to find a linear combination of HSGPA and SAT that best predicts University GPA (UGPA). That is, the problem is to find the values of  $b_1$  and  $b_2$  in the equation shown below that gives the best predictions of UGPA. As in the case of simple linear regression, we define the best predictions as the predictions that minimize the squared errors of prediction.

$$\text{UGPA}' = b_1 \text{HSGPA} + b_2 \text{SAT} + A$$

where  $UGPA'$  is the predicted value of University GPA and A is a constant. For these data, the best prediction equation is shown below:

$$UGPA' = 0.541 \times HSGPA + 0.008 \times SAT + 0.540$$

In other words, to compute the prediction of a student's University GPA, you add up (a) their High-School GPA multiplied by 0.541, (b) their SAT multiplied by 0.008, and (c) 0.540. Table 1 shows the data and predictions for the first five students in the dataset.

Table 1. Data and Predictions.

HSGPA	SAT	UGPA'
3.45	1232	3.38
2.78	1070	2.89
2.52	1086	2.76
3.67	1287	3.55
3.24	1130	3.19

The values of b (b1 and b2) are sometimes called “regression coefficients” and sometimes called “*regression weights*.” These two terms are synonymous.

The *multiple correlation* (R) is equal to the correlation between the predicted scores and the actual scores. In this example, it is the correlation between  $UGPA'$  and  $UGPA$ , which turns out to be 0.79. That is,  $R = 0.79$ . Note that R will never be negative since if there are negative correlations between the predictor variables and the criterion, the regression weights will be negative so that the correlation between the predicted and actual scores will be positive.

## Interpretation of Regression Coefficients

A regression coefficient in multiple regression is the slope of the linear relationship between the criterion variable and the part of a predictor variable that is independent of all other predictor variables. In this example, the regression coefficient for HSGPA can be computed by first predicting HSGPA from SAT and saving the errors of prediction (the differences between HSGPA and  $HSGPA'$ ). These errors of prediction are called “residuals” since they are what is left over in HSGPA after the predictions from SAT are subtracted, and they represent the part

of HSGPA that is independent of SAT. These residuals are referred to as HSGPA.SAT, which means they are the residuals in HSGPA after having been predicted by SAT. The correlation between HSGPA.SAT and SAT is necessarily 0.

The final step in computing the regression coefficient is to find the slope of the relationship between these residuals and UGPA. This slope is the regression coefficient for HSGPA. The following equation is used to predict HSGPA from SAT:

$$\text{HSGPA}' = -1.314 + 0.0036 \times \text{SAT}$$

The residuals are then computed as:

$$\text{HSGPA} - \text{HSGPA}'$$

The linear regression equation for the prediction of UGPA by the residuals is

$$\text{UGPA}' = 0.541 \times \text{HSGPA.SAT} + 3.173$$

Notice that the slope (0.541) is the same value given previously for  $b_1$  in the multiple regression equation.

This means that the regression coefficient for HSGPA is the slope of the relationship between the criterion variable and the part of HSGPA that is *independent* of (uncorrelated with) the other predictor variables. It represents the change in the criterion variable associated with a change of one in the predictor variable when all other predictor variables are held constant. Since the regression coefficient for HSGPA is 0.54, this means that, holding SAT constant, a change of one in HSGPA is associated with a change of 0.54 in UGPA. If two students had the same SAT and differed in HSGPA by 2, then you would predict they would differ in UGPA by  $(2)(0.54) = 1.08$ . Similarly, if they differed by 0.5, then you would predict they would differ by  $(0.5)(0.54) = 0.27$ .

The slope of the relationship between the part of a predictor variable independent of other predictor variables and the criterion is its partial slope. Thus the regression coefficient of 0.541 for HSGPA and the regression coefficient of 0.008 for SAT are partial slopes. Each partial slope represents the relationship between the predictor variable and the criterion holding constant all of the other predictor variables.

It is difficult to compare the coefficients for different variables directly because they are measured on different scales. A difference of 1 in HSGPA is a fairly large difference, whereas a difference of 1 on the SAT is negligible. Therefore, it can be advantageous to transform the variables so that they are on the same scale. The most straightforward approach is to standardize the variables so that they all have a standard deviation of 1. A regression weight for standardized variables is called a “beta weight” and is designated by the Greek letter  $\beta$ . For these data, the beta weights are 0.625 and 0.198. These values represent the change in the criterion (in standard deviations) associated with a change of one standard deviation on a predictor [holding constant the value(s) on the other predictor(s)]. Clearly, a change of one standard deviation on HSGPA is associated with a larger difference than a change of one standard deviation of SAT. In practical terms, this means that if you know a student's HSGPA, knowing the student's SAT does not aid the prediction of UGPA much. However, if you do not know the student's HSGPA, his or her SAT can aid in the prediction since the  $\beta$  weight in the simple regression predicting UGPA from SAT is 0.68. For comparison purposes, the  $\beta$  weight in the simple regression predicting UGPA from HSGPA is 0.78. As is typically the case, the partial slopes are smaller than the slopes in simple regression.

### Partitioning the Sums of Squares

Just as in the case of simple linear regression, the sum of squares for the criterion (UGPA in this example) can be partitioned into the sum of squares predicted and the sum of squares error. That is,

$$SSY = SSY' + SSE$$

which for these data:

$$20.798 = 12.961 + 7.837$$

The sum of squares predicted is also referred to as the “sum of squares explained.” Again, as in the case of simple regression,

$$\text{Proportion Explained} = SSY' / SSY$$

In simple regression, the proportion of variance explained is equal to  $r^2$ ; in multiple regression, the proportion of variance explained is equal to  $R^2$ .

In multiple regression, it is often informative to partition the sums of squares explained among the predictor variables. For example, the sum of squares explained for these data is 12.96. How is this value divided between HSGPA and SAT? One approach that, as will be seen, does not work is to predict UGPA in separate simple regressions for HSGPA and SAT. As can be seen in Table 2, the sum of squares in these separate simple regressions is 12.64 for HSGPA and 9.75 for SAT. If we add these two sums of squares we get 22.39, a value much larger than the sum of squares explained of 12.96 in the multiple regression analysis. The explanation is that HSGPA and SAT are highly correlated ( $r = .78$ ) and therefore much of the variance in UGPA is confounded between HSGPA or SAT. That is, it could be explained by either HSGPA or SAT and is counted twice if the sums of squares for HSGPA and SAT are simply added.

Table 2. Sums of Squares for Various Predictors

Predictors	Sum of Squares
HSGPA	12.64
SAT	9.75
HSGPA and SAT	12.96

Table 3 shows the partitioning of the sums of squares into the sum of squares uniquely explained by each predictor variable, the sum of squares confounded between the two predictor variables, and the sum of squares error. It is clear from this table that most of the sum of squares explained is confounded between HSGPA and SAT. Note that the sum of squares uniquely explained by a predictor variable is analogous to the partial slope of the variable in that both involve the relationship between the variable and the criterion with the other variable(s) controlled.

Table 3. Partitioning the Sum of Squares

Source	Sum of Squares	Proportion
HSGPA (unique)	3.21	0.15
SAT (unique)	0.32	0.02
HSGPA and SAT (Confounded)	9.43	0.45
Error	7.84	0.38
Total	20.8	1

The sum of squares uniquely attributable to a variable is computed by comparing two regression models: the complete model and a reduced model. The complete model is the multiple regression with all the predictor variables included (HSGPA and SAT in this example). A reduced model is a model that leaves out one of the predictor variables. The sum of squares uniquely attributable to a variable is the sum of squares for the complete model minus the sum of squares for the reduced model in which the variable of interest is omitted. As shown in Table 2, the sum of squares for the complete model (HSGPA and SAT) is 12.96. The sum of squares for the reduced model in which HSGPA is omitted is simply the sum of squares explained using SAT as the predictor variable and is 9.75. Therefore, the sum of squares uniquely attributable to HSGPA is  $12.96 - 9.75 = 3.21$ . Similarly, the sum of squares uniquely attributable to SAT is  $12.96 - 12.64 = 0.32$ . The confounded sum of squares in this example is computed by subtracting the sum of squares uniquely attributable to the predictor variables from the sum of squares for the complete model:  $12.96 - 3.21 - 0.32 = 9.43$ . The computation of the confounded sums of squares in analyses with more than two predictors is more complex and beyond the scope of this text.

Since the variance is simply the sum of squares divided by the degrees of freedom, it is possible to refer to the proportion of variance explained in the same way as the proportion of the sum of squares explained. It is slightly more common to refer to the proportion of variance explained than the proportion of the sum of squares explained and, therefore, that terminology will be adopted frequently here.

When variables are highly correlated, the variance explained uniquely by the individual variables can be small even though the variance explained by the variables taken together is large. For example, although the proportions of variance

explained uniquely by HSGPA and SAT are only 0.15 and 0.02 respectively, together these two variables explain 0.62 of the variance. Therefore, you could easily underestimate the importance of variables if only the variance explained uniquely by each variable is considered. Consequently, it is often useful to consider a set of related variables. For example, assume you were interested in predicting job performance from a large number of variables some of which reflect cognitive ability. It is likely that these measures of cognitive ability would be highly correlated among themselves and therefore no one of them would explain much of the variance independent of the other variables. However, you could avoid this problem by determining the proportion of variance explained by all of the cognitive ability variables considered together as a set. The variance explained by the set would include all the variance explained uniquely by the variables in the set as well as all the variance confounded among variables in the set. It would not include variance confounded with variables outside the set. In short, you would be computing the variance explained by the set of variables that is independent of the variables not in the set.

## Inferential Statistics

We begin by presenting the formula for testing the significance of the contribution of a set of variables. We will then show how special cases of this formula can be used to test the significance of  $R^2$  as well as to test the significance of the unique contribution of individual variables.

The first step is to compute two regression analyses: (1) an analysis in which all the predictor variables are included and (2) an analysis in which the variables in the set of variables being tested are **excluded**. The former regression model is called the “complete model” and the latter is called the “reduced model.” The basic idea is that if the reduced model explains much less than the complete model, then the set of variables excluded from the reduced model is important.

The formula for testing the contribution of a group of variables is:

$$F = \frac{\frac{SSQ_c - SSQ_R}{P_c - P_R}}{\frac{SSQ_T - SSQ_c}{N - P_c - 1}} = \frac{MS_{explained}}{MS_{error}}$$

where:

$SSQ_C$  is the sum of squares for the complete model,

$SSQ_R$  is the sum of squares for the reduced model,

$p_C$  is the number of predictors in the complete model,

$p_R$  is the number of predictors in the reduced model,

$SSQ_T$  is the sum of squares total (the sum of squared deviations of the criterion variable from its mean), and

$N$  is the total number of observations

The degrees of freedom for the numerator is  $p_C - p_R$  and the degrees of freedom for the denominator is  $N - p_C - 1$ . If the  $F$  is significant, then it can be concluded that the variables excluded in the reduced set contribute to the prediction of the criterion variable independently of the other variables.

This formula can be used to test the significance of  $R^2$  by defining the reduced model as having no predictor variables. In this application,  $SSQ_R$  and  $p_R = 0$ . The formula is then simplified as follows:

$$F_{(p_C, N-p_C-1)} = \frac{\frac{SSQ_C}{p_C}}{\frac{SSQ_T - SSQ_C}{N - p_C - 1}} = \frac{MS_{\text{explained}}}{MS_{\text{error}}}$$

which for this example becomes:

$$F = \frac{\frac{12.96}{2}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{6.48}{0.08} = 84.35.$$

The degrees of freedom are 2 and 102. The F distribution calculator shows that  $p < 0.001$ .

The reduced model used to test the variance explained uniquely by a single predictor consists of all the variables except the predictor variable in question. For example, the reduced model for a test of the unique contribution of HSGPA contains only the variable SAT. Therefore, the sum of squares for the reduced model is the sum of squares when UGPA is predicted by SAT. This sum of squares is 9.75. The calculations for F are shown below:

$$F_{(1,102)} = \frac{\frac{12.96 - 9.75}{2-1}}{\frac{20.80 - 12.96}{105-2-1}} = \frac{3.212}{0.077} = 41.80.$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that  $p < 0.001$ .

Similarly, the reduced model in the test for the unique contribution of SAT consists of HSGPA.

$$F = \frac{\frac{12.96 - 12.64}{2-1}}{\frac{20.80 - 12.96}{105-2-1}} = \frac{0.322}{0.077} = 4.19.$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that  $p = 0.0432$ .

The significance test of the variance explained uniquely by a variable is identical to a significance test of the regression coefficient for that variable. A regression coefficient and the variance explained uniquely by a variable both reflect the relationship between a variable and the criterion independent of the other variables. If the variance explained uniquely by a variable is not zero, then the regression coefficient cannot be zero. Clearly, a variable with a regression coefficient of zero would explain no variance.

Other inferential statistics associated with multiple regression that are beyond the scope of this text. Two of particular importance are (1) confidence intervals on regression slopes and (2) confidence intervals on predictions for specific observations. These inferential statistics can be computed by standard statistical analysis packages such as R, SPSS, STATA, SAS, and JMP.

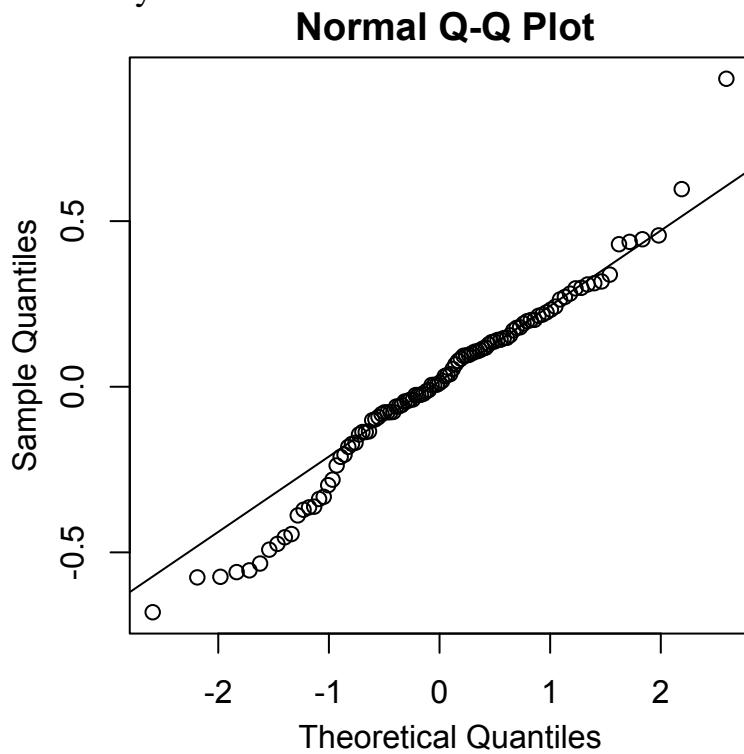
## **Assumptions**

No assumptions are necessary for computing the regression coefficients or for partitioning the sums of squares. However, there are several assumptions made when interpreting inferential statistics. Moderate violations of Assumptions 1-3 do not pose a serious problem for testing the significance of predictor variables. However, even small violations of these assumptions pose problems for confidence intervals on predictions for specific observations.

### 1. Residuals are normally distributed:

As in the case of simple linear regression, the residuals are the errors of prediction. Specifically, they are the differences between the actual scores on the criterion and the predicted scores. A Q-Q plot for the residuals for the example data is shown below. This plot reveals that the actual data values at the lower end of the distribution do not increase as much as would be expected for a normal distribution. It also reveals that the highest value in the data is higher than would be expected for the highest value in a sample of this size from a normal distribution. Nonetheless, the distribution does not deviate greatly from

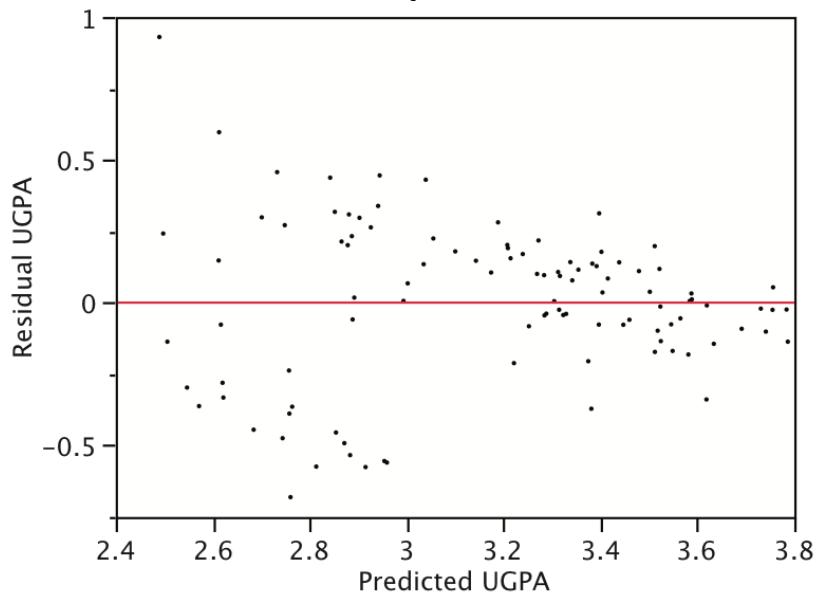
normality.



2. Homoscedasticity:

It is assumed that the variance of the errors of prediction are the same for all predicted values. As can be seen below, this assumption is violated in the example data because the errors of prediction are much larger for observations with low-to-medium predicted scores than for observations with high predicted scores. Clearly, a confidence interval on a low predicted UGPA would

underestimate the uncertainty.



### 3. Linearity:

It is assumed that the relationship between each predictor variable and the criterion variable is linear. If this assumption is not met, then the predictions may systematically overestimate the actual values for one range of values on a predictor variable and underestimate them for another.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 14: Regression Toward the Mean

In a discussion about the Dallas Cowboy football team, there was a comment that the quarterback threw far more interceptions in the first two games than is typical (there were two interceptions per game). The author correctly pointed out that, because of regression toward the mean, performance in the future is expected to improve. However, the author defined regression toward the mean as, "In nerd land, that basically means that things tend to even out over the long run."

## **What do you think?**

Comment on that definition.

That definition is sort of correct, but it could be stated more precisely. Things don't always tend to even out in the long run. If a great player has an average game, then things wouldn't even out (to the average of all players) but would regress toward that player's high mean performance.

## References

Schall, T., & Smith, G. (2000) Do Baseball Players Regress Toward the Mean? *The American Statistician*, 54, 231-235.

## Exercises

### *Prerequisites*

All material presented in the Regression chapter

1. What is the equation for a regression line? What does each term in the line refer to?
2. The formula for a regression equation is  $Y' = 2X + 9$ .
  - a. What would be the predicted score for a person scoring 6 on X?
  - b. If someone's predicted score was 14, what was this person's score on X?
3. What criterion is used for deciding which regression line fits best?
4. What does the standard error of the estimate measure? What is the formula for the standard error of the estimate?
5.
  - a. In a regression analysis, the sum of squares for the predicted scores is 100 and the sum of squares error is 200, what is  $R^2$ ?
  - b. In a different regression analysis, 40% of the variance was explained. The sum of squares total is 1000. What is the sum of squares of the predicted values?
6. For the X,Y data below, compute:
  - a.  $r$  and determine if it is significantly different from zero.
  - b. the slope of the regression line and test if it differs significantly from zero.
  - c. the 95% confidence interval for the slope.

X	Y
4	6
3	7
5	12
11	17
10	9
14	21

7. What assumptions are needed to calculate the various inferential statistics of linear regression?
8. The correlation between years of education and salary in a sample of 20 people from a certain company is .4. Is this correlation statistically significant at the .05 level?
9. A sample of X and Y scores is taken, and a regression line is used to predict Y from X. If  $SSY' = 300$ ,  $SSE = 500$ , and  $N = 50$ , what is:
- $SSY$ ?
  - the standard error of the estimate?
  - $R^2$ ?
10. Using linear regression, find the predicted post-test score for someone with a score of 45 on the pre-test.

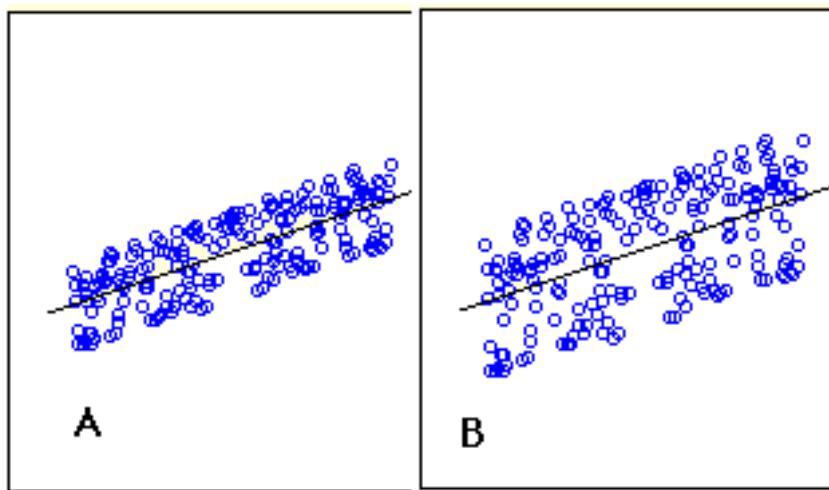
<b>Pre</b>	<b>Post</b>
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	50
55	63
49	57
45	73
57	63
46	46
60	60
65	47
64	73
50	58
74	85
59	44

11. The equation for a regression line predicting the number of hours of TV watched by children (Y) from the number of hours of TV watched by their parents (X) is  $Y' = 4 + 1.2X$ . The sample size is 12.

- a. If the standard error of  $b$  is .4, is the slope statistically significant at the .05 level?
- b. If the mean of  $X$  is 8, what is the mean of  $Y$ ?
12. Based on the table below, compute the regression line that predicts  $Y$  from  $X$ .

$M_x$	$M_y$	$s_x$	$s_y$	$r$
10	12	2.5	3.0	-0.6

13. Does A or B have a larger standard error of the estimate?



14. True/false: If the slope of a simple linear regression line is statistically significant, then the correlation will also always be significant.
15. True/false: If the slope of the relationship between  $X$  and  $Y$  is larger for Population 1 than for Population 2, the correlation will necessarily be larger in Population 1 than in Population 2. Why or why not?
16. True/false: If the correlation is .8, then 40% of the variance is explained.
17. True/false: If the actual  $Y$  score was 31, but the predicted score was 28, then the error of prediction is 3.

*Questions from Case Studies*

Angry Moods (AM) case study

18. (AM) Find the regression line for predicting Anger-Out from Control-Out.

- a. What is the slope?
- b. What is the intercept?
- c. Is the relationship at least approximately linear?
- d. Test to see if the slope is significantly different from 0.
- e. What is the standard error of the estimate?

SAT and GPA (SG) case study

19. (SG) Find the regression line for predicting the overall university GPA from the high school GPA.

- a. What is the slope?
- b. What is the y-intercept?
- c. If someone had a 2.2 GPA in high school, what is the best estimate of his or her college GPA?
- d. If someone had a 4.0 GPA in high school, what is the best estimate of his or her college GPA?

Driving (D) case study

20. (D) What is the correlation between age and how often the person chooses to drive in inclement weather? Is this correlation statistically significant at the .01 level? Are older people more or less likely to report that they drive in inclement weather?

21. (D) What is the correlation between how often a person chooses to drive in inclement weather and the percentage of accidents the person believes occur in inclement weather? Is this correlation significantly different from 0?

22. (D) Use linear regression to predict how often someone rides public transportation in inclement weather from what percentage of accidents that person thinks occur in inclement weather. (Pubtran by Accident)

- (a) Create a scatter plot of this data and add a regression line.
- (b) What is the slope?

- (c) What is the intercept?
- (d) Is the relationship at least approximately linear?
- (e) Test if the slope is significantly different from 0.
- (f) Comment on possible assumption violations for the test of the slope.
- (g) What is the standard error of the estimate?

# 15. Analysis of Variance

- A. Introduction
- B. ANOVA Designs
- C. One-Factor ANOVA (Between-Subjects)
- D. Multi-Factor ANOVA (Between-Subjects)
- E. Unequal Sample Sizes
- F. Tests Supplementing ANOVA
- G. Within-Subjects ANOVA

# Introduction

by David M. Lane

## *Prerequisites*

- Chapter 3: Variance
- Chapter 11: Significance Testing
- Chapter 12: All Pairwise Comparisons among Means

## *Learning Objectives*

1. What null hypothesis is tested by ANOVA
2. Describe the uses of ANOVA

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called “Analysis of Variance” rather than “Analysis of Means.” As you will see, the name is appropriate because inferences about means are made by analyzing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study “Smiles and Leniency,” the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. The chapter “All Pairwise Comparisons among Means” showed how to test differences among means. The results from the Tukey HSD test are shown in Table 1.

Table 1. Six pairwise comparisons.

Comparison	Mi-Mj	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.01
Felt - Miserable	0	0	1
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

Notice that the only significant difference is between the False and Neutral conditions.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is

$$\mu_{\text{false}} = \mu_{\text{felt}} = \mu_{\text{miserable}} = \mu_{\text{neutral}}.$$

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the Tukey HSD test. The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analyses that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

# Analysis of Variance Designs

by David M. Lane

## *Prerequisites*

- Chapter 15: Introduction to ANOVA

## *Learning Objectives*

1. Be able to identify the factors and levels of each factor from a description of an experiment
2. Determine whether a factor is a between-subjects or a within-subjects factor
3. Define factorial design

There are many types of experimental designs that can be analyzed by ANOVA. This section discusses many of these designs and defines several key terms used.

## **Factors and Levels**

The section on variables defined an independent variable as a *variable* manipulated by the experimenter. In the case study “Smiles and Leniency,” the effect of different types of smiles on the leniency showed to a person was investigated. Four different types of smiles (neutral, false, felt, miserable, on leniency) were shown. In this experiment, “Type of Smile” is the independent variable. In describing an ANOVA design, the term factor is a synonym of independent variable. Therefore, “Type of Smile” is the factor in this experiment. Since four types of smiles were compared, the factor “Type of Smile” has four *levels*.

An ANOVA conducted on a design in which there is only one factor is called a *one-way ANOVA*. If an experiment has two factors, then the ANOVA is called a *two-way ANOVA*. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

## Between- and Within-Subjects Factors

In the “Smiles and Leniency” study, the four levels of the factor “Type of Smile” were represented by four separate groups of subjects. When different subjects are used for the levels of a factor, the factor is called a *between-subjects factor* or a *between-subjects variable*. The term “between subjects” reflects the fact that comparisons are between different groups of subjects.

In the “ADHD Treatment” study, every subject was tested with **each** of four dosage levels (0, 0.15, 0.30, 0.60 mg/kg) of a drug. Therefore there was only one group of subjects, and comparisons were not between different groups of subjects but between conditions within the same subjects. When the same subjects are used for the levels of a factor, the factor is called a *within-subjects factor* or a *within-subjects variable*. Within-subjects variables are sometimes referred to as repeated-measures variables since there are repeated measurements of the same subjects.

## Multi-Factor Designs

It is common for designs to have more than one factor. For example, consider a hypothetical study of the effects of age and gender on reading speed in which males and females from the age levels of 8 years, 10 years, and 12 years are tested. There would be a total of six different groups as shown in Table 1.

Table 1. Gender x Age Design

Group	Gender	Age
1	Female	8
2	Female	10
3	Female	12
4	Male	8
5	Male	10
6	Male	12

This design has two factors: age and gender. Age has three levels and gender has two levels. When all combinations of the levels are included (as they are here), the design is called a *factorial design*. A concise way of describing this design is as a Gender (2) x Age (3) factorial design where the numbers in parentheses indicate

the number of levels. Complex designs frequently have more than two factors and may have combinations of between- and within-subjects factors.

# One-Factor ANOVA (Between Subjects)

by David M. Lane

## *Prerequisites*

- Chapter 3: Variance
- Chapter 7: Introduction to Normal Distributions
- Chapter 11: Significance Testing
- Chapter 11: One- and Two-Tailed Tests
- Chapter 12: t Test of Differences Between Groups
- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs

## *Learning Objectives*

1. State what the Mean Square Error (MSE) estimates when the null hypothesis is true and when the null hypothesis is false
2. State what the Mean Square Between (MSB) estimates when the null hypothesis is true and when the null hypothesis is false
3. State the assumptions of a one-way ANOVA
4. Compute MSE
5. Compute MSB
6. Compute F and its two degrees of freedom parameters
7. Describe the shape of the F distribution
8. Explain why ANOVA is best thought of as a two-tailed test even though literally only one tail of the distribution is used
9. State the relationship between the t and F distributions
10. Partition the sums of squares into conditions and error
11. Format data to be used with a computer statistics program

This section shows how ANOVA can be used to analyze a one-factor between-subjects design. We will use as our main example the “Smiles and Leniency” case study. In this study there were four conditions with 34 subjects in each condition. There was one score per subject. The null hypothesis tested by ANOVA is that the population means for all conditions are the same. This can be expressed as follows:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

where  $H_0$  is the null hypothesis and  $k$  is the number of conditions. In the smiles and leniency study,  $k = 4$  and the null hypothesis is

$$H_0: \mu_{\text{false}} = \mu_{\text{felt}} = \mu_{\text{miserable}} = \mu_{\text{neutral}}.$$

If the null hypothesis is rejected, then it can be concluded that at least one of the population means is different from at least one other population mean.

Analysis of variance is a method for testing differences among means by analyzing variance. The test is based on two estimates of the population variance ( $\sigma^2$ ). One estimate is called the mean square error (MSE) and is based on differences among scores within the groups. MSE estimates  $\sigma^2$  regardless of whether the null hypothesis is true (the population means are equal). The second estimate is called the mean square between (MSB) and is based on differences among the sample means. MSB only estimates  $\sigma^2$  if the population means are equal. If the population means are not equal, then MSB estimates a quantity larger than  $\sigma^2$ . Therefore, if the MSB is much larger than the MSE, then the population means are unlikely to be equal. On the other hand, if the MSB is about the same as MSE, then the data are consistent with the hypothesis that the population means are equal.

Before proceeding with the calculation of MSE and MSB, it is important to consider the assumptions made by ANOVA:

1. The populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the values are not independent. The analysis of data with two scores per subject is shown in the section on within-subjects ANOVA later in this chapter.

These assumptions are the same as for a t test of differences between groups except that they apply to two or more groups, not just to two groups.

The means and variances of the four groups in the “Smiles and Leniency” case study are shown in Table 1. Note that there are 34 subjects in each of the four conditions (False, Felt, Miserable, and Neutral).

Table 1. Means and Variances from “Smiles and Leniency” Study

Condition	Mean	Variance
FALSE	5.3676	3.338
Felt	4.9118	2.8253
Miserable	4.9118	2.1132
Neutral	4.1176	2.3191

## Sample Sizes

The first calculations in this section all assume that there is an equal number of observations in each group. Unequal sample size calculations are shown in the section on sources of variation. We will refer to the number of observations in each group as  $n$  and the total number of observations as  $N$ . For these data there are four groups of 34 observations. Therefore  $n = 34$  and  $N = 136$ .

## Computing MSE

Recall that the assumption of homogeneity of variance states that the variance within each of the populations ( $\sigma^2$ ) is the same. This variance,  $\sigma^2$ , is the quantity estimated by MSE and is computed as the mean of the sample variances. For these data, the MSE is equal to 2.6489.

## Computing MSB

The formula for MSB is based on the fact that the variance of the sampling distribution of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{n}$$

where  $n$  is the sample size of each group. Rearranging this formula, we have

$$\sigma^2 = n\sigma_M^2$$

Therefore, if we knew the variance of the sampling distribution of the mean, we could compute  $\sigma^2$  by multiplying it by  $n$ . Although we do not know the variance of the sampling distribution of the mean, we can estimate it with the variance of the

sample means. For the leniency data, the variance of the four sample means is 0.270. To estimate  $\sigma^2$ , we multiply the variance of the sample means (0.270) by  $n$  (the number of observations in each group, which is 34). We find that  $MSB = 9.179$ .

To sum up these steps:

1. Compute the means.
2. Compute the variance of the means.
3. Multiply the variance of the means by  $n$ .

## Recap

If the population means are equal, then both  $MSE$  and  $MSB$  are estimates of  $\sigma^2$  and should therefore be about the same. Naturally, they will not be exactly the same since they are just estimates and are based on different aspects of the data: The  $MSB$  is computed from the sample means and the  $MSE$  is computed from the sample variances.

If the population means are not equal, then  $MSE$  will still estimate  $\sigma^2$  because differences in population means do not affect variances. However, differences in population means affect  $MSB$  since differences among population means are associated with differences among sample means. It follows that the larger the differences among sample means, the larger the  $MSB$ . **In short,  $MSE$  estimates  $\sigma^2$  whether or not the population means are equal, whereas  $MSB$  estimates  $\sigma^2$  only when the population means are equal and estimates a larger quantity when they are not equal.**

## Comparing $MSE$ and $MSB$

The critical step in an ANOVA is comparing  $MSE$  and  $MSB$ . Since  $MSB$  estimates a larger quantity than  $MSE$  only when the population means are not equal, a finding of a larger  $MSB$  than an  $MSE$  is a sign that the population means are not equal. But since  $MSB$  could be larger than  $MSE$  by chance even if the population means are equal,  $MSB$  must be much larger than  $MSE$  in order to justify the conclusion that the population means differ. But how much larger must  $MSB$  be? For the “Smiles and Leniency” data, the  $MSB$  and  $MSE$  are 9.179 and 2.649, respectively. Is that difference big enough? To answer, we would need to know the probability of getting that big a difference or a bigger difference if the population means were all equal. The mathematics necessary to answer this question were

worked out by the statistician R. Fisher. Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

For these data, the F ratio is

$$F = 9.179/2.649 = 3.465.$$

Therefore, the MSB is 3.465 times higher than MSE. Would this have been likely to happen if all the population means were equal? That depends on the sample size. With a small sample size, it would not be too surprising because results from small samples are unstable. However, with a very large sample, the MSB and MSE are almost always about the same, and an F ratio of 3.465 or larger would be very unusual. Figure 1 shows the *sampling distribution* of F for the sample size in the “Smiles and Leniency” study. As you can see, it has a positive skew.

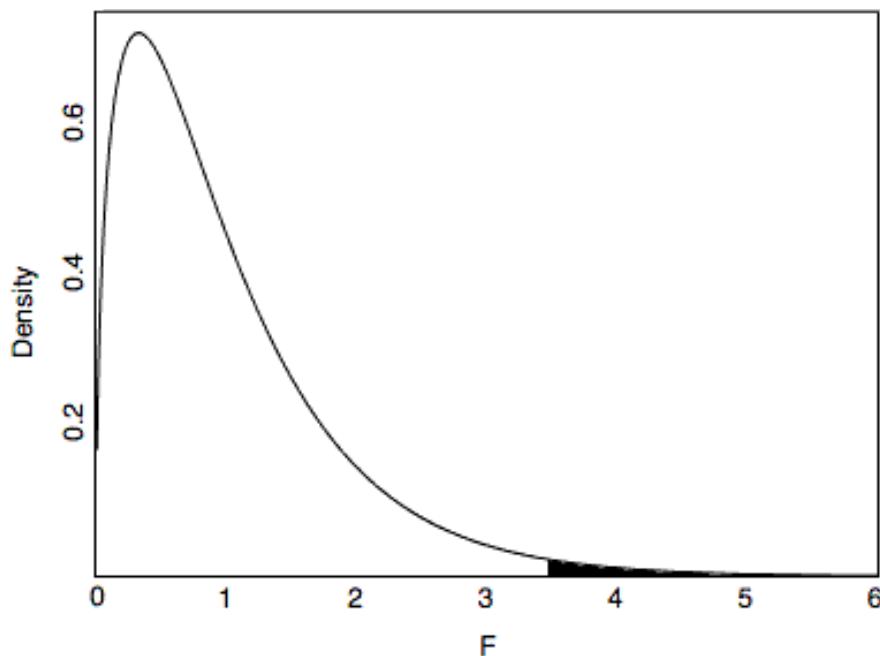


Figure 1. Distribution of F.

From Figure 1, you can see that F ratios of 3.465 or above are unusual occurrences. The area to the right of 3.465 represents the probability of an F that large or larger and is equal to 0.018 and therefore the null hypothesis can be rejected. The conclusion that at least one of the population means is different from at least one of the others is justified.

The shape of the F distribution depends on the sample size. More precisely, it depends on two degrees of freedom (df) parameters: one for the numerator (MSB) and one for the denominator (MSE). Recall that the degrees of freedom for an estimate of variance is equal to the number of observations minus one. Since the MSB is the variance of  $k$  means, it has  $k - 1$  df. The MSE is an average of  $k$  variances, each with  $n-1$  df. Therefore, the df for MSE is  $k(n - 1) = N - k$ . where  $N$  is the total number of observations,  $n$  is the number of observations in each group, and  $k$  is the number of groups. To summarize:

$$df_{\text{numerator}} = k - 1$$

$$df_{\text{denominator}} = N - k$$

For the “Smiles and Leniency” data,

$$df_{\text{numerator}} = k - 1 = 4 - 1 = 3$$

$$df_{\text{denominator}} = N - k = 136 - 4 = 132$$

$$F = 3.465$$

The F distribution calculator shows that  $p = 0.018$ .

### One-Tailed or Two?

Is the probability value from an F ratio a one-tailed or a two-tailed probability? In the literal sense, it is a one-tailed probability since, as you can see in Figure 1, the probability is the area in the right-hand tail of the distribution. However, the F ratio is sensitive to any pattern of differences among means. It is, therefore, a test of a two-tailed hypothesis and is best considered a two-tailed test.

### Relationship to the t test

Since an ANOVA and an independent-groups t test can both test the difference between two means, you might be wondering which one to use. Fortunately, it does not matter since the results will always be the same. When there are only two groups, the following relationship between F and t will always hold:

$$F(1, df_{\text{df}}) = t^2(df)$$

where  $df_{\text{df}}$  is the degrees of freedom for the denominator of the F test and  $df$  is the degrees of freedom for the t test.  $df_{\text{df}}$  will always equal  $df$ .

## Sources of Variation

Why do scores in an experiment differ from one another? Consider the scores of two subjects in the “Smiles and Leniency” study: one from the “False Smile” condition and one from the “Felt Smile” condition. An obvious possible reason that the scores could differ is that the subjects were treated differently (they were in different conditions and saw different stimuli). A second reason is that the two subjects may have differed with regard to their tendency to judge people leniently. A third is that, perhaps, one of the subjects was in a bad mood after receiving a low grade on a test. You can imagine that there are innumerable other reasons why the scores of the two subjects could differ. All of these reasons except the first (subjects were treated differently) are possibilities that were not under experimental investigation and, therefore, all of the differences (variation) due to these possibilities are unexplained. It is traditional to call unexplained variance error even though there is no implication that an error was made. Therefore, the variation in this experiment can be thought of as being either variation due to the condition the subject was in or due to error (the sum total of all reasons the subjects' scores could differ that were not measured).

One of the important characteristics of ANOVA is that it partitions the variation into its various sources. In ANOVA, the term *sum of squares* (SSQ) is used to indicate variation. The total variation is defined as the sum of squared differences between each score and the mean of all subjects. The mean of all subjects is called the grand mean and is designated as GM. (When there is an equal number of subjects in each condition, the grand mean is the mean of the condition means.) The total sum of squares is defined as

$$SSQ_{total} = \sum (X - GM)^2$$

which means to take each score, subtract the grand mean from it, square the difference, and then sum up these squared values. For the “Smiles and Leniency” study,  $SSQ_{total} = 377.19$ .

The sum of squares condition is calculated as shown below.

$$SSQ_{condition} = n \sum (M_1 - GM)^2 + (M_2 - GM)^2 + \cdots + (M_k - GM)^2$$

where  $n$  is the number of scores in **each** group,  $k$  is the number of groups,  $M_1$  is the mean for Condition 1,  $M_2$  is the mean for Condition 2, and  $M_k$  is the mean for Condition  $k$ . For the “Smiles and Leniency” study, the values are:

$$\begin{aligned} SSQ_{\text{condition}} &= 34 [(5.37 - 4.83)^2 + (4.91 - 4.83)^2 + \\ &\quad (4.91 - 4.83)^2 + (4.12 - 4.83)^2] \\ &= 27.5 \end{aligned}$$

If there are unequal sample sizes, the only change is that the following formula is used for the sum of squares condition:

$$SSQ_{\text{condition}} = \sum n_i (M_i - GM)^2 + n_2 (M_2 - GM)^2 + \dots + n_k (M_k - GM)^2$$

where  $n_i$  is the sample size of the  $i^{\text{th}}$  condition.  $SSQ_{\text{total}}$  is computed the same way as shown above.

The sum of squares error is the sum of the squared deviations of each score from its group mean. This can be written as

$$SSQ_{\text{error}} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \dots + \sum (X_{ik} - M_k)^2$$

where  $X_{i1}$  is the  $i^{\text{th}}$  score in group 1 and  $M_1$  is the mean for group 1,  $X_{i2}$  is the  $i^{\text{th}}$  score in group 2 and  $M_2$  is the mean for group 2, etc. For the “Smiles and Leniency” study, the means are: 5.368, 4.912, 4.912, and 4.118. The  $SSQ_{\text{error}}$  is therefore:

$$\begin{aligned} (2.5 - 5.368)^2 + (5.5 - 5.368)^2 + \dots + (6.5 - 4.118)^2 &= \\ 349.65 \end{aligned}$$

The sum of squares error can also be computed by subtraction:

$$SSQ_{\text{error}} = SSQ_{\text{total}} - SSQ_{\text{condition}}$$

$$SSQ_{\text{error}} = 377.189 - 27.535 = 349.65.$$

Therefore, the total sum of squares of 377.19 can be partitioned into  $SSQ_{\text{condition}}$  (27.53) and  $SSQ_{\text{error}}$  (349.66).

Once the sums of squares have been computed, the mean squares (MSB and MSE) can be computed easily. The formulas are:

$$\text{MSB} = SSQ_{\text{condition}}/dfn$$

where dfn is the degrees of freedom numerator and is equal to  $k - 1 = 3$ .

$$\text{MSB} = 27.535/3 = 9.18$$

which is the same value of MSB obtained previously (except for rounding error). Similarly,

$$\text{MSE} = SSQ_{\text{error}}/dfn$$

where dfd is the degrees of freedom for the denominator and is equal to  $N - k$ .

$$dfd = 136 - 4 = 132$$

$$\text{MSE} = 349.66/132 = 2.65$$

which is the same as obtained previously (except for rounding error). Note that the dfd is often called the dfe for *degrees of freedom error*.

The Analysis of Variance Summary Table shown below is a convenient way to summarize the partitioning of the variance. The rounding errors have been corrected.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.465	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The first column shows the sources of variation, the second column shows the degrees of freedom, the third shows the sums of squares, the fourth shows the

mean squares, the fifth shows the F ratio, and the last shows the probability value. Note that the mean squares are always the sums of squares divided by degrees of freedom. The F and p are relevant only to Condition. Although the mean square total could be computed by dividing the sum of squares by the degrees of freedom, it is generally not of much interest and is omitted here.

## Formatting data for Computer Analysis

Most computer programs that compute ANOVAs require your data to be in a specific form. Consider the data in Table 3.

Table 3. Example Data

Group 1	Group 2	Group 3
3	2	8
4	4	5
5	6	5

Here there are three groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 3 is shown in Table 4.

Table 4. Reformatted Data

<b>G</b>	<b>Y</b>
1	3
1	4
1	5
2	2
2	4
2	6
3	8
3	5
3	5

# Multi-Factor Between-Subjects Designs

by David M. Lane

## *Prerequisites*

- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs

## *Learning Objectives*

1. Define main effect, simple effect, interaction, and marginal mean
2. State the relationship between simple effects and interaction
3. Compute the source of variation and df for each effect in a factorial design
4. Plot the means for an interaction
5. Define three-way interaction

## Basic Concepts and Terms

In the “Bias Against Associates of the Obese” case study, the researchers were interested in whether the weight of a companion of a job applicant would affect judgments of a male applicant's qualifications for a job. Two *independent variables* were investigated: (1) whether the companion was obese or of typical weight and (2) whether the companion was a girlfriend or just an acquaintance. One approach could have been to conduct two separate studies, one with each independent variable. However, it is more efficient to conduct one study that includes both independent variables. Moreover, there is a much bigger advantage than efficiency for including two variables in the same study: it allows a test of the *interaction* between the variables. There is an interaction when the effect of one variable differs depending on the *level* of a second variable. For example, it is possible that the effect of having an obese companion would differ depending on the relationship to the companion. Perhaps there is more prejudice against a person with an obese companion if the companion is a girlfriend than if she is just an acquaintance. If so, there would be an interaction between the obesity factor and the relationship factor.

There are three effects of interest in this experiment:

1. Weight: Are applicants judged differently depending on the weight of their companion?
2. Relationship: Are applicants judged differently depending on their relationship with their companion?

3. Weight x Relationship Interaction: Does the effect of weight differ depending on the relationship with the companion?

The first two effects (Weight and Relationship) are both *main effects*. A main effect of an independent variable is the effect of the variable averaging over the levels of the other variable(s). It is convenient to talk about main effects in terms of *marginal means*. A marginal mean for a level of a variable is the mean of the means of all levels of the other variable. For example, the marginal mean for the level “Obese” is the mean of “Girlfriend Obese” and “Acquaintance Obese.” Table 1 shows that this marginal mean is equal to the mean of 5.65 and 6.15, which is 5.90. Similarly, the marginal mean for the level “Typical” is the mean of 6.19 and 6.59, which is 6.39. The main effect of Weight is based on a comparison of these two marginal means. Similarly, the marginal means for “Girlfriend” and “Acquaintance” are 5.92 and 6.37..

Table 1. Means for All Four Conditions

		Companion Weight		Marginal Mean
		Obese	Typical	
Relationship	Girlfriend	5.65	6.19	5.92
	Acquaintance	6.15	6.59	6.37
	Marginal Mean	5.9	6.39	

In contrast to a main effect, which is the effect of a variable averaged across levels of another variable, the simple effect of a variable is the effect of the variable at a single level of another variable. The simple effect of Weight at the level of “Girlfriend” is the difference between the “Girlfriend Typical” and the “Girlfriend Obese” conditions. The difference is  $6.19 - 5.65 = 0.54$ . Similarly, the simple effect of Weight at the level of “Acquaintance” is the difference between the “Acquaintance Typical” and the “Acquaintance Obese” conditions. The difference is  $6.59 - 6.15 = 0.44$ .

Recall that there is an interaction when the effect of one variable differs depending on the level of another variable. This is equivalent to saying that **there is an interaction when the simple effects differ**. In this example, the simple effects of

weight are 0.54 and 0.44. As shown below, these simple effects are not significantly different.

## Tests of Significance

The important questions are not whether there are main effects and interactions in the sample data. Instead, what is important is what the sample data allow you to conclude about the population. This is where Analysis of Variance comes in. ANOVA tests main effects and interactions for *significance*. An ANOVA Summary Table for these data is shown in Table 2.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Weight	1	10.4673	10.4673	6.214	0.0136
Relation	1	8.8144	8.8144	5.233	0.0234
W x R	1	0.1038	0.1038	0.062	0.8043
Error	172	289.7132	1.6844		
Total	175	310.1818			

Consider first the effect of “Weight.” The *degrees of freedom* (df) for “Weight” is 1. The degrees of freedom for a main effect is always equal to the number of levels of the variable minus one. Since there are two levels of the “Weight” variable (typical and obese), the df is  $2 - 1 = 1$ . We skip the calculation of the sum of squares (SSQ) not because it is difficult, but because it is so much easier to rely on computer programs to compute it. The mean square (MS) is the sum of squares divided by the df. The F ratio is computed by dividing the MS for the effect by the MS for error (MSE). For the effect of “Weight,”  $F = 10.4673/1.6844 = 6.214$ . The last column, p, is the probability of getting an F of 6.214 or larger given that there is no effect of weight in the population. The p value is 0.0136 and therefore the null *hypothesis* of no main effect of “Weight” is rejected. The conclusion is that being accompanied by an obese companion lowers judgments of qualifications.

The effect “Relation” is interpreted the same way. The conclusion is that being accompanied by a girlfriend leads to lower ratings than being accompanied by an acquaintance.

The df for an interaction is the product of the df's of variables in the interaction. For the “Weight x Relation” interaction (W x R), the df = 1 since both

Weight and Relation have one df:  $1 \times 1 = 1$ . The p value for the interaction is 0.8043, which is the probability of getting an interaction as big or bigger than the one obtained in the experiment if there were no interaction in the population. Therefore, these data provide no evidence for an interaction. Always keep in mind that the lack of evidence for an effect does not justify the conclusion that there is no effect. In other words, you do not accept the null hypothesis just because you do not reject it.

For “Error,” the degrees of freedom is equal to the total number of observations minus the total number of groups. The sample sizes of the four conditions in this experiment are shown in Table 3. The total number of observations is  $40 + 42 + 40 + 54 = 176$ . Since there are four groups,  $dfe = 176 - 4 = 172$ .

Table 3. Sample Sizes for All Four Conditions

		Companion Weight	
		Obese	Typical
Relationship	Girlfriend	40	42
	Acquaintance	40	54

The final row in the ANOVA Summary Table is “Total.” The degrees of freedom total is equal to the sum of all degrees of freedom. It is also equal to the number of observations minus 1, or  $176 - 1 = 175$ . When there are equal sample sizes, the sum of squares total will equal the sum of all other sums of squares. However, when there are unequal sample sizes, as there are here, this will not generally be true. The reasons for this are complex and are discussed in the section Unequal Sample Sizes.

## Plotting Means

Although the plot shown in Figure 1 illustrates the main effects as well as the interaction (or lack of an interaction), it is called an *interaction plot*. It is important to consider the components of this plot carefully. First, the dependent variable is on the Y-axis. Second, one of the independent variables is on the X-axis. In this case, it is the variable “Weight.” Finally, a separate line is drawn for each level of the other independent variable. It is better to label the lines right on the graph, as shown here, than with a legend.

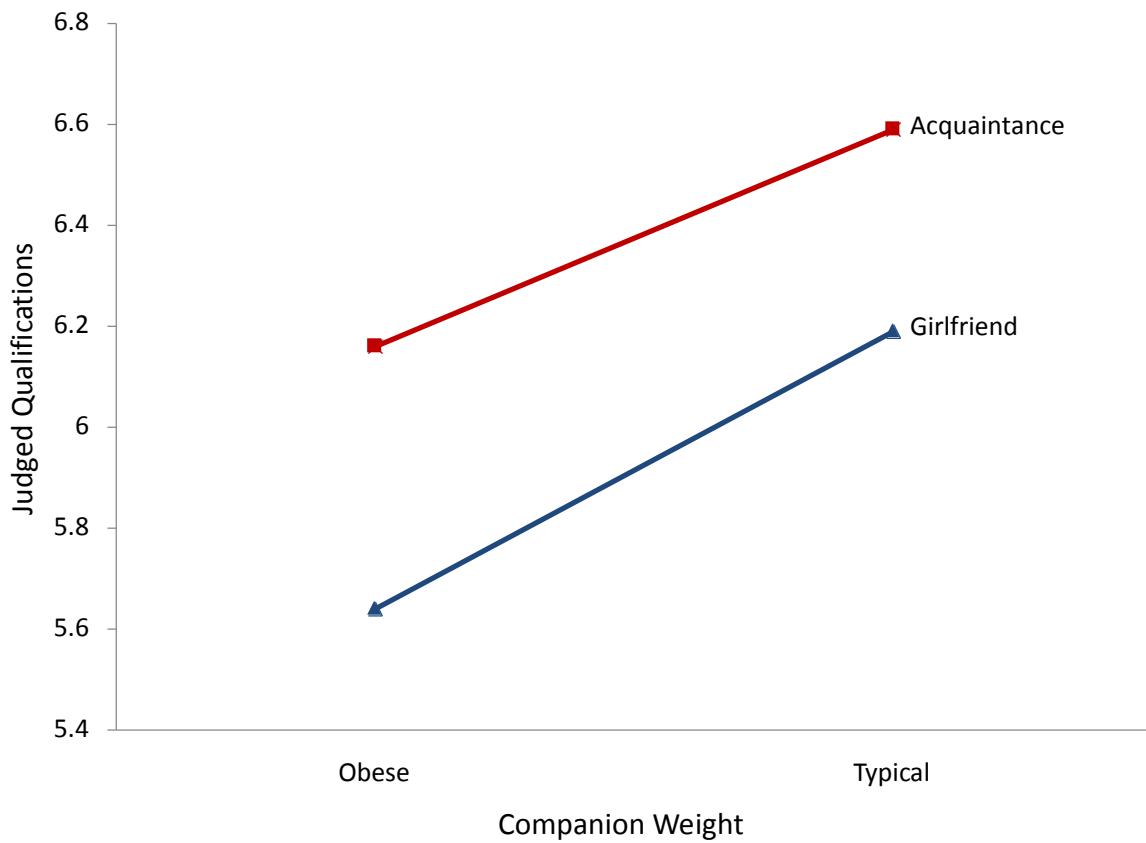


Figure 1. An interaction plot.

If you have three or more levels on the X-axis, you should not use lines unless there is some numeric ordering to the levels. If your variable on the X-axis is a qualitative variable, you can use a plot such as the one in Figure 2. However, as discussed in the section on bar charts, it would be better to replace each bar with a *box plot*.

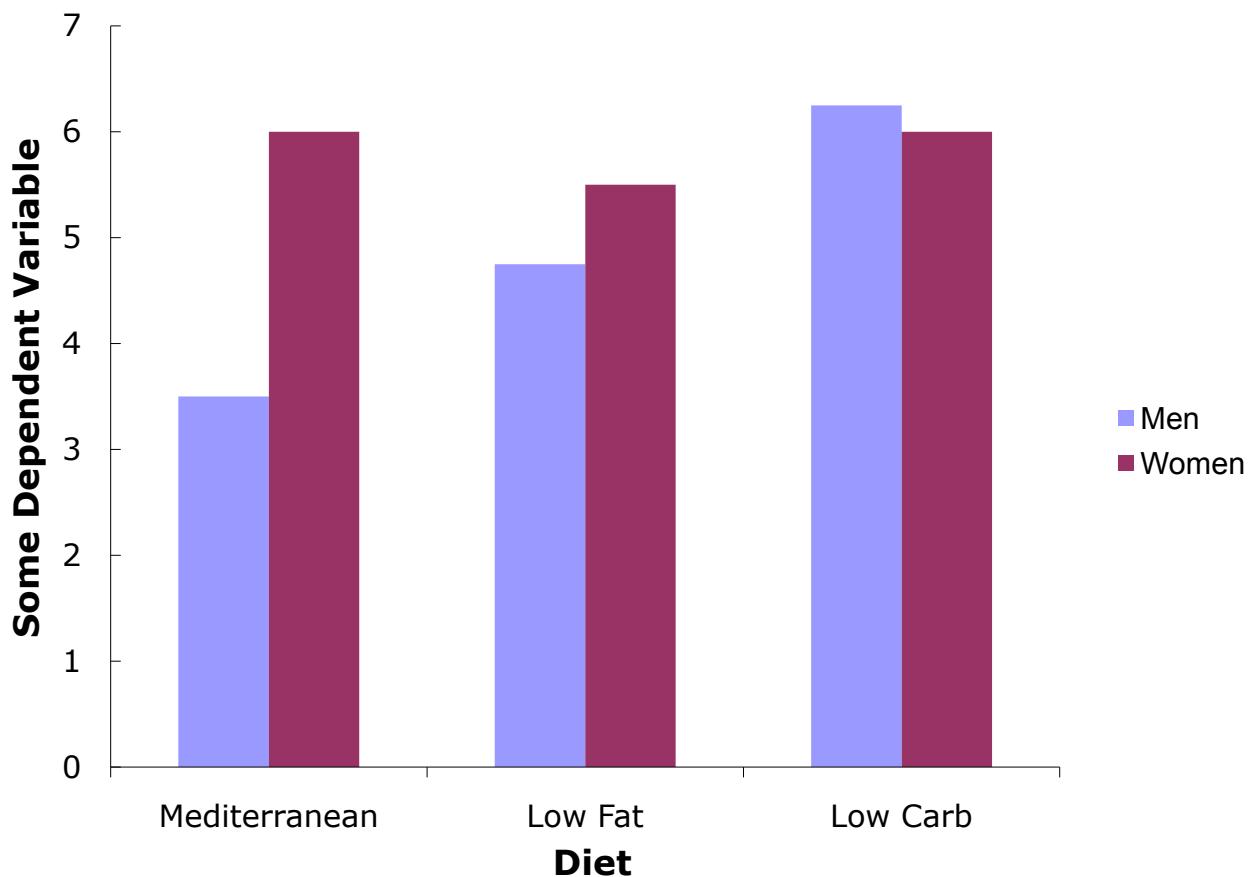


Figure 2. Plot with a qualitative variable on the X-axis.

Figure 3 shows such a plot. Notice how it contains information about the medians, quantiles, and minimums and maximums not contained in Figure 2. Most important, you get an idea about how much the distributions overlap from Figure 3 which you do not get from Figure 2.

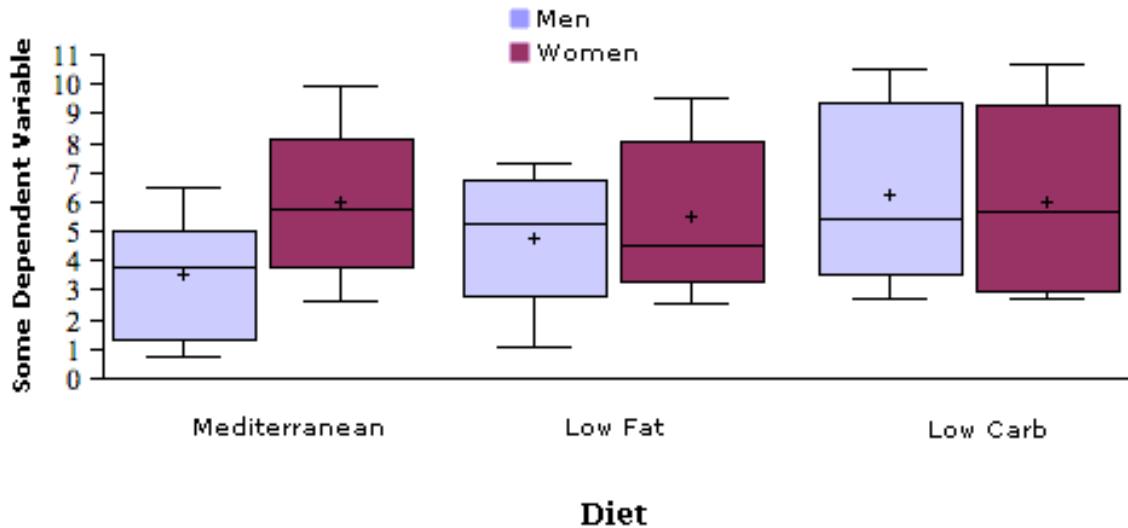


Figure 3. Box plots.

Line graphs are a good option when there are more than two levels of a numeric variable. Figure 4 shows an example. A line graph has the advantage of showing the pattern of interaction clearly. Its disadvantage is that it does not convey the distributional information contained in box plots.

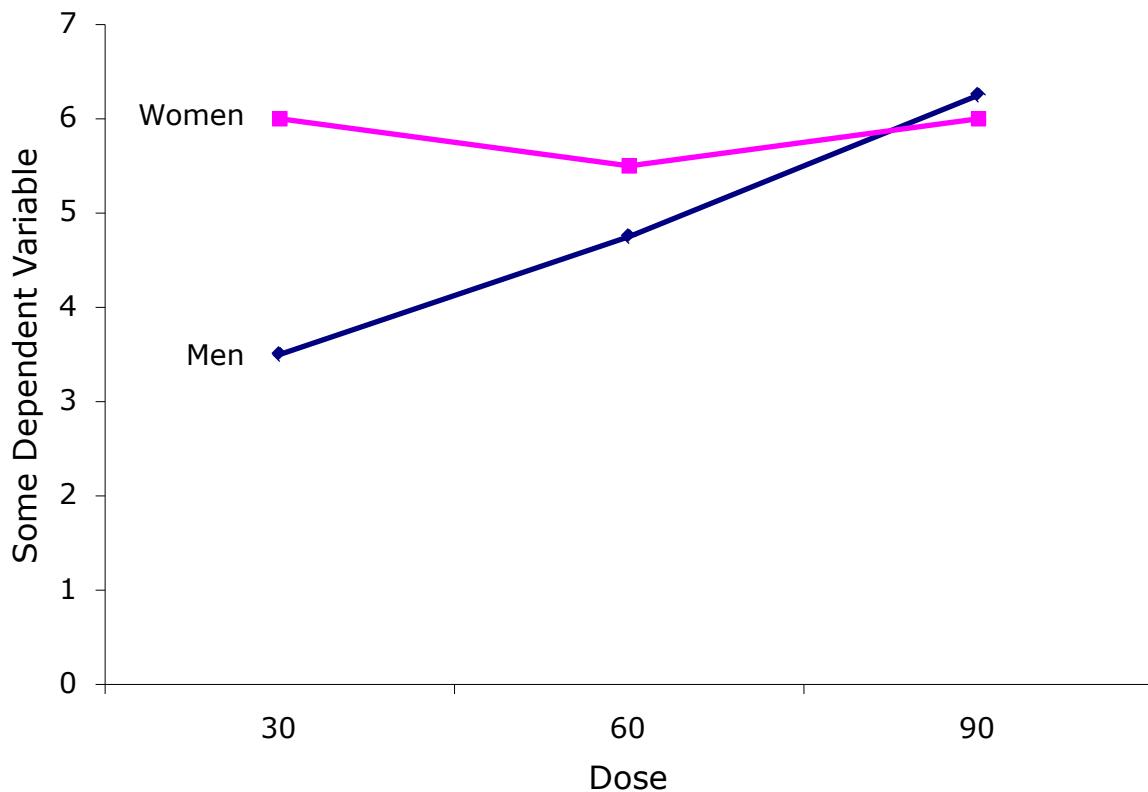


Figure 4. Plot with a quantitative variable on the X-axis.

## An Example with Interaction

The following example was presented in the section on specific comparisons among means. It is also relevant here.

This example uses the made-up data from a hypothetical experiment shown in Table 4. Twelve subjects were selected from a population of high-self-esteem subjects and an additional 12 subjects were selected from a population of low-self-esteem subjects. Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they succeeded and the other half were told they failed. Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects in all.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table 4. Data from Hypothetical Experiment on Attribution

		Esteem	
		High	Low
Outcome	Success	7	6
		8	5
		7	7
		8	4
		9	5
		5	6
	Failure	4	9
		6	8
		5	9
		4	8
		7	7
		3	6

The ANOVA Summary Table for these data is shown in Table 5.

Table 5. ANOVA Summary Table for Made-Up Data

Source	df	SSQ	MS	F	p
Outcome	1	0.0417	0.0417	0.0256	0.8744
Esteem	1	2.0417	2.0417	1.2564	0.2756
O x E	1	35.0417	35.0417	21.5641	0.0002
Error	20	32.5	1.625		
Total	23	69.625			

As you can see, the only significant effect is the Outcome x Esteem (O x E) interaction. The form of the interaction can be seen in Figure 5.

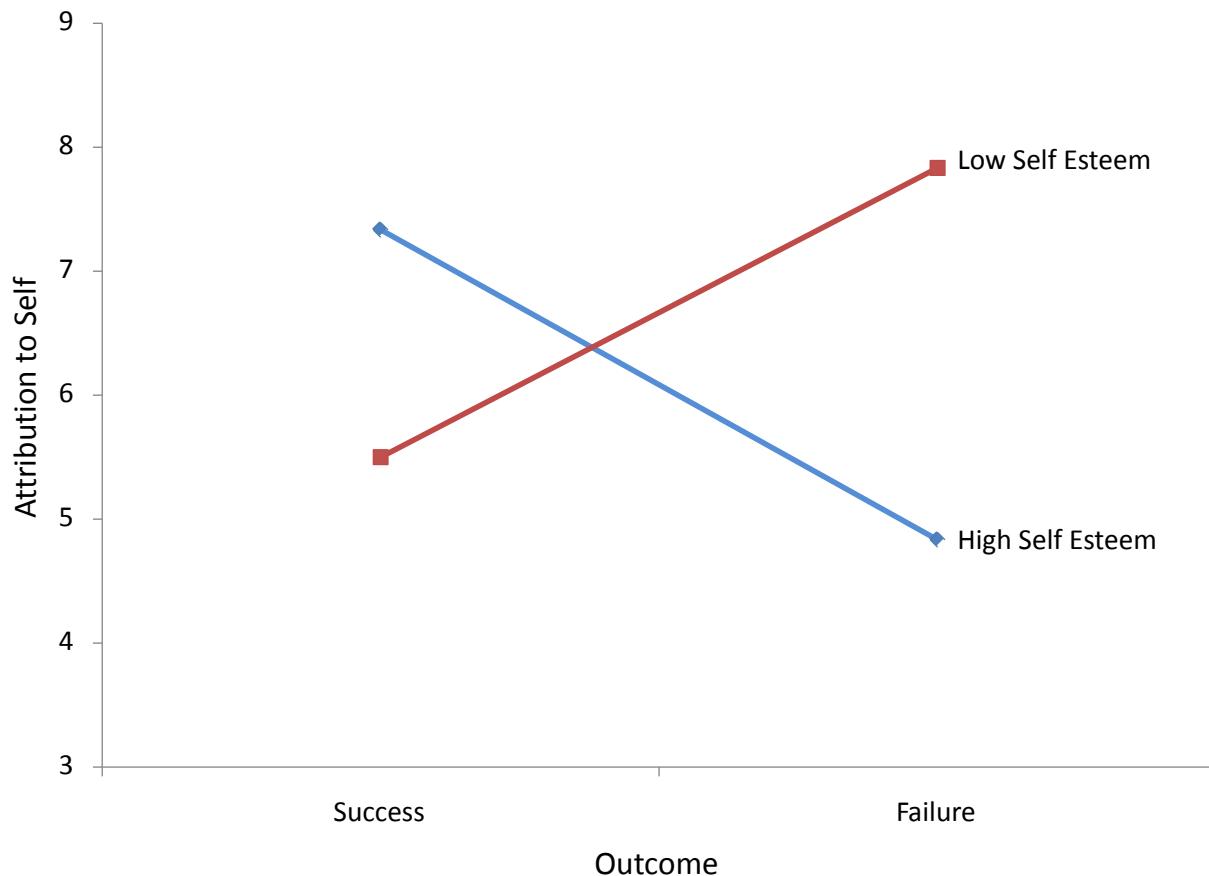


Figure 5. Interaction plot for made-up data.

Clearly the effect of “Outcome” is different for the two levels of “Esteem”: For subjects high in self-esteem, failure led to less attribution to oneself than did success. By contrast, for subjects low in self-esteem, failure led to more attribution to oneself than did success. Notice that the two lines in the graph are not parallel.

**Nonparallel lines indicate interaction. The significance test for the interaction determines whether it is justified to conclude that the lines in the population are not parallel.** Lines do not have to cross for there to be an interaction.

### Three-Factor Designs

Three-factor designs are analyzed in much the same way as two-factor designs. Table 6 shows the analysis of a study described by Franklin and Cooley (2002) investigating three factors on the strength of industrial fans: (1) Hole Shape (Hex or Round), (2) Assembly Method (Staked or Spun), and (3) Barrel Surface (Knurled or Smooth). The dependent variable, Breaking Torque, was measured in foot-pounds. There were eight observations in each of the eight combinations of the three factors.

As you can see in Table 6, there are three main effects, three two-way interactions, and one three-way interaction. The degrees of freedom for the main effects are, as in a two-factor design, equal to the number of levels of the factor minus one. Since all the factors here have two levels, all the main effects have one degree of freedom. The interaction degrees of freedom is always equal to the product of the degrees of freedom of the component parts. This holds for the three-factor interaction as well as for the two-factor interactions. The error degrees of freedom is equal to the number of observations (64) minus the number of groups (8) and equals 56.

Table 6. ANOVA Summary Table for Fan Data

Source	df	SSQ	MS	F	p
Hole	1	8258.27	8258.27	266.68	<0.0001
Assembly	1	13369.14	13369.14	431.73	<0.0001
H x A	1	2848.89	2848.89	92	<0.0001
Barrel	1	35.0417	35.0417	21.5641	<0.0001
H x B	1	594.14	594.14	19.1865	<0.0001
A x B	1	135.14	135.14	4.36	0.0413

H x A x B	1	1396.89	1396.89	45.11	<0.0001
Error	56	1734.12	30.97		
Total	63	221386.91			

A three-way interaction means that the two-way interactions differ as a function of the level of the third variable. The usual way to portray a three-way interaction is to plot the two-way interactions separately. Figure 6 shows the Barrel (Knurled or Smooth) x Assembly (Staked or Spun) separately for the two levels of Hole Shape (Hex or Round). For the Hex Shape, there is very little interaction with the lines being close to parallel with a very slight tendency for the effect of Barrel to be bigger for Staked than for Spun. The two-way interaction for the Round Shape is different: The effect of Barrel is bigger for Spun than for Staked. The finding of a significant three-way interaction indicates that this difference in two-way interactions is significant.

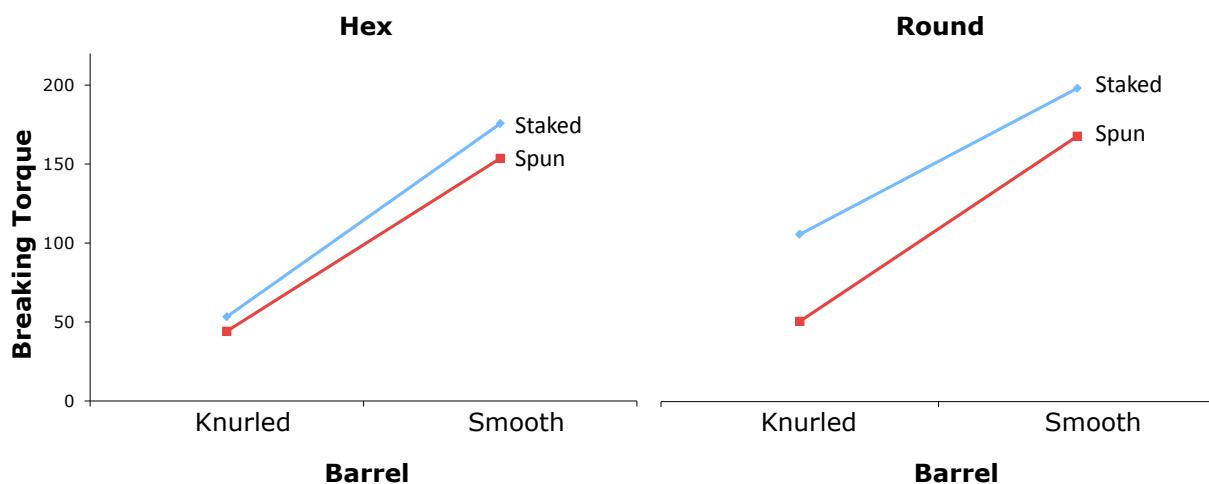


Figure 6. Plot of the three-way interaction.

## Formatting Data for Computer Analysis

The data in Table 4 have been reformatted in Table 7. Note how there is one column to indicate the level of outcome and one column to indicate the level of esteem. The coding is as follows:

High-self-esteem: 1

Low self-esteem: 2

Success: 1

Failure: 2

Table 7. Attribution Data Reformatted

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7
2	2	6

# Unequal Sample Sizes

by David M. Lane

## *Prerequisites*

- Chapter 15: ANOVA Designs
- Chapter 15: Multi-Factor Designs

## *Learning Objectives*

1. State why unequal  $n$  can be a problem
2. Define confounding
3. Compute weighted and unweighted means
4. Distinguish between Type I and Type III sums of squares
5. Describe why the cause of the unequal sample sizes makes a difference in the interpretation

## **The Problem of Confounding**

Whether by design, accident, or necessity, the number of subjects in each of the conditions in an experiment may not be equal. For example, the sample sizes for the “Bias Against Associates of the Obese” case study are shown in Table 1.

Although the sample sizes were approximately equal, the “Acquaintance Typical” condition had the most subjects. Since  $n$  is used to refer to the sample size of an individual group, designs with unequal sample sizes are sometimes referred to as designs with unequal  $n$ .

Table 1. Sample Sizes for “Bias Against Associates of the Obese” Study.

		Companion Weight	
		Obese	Typical
Relationship	Girl Friend	40	42
	Acquaintance	40	54

We consider an absurd design to illustrate the main problem caused by unequal  $n$ . Suppose an experimenter were interested in the effects of diet and exercise on cholesterol. The sample sizes are shown in Table 2.

Table 2. Sample Sizes for “Diet and Exercise” Example.

		Exercise	
		Moderate	None
Diet	Low Fat	5	0
	High Fat	0	5

What makes this example absurd is that there are no subjects in either the “Low-Fat No-Exercise” condition or the “High-Fat Moderate-Exercise” condition. The hypothetical data showing change in cholesterol are shown in Table 3.

Table 3. Data for “Diet and Exercise” Example.

		Exercise		
		Moderate	None	Mean
Diet	Low Fat	-20		-25
		-25		
		-30		
		-35		
		-15		
	High Fat		-20 6 -10 -6 5	-5
	Mean	-25	-5	-15

The last column shows the mean change in cholesterol for the two diet conditions, whereas the last row shows the mean change in cholesterol for the two Exercise conditions. The value of -15 in the lower-right-most cell in the table is the mean of all subjects.

We see from the last column that those on the low-fat diet lowered their cholesterol an average of 25 units, whereas those on the high-fat diet lowered theirs by only an average of 5 units. However, there is no way of knowing whether the difference is due to diet or to exercise since every subject in the low-fat

condition was in the moderate-exercise condition and every subject in the high-fat condition was in the no-exercise condition. Therefore, Diet and Exercise are completely *confounded*. The problem with unequal n is that it causes confounding.

## Weighted and Unweighted Means

The difference between *weighted* and *unweighted means* is a difference critical for understanding how to deal with the confounding resulting from unequal n.

Weighted and unweighted means will be explained using the data shown in Table 4. Here, Diet and Exercise are confounded because 80% of the subjects in the low-fat condition exercised as compared to 20% of those in the high-fat condition. However, there is not complete confounding as there was with the data in Table 3.

The weighted mean for “Low Fat” is computed as the mean of the “Low-Fat Moderate-Exercise” mean and the “Low-Fat No-Exercise” mean, weighted in accordance with sample size. To compute a weighted mean, you multiply each mean by its sample size and divide by N, the total number of observations. Since there are four subjects in the “Low-Fat Moderate-Exercise” condition and one subject in the “Low-Fat No-Exercise” condition, the means are weighted by factors of 4 and 1 as shown below, where  $M_w$  is the weighted mean.

$$M_w = \frac{(4)(-27.5) + (1)(-20)}{5} = -26$$

The weighted mean for the low-fat condition is also the mean of all five scores in this condition. Thus if you ignore the factor “Exercise,” you are implicitly computing weighted means.

The unweighted mean for the low-fat condition ( $M_u$ ) is simply the mean of the two means.

$$M_u = \frac{-27.5 - 20}{2} = -23.75$$

Table 4. Data for Diet and Exercise with Partial Confounding Example

		Exercise			
		Moderate	None	Weighted Mean	Unweighted Mean
Diet	Low Fat	-20	-20	-26	-23.75
		-25			
		-30			
		-35			
		M=-27.5	M=-20.0		
Diet	High Fat	-15	6	-4	-8.125
			-6		
			5		
			-10		
		M=-15.0	M=-1.25		
	Weighted Mean	-25	-5		
	Unweighted Mean	-21.25	-10.625		

One way to evaluate the *main effect* of Diet is to compare the weighted mean for the low-fat diet (-26) with the weighted mean for the high-fat diet (-4). This difference of -22 is called “the effect of diet ignoring exercise” and is misleading since most of the low-fat subjects exercised and most of the high-fat subjects did not. However, the difference between the unweighted means of -15.625 (-23.75 minus -8.125) is not affected by this confounding and is therefore a better measure of the main effect. In short, weighted means ignore the effects of other variables (exercise in this example) and result in confounding; unweighted means control for the effect of other variables and therefore eliminate the confounding.

Statistical analysis programs use different terms for means that are computed controlling for other effects. SPSS calls them *estimated marginal means*, whereas SAS and SAS JMP call them *least squares means*.

## Types of Sums of Squares

When there is unequal  $n$ , the sum of squares total is not equal to the sum of the sums of squares for all the other sources of variation. This is because the confounded sums of squares are not apportioned to any source of variation. For the data in Table 5, the sum of squares for Diet is 390.625, the sum of squares for Exercise is 180.625, and the sum of squares confounded between these two factors is 819.375 (the calculation of this value is beyond the scope of this introductory text). In the ANOVA Summary Table shown in Table 5, this large portion of the sums of squares is not apportioned to any source of variation and represents the “missing” sums of squares. That is, if you add up the sums of squares for Diet, Exercise,  $D \times E$ , and Error, you get 902.625. If you add the confounded sum of squares of 819.375 to this value, you get the total sum of squares of 1722.000. When confounded sums of squares are not apportioned to any source of variation, the sums of squares are called *Type III sums of squares*. Type III sums of squares are, by far, the most common and if sums of squares are not otherwise labeled, it can safely be assumed that they are Type III.

Table 5. ANOVA Summary Table for Type III SSQ

Source	df	SSQ	MS	F	p
Diet	1	390.625	390.625	7.42	0.034
Exercise	1	180.625	180.625	3.43	0.113
$D \times E$	1	15.625	15.625	0.3	0.605
Error	6	315.75	52.625		
Total	9	1722			

When all confounded sums of squares are apportioned to sources of variation, the sums of squares are called *Type I sums of squares*. The order in which the confounded sums of squares are apportioned is determined by the order in which the effects are listed. The first effect gets any sums of squares confounded between it and any of the other effects. The second gets the sums of squares confounded between it and subsequent effects, but not confounded with the first effect, etc. The Type I sums of squares are shown in Table 6. As you can see, with Type I sums of squares, the sum of all sums of squares is the total sum of squares.

Table 6. ANOVA Summary Table for Type I SSQ

Source	df	SSQ	MS	F	p
Diet	1	1210	1210	22.99	0.003
Exercise	1	180.625	180.625	3.43	0.113
D x E	1	15.625	15.625	0.3	0.605
Error	6	315.75	52.625		
Total	9	1722			

In *Type II sums of squares*, sums of squares confounded between main effects are not apportioned to any source of variation, whereas sums of squares confounded between main effects and interactions are apportioned to the main effects. In our example, there is no confounding between the D x E interaction and either of the main effects. Therefore, the Type II sums of squares are equal to the Type III sums of squares.

#### *Which Type of Sums of Squares to Use (optional)*

Type I sums of squares allow the variance confounded between two main effects to be apportioned to one of the main effects. Unless there is a strong argument for how the confounded variance should be apportioned (which is rarely, if ever, the case), Type I sums of squares are not recommended.

There is not a consensus about whether Type II or Type III sums of squares is to be preferred. On the one hand, if there is no interaction, then Type II sums of squares will be more powerful for two reasons: (1) variance confounded between the main effect and interaction is properly assigned to the main effect and (2) weighting the means by sample sizes gives better estimates of the effects. To take advantage of the greater power of Type II sums of squares, some have suggested that if the interaction is not significant, then Type II sums of squares should be used. Maxwell and Delaney (2003) caution that such an approach could result in a Type II error in the test of the interaction. That is, it could lead to the conclusion that there is no interaction in the population when there really is one. This, in turn, would increase the Type I error rate for the test of the main effect. As a result, their general recommendation is to use Type III sums of squares.

Maxwell and Delaney (2003) recognized that some researchers prefer Type II sums of squares when there are strong theoretical reasons to suspect a lack of

interaction and the  $p$  value is much higher than the typical  $\alpha$  level of 0.05. However, this argument for the use of Type II sums of squares is not entirely convincing. As Tukey (1991) and others have argued, it is doubtful that any effect, whether a main effect or an interaction, is exactly 0 in the population. Incidentally, Tukey argued that the role of significance testing is to determine whether a confident conclusion can be made about the direction of an effect, not simply to conclude that an effect is not exactly 0.

Finally, if one assumes that there is no interaction, then an ANOVA model with no interaction term should be used rather than Type II sums of squares in a model that includes an interaction term. (Models without interaction terms are not covered in this book).

There are situations in which Type II sums of squares are justified even if there is strong interaction. This is the case because the hypotheses tested by Type II and Type III sums of squares are different, and the choice of which to use should be guided by which hypothesis is of interest. Recall that Type II sums of squares weight cells based on their sample sizes whereas Type III sums of squares weight all cells the same. Consider Figure 1 which shows data from a hypothetical A(2) x B(2) design. The sample sizes are shown numerically and are represented graphically by the areas of the endpoints.

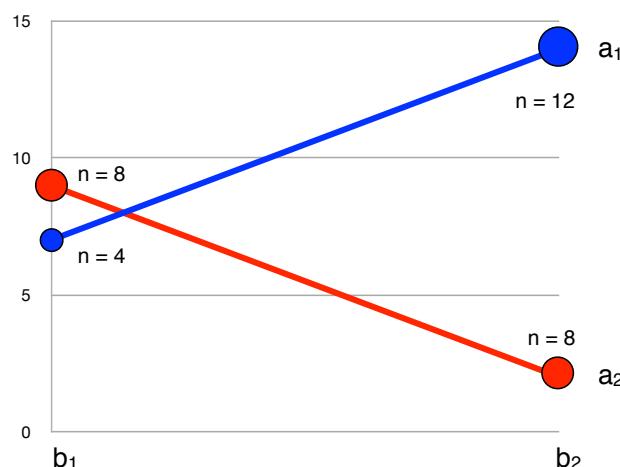


Figure 1. An interaction plot with unequal sample sizes.

First, let's consider the hypothesis for the main effect of B tested by the Type III sums of squares. Type III sums of squares weight the means equally and, for these data, the marginal means for  $b_1$  and  $b_2$  are equal:

$$(b_1a_1 + b_1a_2)/2 = (7 + 9)/2 = 8.$$

$$(b_2a_1 + b_2a_2)/2 = (14 + 2)/2 = 8.$$

Thus, there is no main effect of B when tested using Type III sums of squares.

For Type II sums of squares, the means are weighted by sample size. For  $b_1$ :

$$(4 \times b_1a_1 + 8 \times b_1a_2)/12 =$$

$$(4 \times 7 + 8 \times 9)/12 = 8.33$$

For  $b_2$ :

$$(12 \times b_2a_1 + 8 \times b_2a_2)/20 =$$

$$(12 \times 14 + 8 \times 2)/20 = 9.2.$$

Since the weighted marginal mean for  $b_2$  is larger than the weighted marginal mean for  $b_1$ , there is a main effect of B when tested using Type II sums of squares.

The Type II and Type III analyses are testing different hypotheses. First, let's consider the case in which the differences in sample sizes arise because in the sampling of intact groups, the sample cell sizes reflect the population cell sizes (at least approximately). In this case, it makes sense to weight some means more than others and conclude that there is a main effect of B. This is the result obtained with Type II sums of squares. However, if the sample size differences arose from random assignment, and there just happened to be more observations in some cells than others, then one would want to estimate what the main effects would have been with equal sample sizes and, therefore, weight the means equally. With the means weighted equally, there is no main effect of B, the result obtained with Type III sums of squares.

## Causes of Unequal Sample Sizes

None of the methods for dealing with unequal sample sizes are valid if the experimental treatment is the source of the unequal sample sizes. Imagine an experiment seeking to determine whether publicly performing an embarrassing act would affect one's anxiety about public speaking. In this imaginary experiment, the experimental group is asked to reveal to a group of people the most embarrassing

thing they have ever done. The control group is asked to describe what they had at their last meal. Twenty subjects are recruited for the experiment and randomly divided into two equal groups of 10, one for the experimental treatment and one for the control. Following their descriptions, subjects are given an attitude survey concerning public speaking. This seems like a valid experimental design. However, of the 10 subjects in the experimental group, four withdrew from the experiment because they did not wish to publicly describe an embarrassing situation. None of the subjects in the control group withdrew. Even if the data analysis were to show a significant effect, it would not be valid to conclude that the treatment had an effect because a likely alternative explanation cannot be ruled out; namely, subjects who were willing to describe an embarrassing situation differed from those who were not. Thus, the differential dropout rate destroyed the *random assignment* of subjects to conditions, a critical feature of the experimental design. No amount of statistical adjustment can compensate for this flaw.

## References

- Maxwell, S. E., & Delaney, H. D. (2003) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Second Edition, Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Tukey, J. W. (1991) The philosophy of multiple comparisons, *Statistical Science*, 6, 110-116.

# Tests Supplementing ANOVA

by David M. Lane

## *Prerequisites*

- Chapter 15: One-Factor ANOVA, Multi-Factor ANOVA
- Chapter 15: Pairwise Comparisons Among Means
- Chapter 15: Specific Comparisons Among Means

## *Learning Objectives*

1. Compute Tukey HSD test
2. Describe an interaction in words
3. Describe why one might want to compute simple effect tests following a significant interaction

The *null hypothesis* tested in a one-factor ANOVA is that all the population means are equal. Stated more formally,

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

where  $H_0$  is the null hypothesis and  $k$  is the number of conditions. When the null hypothesis is rejected, all that can be said is that at least one population mean is different from at least one other population mean. The methods for doing more specific tests described in "All Pairwise Comparisons among Means" and in "Specific Comparisons" apply here. Keep in mind that these tests are valid whether or not they are preceded by an ANOVA.

## **Main Effects**

As will be seen, significant *main effects* in multi-factor designs can be followed up in the same way as significant effects in one-way designs. Table 1 shows the data from an imaginary experiment with three *levels* of Factor A and two levels of Factor B.

Table 1. Made-Up Example Data.

	A1	A2	A3	Marginal Means
B1	5	9	5	7.08
	4	8	9	
	6	7	9	
	5	8	8	
	Mean = 5	Mean = 8	Mean = 8.25	
B2	4	8	8	6.5
	3	6	9	
	6	8	7	
	8	5	6	
	Mean = 5.25	Mean = 6.75	Mean = 7.50	
Marginal Means	5.125	7.375	7.875	6.79

Table 2 shows the ANOVA Summary Table for these data. The *significant* main effect of A indicates that, in the population, at least one of the *marginal means* for A is different from at least one of the others.

Table 2. ANOVA Summary Table for Made-Up Example Data.

Source	df	SSQ	MS	F	p
A	2	34.333	17.17	9.29	0.002
B	1	2.042	2.04	1.1	0.307
A x B	2	2.333	1.167	0.63	0.543
Error	18	33.25	1.847		
Total	23	71.958			

The Tukey HSD test can be used to test all pairwise comparisons among means in a one-factor ANOVA as well as comparisons among marginal means in a multi-factor ANOVA. The formula for the equal-sample-size case is shown below.

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

where  $M_i$  and  $M_j$  are marginal means, MSE is the mean square error from the ANOVA, and  $n$  is the number of scores each mean is based upon. For this example,  $MSE = 1.847$  and  $n = 8$  because there are eight scores at each level of A. The probability value can be computed using the Studentized Range Calculator. The degrees of freedom is equal to the degrees of freedom error. For this example,  $df = 18$ . The results of the Tukey HSD test are shown in Table 3. The mean for  $A_1$  is significantly lower than the mean for  $A_2$  and the mean for  $A_3$ . The means for  $A_2$  and  $A_3$  are not significantly different.

Table 3. Pairwise Comparisons Among Marginal Means for A.

Comparison	$M_i - M_j$	Q	p
$A_1 - A_2$	-2.25	-4.68	0.01
$A_1 - A_3$	-2.75	-5.73	0.002
$A_2 - A_3$	-0.5	-1.04	0.746

Specific comparisons among means are also carried out much the same way as shown in the relevant section on testing means. The formula for L is

$$L = \sum c_i M_i$$

where  $c_i$  is the coefficient for the  $i^{\text{th}}$  marginal mean and  $M_i$  is the  $i^{\text{th}}$  marginal mean. For example, to compare  $A_1$  with the average of  $A_2$  and  $A_3$ , the coefficients would be 1, -0.5, -0.5. Therefore,

$$\begin{aligned} L &= (1)(5.125) + (-0.5)(7.375) + (-0.5)(7.875) \\ &= -2.5. \end{aligned}$$

To compute t, use:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}}$$

$$= -4.25$$

where MSE is the mean square error from the ANOVA and n is the number of scores each marginal mean is based on (eight in this example). The degrees of freedom is the degrees of freedom error from the ANOVA and is equal to 18. Using the Online Calculator, we find that the two-tailed probability value is 0.0005. Therefore, the difference between A<sub>1</sub> and the average of A<sub>2</sub> and A<sub>3</sub> is significant.

Important issues concerning multiple comparisons and *orthogonal comparisons* are discussed in the Specific Comparisons section in the Testing Means chapter.

## Interactions

The presence of a significant interaction makes the interpretation of the results more complicated. Since an interaction means that the *simple effects* are different, the main effect as the mean of the simple effects does not tell the whole story. This section discusses how to describe interactions, proper and improper uses of simple effects tests, and how to test components of interactions.

## Describing Interactions

A crucial first step in understanding a significant interaction is constructing an *interaction plot*. Figure 1 shows an interaction plot from data presented in the section on Multi-Factor ANOVA.

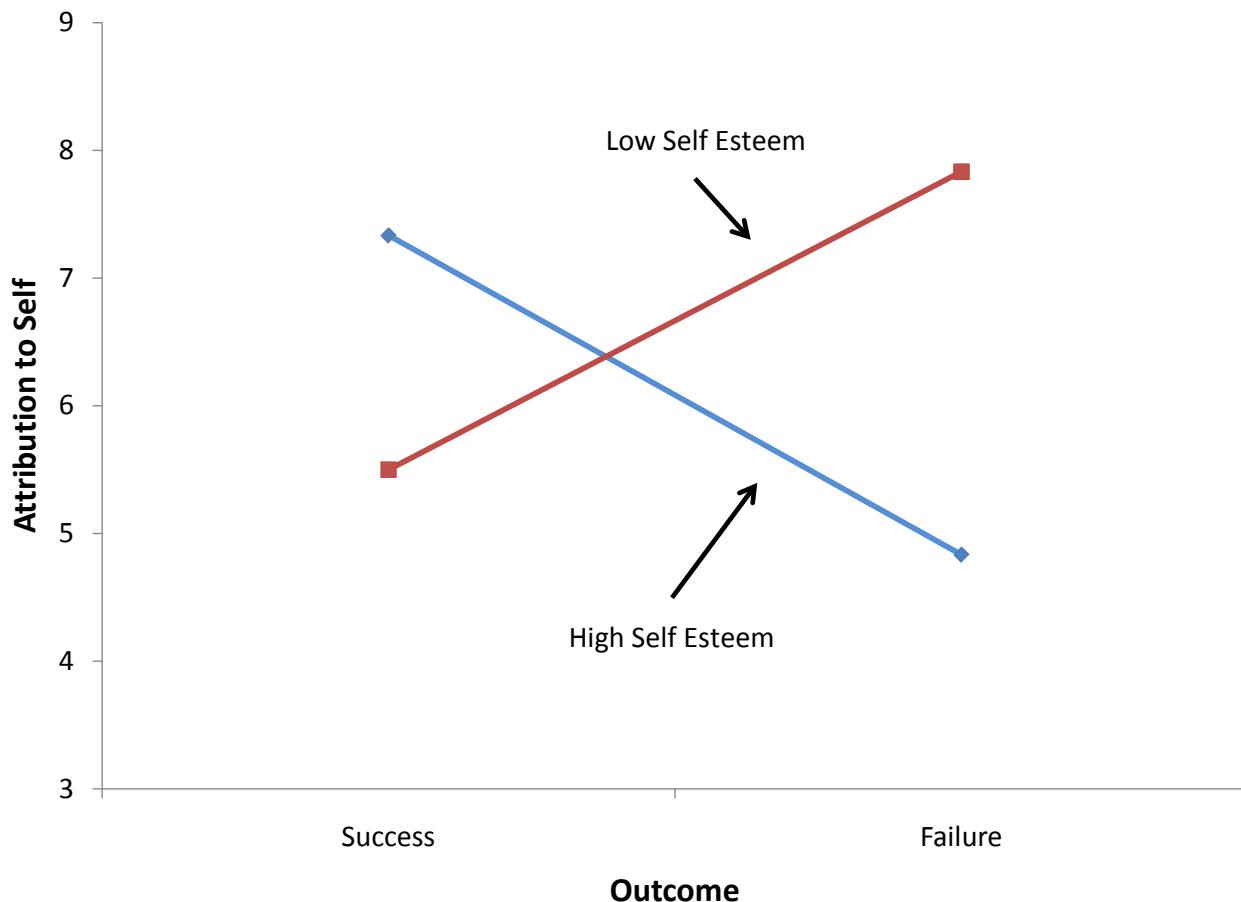


Figure 1. Interaction plot for made-up data.

The second step is to describe the interaction in a clear and understandable way. This is often done by describing how the *simple effects* differed. Since this should be done using as little jargon as possible, the expression “simple effect” need not appear in the description. An example is as follows:

The effect of Outcome differed depending on the subject's self-esteem. The difference between the attribution to self following success and the attribution to self following failure was larger for high-self-esteem subjects (mean difference = 2.50) than for low-self-esteem subjects (mean difference = -2.33).

No further analyses are helpful in understanding the interaction since the interaction means only that the simple effects differ. The interaction's significance indicates that the simple effects differ from each other, but provides no information about whether they differ from zero.

## Simple Effect Tests

It is not necessary to know whether the simple effects differ from zero in order to understand an interaction because the question of whether simple effects differ from zero has nothing to do with interaction except that if they are both zero there is no interaction. It is not uncommon to see research articles in which the authors report that they analyzed simple effects in order to explain the interaction. However, this is not a valid approach since an interaction does not depend on the analysis of the simple effects.

However, there is a reason to test simple effects following a significant interaction. Since an interaction indicates that simple effects differ, it means that the main effects are not general. In the made-up example, the main effect of Outcome is not very informative, and the effect of outcome should be considered separately for high- and low-self-esteem subjects.

As will be seen, the simple effects of Outcome are significant and in opposite directions: Success significantly increases attribution to self for high-self-esteem subjects and significantly lowers attribution to self for low-self-esteem subjects. This is a very easy result to interpret.

What would the interpretation have been if neither simple effect had been significant? On the surface, this seems impossible: How can the simple effects both be zero if they differ from each other significantly as tested by the interaction? The answer is that a non-significant simple effect does not mean that the simple effect is zero: the null hypothesis should not be accepted just because it is not rejected.

(See section on Interpreting Non-Significant Results)

If neither simple effect is significant, the conclusion should be that the simple effects differ, and that at least one of them is not zero. However, no conclusion should be drawn about which simple effect(s) is/are not zero.

Another error that can be made by mistakenly accepting the null hypothesis is to conclude that two simple effects are different because one is significant and the other is not. Consider the results of an imaginary experiment in which the researcher hypothesized that addicted people would show a larger increase in brain activity following some treatment than would non-addicted people. In other words, the researcher hypothesized that addiction status and treatment would interact. The results shown in Figure 2 are very much in line with the hypothesis. However, the test of the interaction resulted in a *probability value* of 0.08, a value not quite low enough to be significant at the conventional 0.05 level. The proper conclusion is

that the experiment supports the researcher's hypothesis, but not strongly enough to allow a confident conclusion.

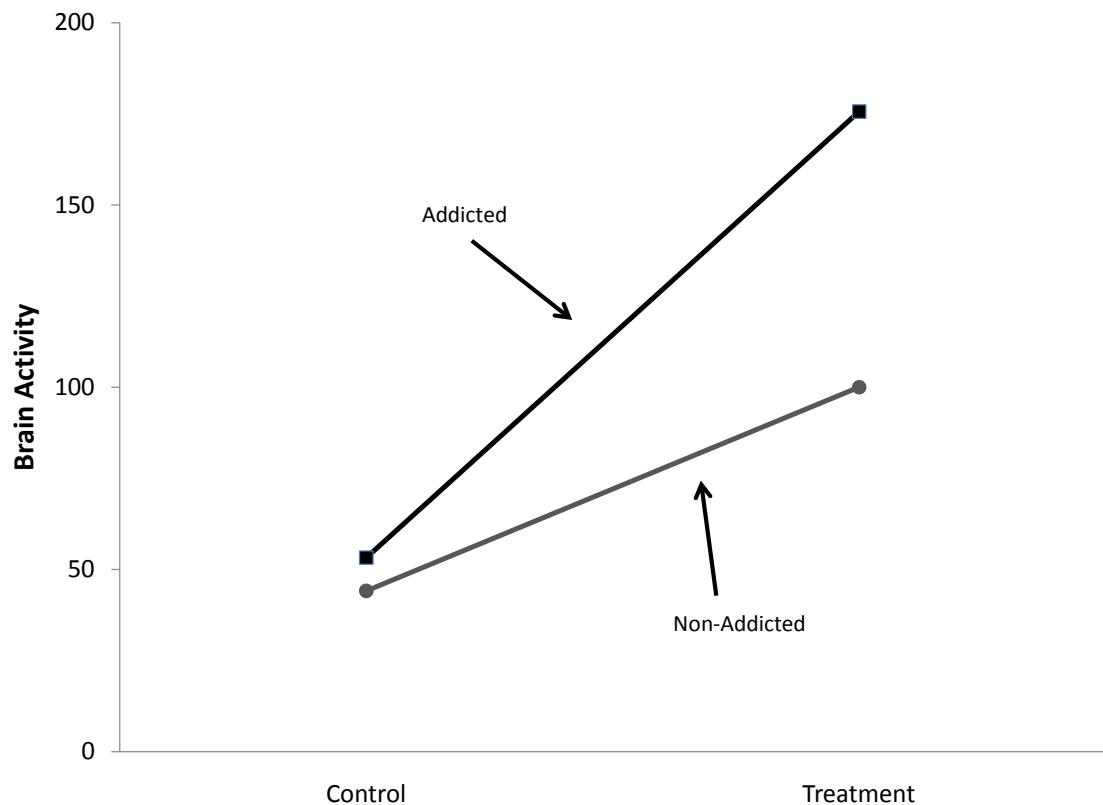


Figure 2. Made-up data with one significant simple effect.

Unfortunately, the researcher was not satisfied with such a weak conclusion and went on to test the simple effects. It turned out that the effect of Treatment was significant for the Addicted group ( $p = 0.02$ ) but not significant for the Non-Addicted group ( $p = 0.09$ ). The researcher then went on to conclude that since there is an effect of Treatment for the Addicted group but not for the Non-Addicted group, the hypothesis of a greater effect for the former than for the latter group is demonstrated. This is faulty logic, however, since it is based on accepting the null hypothesis that the simple effect of Treatment is zero for the Non-Addicted group just because it is not significant.

### **Components of Interaction (optional)**

Figure 3 shows the results of an imaginary experiment on diet and weight loss. A control group and two diets were used for both overweight teens and overweight adults.

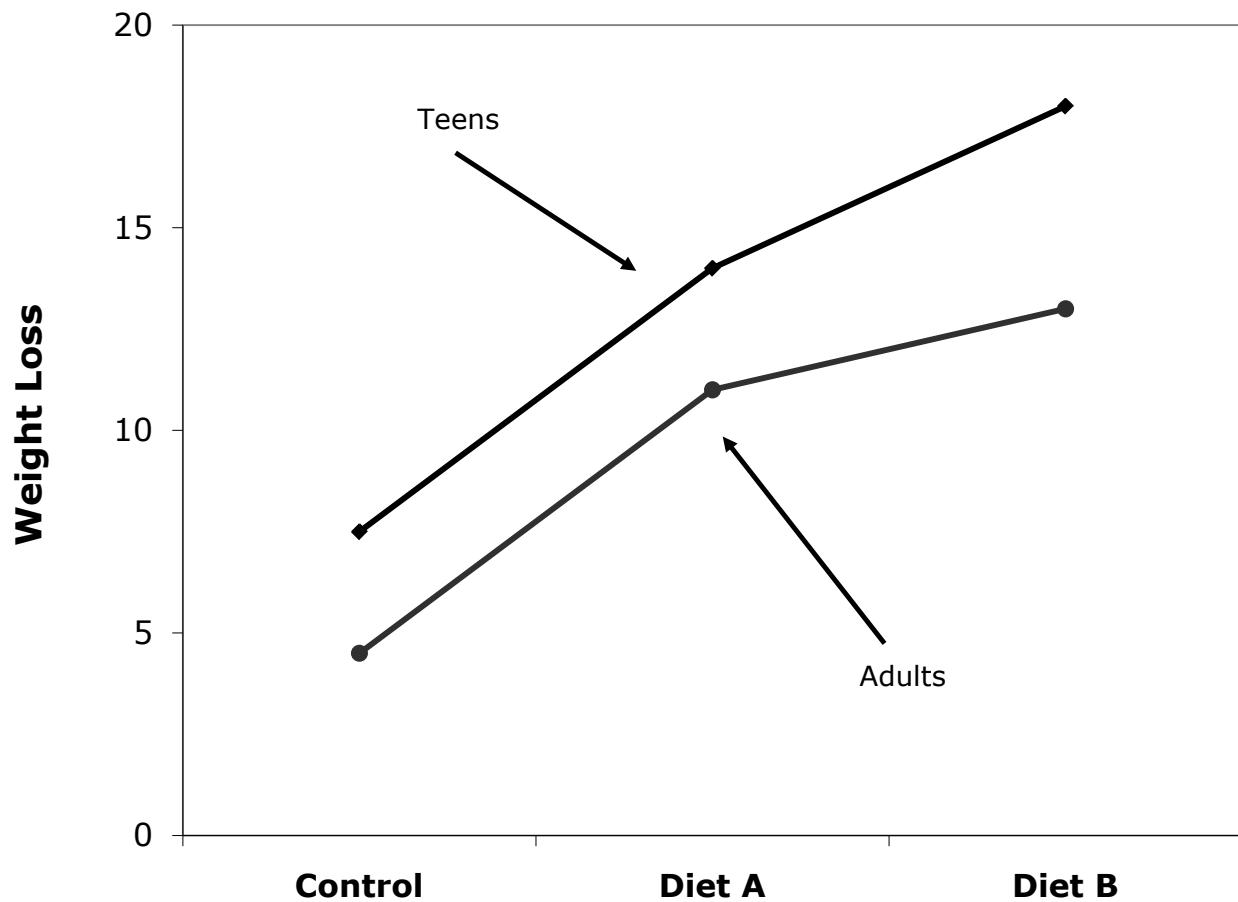


Figure 3. Made-up Data for Diet Study.

The difference between Diet A and the Control diet was essentially the same for teens and adults, whereas the difference between Diet B and Diet A was much larger for the teens than it was for the adults. Over one portion of the graph the lines are parallel, whereas over another portion they are not. It is possible to test these portions or components of interactions using the method of specific comparisons discussed previously. The test of the difference between Teens and Adults on the difference between Diets A and B could be tested with the coefficients shown in Table 4. Naturally, the same consideration regarding multiple comparisons and *orthogonal comparisons* that apply to other comparisons among means also apply to comparisons involving components of interactions.

Table 4. Coefficients for a Component of the Interaction.

Age Group	Diet	Coefficient
Teen	Control	0
Teen	A	1
Teen	B	-1
Adult	Control	0
Adult	A	-1
Adult	B	1

# Within-Subjects ANOVA

by David M. Lane

## *Prerequisites*

- Chapter 12: Difference Between Two Means (Correlated Pairs)
- Chapter 15: Additional Measures of Central Tendency
- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs, Multi-Factor ANOVA

## *Learning Objectives*

1. Define a within-subjects factor
2. Explain why a within-subjects design can be expected to have more power than a between-subjects design
3. Be able to create the Source and df columns of an ANOVA summary table for a one-way within-subjects design
4. Explain error in terms of interaction
5. Discuss the problem of carryover effects
6. Be able to create the Source and df columns of an ANOVA summary table for a design with one between-subjects and one within-subjects variable
7. Define sphericity
8. Describe the consequences of violating the assumption of sphericity
9. Discuss courses of action that can be taken if sphericity is violated

*Within-subjects factors* involve comparisons of the same subjects under different conditions. For example, in the “ADHD Treatment” study, each child's performance was measured four times, once after being on each of four drug doses for a week. Therefore, each subject's performance was measured at each of the four *levels* of the *factor* “Dose.” Note the difference from *between-subjects factors* for which each subject's performance is measured only once and the comparisons are among different groups of subjects. A within-subjects factor is sometimes referred to as a *repeated-measures factor* since repeated measurements are taken on each subject. An experimental design in which the independent variable is a *within-subjects factor* is called a within-subjects design.

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ from one another. In an experiment on problem

solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more *power* than between-subjects designs.

## One-Factor Designs

Let's consider how to analyze the data from the “ADHD Treatment” case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. For now, we will be concerned only with testing the difference between the mean in the placebo condition (the lowest dosage, D0) and the mean in the highest dosage condition (D60). The details of the computations are relatively unimportant since they are almost universally done by computers. Therefore we jump right to the ANOVA Summary table shown in Table 1.

Table 1. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	5781.98	251.39		
Dosage	1	295.02	295.02	10.38	0.004
Error	23	653.48	28.41		
Total	47	6730.48			

The first source of variation, “Subjects,” refers to the differences among subjects. If all the subjects had exactly the same mean (across the two dosages), then the sum of squares for subjects would be zero; the more subjects differ from each other, the larger the sum of squares subjects.

Dosage refers to the differences between the two dosage levels. If the means for the two dosage levels were equal, the sum of squares would be zero. The larger the difference between means, the larger the sum of squares.

The error reflects the degree to which the effect of dosage is different for different subjects. If subjects all responded very similarly to the drug, then the error would be very low. For example, if all subjects performed moderately better

with the high dose than they did with the placebo, then the error would be low. On the other hand, if some subjects did better with the placebo while others did better with the high dose, then the error would be high. It should make intuitive sense that the less consistent the effect of dosage, the larger the dosage effect would have to be in order to be significant. The degree to which the effect of dosage differs depending on the subject is the Subjects x Dosage interaction. Recall that an interaction occurs when the effect of one variable differs depending on the level of another variable. In this case, the size of the error term is the extent to which the effect of the variable “Dosage” differs depending on the level of the variable “Subjects.” Note that each subject is a different level of the variable “Subjects.”

Other portions of the summary table have the same meaning as in between-subjects ANOVA. The F for dosage is the mean square for dosage divided by the mean square error. For these data, the F is significant with  $p = 0.004$ . Notice that this F test is equivalent to the t test for correlated pairs, with  $F = t^2$ .

Table 2 shows the ANOVA Summary Table when all four doses are included in the analysis. Since there are now four dosage levels rather than two, the df for dosage is three rather than one. Since the error is the Subjects x Dosage interaction, the df for error is the df for “Subjects” (23) times the df for Dosage (3) and is equal to 69.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	9065.49	394.15		
Dosage	3	557.61	185.87	5.18	0.003
Error	69	2476.64	35.89		
Total	95	12099.74			

## Carryover Effects

Often performing in one condition affects performance in a subsequent condition in such a way as to make a within-subjects design impractical. For example, consider an experiment with two conditions. In both conditions subjects are presented with pairs of words. In Condition A, subjects are asked to judge whether the words have similar meaning whereas in Condition B, subjects are asked to judge whether they sound similar. In both conditions, subjects are given a surprise

memory test at the end of the presentation. If Condition were a within-subjects variable, then there would be no surprise after the second presentation and it is likely that the subjects would have been trying to memorize the words.

Not all carryover effects cause such serious problems. For example, if subjects get fatigued by performing a task, then they would be expected to do worse on the second condition they were in. However, as long as the order of presentation is counterbalanced so that half of the subjects are in Condition A first and Condition B second, the fatigue effect itself would not invalidate the results, although it would add noise and reduce power. The carryover effect is symmetric in that having Condition A first affects performance in Condition B to the same degree that having Condition B first affects performance in Condition A.

Asymmetric carryover effects cause more serious problems. For example, suppose performance in Condition B were much better if preceded by Condition A, whereas performance in Condition A was approximately the same regardless of whether it was preceded by Condition B. With this kind of carryover effect, it is probably better to use a between-subjects design.

### **One Between- and One Within-Subjects Factor**

In the “Stroop Interference” case study, subjects performed three tasks: naming colors, reading color words, and naming the ink color of color words. Some of the subjects were males and some were females. Therefore, this design had two factors: gender and task. The ANOVA Summary Table for this design is shown in Table 3.

Table 3. ANOVA Summary Table for Stroop Experiment

Source	df	SSQ	MS	F	p
Gender	1	83.32	83.32	1.99	0.165
Error	45	1880.56	41.79		
Task	2	9525.97	4762.99	228.06	<0.001
Gender x Task	2	55.85	27.92	1.34	0.268
Error	90	1879.67	20.89		

The computations for the sums of squares will not be covered since computations are normally done by software. However, there are some important things to learn

from the summary table. First, notice that there are two error terms: one for the between-subjects variable Gender and one for both the within-subjects variable Task and the interaction of the between-subjects variable and the within-subjects variable. Typically, the mean square error for the between-subjects variable will be higher than the other mean square error. In this example, the mean square error for Gender is about twice as large as the other mean square error.

The degrees of freedom for the between-subjects variable is equal to the number of levels of the between-subjects variable minus one. In this example, it is one since there are two levels of gender. Similarly, the degrees of freedom for the within-subjects variable is equal to the number of levels of the variable minus one. In this example, it is two since there are three tasks. The degrees of freedom for the interaction is the product of the degrees of freedom for the two variables. For the Gender x Task interaction, the degrees of freedom is the product of degrees of freedom Gender (which is 1) and the degrees of freedom Task (which is 2) and is equal to 2.

## Assumption of Sphericity

Within-subjects ANOVA makes a restrictive assumption about the variances and the correlations among the dependent variables. Although the details of the assumption are beyond the scope of this book, it is approximately correct to say that it is assumed that all the correlations are equal and all the variances are equal. Table 4 shows the correlations among the three dependent variables in the Stroop Interference case study.

Table 4. Correlations Among Dependent Variables

	word reading	color naming	interference
word reading	1	0.7013	0.1583
color naming	0.7013	1	0.2382
interference	0.1583	0.2382	1

Note that the correlation between the word reading and the color naming variables of 0.7013 is much higher than the correlation between either of these variables with the interference variable. Moreover, as shown in Table 5, the variances among the variables differ greatly.

Table 5. Variances.

Variable	Variance
word reading	15.77
color naming	13.92
interference	55.07

Naturally the assumption of sphericity, like all assumptions, refers to populations not samples. However, it is clear from these sample data that the assumption is not met in the population.

### **Consequences of Violating the Assumption of Sphericity**

Although ANOVA is robust to most violations of its assumptions, the assumption of sphericity is an exception: Violating the assumption of sphericity leads to a substantial increase in the Type I error rate. Moreover, this assumption is rarely met in practice. Although violations of this assumption had at one time received little attention, the current consensus of data analysts is that it is no longer considered acceptable to ignore them.

### **Approaches to Dealing with Violations of Sphericity**

If an effect is highly significant, there is a conservative test that can be used to protect against an inflated Type I error rate. This test consists of adjusting the degrees of freedom for all within-subjects variables as follows: The degrees of freedom numerator and denominator are divided by the number of scores per subject minus one. Consider the effect of Task shown in Table 3. There are three scores per subject and therefore the degrees of freedom should be divided by two. The adjusted degrees of freedom are:

$$(2)(1/2) = 1 \text{ for the numerator and}$$

$$(90)(1/2) = 45 \text{ for the denominator}$$

The probability value is obtained using the F probability calculator with the new degrees of freedom parameters. The probability of an F of 228.06 of larger with 1 and 45 degrees of freedom is less than 0.001. Therefore, there is no need to worry about the assumption violation in this case.

Possible violation of sphericity does make a difference in the interpretation of the analysis shown in Table 2. The probability value of an F of 5.18 with 1 and 23 degrees of freedom is 0.032, a value that would lead to a more cautious conclusion than the p value of 0.003 shown in Table 2.

The correction described above is very conservative and should only be used when, as in Table 3, the probability value is very low. A better correction, but one that is very complicated to calculate, is to multiply the degrees of freedom by a quantity called  $\epsilon$  (the Greek letter epsilon). There are two methods of calculating  $\epsilon$ . The correction called the Huynh-Feldt (or H-F) is slightly preferred to the one called the Greenhouse Geisser (or G-G), although both work well. The G-G correction is generally considered a little too conservative.

A final method for dealing with violations of sphericity is to use a multivariate approach to within-subjects variables. This method has much to recommend it, but it is beyond the scope of this text.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 15: Multi-factor ANOVA

A research design to compare three drugs for the treatment of Alzheimer's disease is [described here](#). For the first two years of the study, researchers will follow the subjects with scans and memory tests.

## **What do you think?**

Assume the data were analyzed as a two-factor design with pre-post testing as one factor and the three drugs as the second factor. What term in an ANOVA would reflect whether the pre-post change was different for the three drugs??

It would be the interaction of the two factors since the question is whether the effect of one factor (pre-post) differs as a function of the level of a second factor (drug).

## Exercises

### *Prerequisites*

- All material presented in the ANOVA Chapter

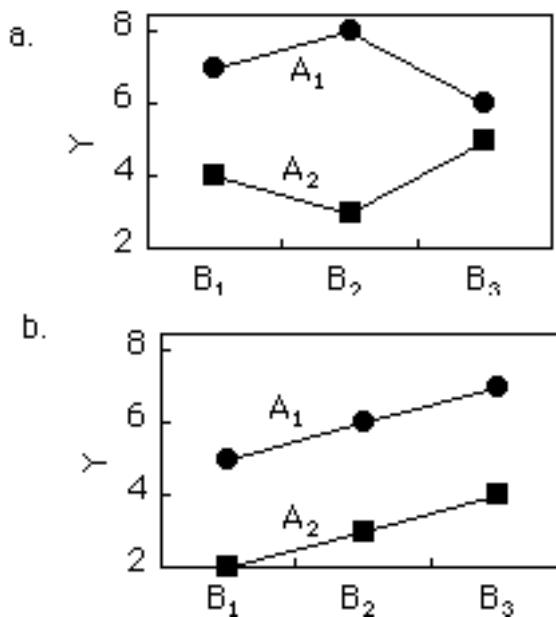
1. What is the null hypothesis tested by analysis of variance?
2. What are the assumptions of between-subjects analysis of variance?
3. What is a between-subjects variable?
4. Why not just compute t-tests among all pairs of means instead computing an analysis of variance?
5. What is the difference between “N” and “n”?
6. How is it that estimates of variance can be used to test a hypothesis about means?
7. Explain why the variance of the sample means has to be multiplied by “n” in the computation of  $MS_{between}$ .
8. What kind of skew does the F distribution have?
9. When do  $MS_{between}$  and  $MS_{error}$  estimate the same quantity?
10. If an experiment is conducted with 5 conditions and 6 subjects in each condition, what are  $df_n$  and  $df_e$ ?
11. How is the shape of the F distribution affected by the degrees of freedom?
12. What are the two components of the total sum of squares in a one-factor between-subjects design?
13. How is the mean square computed from the sum of squares?
14. An experimenter is interested in the effects of two independent variables on self-esteem. What is better about conducting a factorial experiment than conducting two separate experiments, one for each independent variable?

15. An experiment is conducted on the effect of age (5 yr, 10 yr and 15 yr) and treatment condition (experimental versus control) on reading speed. Which statistical term (main effect, simple effect, interaction, specific comparison) applies to each of the descriptions of effects.
- The effect of the treatment was larger for 15-year olds than it was for 5- or 10-year olds.
  - Overall, subjects in the treatment condition performed faster than subjects in the control condition.
  - The age effect was significant under the treatment condition.
  - The difference between the 15- year olds and the average of the 5- and 10- year olds was significant.
  - As they grow older, children read faster.
16. An A(3) x B(4) factorial design with 6 subjects in each group is analyzed. Give the source and degrees of freedom columns of the analysis of variance summary table.
17. The following data are from a hypothetical study on the effects of age and time on scores on a test of reading comprehension. Compute the analysis of variance summary table.

	<b>12-year olds</b>	<b>16-year olds</b>
<b>30 minutes</b>	66	74
	68	71
	59	67
	72	82
	46	76
<b>60 minutes</b>	69	95
	61	92
	69	95
	73	98
	61	94

18. Define “Three-way interaction”

19. Define interaction in terms of simple effects.
20. Plot an interaction for an  $A(2) \times B(2)$  design in which the effect of B is greater at  $A_1$  than it is at  $A_2$ . The dependent variable is “Number correct.” Make sure to label both axes.
21. Following are two graphs of population means for  $2 \times 3$  designs. For each graph, indicate which effect(s) (A, B, or  $A \times B$ ) are nonzero.



22. The following data are from an  $A(2) \times B(4)$  factorial design.

	B1	B2	B3	B4
A1	1	2	3	4
	3	2	4	5
	4	4	2	6
	5	5	6	8
A2	1	2	4	8
	1	3	6	9
	2	2	7	9
	2	4	8	8

- a. Compute an analysis of variance.
- b. Test differences among the four levels of B using the Bonferroni correction.
- c. Test the linear component of trend for the effect of B.

- d. Plot the interaction.
  - e. Describe the interaction in words.
23. Why are within-subjects designs usually more powerful than between-subjects design?
24. What source of variation is found in an ANOVA summary table for a within-subjects design that is not in an ANOVA summary table for a between-subjects design. What happens to this source of variation in a between-subjects design?
25. The following data contain three scores from each of five subjects. The three scores per subject are their scores on three trials of a memory task.

4	6	7
3	7	7
2	8	5
1	4	7
4	6	9

- a. Compute an ANOVA
  - b. Test all pairwise differences between means using the Bonferroni test at the .01 level.
  - c. Test the linear and quadratic components of trend for these data.
26. Give the source and df columns of the ANOVA summary table for the following experiments:
- a. Twenty two subjects are each tested on a simple reaction time task and on a choice reaction time task.
  - b. Twelve male and 12 female subjects are each tested under three levels of drug dosage: 0 mg, 10 mg, and 20 mg.
  - c. Twenty subjects are tested on a motor learning task for three trials a day for two days.
  - d. An experiment is conducted in which depressed people are either assigned to a drug therapy group, a behavioral therapy group, or a control group. Ten

subjects are assigned to each group. The level of measured once a month for four months.

*Questions from Case Studies*

Stroop Interference (S) case study

27. (S) The dataset Stroop Interference has the scores (times) for males and females on each of three tasks.

- a. Do a Gender (2) x Task (3) analysis of variance.
- b. Plot the interaction.

ADHD Treatment (AT) case study

28. (AT) The dataset ADHD Treatment has four scores per subject. a. Is the design between-subjects or within-subjects? b. Create an ANOVA summary table.

29. (AT) Using the Anger Expression Index from the Angry Moods study as the dependent variable, perform a 2x2 ANOVA with gender and sports participation as the two factors. Do athletes and non-athletes differ significantly in how much anger they express? Do the genders differ significantly in Anger Expression Index? Is the effect of sports participation significantly different for the two genders?

Weapons and Aggression (WA) case study

30. (WA) Using the Weapons and Aggression data, Compute a 2x2 ANOVA with the following two factors: prime type (was the first word a weapon or not?) and word type (was the second word aggressive or non-aggressive?). Consider carefully whether the variables are between-subject or within-subjects variables.

“Smiles and Leniency” (SL) case study

31. (SL) Compute the ANOVA summary table for the smiles and leniency data.

# 16. Transformations

- A. Log
- B. Tukey's Ladder of Powers
- C. Box-Cox Transformations
- D. Exercises

The focus of statistics courses is the exposition of appropriate methodology to analyze data to answer the question at hand. Sometimes the data are given to you, while other times the data are collected as part of a carefully-designed experiment. Often the time devoted to statistical analysis is less than 10% of the time devoted to data collection and preparation. If aspects of the data preparation fail, then the success of the analysis is in jeopardy. Sometimes errors are introduced into the recording of data. Sometimes biases are inadvertently introduced in the selection of subjects or the mis-calibration of monitoring equipment.

In this chapter, we focus on the fact that many statistical procedures work best if individual variables have certain properties. The measurement scale of a variable should be part of the data preparation effort. For example, the correlation coefficient does not require the variables have a normal shape, but often relationships can be made clearer by re-expressing the variables. An economist may choose to analyze the logarithm of prices if the relative price is of interest. A chemist may choose to perform a statistical analysis using the inverse temperature as a variable rather than the temperature itself. But note that the inverse of a temperature will differ depending on whether it is measured in  $^{\circ}\text{F}$ ,  $^{\circ}\text{C}$ , or  $^{\circ}\text{K}$ .

The introductory chapter covered linear transformations. These transformations normally do not change statistics such as Pearson's  $r$ , although they do affect the mean and standard deviation. The first section here is on log transformations which are useful to reduce skew. The second section is on Tukey's ladder of powers. You will see that log transformations are a special case of the ladder of powers. Finally, we cover the relatively advanced topic of the Box-Cox transformation.

# Log Transformations

by David M. Lane

## *Prerequisites*

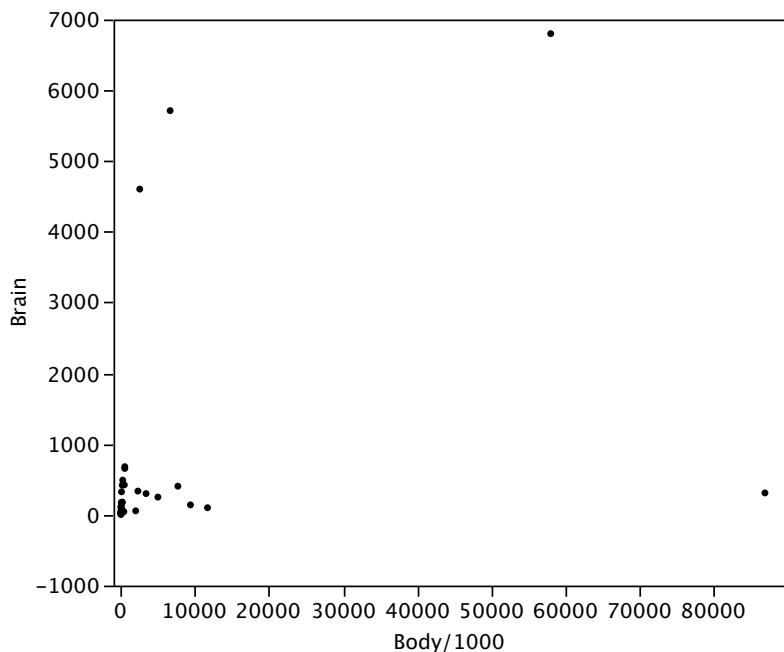
- Chapter 1: Logarithms
- Chapter 1: Shapes of Distributions
- Chapter 3: Additional Measures of Central Tendency
- Chapter 4: Introduction to Bivariate Data

## *Learning Objectives*

1. State how a log transformation can help make a relationship clear
2. Describe the relationship between logs and the geometric mean

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

Figure 1 shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.



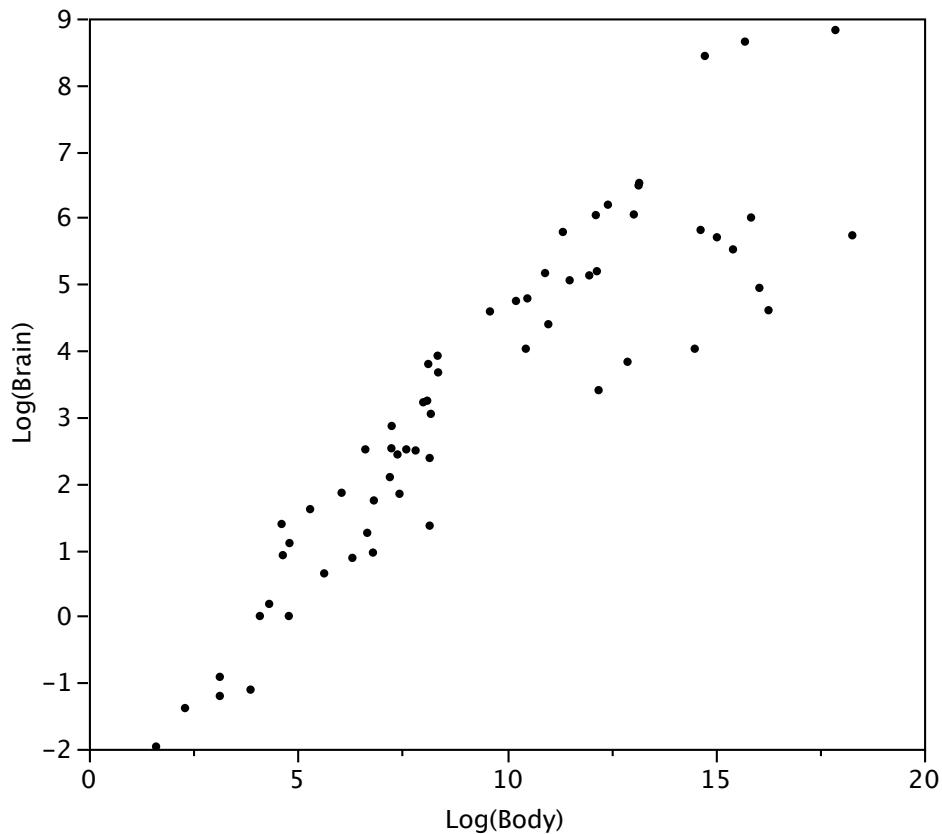


Figure 1. Scatter plots of brain weight as a function of body weight in terms of both raw data (upper panel) and log-transformed data (lower panel).

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel.

The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

Table 1 shows the logs (base 10) of the numbers 1, 10, and 100. The arithmetic mean of the three logs is

$$(0 + 1 + 2) / 3 = 1$$

The anti-log of this arithmetic mean of 1 is:

$$10^1 = 10$$

which is the geometric mean:

$$(1 \times 10 \times 100)^{.3333} = 10.$$

Table 1. Logarithms.

<b>X</b>	<b>Log<sub>10</sub>(X)</b>
1	0
10	1
100	2

Therefore, if the arithmetic means of two sets of log-transformed data are equal then the geometric means are equal.

# Tukey Ladder of Powers

by David W. Scott

## *Prerequisites*

- Chapter 1: Logarithms
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 12: Independent Groups t Test
- Chapter 13: Introduction to Power
- Chapter 16: Tukey Ladder of Powers

## *Learning Objectives*

1. Give the Tukey ladder of transformations
2. Find a transformation that reveals a linear relationship
3. Find a transformation to approximate a normal distribution

## Introduction

We assume we have a collection of bivariate data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

and that we are interested in the relationship between variables x and y. Plotting the data on a scatter diagram is the first step. As an example, consider the population of the United States for the 200 years before the Civil War. Of course, the decennial census began in 1790. These data are plotted two ways in Figure 1. Malthus predicted that geometric growth of populations coupled with arithmetic growth of grain production would have catastrophic results. Indeed the US population followed an exponential curve during this period.

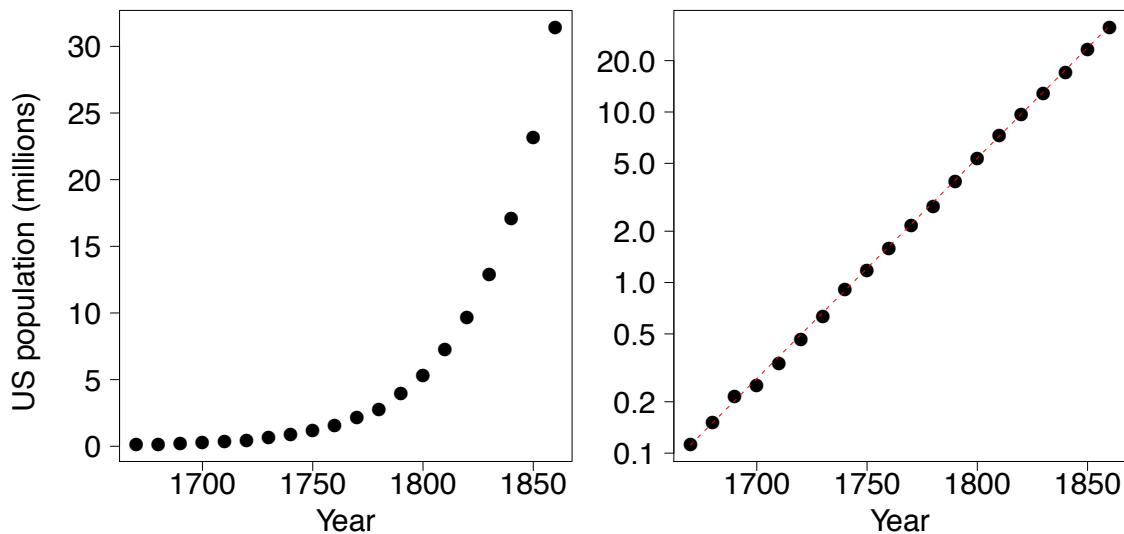


Figure 1. The US population from 1670 - 1860. The Y-axis on the right panel is on a log scale.

### Tukey's Transformation Ladder

Tukey (1977) describes an orderly way of re-expressing variables using a power transformation. You may be familiar with polynomial regression (a form of multiple regression) in which the simple linear model  $y = b_0 + b_1X$  is extended with terms such as  $b_2X^2 + b_3X^3 + b_4X^4$ . Alternatively, Tukey suggests exploring simple relationships such as

$$y = b_0 + b_1X^\lambda \text{ or } y^\lambda = b_0 + b_1X \quad (\text{Equation 1})$$

where  $\lambda$  is a parameter chosen to make the relationship as close to a straight line as possible. Linear relationships are special, and if a transformation of the type  $x^\lambda$  or  $y^\lambda$  works as in Equation (1), then we should consider changing our measurement scale for the rest of the statistical analysis.

There is no constraint on values of  $\lambda$  that we may consider. Obviously choosing  $\lambda = 1$  leaves the data unchanged. Negative values of  $\lambda$  are also reasonable. For example, the relationship

$$y = b_0 + b_1/x$$

would be represented by  $\lambda = -1$ . The value  $\lambda = 0$  has no special value, since  $X^0 = 1$ , which is just a constant. Tukey (1977) suggests that it is convenient to simply define the transformation when  $\lambda = 0$  to be the logarithm function rather than the

constant 1. We shall revisit this convention shortly. The following table gives examples of the Tukey ladder of transformations.

Table 1. Tukey's Ladder of Transformations

$\lambda$	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	$x$	$x^2$

If  $x$  takes on negative values, then special care must be taken so that the transformations make sense, if possible. We generally limit ourselves to variables where  $x > 0$  to avoid these considerations. For some dependent variables such as the number of errors, it is convenient to add 1 to  $x$  before applying the transformation.

Also, if the transformation parameter  $\lambda$  is negative, then the transformed variable  $x^\lambda$  is reversed. For example, if  $x$  is increasing, then  $1/x$  is decreasing. We choose to redefine the Tukey transformation to be  $-(x^\lambda)$  if  $\lambda < 0$  in order to preserve the order of the variable after transformation. Formally, the Tukey transformation is defined as

$$\tilde{x}_\lambda = \begin{cases} x^\lambda & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -(x^\lambda) & \text{if } \lambda < 0 \end{cases} \quad (2)$$

In Table 2 we reproduce Table 1 but using the modified definition when  $\lambda < 0$ .

Table 2. Modified Tukey's Ladder of Transformations

$\lambda$	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	$x$	$x^2$

## The Best Transformation for Linearity

The goal is to find a value of  $\lambda$  that makes the scatter diagram as linear as possible. For the US population, the logarithmic transformation applied to  $y$  makes the relationship almost perfectly linear. The red dashed line in the right frame of Figure 1 has a slope of about 1.35; that is, the US population grew at a rate of about 35% per decade.

The logarithmic transformation corresponds to the choice  $\lambda = 0$  by Tukey's convention. In Figure 2, we display the scatter diagram of the US population data for  $\lambda = 0$  as well as for other choices of  $\lambda$ .

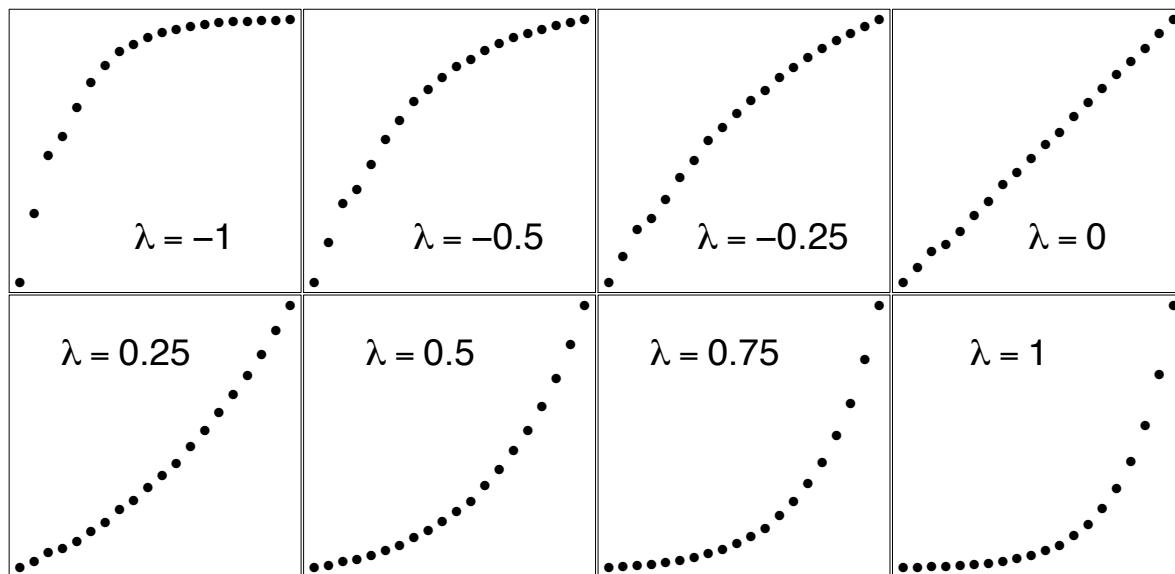


Figure 2. The US population from 1670 to 1860 for various values of  $\lambda$ .

The raw data are plotted in the bottom right frame of Figure 2 when  $\lambda = 1$ . The logarithmic fit is in the upper right frame when  $\lambda = 0$ . Notice how the scatter diagram smoothly morphs from convex to concave as  $\lambda$  increases. Thus intuitively there is a unique best choice of  $\lambda$  corresponding to the “most linear” graph.

One way to make this choice objective is to use an objective function for this purpose. One approach might be to fit a straight line to the transformed points and try to minimize the residuals. However, an easier approach is based on the fact that the correlation coefficient,  $r$ , is a measure of the linearity of a scatter diagram. In particular, if the points fall on a straight line then their correlation will be  $r = 1$ . (We need not worry about the case when  $r = -1$  since we have defined the Tukey transformed variable  $x_\lambda$  to be positively correlated with  $x$  itself.)

In Figure 3, we plot the correlation coefficient of the scatter diagram  $(x, \tilde{y}_\lambda)$  as a function of  $\lambda$ . It is clear that the logarithmic transformation ( $\lambda = 0$ ) is nearly optimal by this criterion.

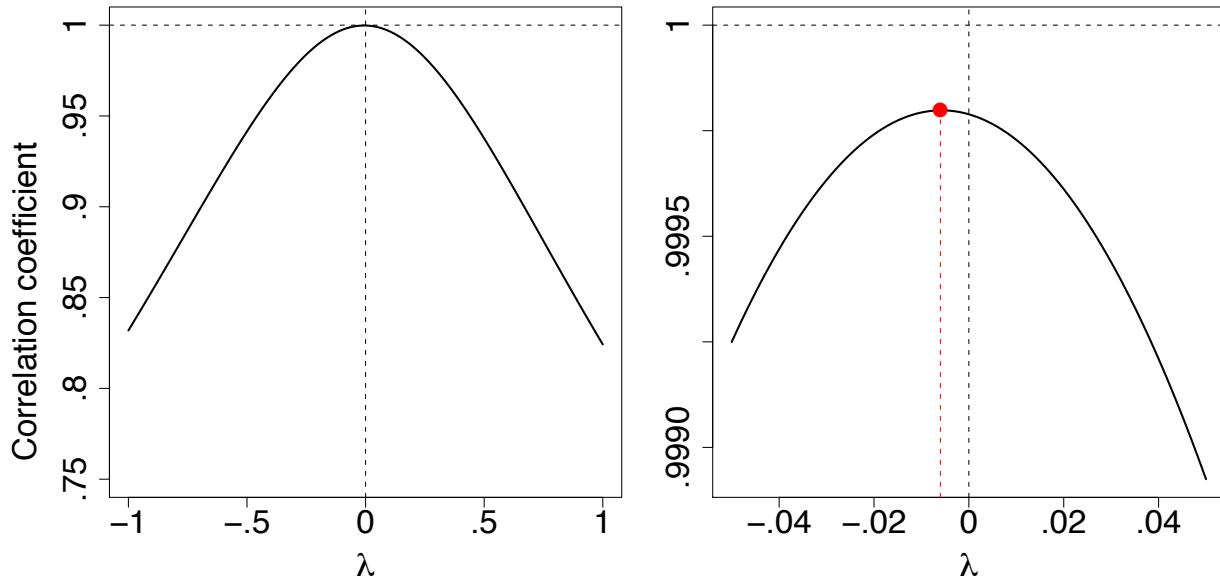


Figure 3. Graph of US population correlation coefficient as function of  $\lambda$ .

Is the US population still on the same exponential growth pattern? In Figure 4 we display the US population from 1630 to 2000 using the transformation and fit used in the right frame of Figure 1. Fortunately, the exponential growth (or at least its rate) was not sustained into the Twentieth Century. If it had, the US population in the year 2000 would have been over 2 billion (2.07 to be exact), larger than the population of China.

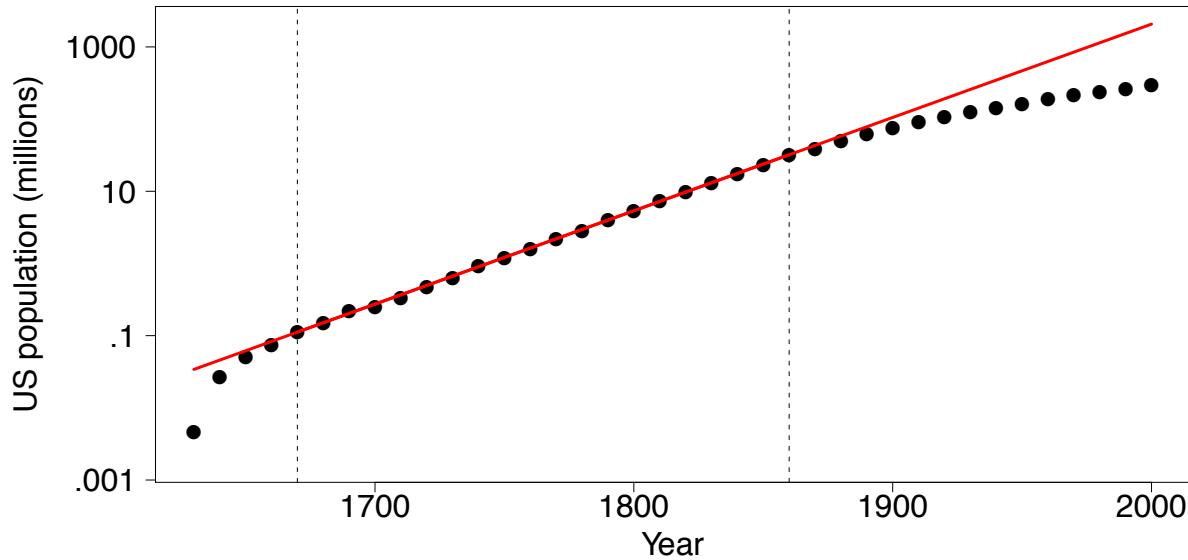


Figure 4. Graph of US population 1630-2000 with  $\lambda = 0$ .

We can examine the decennial census population figures of individual states as well. In Figure 5 we display the population data for the state of New York from 1790 to 2000, together with an estimate of the population in 2008. Clearly something unusual happened starting in 1970. (This began the period of mass migration to the West and South as the rust belt industries began to shut down.) Thus, we compute the best  $\lambda$  value using the data from 1790-1960 in the middle frame of Figure 5. The right frame displays the transformed data, together with the linear fit for the 1790-1960 period. The value of  $\lambda = 0.41$  is not obvious and one might reasonably choose to use  $\lambda = 0.50$  for practical reasons.

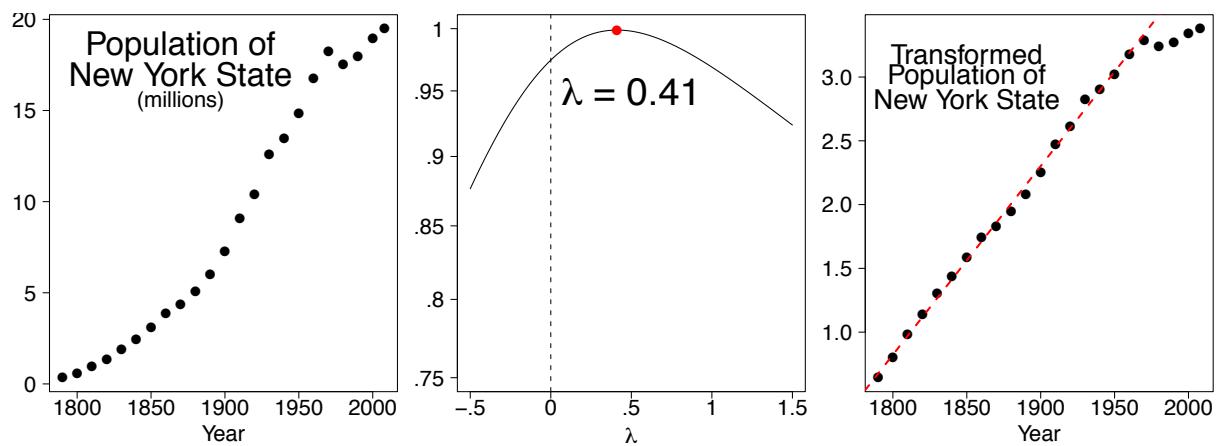


Figure 5. Graphs related to the New York state population 1790-2008.

If we look at one of the younger states in the West, the picture is different. Arizona has attracted many retirees and immigrants. Figure 6 summarizes our findings. Indeed, the growth of population in Arizona is logarithmic, and appears to still be logarithmic through 2005.

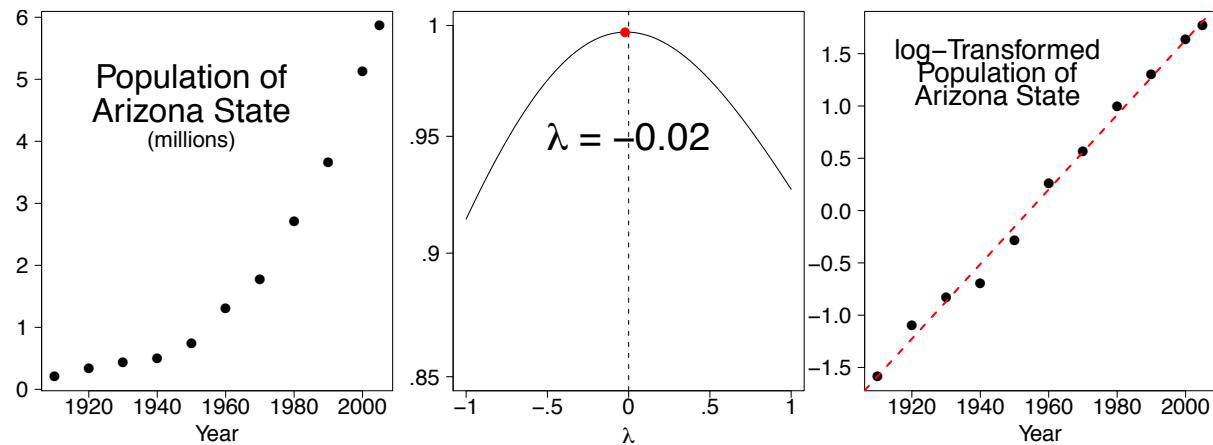


Figure 6. Graphs related to the Arizona state population 1910-2005.

## Reducing Skew

Many statistical methods such as t tests and the analysis of variance assume normal distributions. Although these methods are relatively robust to violations of normality, transforming the distributions to reduce skew can markedly increase their power.

As an example, the data in the “Stereograms” case study is very skewed. A t test of the difference between the two conditions using the raw data results in a p value of 0.056, a value not conventionally considered significant. However, after a log transformation ( $\lambda = 0$ ) that reduces the skew greatly, the p value is 0.023 which is conventionally considered significant.

The demonstration in Figure 7 shows distributions of the data from the Stereograms case study as transformed with various values of  $\lambda$ . Decreasing  $\lambda$  makes the distribution less positively skewed. Keep in mind that  $\lambda = 1$  is the raw data. Notice that there is a slight positive skew for  $\lambda = 0$  but much less skew than found in the raw data ( $\lambda = 1$ ). Values of below 0 result in negative skew.

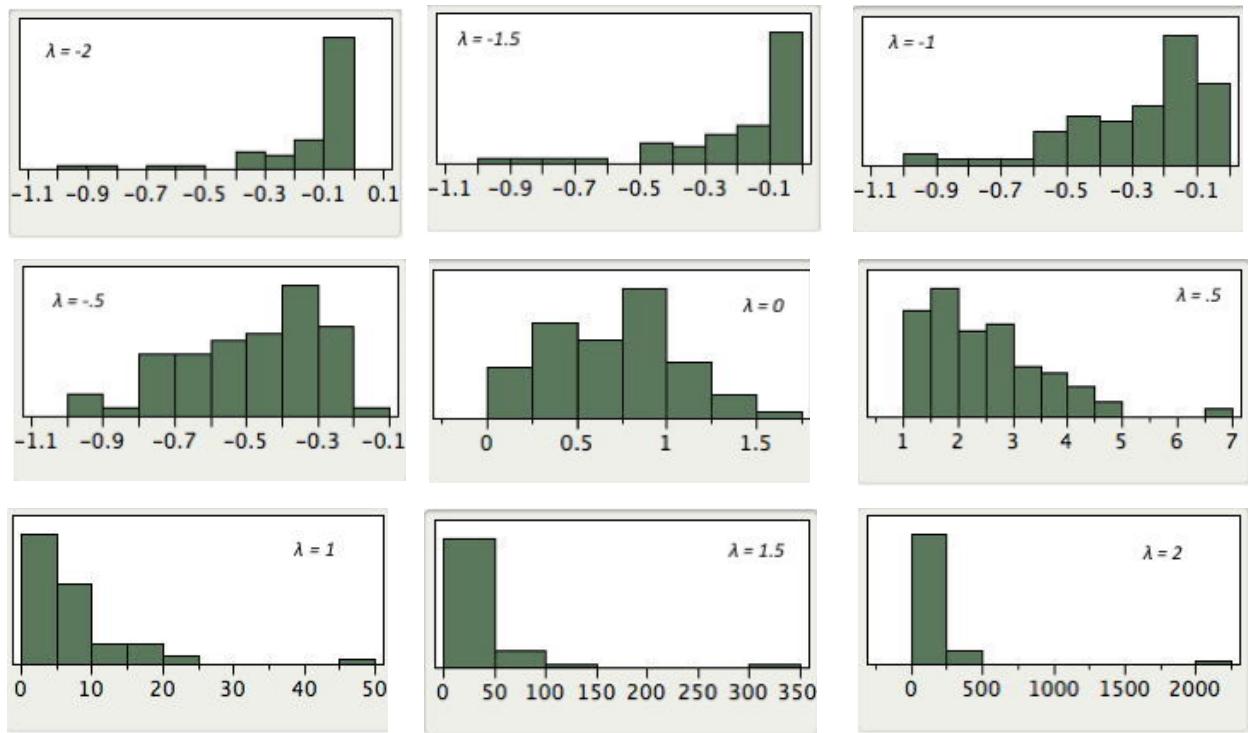


Figure 7. Distribution of data from the Stereogram case study for various values of  $\lambda$ .

# Box-Cox Transformations

by David Scott

## *Prerequisites*

This section assumes a higher level of mathematics background than most other sections of this work.

- Chapter 1: Logarithms
- Chapter 3: Additional Measures of Central Tendency (Geometric Mean)
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 16: Tukey Ladder of Powers

George Box and Sir David Cox collaborated on one paper (Box, 1964). The story is that while Cox was visiting Box at Wisconsin, they decided they should write a paper together because of the similarity of their names (and that both are British). In fact, Professor Box is married to the daughter of Sir Ronald Fisher.

The Box-Cox transformation of the variable  $x$  is also indexed by  $\lambda$ , and is defined as

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda} . \quad (\text{Equation 1})$$

At first glance, although the formula in Equation (1) is a scaled version of the Tukey transformation  $x^\lambda$ , this transformation does not appear to be the same as the Tukey formula in Equation (2). However, a closer look shows that when  $\lambda < 0$ , both  $x_\lambda$  and  $x'_\lambda$  change the sign of  $x^\lambda$  to preserve the ordering. Of more interest is the fact that when  $\lambda = 0$ , then the Box-Cox variable is the indeterminate form 0/0. Rewriting the Box-Cox formula as

$$x'_\lambda = \frac{e^{\lambda \log(x)} - 1}{\lambda} \approx \frac{(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \dots) - 1}{\lambda} \rightarrow \log(x)$$

as  $\lambda \rightarrow 0$ . This same result may also be obtained using l'Hôpital's rule from your calculus course. This gives a rigorous explanation for Tukey's suggestion that the

log transformation (which is not an example of a polynomial transformation) may be inserted at the value  $\lambda = 0$ .

Notice with this definition of  $x'_\lambda$  that  $x = 1$  always maps to the point  $x'_\lambda = 0$  for all values of  $\lambda$ . To see how the transformation works, look at the examples in Figure 1. In the top row, the choice  $\lambda = 1$  simply shifts  $x$  to the value  $x-1$ , which is a straight line. In the bottom row (on a semi-logarithmic scale), the choice  $\lambda = 0$  corresponds to a logarithmic transformation, which is now a straight line. We superimpose a larger collection of transformations on a semi-logarithmic scale in Figure 2.

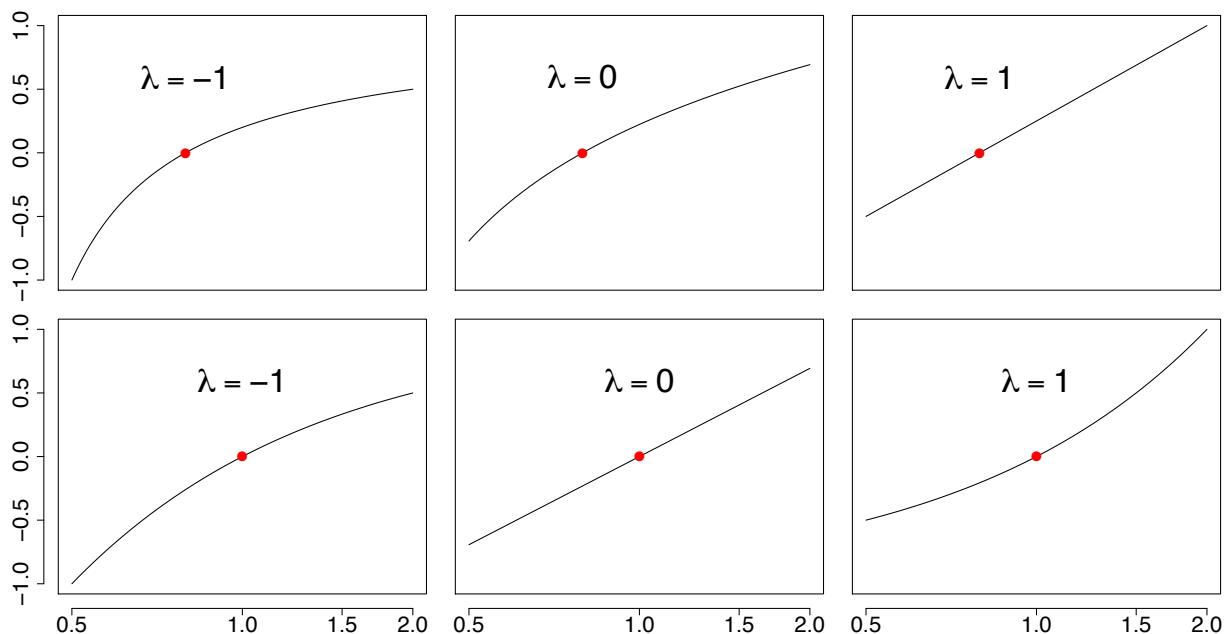


Figure 1. Examples of the Box-Cox transformation  $x'_\lambda$  versus  $x$  for  $\lambda = -1, 0$ ,

1. In the second row,  $x'_\lambda$  is plotted against  $\log(x)$ . The red point is at  $(1, 0)$ .

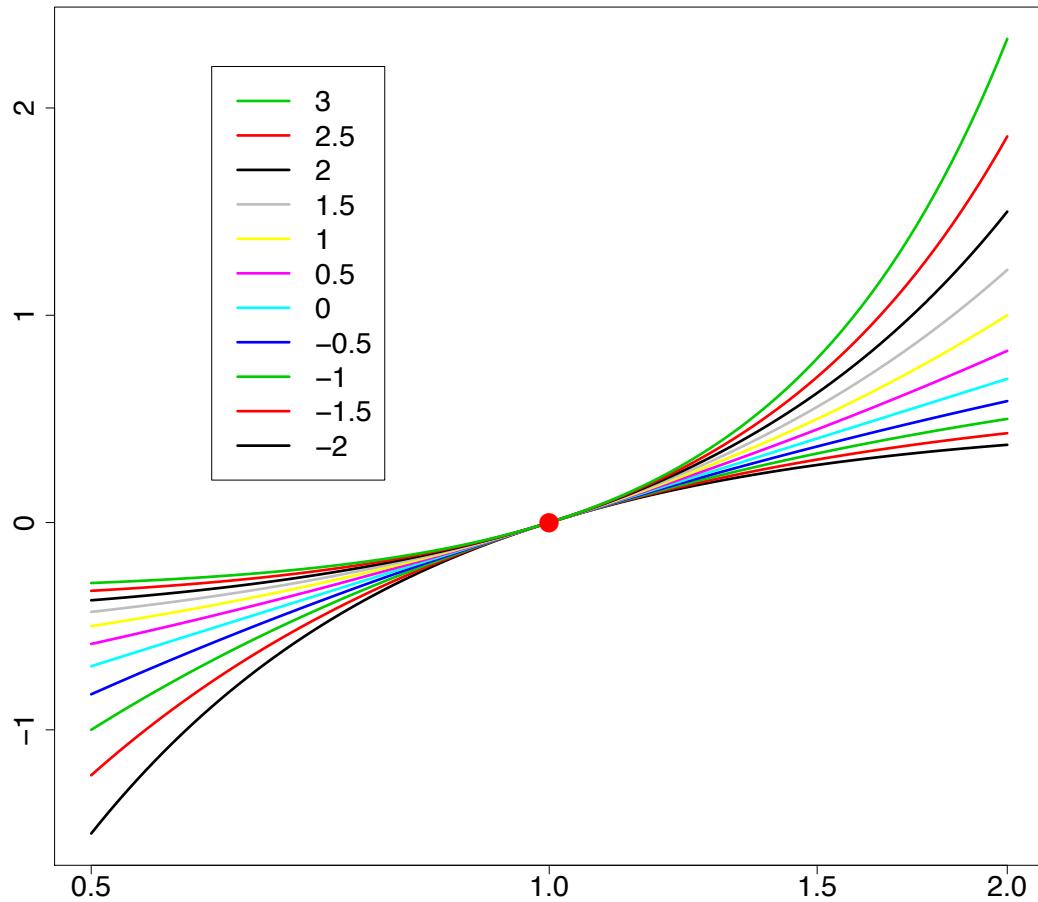


Figure 2. Examples of the Box-Cox transformation versus  $\log(x)$  for  $-2 < \lambda < 3$ . The bottom curve corresponds to  $\lambda = -2$  and the upper to  $\lambda = 3$ .

### Transformation to Normality

Another important use of variable transformation is to eliminate skewness and other distributional features that complicate analysis. Often the goal is to find a simple transformation that leads to normality. In the article on q-q plots, we discuss how to assess the normality of a set of data,

$$x_1, x_2, \dots, x_n.$$

Data that are normal lead to a straight line on the q-q plot. Since the correlation coefficients maximized when a scatter diagram is linear, we can use the same approach above to find the most normal transformation.

Specifically, we form the  $n$  pairs

$$\left( \Phi^{-1} \left( \frac{i - 0.5}{n} \right), x_{(i)} \right), \quad \text{for } i = 1, 2, \dots, n,$$

where  $\Phi^{-1}$  is the inverse CDF of the normal density and  $x_{(i)}$  denotes the  $i^{\text{th}}$  sorted value of the data set. As an example, consider a large sample of British household incomes taken in 1973, normalized to have mean equal to one ( $n = 7,125$ ). Such data are often strongly skewed, as is clear from Figure 3. The data were sorted and paired with the 7125 normal quantiles. The value of  $\lambda$  that gave the greatest correlation ( $r = 0.9944$ ) was  $\lambda = 0.21$ .

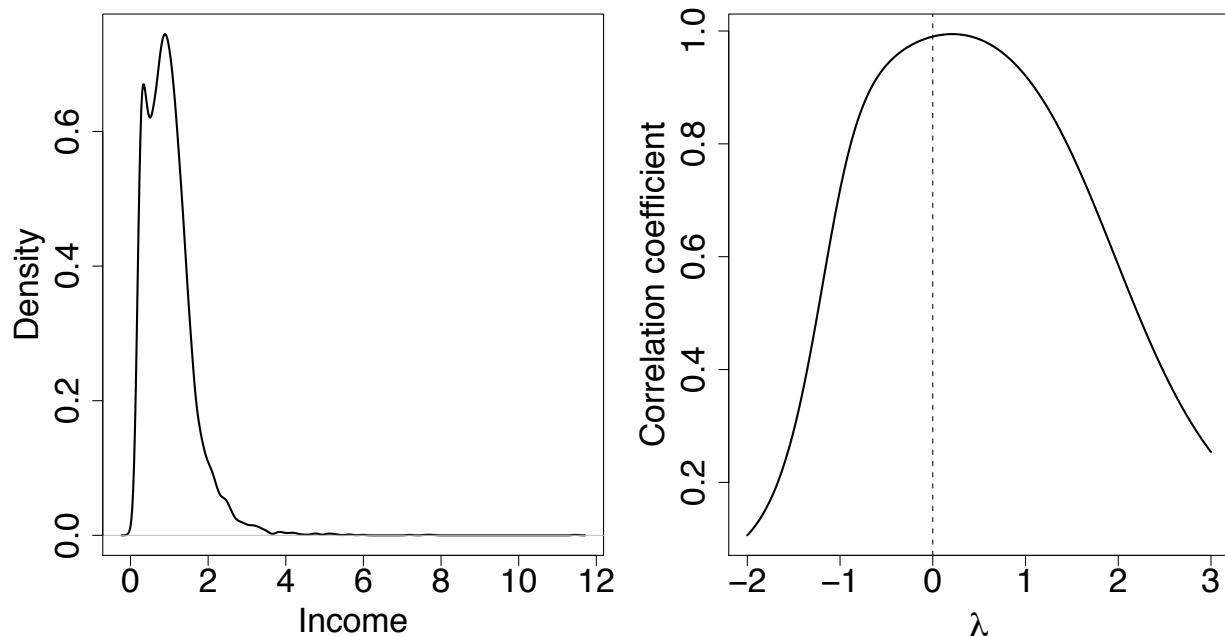


Figure 3. (L) Density plot of the 1973 British income data. (R) The best value of  $\lambda$  is 0.21.

The kernel density plot of the optimally transformed data is shown in the left frame of Figure 4. While this figure is much less skewed than in Figure 3, there is clearly an extra “component” in the distribution that might reflect the poor. Economists often analyze the logarithm of income corresponding to  $\lambda = 0$ ; see Figure 4. The correlation is only  $r = 0.9901$  in this case, but for convenience, the log-transform probably will be preferred.

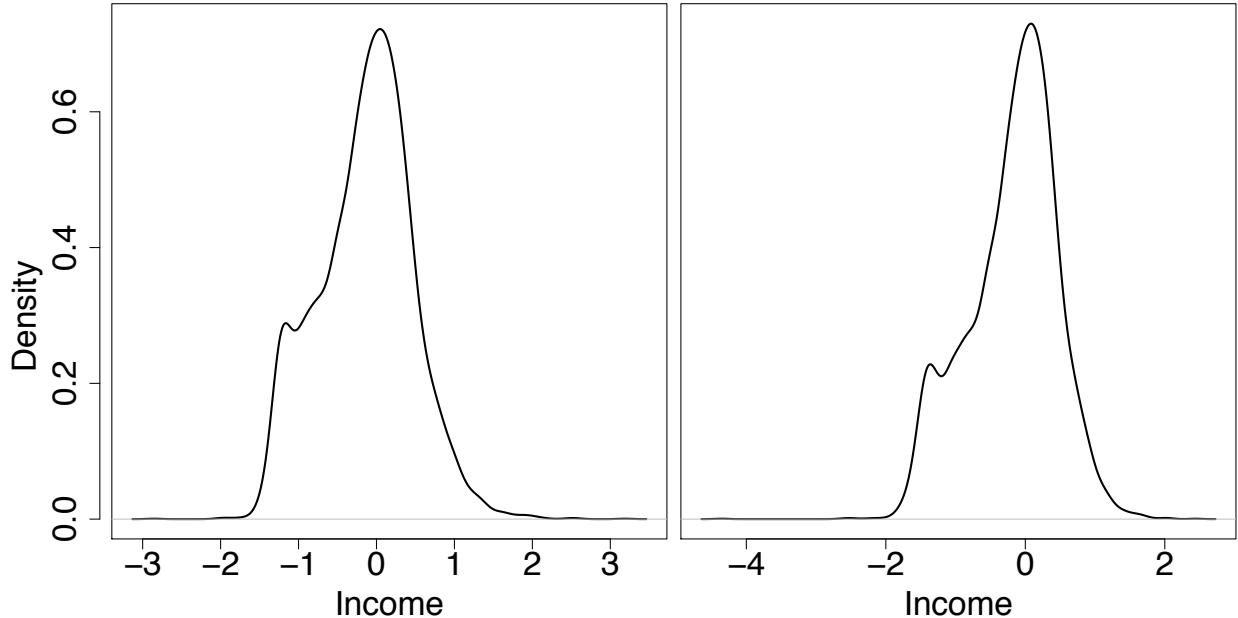


Figure 4. (L) Density plot of the 1973 British income data transformed with  $\lambda = 0.21$ . (R) The log-transform with  $\lambda = 0$ .

### Other Applications

Regression analysis is another application where variable transformation is frequently applied. For the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

and fitted model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p ,$$

each of the predictor variables  $x_j$  can be transformed. The usual criterion is the variance of the residuals, given by

$$\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 .$$

Occasionally, the response variable  $y$  may be transformed. In this case, care must be taken because the variance of the residuals is not comparable as  $\lambda$  varies. Let  $\bar{g}_y$  represent the geometric mean of the response variables.

$$\bar{g}_y = \left( \prod_{i=1}^n y_i \right)^{1/n}$$

Then the transformed response is defined as

$$y'_\lambda = \frac{y^\lambda - 1}{\lambda \cdot \bar{g}_y^{\lambda-1}}$$

When  $\lambda = 0$  (the logarithmic case),

$$y'_0 = \bar{g}_y \cdot \log(y)$$

For more examples and discussions, see Kutner, Nachtsheim, Neter, and Li (2004).

# Statistical Literacy

by David M. Lane

## Prerequisites

- Chapter 16: Logarithms

Many financial web pages give you the option of using a linear or a logarithmic Y-axis. An example from Google Finance is shown below.



## What do you think?

To get a straight line with the linear option chosen, the price would have to go up the same amount every time period. What would result in a straight line with the logarithmic option chosen?

The price would have to go up the same proportion every time period. For example, go up 0.1% every day.

## References

- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*, McGraw-Hill/Irwin, Homewood, IL.
- Tukey, J. W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA.

## Exercises

### *Prerequisites*

#### All Content in This Chapter

1. When is a log transformation valuable?
2. If the arithmetic mean of  $\log_{10}$  transformed data were 3, what would be the geometric mean?
3. Using Tukey's ladder of transformation, transform the following data using a  $\lambda$  of 0.5: 9, 16, 25
4. What value of  $\lambda$  in Tukey's ladder decreases skew the most?
5. What value of  $\lambda$  in Tukey's ladder increases skew the most?
6. In the [ADHD](#) case study, transform the data in the placebo condition (D0) with  $\lambda$ 's of .5, 0, -.5, and -1. How does the skew in each of these compare to the skew in the raw data. Which transformation leads to the least skew?

# 17. Chi Square

- A. Chi Square Distribution
- B. One-Way Tables
- C. Contingency Tables
- D. Exercises

Chi Square is a distribution that has proven to be particularly useful in statistics. The first section describes the basics of this distribution. The following two sections cover the most common statistical tests that make use of the Chi Square distribution. The section “One-Way Tables” shows how to use the Chi Square distribution to test the difference between theoretically expected and observed frequencies. The section “Contingency Tables” shows how to use Chi Square to test the association between two nominal variables. This use of Chi Square is so common that it is often referred to as the “Chi Square Test.”

# Chi Square Distribution

by David M. Lane

## Prerequisites

- Chapter 1: Distributions
- Chapter 7: Standard Normal Distribution
- Chapter 10: Degrees of Freedom

## Learning Objectives

1. Define the Chi Square distribution in terms of squared normal deviates
2. Describe how the shape of the Chi Square distribution changes as its degrees of freedom increase

A *standard normal deviate* is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates. The *degrees of freedom* of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as  $\chi^2(1)$ , is simply the distribution of a single normal deviate squared. The area of a Chi Square distribution below 4 is the same as the area of a standard normal distribution below 2, since 4 is  $2^2$ .

Consider the following problem: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher? Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a Chi Square (with 2 df) being six or higher is 0.050.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom. As the degrees of freedom increases, the Chi Square distribution approaches a normal distribution. Figure 1 shows density functions for three Chi Square distributions. Notice how the skew decreases as the degrees of freedom increase.

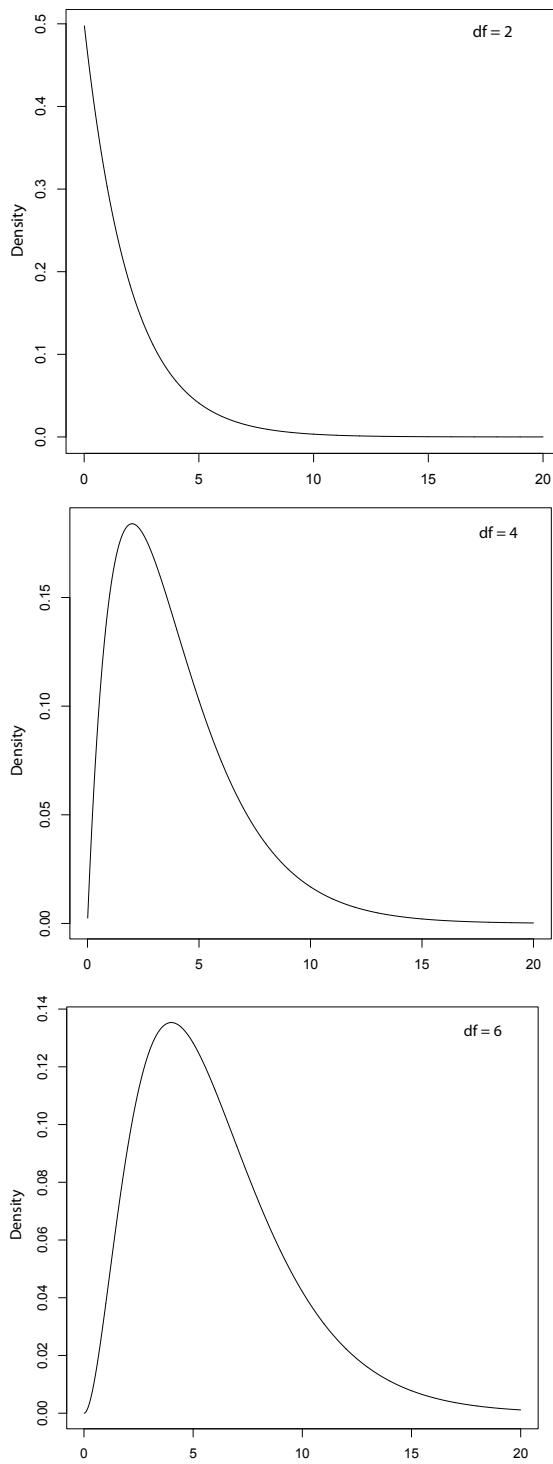


Figure 1. Chi Square distributions with 2, 4, and 6 degrees of freedom.

The Chi Square distribution is very important because many test statistics are approximately distributed as Chi Square. Two of the more common tests using the

Chi Square distribution are tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between categorical variables (contingency tables). Numerous other tests beyond the scope of this work are based on the Chi Square distribution.

# One-Way Tables (Testing Goodness of Fit)

by David M. Lane

## *Prerequisites*

- Chapter 5: Basic Concepts of Probability
- Chapter 11: Significance Testing
- Chapter 17: Chi Square Distribution

## *Learning Objectives*

1. Describe what it means for there to be theoretically-expected frequencies
2. Compute expected frequencies
3. Compute Chi Square
4. Determine the degrees of freedom

The Chi Square distribution can be used to test whether observed data differ significantly from theoretical expectations. For example, for a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table 1 were obtained by rolling a six-sided die 36 times. However, as can be seen in Table 1, some outcomes occurred more frequently than others. For example, a “3” came up nine times, whereas a “4” came up only two times. Are these data consistent with the hypothesis that the die is a fair die? Naturally, we do not expect the sample frequencies of the six possible outcomes to be the same since chance differences will occur. So, the finding that the frequencies differ does not mean that the die is not fair. One way to test whether the die is fair is to conduct a significance test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Table 1. Outcome Frequencies from a Six-Sided Die

Outcome	Frequency
1	8
2	5
3	9

4	2
5	7
6	5

The first step in conducting the significance test is to compute the expected frequency for each outcome given that the null hypothesis is true. For example, the expected frequency of a “1” is 6, since the probability of a “1” coming up is  $1/6$  and there were a total of 36 rolls of the die.

$$\text{Expected frequency} = (1/6)(36) = 6$$

Note that the expected frequencies are expected only in a theoretical sense. We do not really “expect” the observed frequencies to match the “expected frequencies” exactly.

The calculation continues as follows. Letting  $E$  be the expected frequency of an outcome and  $O$  be the observed frequency of that outcome, compute

$$\frac{(E - O)^2}{E}$$

for each outcome. Table 2 shows these calculations.

Table 2. Outcome Frequencies from a Six-Sided Die

Outcome	E	O	$(E-O)^2/E$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.5
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Next we add up all the values in Column 4 of Table 2.

$$\sum \frac{(E - O)^2}{E} = 5.33$$

This sampling distribution of

$$\sum \frac{(E - O)^2}{E}$$

is approximately distributed as Chi Square with  $k-1$  degrees of freedom, where  $k$  is the number of categories. Therefore, for this problem the test statistic is

$$x_5^2 = 5.333$$

which means the value of Chi Square with 5 degrees of freedom is 5.333.

From a Chi Square calculator it can be determined that the probability of a Chi Square of 5.333 or larger is 0.377. Therefore, the null hypothesis that the die is fair cannot be rejected.

The Chi Square test can also be used to test other deviations between expected and observed frequencies. The following example shows a test of whether the variable “University GPA” in the SAT and College GPA case study is normally distributed.

The first column in Table 3 shows the normal distribution divided into five ranges. The second column shows the proportions of a normal distribution falling in the ranges specified in the first column. The expected frequencies (E) are calculated by multiplying the number of scores (105) by the proportion. The final column shows the observed number of scores in each range. It is clear that the observed frequencies vary greatly from the expected frequencies. Note that if the distribution were normal, then there would have been only about 35 scores between 0 and 1, whereas 60 were observed.

Table 3. Expected and Observed Scores for 105 University GPA Scores.

Range	Proportion	E	O
Above 1	0.159	16.695	9
0 to 1	0.341	35.805	60
-1 to 0	0.341	35.805	17
Below -1	0.159	16.695	19

The test of whether the observed scores deviate significantly from the expected scores is computed using the familiar calculation.

$$\chi_3^2 = \sum \frac{(E - O)^2}{E} = 30.09$$

The subscript “3” means there are three degrees of freedom. As before, the degrees of freedom is the number of outcomes minus one, which is three in this example. The Chi Square distribution calculator shows that  $p < 0.001$  for this Chi Square. Therefore, the null hypothesis that the scores are normally distributed can be rejected.

# Contingency Tables

by David M. Lane

## Prerequisites

- Chapter 17: Chi Square Distribution
- Chapter 17: One-Way Tables

## Learning Objectives

1. State the null hypothesis tested concerning contingency tables
2. Compute expected cell frequencies
3. Compute Chi Square and df

This section shows how to use Chi Square to test the relationship between nominal variables for significance. For example, Table 1 shows the data from the Mediterranean Diet and Health case study.

Table 1. Frequencies for Diet and Health Study

Diet	Outcome				
	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

The question is whether there is a significant relationship between diet and outcome. The first step is to compute the expected frequency for each cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the AHA Diet/Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome, then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect  $(0.0364)(303) = 11.02$  cancers on the AHA diet. Similarly, we would expect  $(0.0364)(302) = 10.98$  cancers on the Mediterranean diet. In

general, the expected frequency for a cell in the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column is equal to

$$E_{i,j} = \frac{T_i T_j}{T}$$

where  $E_{i,j}$  is the expected frequency for cell  $i,j$ ,  $T_i$  is the total for the  $i^{\text{th}}$  row,  $T_j$  is the total for the  $j^{\text{th}}$  column, and  $T$  is the total number of observations. For the AHA Diet/Cancers cell,  $i = 1, j = 1$ ,  $T_i = 303$ ,  $T_j = 22$ , and  $T = 605$ . Table 2 shows the expected frequencies (in parenthesis) for each cell in the experiment. Table 2 shows the expected frequencies (in parenthesis) for each cell in the experiment.

Table 2. Observed and Expected Frequencies for Diet and Health Study

Diet	Outcome				
	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

The significance test is conducted by computing Chi Square as follows.

$$\chi^2_3 = \sum \frac{(E - O)^2}{E} = 16.55$$

The degrees of freedom is equal to  $(r-1)(c-1)$ , where  $r$  is the number of rows and  $c$  is the number of columns. For this example, the degrees of freedom is  $(2-1)(4-1) = 3$ . The Chi Square calculator can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is equal to 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

A key assumption of this Chi Square test is that each subject contributes data to only one cell. Therefore, the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment. Consider an experiment in

which each of 16 subjects attempted two anagram problems. The data are shown in Table 3.

Table 3. Anagram problem data

	Anagram 1	Anagram 2
Solved	10	4
Did not Solve	6	12

It would not be valid to use the Chi Square test on these data since each subject contributed data to two cells: one cell based on their performance on Anagram 1 and one cell based on their performance on Anagram 2. The total of the cell frequencies in the table is 32, but the total number of subjects is only 16.

The formula for Chi Square yields a statistic that is only approximately a Chi Square distribution. In order for the approximation to be adequate, the total number of subjects should be at least 20. Some authors claim that the correction for continuity should be used whenever an expected cell frequency is below 5. Research in statistics has shown that this practice is not advisable. For example, see: Bradley, Bradley, McGrath, & Cutcomb (1979). The correction for continuity when applied to  $2 \times 2$  contingency tables is called the Yates correction.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 17: Contingency Tables

An experiment was conducted to test whether the spice saffron can inhibit liver cancer. Two groups of rats were tested. Both groups were injected with chemicals known to increase the chance of liver cancer. The experimental group was fed saffron ( $n = 24$ ) whereas the control group was not ( $n = 8$ ). The experiment is [described here](#).

Only 4 of the 24 subjects in the saffron group developed cancer as compared to 6 of the 8 subjects in the control group.

## What do you think?

What method could be used to test whether this difference between the experimental and control groups is statistically significant?

The Chi Square test of contingency tables could be used. It yields a Chi Squared ( $df = 1$ ) of 9.50 which has an associated p of 0.002.  $\chi^2$

## **References**

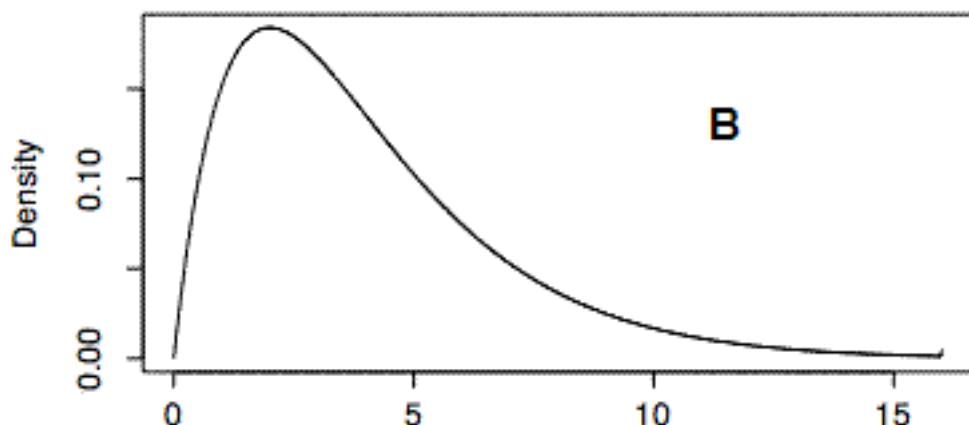
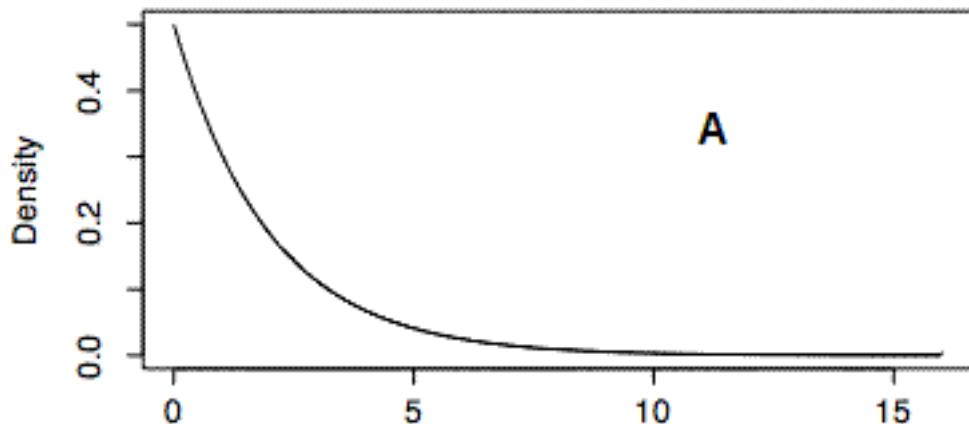
Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979) Type I error rate of the chi square test of independence in  $r \times c$  tables that have small expected frequencies. *Psychological Bulletin, 86*, 1200-1297.

## Exercises

### Prerequisites

- All material presented in the Chi Square Chapter

1. Which of the two Chi Square distributions shown below (A or B) has the larger degrees of freedom? How do you know?



2. Twelve subjects were each given two flavors of ice cream to taste and then were asked whether they liked them. Two of the subjects liked the first flavor and nine of them liked the second flavor. Is it valid to use the Chi Square test to determine whether this difference in proportions is significant? Why or why not?
3. A die is suspected of being biased. It is rolled 25 times with the following result:

Outcome	Frequency
1	9
2	4
3	1
4	8
5	3
6	0

Conduct a significance test to see if the die is biased. (a) What Chi Square value do you get and how many degrees of freedom does it have? (b) What is the p value?

4. A recent experiment investigated the relationship between smoking and urinary incontinence. Of the 322 subjects in the study who were incontinent, 113 were smokers, 51 were former smokers, and 158 had never smoked. Of the 284 control subjects who were not in- continent, 68 were smokers, 23 were former smokers, and 193 had never smoked.

- a. Create a table displaying this data.
- b. What is the expected frequency in each cell?
- c. Conduct a significance test to see if there is a relationship between smoking and incontinence. What Chi Square value do you get? What p value do you get?
- d. What do you conclude?

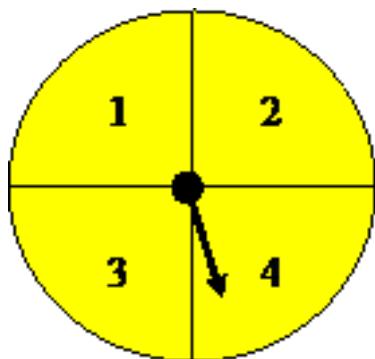
5. At a school pep rally, a group of sophomore students organized a free raffle for prizes. They claim that they put the names of all of the students in the school in the basket and that they randomly drew 36 names out of this basket. Of the prize winners, 6 were freshmen, 14 were sophomores, 9 were juniors, and 7 were seniors. The results do not seem that random to you. You think it is a little fishy that sophomores organized the raffle and also won the most prizes. Your school is composed of 30% freshmen, 25% sophomores, 25% juniors, and 20% seniors.

- a. What are the expected frequencies of winners from each class?
- b. Conduct a significance test to determine whether the winners of the prizes were distributed throughout the classes as would be expected based on the percentage of students in each group. Report your Chi Square and p values.
- c. What do you conclude?

6. Some parents of the West Bay little leaguers think that they are noticing a pattern. There seems to be a relationship between the number on the kids' jerseys and their position. These parents decide to record what they see. The hypothetical data appear below. Conduct a Chi Square test to determine if the parents' suspicion that there is a relationship between jersey number and position is right. Report your Chi Square and p values.

	Infield	Outfield	Pitcher	Total
<b>0-9</b>	12	5	5	22
<b>10-19</b>	5	10	2	17
<b>20+</b>	4	4	7	15
<b>Total</b>	21	19	14	54

7. True/false: A Chi Square distribution with 2 df has a larger mean than a Chi Square distribution with 12 df.
8. True/false: A Chi Square test is often used to determine if there is a significant relationship between two continuous variables.
9. True/false: Imagine that you want to determine if the spinner shown below is biased. You spin it 50 times and write down how many times the arrow lands in each section. You will reject the null hypothesis at the .05 level and determine that this spinner is biased if you calculate a Chi Square value of 7.82 or higher.



*Questions from Case Studies*

SAT and GPA (SG) case study

10. (SG) Answer these items to determine if the math SAT scores are normally distributed. You may want to first standardize the scores.
- If these data were normally distributed, how many scores would you expect there to be in each of these brackets: (i) smaller than 1 SD below the mean, (ii) in between the mean and 1 SD below the mean, (iii) in between the mean and 1 SD above the mean, (iv) greater than 1 SD above the mean?
  - How many scores are actually in each of these brackets?
  - Conduct a Chi Square test to determine if the math SAT scores are normally distributed based on these expected and observed frequencies.

### Diet and Health (DH) case study

11. (DH) Conduct a Pearson Chi Square test to determine if there is any relationship between diet and outcome. Report the Chi Square and p values and state your conclusions.

The following questions are from ARTIST (reproduced with permission)



12. A study compared members of a medical clinic who filed complaints with a random sample of members who did not complain. The study divided the complainers into two subgroups: those who filed complaints about medical treatment and those who filed nonmedical complaints. Here are the data on the total number in each group and the number who voluntarily left the medical clinic. Set up a two-way table. Analyze these data to see if there is a relationship between complaint (no, yes - medical, yes - nonmedical) and leaving the clinic (yes or no).

	No Complaint	Medical Complaint	Nonmedical Complaint
Total	743	199	440
Left	22	26	28

13. Imagine that you believe there is a relationship between a person's eye color and where he or she prefers to sit in a large lecture hall. You decide to collect data from a random sample of individuals and conduct a chi-square test of independence. What would your two-way table look like? Use the information to construct such a table, and be sure to label the different levels of each category.
14. A geologist collects hand-specimen sized pieces of limestone from a particular area. A qualitative assessment of both texture and color is made with the following results. Is there evidence of association between color and texture for these limestones? Explain your answer.

Texture	Colour		
	Light	Medium	Dark
Fine	4	20	8
Medium	5	23	12
Coarse	21	23	4

15. Suppose that college students are asked to identify their preferences in political affiliation (Democrat, Republican, or Independent) and in ice cream (chocolate, vanilla, or straw- berry). Suppose that their responses are represented in the following two-way table (with some of the totals left for you to calculate).

	Chocolate	Vanilla	Strawberry	Total
Democrat	26	43	13	82
Republican	45	12	8	65
Independent	9	13	4	26
Total	80	68	25	173

- What proportion of the respondents prefer chocolate ice cream?
- What proportion of the respondents are Independents?
- What proportion of Independents prefer chocolate ice cream?
- What proportion of those who prefer chocolate ice cream are Independents?

- e. Analyze the data to determine if there is a relationship between political party preference and ice cream preference.
16. NCAA collected data on graduation rates of athletes in Division I in the mid-1980s. Among 2,332 men, 1,343 had not graduated from college, and among 959 women, 441 had not graduated.
- Set up a two-way table to examine the relationship between gender and graduation.
  - Identify a test procedure that would be appropriate for analyzing the relationship between gender and graduation. Carry out the procedure and state your conclusion

# 18. Distribution-Free Tests

by David M. Lane

## A. Benefits of Distribution-Free Tests

## B. Randomization Tests

1. Two Means
2. Two or More Means
3. Randomization Tests: Association (Pearson's  $r$ )
4. Contingency Tables (Fisher's Exact Test)

## C. Rank Randomization Tests

1. Two Means (Mann-Whitney U, Wilcoxon Rank Sum)
2. Two or More Means (Kruskal-Wallis)
3. Association (Spearman's  $\rho$ )

## D. Exercises

# Benefits

by David M. Lane

## *Prerequisites*

- Chapter 7: Normal Distributions
- Chapter 3: Shapes of Distributions
- Chapter 13: Introduction to Power
- Chapter 16: Transformations

## *Learning Objectives*

1. State how distribution-free tests can avoid an inflated Type I error rate
2. State how distribution-free tests can affect power

Most tests based on the normal distribution are said to be robust when the assumption of normality is violated. To the extent to which actual *probability values* differ from nominal probability values, the actual probability values tend to be higher than the nominal p values. For example, if the probability of a difference as extreme or more extreme were 0.04, the test might report that the probability value is 0.06. Although this sounds like a good thing because the *Type I error* rate is lower than the nominal rate, it has a serious downside: reduced *power*. When the *null hypothesis* is false, the probability of rejecting the null hypothesis can be substantially lower than it would have been if the distributions were distributed normally.

Tests assuming normality can have particularly low power when there are extreme values or outliers. A contributing factor is the sensitivity of the mean to extreme values. Although transformations can ameliorate this problem in some situations, they are not a universal solution.

Tests assuming normality often have low power for leptokurtic distributions. Transformations are generally less effective for reducing kurtosis than for reducing.

Because distribution-free tests do not assume normality, they can be less susceptible to non-normality and extreme values. Therefore, they can be more powerful than the standard tests of means that assume normality.

# Randomization Tests: Two Conditions

by David M. Lane

## Prerequisites

- Chapter 18: Permutations and Combinations
- Chapter 11: One- and Two-Tailed Tests

## Learning Objectives

1. Explain the logic of randomization tests
2. Compute a randomization test of the difference between independent groups

The data in Table 1 are from a fictitious experiment comparing an experimental group with a control group. The scores in the Experimental Group are generally higher than those in the Control Group with the Experimental Group mean of 14 being considerably higher than the Control Group mean of 4. Would a difference this large or larger be likely if the two treatments had identical effects? The approach taken by randomization tests is to consider all possible ways the values obtained in the experiment could be assigned to the two groups. Then, the location of the actual data within the list is used to assess how likely a difference that large or larger would occur by chance.

Table 1. Fictitious data.

Experimental	Control
7	0
8	2
11	5
30	9

First, consider all possible ways the 8 values could be divided into two sets of 4. We can apply the formula from the section on *Permutations and Combinations* for the number of combinations of n items taken r at a time and find that there are 70 ways.

$$C_r^n = \frac{n!}{(n-r)! r!} = \frac{8!}{(8-4)! 4!} = 70$$

Of these 70 ways of dividing the data, how many result in a difference between means of 10 or larger? From Table 1 you can see that there are two rearrangements that would lead to a bigger difference than 10: (a) the score of 7 could have been in the Control Group with the score of 9 in the Experimental Group and (b) the score of 8 could have been in the Control Group with the score of 9 in the Experimental Group. Therefore, including the actual data, there are 3 ways to produce a difference as large or larger than the one obtained. This means that if assignments to groups were made randomly, the probability of this large or a larger advantage of the Experimental Group is  $3/70 = 0.0429$ . Since only one direction of difference is considered (Experimental larger than Control), this is a one-tailed probability. The *two-tailed* probability is 0.0857 since there are 6/70 ways to arrange the data so that the absolute value of the difference between groups is as large or larger than the one obtained.

Clearly, this type of analysis would be very time consuming for even moderate sample sizes. Therefore, it is most useful for very small sample sizes.

An alternate approach made practical by computer software is to randomly divide the data into groups thousands of times and count the proportion of times the difference is as big or bigger than that found with the actual data. If the number of times the data are divided randomly is very large, then this proportion will be very close to the proportion you would get if you listed all possible ways the data could be divided.

# Randomization Tests: Two or More Conditions

by David M. Lane

## *Prerequisites*

- Chapter 18: Randomization Tests (two conditions)

## *Learning Objectives*

1. Compute a randomization test for differences among more than two conditions.

The method of randomization for testing differences among more than two means is essentially very similar to the method when there are exactly two means. Table 1 shows the data from a fictitious experiment with three groups.

Table 1. Fictitious data.

T1	T2	Control
7	14	0
8	19	2
11	21	5
12	122	9

The first step in a randomization test is to decide on a test statistic. Then we compute the proportion of the possible arrangements of the data for which that test statistic is as large as or larger than the arrangement of the actual data. When comparing several means, it is convenient to use the F ratio. The F ratio is computed not to test for significance directly, but as a measure of how different the groups are. For these data, the F ratio for a one-way ANOVA is 2.06.

The next step is to determine how many arrangements of the data result in as large or larger F ratios. There are 6 arrangements that lead to the same F of 2.06: the six arrangements of the three columns. One such arrangement is shown in Table 2. The six are:

- (1) T1, T2, Control
- (2) T1, Control, T2
- (3) T2, T1, Control
- (4) T2, Control, T1
- (5) Control, T1, T2
- (6) Control, T2, T1

For each of the 6 arrangements there are two changes that lead to a higher F ratio: swapping the 7 for the 9 (which gives an F of 2.08) and swapping the 8 for the 9 (which gives an F of 2.07). The former of these two is shown in Table 3.

Table 2. Fictitious data with data for T1 and T2 swapped

T1	Control	T2
7	14	0
8	19	2
11	21	5
12	122	9

Table 3. Data from Table 1 with the 7 and the 9 swapped.

T1	T2	Control
9	14	0
8	19	2
11	21	5
12	122	7

Thus, there are six arrangements, each with two swaps that lead to a larger F ratio. Therefore, the number of arrangements with an F as large or larger than the actual arrangement is 6 (for the arrangements with the same F) + 12 (for the arrangements with a larger F), which makes 18 in all.

The next step is to determine the total number of possible arrangements. This can be computed from the following formula:

$$Arrangements = (n!)^k = (4!)^3 = 13,824$$

where n is the number of observations in each group (assumed to be the same for all groups), and k is the number of groups. Therefore, the proportion of arrangements with an F as large or larger than the F of 2.06 obtained with the data is

$$18/13,824 = 0.0013.$$

Thus, if there were no treatment effect, it is very unlikely that an F as large or larger than the one obtained would be found.

# Randomization Tests: Association (Pearson's r)

by David M. Lane

## Prerequisites

- Chapter 14: Inferential Statistics for b and r

## Learning Objectives

1. Compute a randomization test for Pearson's r.

A significance test for Pearson's r is described in the section *inferential statistics for b and r*. The significance test described in that section assumes normality. This section describes a method for testing the significance of r that makes no distributional assumptions.

Table 1. Example data.

X	Y
1	1
2.4	2
3.8	2.3
4	3.7
11	2.5

The approach is to consider the X variable fixed and compare the correlation obtained in the actual data to the correlations that could be obtained by rearranging the Y variable. For the data shown in Table 1, the correlation between X and Y is 0.385. There is only one arrangement of Y that would produce a higher correlation. This arrangement is shown in Table 2 and the r is 0.945. Therefore, there are two arrangements of Y that lead to correlations as high or higher than the actual data.

Table 2. The example data arranged to give the highest r.

X	Y
1	1
2.4	2
3.8	2.3

4	2.5
11	3.7

The next step is to calculate the number of possible arrangements of Y. The number is simply  $N!$  where N is the number of pairs of scores. Here, the number of arrangements is  $5! = 120$ . Therefore, the probability value is  $2/120 = 0.017$ . Note that this is a one-tailed probability since it is the proportion of arrangements that give an  $r$  as large or larger. For the two-tailed probability, you would also count arrangements for which the value of  $r$  were less than or equal to -0.385. In randomization tests, the two-tailed probability is not necessarily double the one-tailed probability.

# Randomization Tests: Contingency Tables: (Fisher's Exact Test)

by David M. Lane

## *Prerequisites*

- Chapter 17: Contingency Tables

## *Learning Objectives*

1. State the situation when Fisher's exact test can be used
2. Calculate Fisher's exact test
3. Describe how conservative the Fisher exact test is relative to a Chi Square test

The chapter on Chi Square showed one way to test the relationship between two nominal variables. A special case of this kind of relationship is the difference between proportions. This section shows how to compute a significance test for a difference in proportions using a randomization test. Suppose, in a fictitious experiment, 4 subjects in an Experimental Group and 4 subjects in a Control Group are asked to solve an anagram problem. Three of the 4 subjects in the Experimental Group and none of the subjects in the Control Group solved the problem. Table 1 shows the results in a contingency table.

Table 1. Anagram Problem Data.

	Experimental	Control	Total
Solved	3	0	3
Did not Solve	1	4	5
Total	4	4	8

The significance test we are going to perform is called the Fisher Exact Test. The basic idea is to take the row totals and column totals as “given” and add the probability of obtaining the pattern of frequencies obtained in the experiment and the probabilities of all other patterns that reflect a greater difference between conditions. The formula for obtaining any given pattern of frequencies is:

$$\frac{n! (N - n)! R! (N - R)!}{r! (n - r)! (R - r)! (N - n - R + r)! N!}$$

where N is the total sample size (8), n is the sample size for the first group (4), r is the number of successes in the first group (3), and R is the total number of successes (3). For this example, the probability is

$$\frac{4! (8 - 4)! 3! (8 - 3)!}{3! (4 - 3)! (3 - 3)! (8 - 4 - 3 + 3)! 8!} = 0.0714$$

Since more extreme outcomes do not exist given the row and column totals, the p value is 0.0714. This is a one-tailed probability since it only considers outcomes as extreme or more extreme favoring the Experimental Group. An equally extreme outcome favoring the Control Group is shown in Table 2, which also has a probability of 0.0714. Therefore, the two-tailed probability is 0.1428. Note that in the Fisher Exact Test, the two-tailed probability is not necessarily double the one-tailed probability.

Table 2. Anagram Problem Favoring Control Group.

	Experimental	Control	Total
Solved	0	3	3
Did not Solve	4	1	5
Total	4	4	8

The Fisher Exact Test is “exact” in the sense that it is not based on a statistic that is approximately distributed as, for example, Chi Square. However, because it assumes that both marginal totals are fixed, it can be considerably less powerful than the Chi Square test. Even though the Chi Square test is an approximate test, the approximation is quite good in most cases and tends to have too low a Type I error rate more often than too high a Type I error rate.

# Rank Randomization: Two Conditions (Mann-Whitney U, Wilcoxon Rank Sum)

by David M. Lane

## *Prerequisites*

- Chapter 5: Permutations and Combinations
- Chapter 17: Randomization Tests for Two Conditions

## *Learning Objectives*

1. State the difference between a randomization test and a rank randomization test
2. Describe why rank randomization tests are more common
3. Be able to compute a Mann-Whitney U test

The major problem with randomization tests is that they are very difficult to compute. Rank randomization tests are performed by first converting the scores to ranks and then computing a randomization test. The primary advantage of rank randomization tests is that there are tables that can be used to determine significance. The disadvantage is that some information is lost when the numbers are converted to ranks. Therefore, rank randomization tests are generally less powerful than randomization tests based on the original numbers.

There are several names for rank randomization tests for differences in central tendency. The two most common are the Mann-Whitney U test and the Wilcoxon Rank Sum Test

Consider the data shown in Table that were used as an example in the section on *randomization tests*.

Table 1. Fictitious data.

Experimental	Control
7	0
8	2
11	5
30	9

A rank randomization test on these data begins by converting the numbers to ranks.

Table 2. Fictitious data converted to ranks. Rank sum = 24.

Experimental	Control
4	1
5	2
7	3
8	6

The probability value is determined by computing the proportion of the possible arrangements of these ranks that result in a difference between ranks of as large or larger than those in the actual data (Table 2). Since the sum of the ranks (the numbers 1-8) is a constant (36 in this case), we can use the computational shortcut of finding the proportion of arrangements for which the sum of the ranks in the Experimental Group is as high or higher than the sum here ( $4 + 5 + 7 + 8 = 24$ ).

First, consider how many ways the 8 values could be divided into two sets of 4. We can apply the formula from the section on *Permutations and Combinations* for the number of combinations of  $n$  items taken  $r$  at a time ( $n$  = the total number of observations;  $r$  = the number of observations in the first group) and find that there are 70 ways.

$$C_r^n = \frac{n!}{(n-r)! r!} = \frac{8!}{(8-4)! 4!} = 70$$

Of these 70 ways of dividing the data, how many result in a sum of ranks of 24 or more? Tables 3-5 show three rearrangements that would lead to a rank sum of 24 or larger.

Table 3. Rearrangement of data converted to ranks. Rank sum = 26.

Experimental	Control
6	1
5	2
7	3
8	4

Table 4. Rearrangement of data converted to ranks. Rank sum = 25.

Experimental	Control
4	1
6	2
7	3
8	5

Therefore, the actual data represent 1 arrangement with a rank sum of 24 or more and the 3 arrangements represent three others. Therefore, there are 4 arrangements with a rank sum of 24 or more. This makes the probability equal to  $4/70 = 0.057$ . Since only one direction of difference is considered (Experimental larger than Control), this is a *one-tailed* probability. The *two-tailed* probability is  $(2)(0.057) = 0.114$ , since there are  $8/70$  ways to arrange the data so that the sum of the ranks is either (a) as large or larger or (b) as small or smaller than the sum found for the actual data.

The beginning of this section stated that rank randomization tests were easier to compute than randomization tests because tables are available for rank randomization tests. Table 6 can be used to obtain the critical values for equal sample sizes of 4-10.

For the present data, both  $n_1$  and  $n_2 = 4$  so, as can be determined from the table, the rank sum for the Experimental Group must be at least 25 for the difference to be significant at the 0.05 level (one-tailed). Since the sum of ranks equals 24, the probability value is somewhat above 0.05. In fact, by counting the arrangements with the sum of ranks greater than or equal to 24, we found that the probability value is 0.057. Naturally a table can only give the critical value rather than the p value itself. However, with a larger sample size, such as 10 subjects per group, it becomes very time consuming to count all arrangements equalling or exceeding the rank sum of the data. Therefore, for practical reasons, the critical value sometimes suffices.

Table 5. Rearrangement of data converted to ranks. Rank sum = 24.

Experimental	Control
3	1
6	2
7	4
8	5

Table 6. Critical values.

One-Tailed Test							
Rank Sum for Higher Group							
n1	n2	0.2	0.1	0.05	0.025	0.01	0.005
4	4	22	23	25	26	.	.
5	5	33	35	36	38	39	40
6	6	45	48	50	52	54	55
7	7	60	64	66	69	71	73
8	8	77	81	85	87	91	93
9	9	96	101	105	109	112	115
10	10	117	123	128	132	136	139

For larger sample sizes than covered in the tables, you can use the following expression that is approximately normally distributed for moderate to large sample sizes.

$$Z = \frac{W_a - n_a(n_a + n_b + 1)/2}{\sqrt{n_a n_b (n_a + n_b + 1)/12}}$$

where:

$W_a$  is the sum of the ranks for the first group  
 $n_a$  is the sample size for the first group  
 $n_b$  is the sample size for the second group  
 $Z$  is the test statistic

The probability value can be determined from  $Z$  using the *normal distribution calculator*.

The data from the *Stereograms Case Study* can be analyzed using this test. For these data, the sum of the ranks for Group 1 (  $W_a$  ) is 1911, the sample size for Group 1 (  $n_a$  ) is 43, and the sample size for Group 2 (  $n_b$  ) is 35. Plugging these values into the formula results in a Z of 2.13, which has a two-tailed p of 0.033.

# Rank Randomization: Two or More Conditions (Kruskal-Wallis)

by David M. Lane

## *Prerequisites*

- Chapter 17: Chi Square Distribution
- Chapter 18: Randomization Test for Two or More Conditions
- Chapter 18: Rand Randomization (Two Groups)

## *Learning Objectives*

1. Compute the Kruskal-Wallis test

The Kruskal-Wallis test is a rank-randomization test that extends the Wilcoxon test to designs with more than two groups. It tests for differences in central tendency in designs with one between-subjects variable. The test is based on a statistic  $H$  that is approximately distributed as Chi Square. The formula for  $H$  is shown below:

$$H = -3(N + 1) + \frac{12}{N(N + 1)} \sum_{i=1}^k \frac{T_i^2}{n_i}$$

where

$N$  is the total number of observations,  
 $T_i$  is the sum of ranks for the  $i^{\text{th}}$  group,  
 $n_i$  is the sample size for the  $i^{\text{th}}$  group,  
 $k$  is the number of groups.

The first step is to convert the data to ranks (ignoring group membership) and then find the sum of the ranks for each group. Then, compute  $H$  using the formula above. Finally, the significance test is done using a Chi Square distribution with  $k-1$  degrees of freedom.

For the “Smiles and Leniency” case study, the sum of the ranks for the four conditions are:

False:	2732.0
Felt:	2385.5

Miserable: 2424.5  
Neutral: 1776.0

Note that since there are “ties” in the data, the mean rank of the ties is used. For example, there were 10 scores of 2.5 which tied for ranks 4-13. The average of the numbers 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 is 8.5. Therefore, all values of 2.5 were assigned ranks of 8.5.

The sample size for each group is 34.

$$H = -3(136 + 1) + \frac{12}{(136)(137)} \left( \frac{2732^2}{34} + \frac{2385.5^2}{34} + \frac{2424.5^2}{34} + \frac{1776^2}{34} \right) = 9.28$$

Using the *Chi Square Calculator* ([external link](#); requires Java) for Chi Square = 9.28 with  $4-1 = 3$  df results in a p value of 0.028. Thus the null hypothesis of no leniency effect can be rejected.

# Rank Randomization for Association (Spearman's $\rho$ )

by David M. Lane

## Prerequisites

- Chapter 4: Values of Pearson's  $r$
- Chapter 18: Randomization Test for Pearson's  $r$

## Learning Objectives

1. Compute Spearman's  $\rho$
2. Test Spearman's  $\rho$  for significance

The rank randomization test for association is equivalent to the *randomization test for Pearson's  $r$*  except that the numbers are converted to ranks before the analysis is done. Table 1 shows 5 values of  $X$  and  $Y$ . Table 2 shows these same data converted to ranks (separately for  $X$  and  $Y$ ).

Table 1. Example data.

X	Y
1	1
2.4	2
3.8	2.3
4	3.7
11	2.5

Table 2. Ranked data.

X	Y
1	1
2	2
3	3
4	5
5	4

The approach is to consider the X variable fixed and compare the correlation obtained in the actual ranked data to the correlations that could be obtained by rearranging the Y variable. For the data shown in Table 2, the correlation between X and Y is 0.90. The correlation of ranks is called “Spearman's  $\rho$ .”

There is only one arrangement of Y that produces a higher correlation than 0.90: A correlation of 1.0 results if the fourth and fifth observations' Y values are switched (see Table 3). There are also three other arrangements that produce an  $r$  of 0.90 (see Tables 4, 5, and 6). Therefore, there are five arrangements of Y that lead to correlations as high or higher than the actual ranked data (Tables 2 through 6).

The next step is to calculate the number of possible arrangements of Y. The number is simply  $N!$ , where N is the number of pairs of scores. Here, the number of arrangements is  $5! = 120$ . Therefore, the probability value is  $5/120 = 0.042$ . Note that this is a one-tailed probability since it is the proportion of arrangements that give a correlation as large or larger. The two-tailed probability is 0.084.

Since it is hard to count up all the possibilities when the sample size is even moderately large, it is convenient to have a table of critical values.

From the table shown below, you can see that the critical value for a one-tailed test with 5 observations at the 0.05 level is 0.90. Since the correlation for the sample data is 0.90, the association is significant at the 0.05 level (one-tailed). As shown above, the probability value is 0.042. Since the critical value for a two-tailed test is 1.0, Spearman's  $\rho$  is not significant in a two-tailed test.

N	.05 2-tail	.01 2-tail	.05 1-tail	.01 1-tail
5	1		0.9	1
6	0.886	1	0.829	0.943
7	0.786	0.929	0.714	0.893
8	0.738	0.881	0.643	0.833
9	0.7	0.833	0.6	0.783
10	0.648	0.794	0.564	0.745
11	0.618	0.755	0.536	0.709
12	0.587	0.727	0.503	0.671
13	0.56	0.703	0.484	0.648
14	0.538	0.675	0.464	0.622
15	0.521	0.654	0.443	0.604
16	0.503	0.635	0.429	0.582
17	0.485	0.615	0.414	0.566
18	0.472	0.6	0.401	0.55
19	0.46	0.584	0.391	0.535

20	0.447	0.57	0.38	0.52
21	0.435	0.556	0.37	0.508
22	0.425	0.544	0.361	0.496
23	0.415	0.532	0.353	0.486
24	0.406	0.521	0.344	0.476
25	0.398	0.511	0.337	0.466
26	0.39	0.501	0.331	0.457
27	0.382	0.491	0.324	0.448
28	0.375	0.483	0.317	0.44
29	0.368	0.475	0.312	0.433
30	0.362	0.467	0.306	0.425
31	0.356	0.459	0.301	0.418
32	0.35	0.452	0.296	0.412
33	0.345	0.446	0.291	0.405
34	0.34	0.439	0.287	0.399
35	0.335	0.433	0.283	0.394
36	0.33	0.427	0.279	0.388
37	0.325	0.421	0.275	0.383
38	0.321	0.415	0.271	0.378
39	0.317	0.41	0.267	0.373
40	0.313	0.405	0.264	0.368
41	0.309	0.4	0.261	0.364
42	0.305	0.395	0.257	0.359
43	0.301	0.391	0.254	0.355
44	0.298	0.386	0.251	0.351
45	0.294	0.382	0.248	0.347
46	0.291	0.378	0.246	0.343
47	0.288	0.374	0.243	0.34
48	0.285	0.37	0.24	0.336
49	0.282	0.366	0.238	0.333
50	0.279	0.363	0.235	0.329

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 1: Levels of Measurement
- Chapter 18: Benefits
- Chapter 18: Rank Randomization for Two Conditions,

Cardiac troponins are markers of myocardial damage. The levels of troponin in subjects with and without signs of right ventricular strain in the electrocardiogram were compared in the experiment [described here](#).

The Wilcoxon rank sum test was used to test for significance. The troponin concentration in patients with signs of right ventricular strain was higher (median = 0.03 ng/ml) than in patients without right ventricular strain (median < 0.01 ng/ml),  $p < 0.001$ .

## What do you think?

Why might the authors have used the Wilcoxon test rather than a t test? Do you think the conclusions would have been different?

Perhaps the distributions were very non-normal. Typically a transformation can be done to make a distribution more normal but that is not always the case. It is almost certain the same conclusion would have been reached, although it would have been described in terms of mean differences instead of median differences.

## Exercises

### Prerequisites

All of this chapter

1. For the following data, how many ways could the data be arranged (including the original arrangement) so that the advantage of the Experimental Group mean over the Control Group mean is as large or larger than the original arrangement.

Experimental	Control
5	1
10	2
15	3
16	4
17	9

2. For the data in Problem 1, how many ways can the data be rearranged?
3. What is the one-tailed probability for a test of the difference.
4. For the following data, how many ways can the data be rearranged?

T1	T2	Control
7	14	0
8	19	2
11	21	5

5. In general, are rank randomization tests or randomization tests more powerful?
6. What is the advantage of rank randomization tests over randomization tests?
7. Test whether the differences among conditions for the data in Problem 1 is significant (one tailed) at the .01 level using a rank randomization test.

### Questions from Case Studies

SAT and GPA (SG) case study

8. (SG) Compute Spearman's  $\rho$  for the relationship between UGPA and SAT.

Stereograms (S) case study

9. (S) Test the difference in central tendency between the two conditions using a rank-randomization test (with the normal approximation) with a one-tailed test. Give the Z and the p.

Smiles and Leniency (SL) case study

10. (SL) Test the difference in central tendency between the four conditions using a rank-randomization test (with the normal approximation). Give the Chi Square and the p.

# 19. Effect Size

- A. Proportions
- B. Difference between Means
- C. Variance Explained
- D. Exercises

Researchers often seek to learn more than whether the variable under investigation has an effect and/or the direction of the effect. This is particularly true for research that has practical applications. For example, an investigation of the efficacy of a pain-relief drug would seek to determine the extent of the relief and not merely whether there was any relief. Similarly, a study of a test-preparation course's efficacy would seek to determine how much the course raises students' test scores. Finally, a study of the relationship between exercise and blood pressure would seek to determine how much blood pressure decreases for a given amount of exercise. In all of these examples, a significance test would not be sufficient since it would only provide the researcher with information about the existence and direction of the effect. It would not provide any information about the size of the effect.

Before we proceed with a discussion of how to measure effect size, it is important to consider that for some research it is the presence or absence of an effect rather than its size that is important. A controversial example is provided by Bem (2011) who investigated precognition. Bem found statistically significant evidence that subjects' responses are affected by future events. That is, he rejected the null hypothesis that there is no effect. The important question is not the size of the effect but, rather, whether it exists at all. It would be truly remarkable if future events affect present responses even a little. It is important to note that subsequent research (Ritchie, Wiseman, & French, 2012) has failed to replicate Bem's results and the likelihood that the precognition effects he described are real is very low.

# Proportions

by David M. Lane

## *Prerequisites*

- none

## *Learning Objectives*

1. Compute absolute risk reduction
2. Compute relative risk reduction
3. Compute number needed to treat

Often the interpretation of a proportion is self-evident. For example, the obesity rate for white non-Hispanic adults living in the United States was estimated by a study conducted between 2006 and 2008 to be 24%. This value of 24% is easily interpretable and indicates the magnitude of the obesity problem in this population.

Often the question of interest involves the comparison of two outcomes. For example, consider the analysis of proportions in the case study “Mediterranean Diet and Health.” In this study, one group of people followed the diet recommended by the American Heart Association (AHA), whereas a second group followed the “Mediterranean Diet.” One interesting comparison is between the proportions of people who were healthy throughout the study as a function of diet. It turned out that 0.79 of the people who followed the AHA diet and 0.90 of those who followed the Mediterranean diet were healthy. How is the effect size of diet best measured?

We will take the perspective that we are assessing the benefits of switching from the AHA diet to the Mediterranean diet. One way to assess the benefits is to compute the difference between the proportion who were not healthy on the AHA diet (0.21) with the proportion who were not healthy on the Mediterranean diet (0.10). Therefore, the difference in proportions is:

$$0.21 - 0.10 = 0.11.$$

This measure of the benefit is called the *Absolute Risk Reduction* (ARR).

To define ARR more formally, let C be the proportion of people in the control group with the ailment of interest and T be the proportion in the treatment group. ARR can then be defined as:

$$ARR = C - T$$

Alternatively, one could measure the difference in terms of percentages. For our example, the proportion of non-healthy people on the Mediterranean diet (0.10) is 52% lower than the proportion of non-healthy people on the AHA diet (0.21). This value is computed as follows:

$$(0.21 - 0.10) / 0.21 \times 100 = 52\%$$

This measure of the benefit is called the Relative Risk Reduction (RRR). The general formula for RRR is:

$$RRR = (C - T) / C \times 100$$

where C and T are defined as before.

A third commonly used measure is the “odds ratio.” For our example, the odds of being healthy on the Mediterranean diet are 90:10 = 9:1; the odds on the AHA diet are 79:21 = 3.76:1. The ratio of these two odds is 9/3.76 = 2.39. Therefore, the odds of being healthy on the Mediterranean diet is 2.39 times the odds of being healthy on the AHA diet. Note that the odds ratio is the ratio of the odds and not the ratio of the probabilities.

A fourth measure is the number of people who need to be treated in order to prevent one person from having the ailment of interest. In our example, being treated means changing from the AHA diet to the Mediterranean diet. The number who need to be treated can be defined as

$$N = 1/ARR$$

For our example,

$$N = 1/0.11 = 9$$

Therefore, one person who would otherwise not be healthy would be expected to stay healthy for every nine people changing from the AHA diet to the Mediterranean diet.

The obvious question is which of these measures is the best one. Although each measure has its proper uses, the RRR measure can exaggerate the importance of an effect, especially when the absolute risks are low. For example, if a drug

reduced the risk of a certain disease from 1 in 1,000,000 to 1 in 2,000,000, the RRR is 50%. However, since the ARR is only 0.0000005, the practical reduction in risk is minimal.

# Difference Between Two Means

by David M. Lane

## *Prerequisites*

- Chapter 3: Measures of Variability
- Chapter 12: Differences between Two Means (Independent Groups)
- Chapter 16: Chapter Log Transformations

## *Learning Objectives*

1. State how the inherent meaningfulness of the scales affects the type of measure that should be used
2. Compute  $g$
3. Compute  $d$
4. State the effect of the variability of subjects on the size of standardized measures

When the units of a measurement scale are meaningful in their own right, then the difference between means is a good and easily interpretable measure of effect size. For example, a study conducted by Holbrook, Crowther, Lotter, Cheng and King in 2000 investigated the effectiveness of benzodiazepine for the treatment of insomnia. These researchers found that, compared to a placebo, this drug increased total sleep duration by a mean of 61.8 minutes. This difference in means shows clearly the degree to which benzodiazepine is effective. (It is important to note that the drug was found to sometimes have adverse side effects.)

When the dependent variable is measured on a ratio scale, it is often informative to consider the proportional difference between means in addition to the absolute difference. For example, if in the Holbrook et al. study the mean total sleep time for the placebo group were 120 minutes, then the 61.8-minute increase would represent a 51% increase in sleep time. On the other hand, if the mean sleep time for the placebo were 420 minutes, then the 61.8-minute increase would represent a 15% increase in sleep time.

It is interesting to note that if a log transformation is applied to the dependent variable, then equal percent changes on the original scale will result in equal absolute changes on the log scale. For example, suppose the mean sleep time

increased 10% from 400 minutes to 440 in one condition and 10% from 300 to 330 minutes in a second condition. If we take the log base 10 of these values, we find that  $\text{Log}(440) - \text{Log}(400) = 2.643 - 2.602 = 0.041$  and, similarly,  $\text{Log}(330) - \text{Log}(300) = 2.518 - 2.477 = 0.041$ .

Many times the dependent variable is measured on a scale that is not inherently meaningful. For example, in the “Animal Research” case study, attitudes toward animal research were measured on a 7-point scale. The mean rating of women on whether animal research is wrong was 1.47 scale units higher than the mean rating of men. However, it is not clear whether this 1.47-unit difference should be considered a large effect or a small effect, since it is not clear exactly what this difference means.

When the scale of a dependent variable is not inherently meaningful, it is common to consider the difference between means in standardized units. That is, effect size is measured in terms of the number of standard deviations the means differ by. Two commonly used measures are Hedges'  $g$  and Cohen's  $d$ . Both of these measures consist of the difference between means divided by the standard deviation. They differ only in that Hedges'  $g$  uses the version of the standard deviation formula in which you divide by  $N-1$ , whereas Cohen's  $d$  uses the version in which you divide by  $N$ . The two formulas are given below.

$$g = \frac{M_1 - M_2}{\sqrt{MSE}}$$

$$d = g \sqrt{\frac{N}{N-2}}$$

where  $M_1$  is the mean of the first group,  $M_2$  is the mean of the second group,  $MSE$  is the mean square error, and  $N$  is the total number of observations.

Standardized measures such as Cohen's  $d$  and Hedges'  $g$  have the advantage that they are scale free. That is, since the dependent variable is standardized, the original units are replaced by standardized units and are interpretable even if the original scale units do not have clear meaning. Consider the Animal Research case study in which attitudes were measured on a 7-point scale. On a rating of whether animal research is wrong, the mean for women was 5.353, the mean for men was

3.882, and MSE was 2.864. Hedges'  $g$  can be calculated to be 0.87. It is more meaningful to say that the means were 0.87 standard deviations apart than 1.47 scale units apart since the scale units are not well defined.

It is natural to ask what constitutes a large effect. Although there is no objective answer to this question, the guidelines suggested by Cohen (1988) stating that an effect size of 0.2 is a small effect, an effect size of 0.5 is a medium effect, and an effect size of 0.8 is a large effect have been widely adopted. Based on these guidelines, the effect size of 0.87 is a large effect.

It should be noted, however, that these guidelines are somewhat arbitrary and have not been universally accepted. For example, Lenth (2001) argued that other important factors are ignored if Cohen's definition of effect size is used to choose a sample size to achieve a given level of power.

## Interpretational Issues

It is important to realize that the importance of an effect depends on the context. For example, a small effect can make a big difference if only extreme observations are of interest. Consider a situation in which a test is used to select students for a highly selective program. Assume that there are two types of students (red and blue) and that the mean for the red students is 52, the mean for the blue students is 50, both distributions are normal, and the standard deviation for each distribution is 10. The difference in means is therefore only 0.2 standard deviations and would generally be considered to be a small difference. Now assume that only students who scored 70 or higher would be selected for the program. Would there be a big difference between the proportion of blue and red students who would be able to be accepted into the program? It turns out that the proportion of red students who would qualify is 0.036 and the proportion of blue students is 0.023. Although this difference is small in absolute terms, the ratio of red to blue students who qualify is 1.6:1. This means that if 100 students were to be accepted and if equal numbers of randomly-selected red and blue students applied, 62% would be red and 38% would be blue. In most contexts this would be considered an important difference.

When the effect size is measured in standard deviation units as it is for Hedges'  $g$  and Cohen's  $d$ , it is important to recognize that the variability in the subjects has a large influence on the effect size measure. Therefore, if two experiments both compared the same treatment to a control but the subjects were much more homogeneous in Experiment 1 than in Experiment 2, then a standardized effect size measure would be much larger in the former experiment

than in the latter. Consider two hypothetical experiments on the effect of an exercise program on blood pressure. Assume that the mean effect on systolic blood pressure of the program is 10mmHg and that, due to differences in the subject populations sampled in the two experiments, the standard deviation was 20 in Experiment 1 and 30 in Experiment 2. Under these conditions, the standardized measure of effect size would be 0.50 in Experiment 1 and 0.33 in Experiment 2. This standardized difference in effect size occurs even though the effectiveness of the treatment is exactly the same in the two experiments.

# Proportion of Variance Explained

by David M. Lane

## *Prerequisites*

- Chapter 15: One-Factor ANOVA (Between Subjects)
- Chapter 14: Partitioning Sums of Squares
- Chapter 14: Multiple Regression

## *Learning Objectives*

1. State the difference in bias between  $\eta^2$  and  $\omega^2$
2. Compute  $\eta^2$
3. Compute  $\omega^2$
4. Distinguish between  $\omega^2$  and partial  $\omega^2$
5. State the bias in  $R^2$  and what can be done to reduce it

Effect sizes are often measured in terms of the proportion of variance explained by a variable. In this section, we discuss this way to measure effect size in both ANOVA designs and in correlational studies.

## **ANOVA Designs**

Responses of subjects will vary in just about every experiment. Consider, for example, the “*Smiles and Leniency*” case study. A histogram of the dependent variable “leniency” is shown in Figure 1. It is clear that the leniency scores vary considerably. There are many reasons why the scores differ. One, of course, is that subjects were assigned to four different smile conditions and the condition they were in may have affected their leniency score. In addition, it is likely that some subjects are generally more lenient than others, thus contributing to the differences among scores. There are many other possible sources of differences in leniency ratings including, perhaps, that some subjects were in better moods than other subjects and/or that some subjects reacted more negatively than others to the looks or mannerisms of the stimulus person. You can imagine that there are innumerable other reasons why the scores of the subjects could differ.

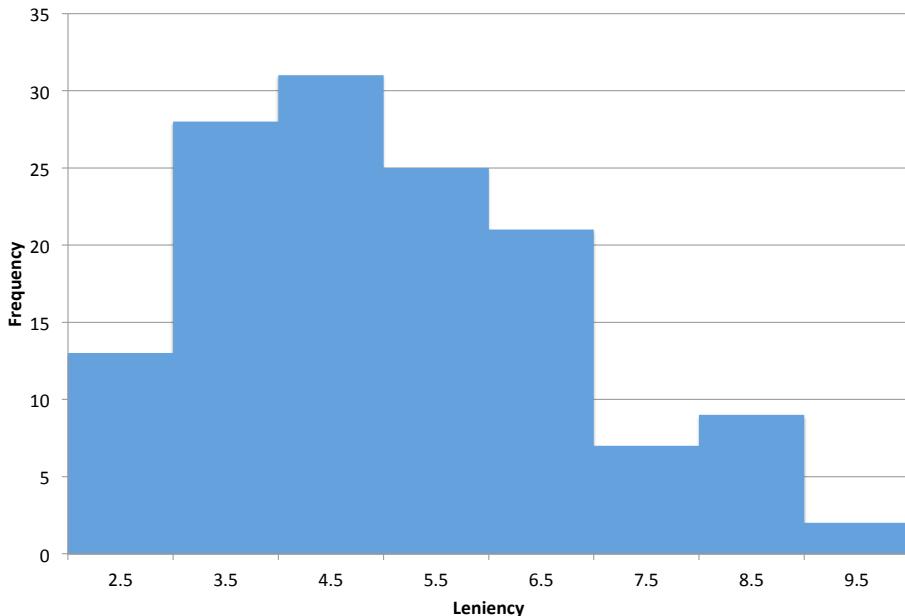


Figure 1. Distribution of leniency scores.

One way to measure the effect of conditions is to determine the proportion of the variance among subjects' scores that is attributable to conditions. In this example, the variance of scores is 2.794. The question is how this variance compares with what the variance would have been if every subject had been in the same treatment condition. We estimate this by computing the variance within each of the treatment conditions and taking the mean of these variances. For this example, the mean of the variances is 2.649. Since the mean variance within the smile conditions is not that much less than the variance ignoring conditions, it is clear that "Smile Condition" is not responsible for a high percentage of the variance of the scores. The most convenient way to compute the proportion explained is in terms of the sum of squares "conditions" and the sum of squares total. The computations for these sums of squares are shown in the chapter on ANOVA. For the present data, the sum of squares for "Smile Condition" is 27.544 and the sum of squares total is 377.189. Therefore, the proportion explained by "Smile Condition" is:

$$27.544 / 377.189 = 0.073.$$

Thus, 0.073 or 7.3% of the variance is explained by "Smile Condition."

An alternative way to look at the variance explained is as the proportion reduction in error. The sum of squares total (377.189) represents the variation when "Smile Condition" is ignored and the sum of squares error (377.189 - 27.544 =

349.654) is the variation left over when “Smile Condition” is accounted for. The difference between 377.189 and 349.654 is 27.535. The reduction in error of 27.535 represents a proportional reduction of  $27.535/377.189 = 0.073$ , the same value as computed in terms of proportion of variance explained.

This measure of effect size, whether computed in terms of variance explained or in terms of percent reduction in error, is called  $\eta^2$  where  $\eta$  is the Greek letter eta. Unfortunately,  $\eta^2$  tends to overestimate the variance explained and is therefore a biased estimate of the proportion of variance explained. As such, it is not recommended (despite the fact that it is reported by a leading statistics package).

An alternative measure,  $\omega^2$  (omega squared), is unbiased and can be computed from

$$\omega^2 = \frac{SSQ_{condition} - (k - 1)MSE}{SSQ_{total} + MSE}$$

where MSE is the mean square error and k is the number of conditions. For this example,  $k = 4$  and  $\omega^2 = 0.052$ .

It is important to be aware that both the variability of the population sampled and the specific levels of the independent variable are important determinants of the proportion of variance explained. Consider two possible designs of an experiment investigating the effect of alcohol consumption on driving ability. As can be seen in Table 1, Design 1 has a smaller range of doses and a more diverse population than Design 2. What are the implications for the proportion of variance explained by Dose? Variation due to Dose would be greater in Design 2 than Design 1 since alcohol is manipulated more strongly than in Design 1. However, the variance in the population should be greater in Design 1 since it includes a more diverse set of drivers. Since with Design 1 the variance due to Dose would be smaller and the total variance would be larger, the proportion of variance explained by Dose would be much less using Design 1 than using Design 2. Thus, the proportion of variance explained is not a general characteristic of the independent variable. Instead, it is dependent on the specific levels of the independent variable used in the experiment and the variability of the population sampled.

Table 1. Design Parameters

Design	Doses	Population
1	0.00	All Drivers between 16 and 80 Years
	0.30	
	0.60	
2	0.00 0.50 1.00	Experienced Drivers between 25 and 30 Years

## Factorial Designs

In one-factor designs, the sum of squares total is the sum of squares condition plus the sum of squares error. The proportion of variance explained is defined relative to sum of squares total. In an A x B design, there are three sources of variation (A, B, A x B) in addition to error. The proportion of variance explained for a variable (A, for example) could be defined relative to the sum of squares total ( $SSQ_A + SSQ_B + SSQ_{AXB} + SSQ_{error}$ ) or relative to  $SSQ_A + SSQ_{error}$ .

To illustrate with an example, consider a hypothetical experiment on the effects of age (6 and 12 years) and of methods for teaching reading (experimental and control conditions). The means are shown in Table 2. The standard deviation of each of the four cells (Age x Treatment combinations) is 5. (Naturally, for real data, the standard deviations would not be exactly equal and the means would not be whole numbers.) Finally, there were 10 subjects per cell resulting in a total of 40 subjects.

Table 2. Condition Means

	Treatment	
Age	Experimental	Control
6	40	42
12	50	56

The sources of variation, degrees of freedom, and sums of squares from the analysis of variance summary table as well as four measures of effect size are shown in Table 3. Note that the sum of squares for age is very large relative to the

other two effects. This is what would be expected since the difference in reading ability between 6- and 12-year-olds is very large relative to the effect of condition.

Table 3. ANOVA Summary Table

Source	df	SSQ	$\eta^2$	partial $\eta^2$	$\omega^2$	partial $\omega^2$
Age	1	1440	0.567	0.615	0.552	0.586
Condition	1	160	0.063	0.151	0.053	0.119
A x C	1	40	0.016	0.043	0.006	0.015
Error	36	900				
Total	39	2540				

First, we consider the two methods of computing  $\eta^2$ , labeled  $\eta^2$  and partial  $\eta^2$ . The value of  $\eta^2$  for an effect is simply the sum of squares for this effect divided by the sum of squares total. For example, the  $\eta^2$  for Age is  $1440/2540 = 0.567$ . As in a one-factor design,  $\eta^2$  is the proportion of the total variation explained by a variable. Partial  $\eta^2$  for Age is  $SSQ_{Age}$  divided by  $(SSQ_{Age} + SSQ_{error})$  which is  $1440/2340 = 0.615$ .

As you can see, the partial  $\eta^2$  is larger than  $\eta^2$ . This is because the denominator is smaller for the partial  $\eta^2$ . The difference between  $\eta^2$  and partial  $\eta^2$  is even larger for the effect of condition. This is because  $SSQ_{Age}$  is large and it makes a big difference whether or not it is included in the denominator.

As noted previously, it is better to use  $\omega^2$  than  $\eta^2$  because  $\eta^2$  has a positive bias. You can see that the values for  $\omega^2$  are smaller than for  $\eta^2$ . The calculations for  $\omega^2$  are shown below:

$$\omega^2 = \frac{SSQ_{effect} - df_{effect}MS_{error}}{SSQ_{total} + MS_{error}}$$

$$\omega_{partial}^2 = \frac{SSQ_{effect} - df_{effect}MS_{error}}{SSQ_{effect} + (N - df_{effect})MS_{error}}$$

where N is the total number of observations.

The choice of whether to use  $\omega^2$  or the partial  $\omega^2$  is subjective; neither one is correct or incorrect. However, it is important to understand the difference and, if you are using computer software, to know which version is being computed. (Beware, at least one software package labels the statistics incorrectly).

## Correlational Studies

In the section “*Partitioning the Sums of Squares*” in the *Regression chapter*, we saw that the sum of squares for Y (the criterion variable) can be partitioned into the sum of squares explained and the sum of squares error. The proportion of variance explained in multiple regression is therefore:

$$SSQ_{\text{explained}} / SSQ_{\text{total}}$$

In simple regression, the proportion of variance explained is equal to  $r^2$ ; in multiple regression, it is equal to  $R^2$ .

In general,  $R^2$  is analogous to  $\eta^2$  and is a biased estimate of the variance explained. The following formula for adjusted  $R^2$  is analogous to  $\omega^2$  and is less biased (although not completely unbiased):

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where N is the total number of observations and p is the number of predictor variables.

## References

- Bem, D. J. (201). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (second ed.). Lawrence Erlbaum Associates.
- Lenth, R. V. (2001) Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, 55, 187-193.
- Ritchie, S. J., Wiseman R., and French, C. C. (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE* 7.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 19:

[This article](#) describes some health effects of drinking coffee. Among the key findings were (a) women who drank four or more cups a day reduced their risk of endometrial cancer by 25% compared with those who drank less than one cup a day and (b) men who drank six or more cups had a 60% lower risk of developing the most deadly form of prostate cancer than those who drank less than one cup a day.

## What do you think?

What is the technical term for the measure of risk reduction reported? What measures of risk reduction cannot be determined from the article? What additional information would have been helpful for assessing risk reduction?

This is called the "relative risk reduction." The article does not provide information necessary to compute the absolute risk reduction, the odds ratio, or the number needed to treat. It would have been helpful if the article had reported the proportion of women drinking less than one cup a day who developed endometrial cancer as well as the analogous statistic for men and prostate cancer.

## Exercises

### *Prerequisites*

All content in this chapter

1. If the probability of a disease is .34 without treatment and .22 with treatment then what is the
  - (a) absolute risk reduction
  - (b) relative risk reduction
  - (c) Odds ratio
  - (d) Number needed to treat
2. When is it meaningful to compute the proportional difference between means?
3. The mean for an experimental group is 12, the mean for the control group were 8, the MSE from the ANOVA is 16, and N, the number of observations is 20, compute  $g$  and  $d$ .
4. Two experiments investigated the same variables but one of the experiment had subject who differed greatly from each other whereas the subjects in the other experiment were relatively homogeneous. Which experiment would likely have the larger value of  $g$ ?
5. Why is  $\omega^2$  preferable to  $\eta^2$ ?
6. What is the difference between  $\eta^2$  and partial  $\eta^2$ ?

### *Questions from Case Studies*

#### *Teacher Ratings (TR)*

7. (TR) What are the values of  $d$  and  $g$ ?
8. (TR) What are the values of  $\omega^2$  and  $\eta^2$ ?

#### *Smiles and Leniency (SL)*

9. (SL) What are the values of  $\omega^2$  and  $\eta^2$ ?

## Obesity and Bias (OB)

10. For compute  $\omega^2$  and partial  $\omega^2$  for the effect of “Weight” in a “Weight x Relatedness” ANOVA.

# 20. Case Studies

The case studies give examples of practical applications of statistical analyses. Many of the case studies contain the actual raw data. Some contain discussions of how the the data were analyzed.

**All links below are external links.**

1. [Angry Moods](#)
2. [Flatulence](#)
3. [Physicians Reactions to Patient Size](#)
4. [Teacher Ratings](#)
5. [Mediterranean Diet and Health](#)
6. [Smiles and Leniency](#)
7. [Animal Research](#)
8. [ADHD Treatment](#)
9. [Weapons and Aggression](#)
10. [SAT and College GPA](#)
11. [Stereograms](#)
12. [Driving](#)
13. [Stroop Interference](#)
14. [TV Violence](#)
15. [Bias Against Associates of the Obese](#)
16. [Shaking and Stirring Martinis](#)
17. [Adolescent Lifestyle Choices](#)
18. [Chocolate and Body Weight](#)
19. [Bedroom TV and Hispanic Children](#)
20. [Weight and Sleep Apnea](#)
21. [Misusing SEM](#)
22. [School Gardens and Vegetable Consumption](#)
23. [TV and Hypertension](#)
24. [Dietary Supplements](#)
25. [Young People and Binge Drinking](#)

- 26. Sugar Consumption in the US Diet
- 27. Nutrition Information Sources and Older Adults
- 28. Mind Set Exercise and the Placebo Effect
- 29. Predicting Present and Future Affect
- 30. Exercise and Memory
- 31. Parental Recognition of Child Obesity
- 32. Educational Attainment and Racial, Ethnic, and Gender Disparity

# 21. Glossary

## ***a priori* Comparison**

A comparison that is planned before (*a priori*) conducting the experiment or at least before the data are examined.

## **Absolute Deviation**

The absolute value of the difference between two numbers. The absolute deviation between 5 and 3 is 2; between 3 and 5 is 2; and between -4 and 2 it is 6.

## **Alternative Hypothesis**

In hypothesis testing, the null hypothesis and an alternative hypothesis are put forward. If the data are sufficiently strong to reject the null hypothesis, then the null hypothesis is rejected in favor of an alternative hypothesis. For instance, if the null hypothesis were that  $\mu_1 = \mu_2$  then the alternative hypothesis (for a two-tailed test) would be  $\mu_1 \neq \mu_2$ .

## **Analysis of Variance**

Analysis of variance is a method for testing hypotheses about means. It is the most widely-used method of statistical inference for the analysis of experimental data.

## **Antilog**

Taking the anti-log of a number undoes the operation of taking the log. Therefore, since  $\text{Log}_{10}(1000) = 3$ , the antilog<sub>10</sub> of 3 is 1,000. Taking the antilog of X raises the base of the logarithm in question to X.

## **Average**

- (i) The (arithmetic) mean
- (ii) Any measure of central tendency

## **Bar Chart**

A graphical method of presenting data. A bar is drawn for each level of a variable. The height of each bar contains the value of the variable. Bar charts are useful for displaying things such as frequency counts and percent increases. They are not recommended for displaying means (despite the widespread practice) since box plots present more information in the same amount of space.

## **Base Rate**

The true proportion of a population having some condition, attribute or disease. For example, the proportion of people with schizophrenia is about 0.01. It is very important to consider the base rate when classifying people. As the saying goes, “if you hear hoofs, think horse not zebra” since you are more likely to encounter a horse than a zebra (at least in most places.)

## **Bayes' Theorem**

Bayes' theorem considers both the prior probability of an event and the diagnostic value of a test to determine the posterior probability of the event. The theorem is shown below:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')}$$

where  $P(D|T)$  is the posterior probability of condition D given test result T,  $P(T|D)$  is the conditional probability of T given D,  $P(D)$  is the prior probability of D,  $P(T|D')$  is the conditional probability of T given not D, and  $P(D')$  is the probability of not D'.

## **Beta weight**

A standardized regression coefficient.

## **Between-Subjects Factor/Variable**

Between-subject variables are independent variables or factors in which a different group of subjects is used for each level of the variable. If an experiment is conducted comparing four methods of teaching vocabulary and if a different group of subjects is used for each of the four teaching methods, then teaching method is a between-subjects variable.

## **Bias**

1. A sampling method is biased if each element does not have an equal chance of being selected. A sample of internet users found reading an online statistics book would be a biased sample of all internet users. A random sample is unbiased. Note that possible bias refers to the sampling method, not the result. An unbiased method could, by chance, lead to a very non-representative sample.
2. An estimator is biased if it systematically overestimates or underestimates the parameter it is estimating. In other words, it is biased if the mean of the sampling distribution of the statistic is not the parameter it is estimating. The sample mean is an unbiased estimate of the population mean. The mean squared deviation of sample scores from their mean is a biased estimate of the variance since it tends to underestimate the population variance.

## **Bimodal Distribution**

A distribution with two distinct peaks.

## **Binomial Distribution**

A probability distribution for independent events for which there are only two possible outcomes such as a coin flip. If one of the two outcomes is defined as a success, then the probability of exactly x successes out of N trials (events) is given by:

$$P(x) = \frac{N!}{x!(N-x)!} \pi^x (1-\pi)^{N-x}$$

## Bin Width

Also known as the class interval, the bin width is a division of data for use in a histogram. For instance, it is possible to partition scores on a 100 point test into class intervals of 1-25, 26-49, 50-74 and 75-100.

## Bivariate

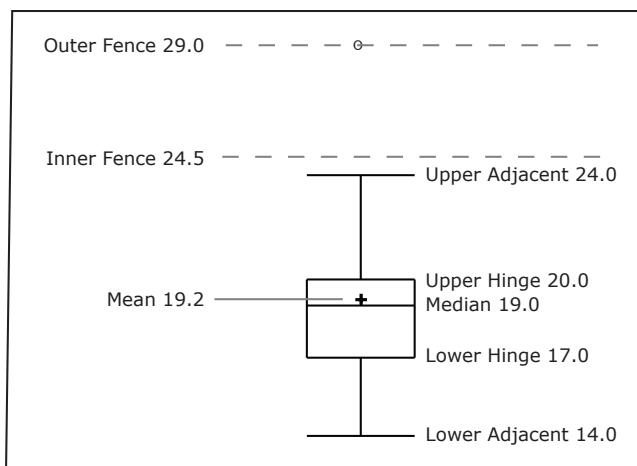
Bivariate data is data for which there are two variables for each observation. That is, two scores per subject.

## Bonferroni Correction

In general, to keep the familywise error rate (FER) at or below .05, the per-comparison error rate (PCER) should be:  $PCER = .05/c$  where  $c$  is the number of comparisons. More generally, to insure that the FER is less than or equal to alpha, use  $PCER = \alpha/c$ .

## Box Plot

One of the more effective graphical summaries of a data set, the box plot generally shows mean, median, 25th and 75th percentiles, and outliers. A standard box plot is composed of the median, upper hinge, lower hinge, higher adjacent value, lower adjacent value, outside values, and far out values. An example is shown below. Parallel box plots are very useful for comparing distributions.



## Central Tendency

There are many measures of the center of a distribution. These are called measures of central tendency. The most common are the mean, median, and, mode. Others include the trimean, trimmed mean, and geometric mean.)

## Class Frequency

One of the components of a histogram, the class frequency is the number of observations in each class interval. See also: relative frequency.

## **Class Interval**

Also known as bin width, the class interval is a division of data for use in a histogram. For instance, it is possible to partition scores on a 100 point test into class intervals of 1-25, 26-49, 50-74 and 75-100.

## **Conditional Probability**

The probability that event A occurs given that event B has already occurred is called the conditional probability of A given B. Symbolically, this is written as  $P(A|B)$ . The probability it rains on Monday given that it rained on Sunday would be written as  $P(\text{Rain on Monday} | \text{Rain on Sunday})$ .

## **Confidence Interval**

A confidence interval is a range of scores likely to contain the parameter being estimated. Intervals can be constructed to be more or less likely to contain the parameter: 95% of 95% confidence intervals contain the estimated parameter whereas 99% of 99% confidence intervals contain the estimated parameter. The wider the confidence interval, the more uncertainty there is about the value of the parameter.

## **Confounding**

Two or more variables are confounded if their effects cannot be separated because they vary together. For example, if a study on the effect of light inadvertently manipulated heat along with light, then light and heat would be confounded.

## **Cook's D**

Cook's D is a measure of the influence of an observation in regression and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

## **Constant**

A value that does not change. Values such as  $\pi$ , or the mass of the Earth are constants.

## **Continuous Variables**

Variables that can take on any value in a certain range. Time and distance are continuous; gender, SAT score and “time rounded to the nearest second” are not. Variables that are not continuous are known as discrete variables. No measured variable is truly continuous; however, discrete variables measured with enough precision can often be considered continuous for practical purposes.

## **Counterbalance**

Counterbalancing is a method of avoiding confounding among variables. Consider an experiment in which subjects are tested on both an auditory reaction time task (in which subjects respond to an auditory stimulus) and a visual reaction time task (in which subjects respond to a visual stimulus). Half of the subjects are given the visual task first and the

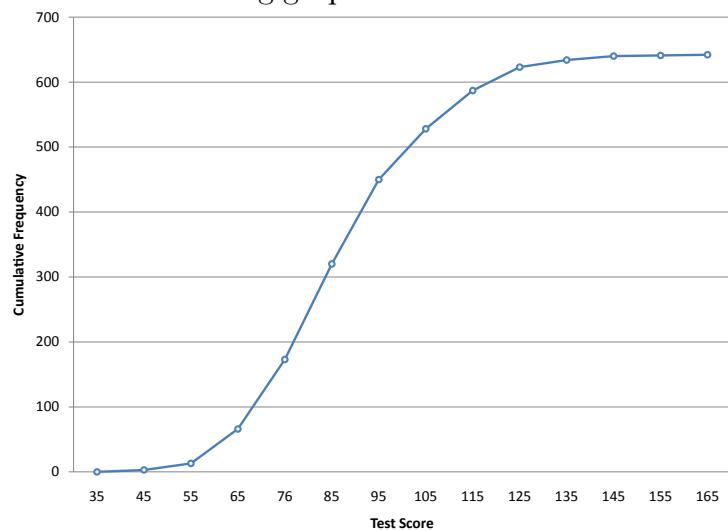
other half of the subjects are given the auditory task first. That way, there is no confounding of order of presentation and task.

### **Criterion Variable**

In regression analysis (such as linear regression) the criterion variable is the variable being predicted. In general, the criterion variable is the dependent variable.

### **Cumulative Frequency Distribution**

A distribution showing the number of observations less than or equal to values on the X-axis. The following graph shows a cumulative distribution for scores on a test.



### **Dependent Variable**

A variable that measures the experimental outcome. In most experiments, the effects of the independent variable on the dependent variables are observed. For example, if a study investigated the effectiveness of an experimental treatment for depression, then the measure of depression would be the dependent variable.

### **Descriptive Statistics**

1. The branch of statistics concerned with describing and summarizing data.
2. A set of statistics such as the mean, standard deviation, and skew that describe a distribution.

### **Deviation Scores**

Scores that are expressed as differences (deviations) from some value, usually the mean. To convert data to deviation scores typically means to subtract the mean score from each other score. Thus, the values 1, 2, and 3 in deviation-score form would be computed by subtracting the mean of 2 from each value and would be -1, 0, 1.

## **Degrees of Freedom**

The degrees of freedom of an estimate is the number of independent pieces of information that go into the estimate. In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. For example, to estimate the population variance, one must first estimate the population mean. Therefore, if the estimate of variance is based on  $N$  observations, there are  $N-1$  degrees of freedom.

## **Discrete Variables**

Variables that can only take on a finite number of values are called “discrete variables.” All qualitative variables are discrete. Some quantitative variables are discrete, such as performance rated as 1,2,3,4, or 5, or temperature rounded to the nearest degree. Sometimes, a variable that takes on enough discrete values can be considered to be continuous for practical purposes. One example is time to the nearest millisecond.

## **Distribution**

The distribution of empirical data is called a frequency distribution and consists of a count of the number of occurrences of each value. If the data are continuous, then a grouped frequency distribution is used. Typically, a distribution is portrayed using a frequency polygon or a histogram.

Mathematical equations are often used to define distributions. The normal distribution is, perhaps, the best known example. Many empirical distributions are approximated well by mathematical distributions such as the normal distribution.

## **Expected Value**

The expected value of a statistic is the mean of the sampling distribution of the statistic. It can be loosely thought of as the long-run average value of the statistic.

## **Factor (Independent Variable)**

Variables that are manipulated by the experimenter, as opposed to dependent variables. Most experiments consist of observing the effect of the independent variable(s) on the dependent variable(s).

## **Factorial Design**

In a factorial design, each level of each independent variable is paired with each level of each other independent variable. Thus, a  $2 \times 3$  factorial design consists of the 6 possible combinations of the levels of the independent variables.

## **False Positive**

A false positive occurs when a diagnostic procedure returns a positive result while the true state of the subject is negative. For example, if a test for strep says the patient has strep when in fact he or she does not, then the error in diagnosis would be called a false

positive. In some contexts, a false positive is called a false alarm. The concept is similar to a Type I error in significance testing.

### **Familywise Error Rate**

When a series of significance tests is conducted, the familywise error rate (FER) is the probability that one or more of the significance tests results in a Type I error.

### **Far Out Value**

One of the components of a box plot, far out values are those that are more than 2 steps beyond the nearest hinge. They are beyond an outer fence.

### **Favorable Outcome**

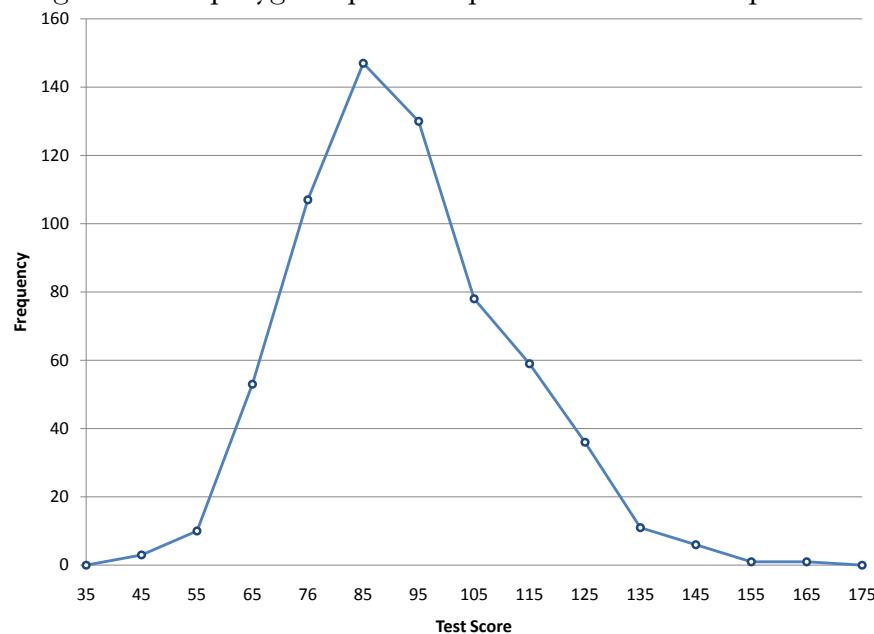
A favorable outcome is the outcome of interest. For example one could define a favorable outcome in the flip of a coin as a head. The term “favorable outcome” does not necessarily mean that the outcome is desirable – in some experiments, the favorable outcome could be the failure of a test, or the occurrence of an undesirable event.

### **Frequency Distribution**

For a discrete variable, a frequency distribution consists of the distribution of the number of occurrences for each value of the variable. For a continuous variable, it is the number of occurrences for a variety of ranges of variables.

### **Frequency Polygon**

A frequency polygon is a graphical representation of a distribution. It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the polygon's points represent the class frequencies.



## **Frequency Table**

A table containing the number of occurrences in each class of data; for example, the number of each color of M&Ms in a bag. Frequency tables often used to create histograms and frequency polygons. When a frequency table is created for a quantitative variable, a grouped frequency table is generally used.

## **Grouped Frequency Table**

A grouped frequency table shows the number of values for various ranges of scores. Below is shown a grouped frequency table for response times (in milliseconds) for a simple motor task.

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

## **Geometric Mean**

The geometric mean is a measure of central tendency. The geometric mean of  $n$  numbers is obtained by multiplying all of them together, and then taking the  $n$ th root of them. For example, for the numbers 1, 10, and 100, the product of all the numbers is:  $1 \times 10 \times 100 = 1,000$ . Since there are three numbers, we take the cubed root of the product (1,000) which is equal to 10.

## **Grouped Frequency Distribution**

A grouped frequency distribution is a frequency distribution in which frequencies are displayed for ranges of data rather than for individual values. For example, the distribution of heights might be calculated by defining one-inch ranges. The frequency of individuals with various heights rounded off to the nearest inch would then be tabulated.

## **Harmonic Mean**

The harmonic mean of  $n$  numbers ( $x_1$  to  $x_n$ ) is computed using the formula

$$n_h = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n}$$

where  $n_h$  is the harmonic mean. Often the harmonic mean of sample sizes is computed.

## **Histogram**

A histogram is a graphical representation of a distribution . It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the bars represent the class frequencies.

## **History Effect**

A problem of confounding where the passage of time, and not the variable of interest, is responsible for observed effects. See also: third variable problem.

## **Homogeneity of Variance**

The assumption that the variances of all the populations are equal.

## **Homoscedasticity**

In linear regression, the assumption that the variance around the regression line is the same for all values of the predictor variable.

## **H-Spread**

One of the components of a box plot, the H-spread is the difference between the upper hinge and the lower hinge.

## **Independence**

Two variables are said to be independent if the value of one variable provides no information about the value of the other variable. These two variables would be uncorrelated so that Pearson's r would be 0.

Two events are independent if the probability the second event occurring is the same regardless of whether or not the first event occurred.

## **Independent Events**

Events A and B are independent events if the probability of Event B occurring is the same whether or not Event A occurs. For example, if you throw two dice, the probability that the second die comes up 1 is independent of whether the first die came up 1.

Formally, this can be stated in terms of conditional probabilities:  $P(A | B) = P(A)$  and  $P(B | A) = P(B)$ .

## **Independent Variable (Factor)**

Variables that are manipulated by the experimenter, as opposed to dependent variables.

Most experiments consist of observing the effect of the independent variable(s) on the dependent variable(s).

## **Inferential Statistics**

The branch of statistics concerned with drawing conclusions about a population from a sample. This is generally done through random sampling, followed by inferences made about central tendency, or any of a number of other aspects of a distribution.

## **Influence**

Influence refers to the degree to which a single observation in regression influences the estimation of the regression parameters. It is often measured in terms how much the predicted scores for other observations would differ if the observation in question were not included.

## **Inner Fence**

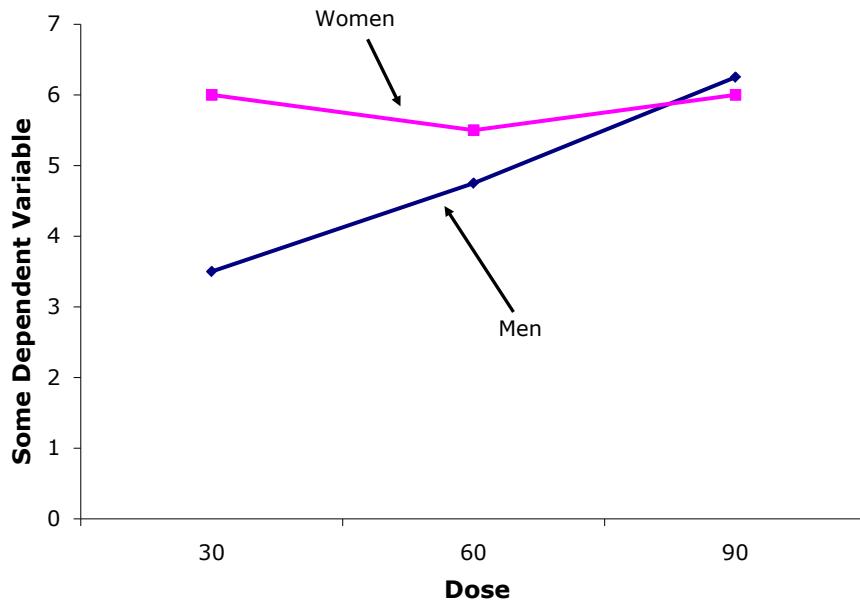
In a box plot, the lower inner fence is one step below the lower hinge while the upper inner fence is one step above the upper hinge.

## **Interaction**

Two independent variables interact if the effect of one of the variables differs depending on the level of the other variable.

## **Interaction Plot**

An interaction plot displays the levels of one variable on the X axis and has a separate line for the means of each level of the other variable. The Y axis is the dependent variable. A look at this graph shows that the effect of dosage is different for males than it is for females.



## **Interquartile Range**

The Interquartile Range (IQR) is the 75th percentile minus the 25th percentile. It is a robust measure of variability.

## **Interval Estimate**

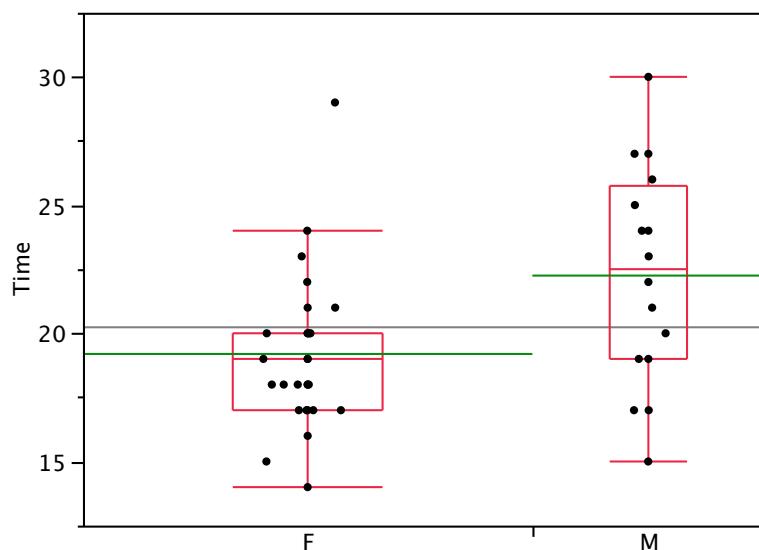
An interval estimate is a range of scores likely to contain the estimated parameter. It can be used synonymously with "confidence interval."

## Interval Scale

One of four commonly used levels of measurement, an interval scale is a numerical scale in which intervals have the same meaning throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10 degree interval has the same physical meaning (in terms of the kinetic energy). Unlike ratio scales, interval scales do not have a true zero point.

## Jitter

When points in a graph are jittered, they are moved horizontally so that all the points can be seen and none are hidden due to overlapping values. An example is shown below:



## Kurtosis

Kurtosis measures how fat or thin the tails of a distribution are relative to a normal distribution. It is commonly defined as:

$$\sum \frac{(X - \mu)^4}{N\sigma^4} - 3$$

Distributions with long tails are called leptokurtic; distributions with short tails are called platykurtic. Normal distributions have zero kurtosis.

## Leptokurtic

A distribution with long tails relative to a normal distribution is leptokurtic.

## Level

When a factor consists of various treatment conditions, each treatment condition is considered a level of that factor. For example, if the factor were drug dosage, and three doses were tested, then each dosage would be one level of the factor and the factor would have three levels.

## Levels of Measurement

Measurement scales differ in their level of measurement. There are four common levels of measurement:

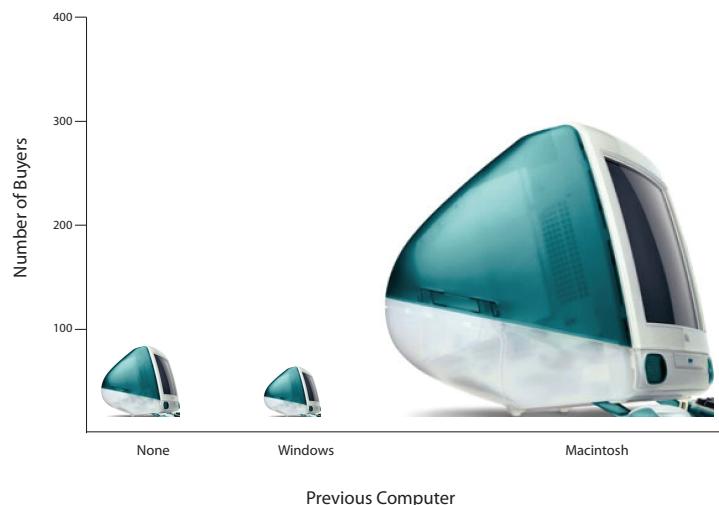
1. Nominal scales are only labels.
2. Ordinal Scales are ordered but are not truly quantitative. Equal intervals on the ordinal scale do not imply equal intervals on the underlying trait.
3. Interval scales are ordered and equal intervals equal intervals on the underlying trait. However, interval scales do not have a true zero point.
4. Ratio scales are interval scales that do have a true zero point. With ratio scales, it is sensible to talk about one value being twice as large as another, for example.

## Leverage

Leverage is a factor affecting the influence of an observation in regression. Leverage is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation.

## Lie Factor

Many problems can arise when fancy graphs are used over plain ones. Distortions can occur when the heights of objects are used to indicate the value because most people will pay attention to the areas of the objects rather than their height. The lie factor is the ratio of the effect apparent in the graph to actual effect in the data; if it deviates by more than 0.05 from 1, the graph is generally unacceptable. The lie factor in the following graph is almost 6.



## Lies

There are three types of lies:

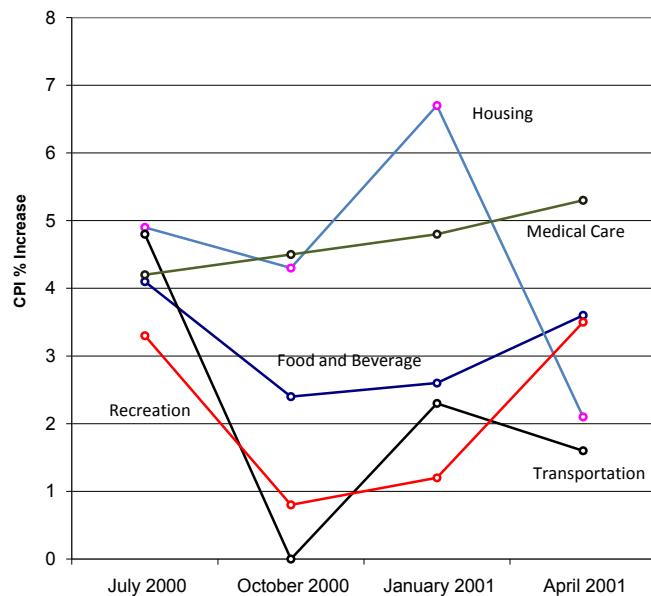
1. regular lies

2. damned lies
3. statistics

This is according to Benjamin Disraeli as quoted by Mark Twain.

### **Line Graph**

Essentially a bar graph in which the height of each bar is represented by a single point, with each of these points connected by a line. Line graphs are best used to show change over time, and should not be used if your X-axis is not an ordered variable. An example is shown below.



### **Linear Combination**

A linear combination of variables is a way of creating a new variable by combining other variables. A linear combination is one in which each variable is multiplied by a coefficient and the products summed. For example, if

$$Y = 3X_1 + 2X_2 + .5X_3$$

then  $Y$  is a linear combination of the variables  $X_1$ ,  $X_2$ , and  $X_3$ .

### **Linear Regression**

Linear regression is a method for predicting a criterion variable from one or more predictor variables. In simple regression, the criterion is predicted from a single predictor variable and the best-fitting straight line is of the form

$$Y' = bX + A$$

where  $Y'$  is the predicted score,  $X$  is the predictor variable,  $b$  is the slope, and  $A$  is the  $Y$  intercept. Typically, the criterion for the “best fitting” line is the line for which the sum of

the squared errors of prediction is minimized. In multiple regression, the criterion is predicted from two or more predictor variables.

### **Linear Relationship**

There is a perfect linear relationship between two variables if a scatterplot of the points falls on a straight line. The relationship is linear even if the points diverge from the line as long as the divergence is random rather than being systematic.

### **Linear Transformation**

A linear transformation is any transformation of a variable that can be achieved by multiplying it by a constant, and then adding a second constant. If  $Y$  is the transformed value of  $X$ , then  $Y = aX + b$ . The transformation from degrees Fahrenheit to degrees Centigrade is linear and is done using the formula:

$$C = 0.55556F - 17.7778.$$

### **Logarithm**

The logarithm of a number is the power the base of the logarithm has to be raised to in order to equal the number. If the base of the logarithm is 10 and the number is 1,000, then the log is 3 since 10 has to be raised to the 3rd power to equal 1,000.

### **Lower Adjacent Value**

A component of a box plot, the lower adjacent value is smallest value in the data above the inner lower fence.

### **Lower Hinge**

A component of a box plot, the lower hinge is the 25th percentile. The upper hinge is the 75th percentile.

### **Main Effect**

A main effect of an independent variable is the effect of the variable averaging over all levels of the other variable(s). For example, in a design with age and gender as factors, the main effect of gender would be the difference between the genders averaging across all ages used in the experiment.

### **Margin of Error**

When a statistic is used to estimate a parameter, it is common to compute a confidence interval. The margin of error is the difference between the statistic and the endpoints of the interval. For example, if the statistic were 0.6 and the confidence interval ranged from 0.4 to 0.8, then the margin of error would be 0.20. Unless otherwise specified, the 95% confidence interval is used.

## Marginal Mean

In a design with two factors, the marginal means for one factor are the means for that factor averaged across all levels of the other factor. In the table shown below, the two factors are “Relationship” and “Companion Weight.” The marginal means for each of the two levels of Relationship (Girl Friend and Acquaintance) are computed by averaging across the two levels of Companion Weight. Thus, the marginal mean for Acquaintance of 6.37 is the mean of 6.15 and 6.59.

		Companion Weight		
		Obese	Typical	Marginal Mean
Relationship	Girl Friend	5.65	6.19	<b>5.92</b>
	Acquaintance	6.15	6.59	<b>6.37</b>
	Marginal Mean	<b>5.9</b>	<b>6.39</b>	

## Mean

Also known as the arithmetic mean, the mean is typically what is meant by the word “average.” The mean is perhaps the most common measure of central tendency. The mean of a variable is given by (the sum of all its values)/(the number of values). For example, the mean of 4, 8, and 9 is 7. The sample mean is written as  $M$ , and the population mean as the Greek letter mu ( $\mu$ ). Despite its popularity, the mean may not be an appropriate measure of central tendency for skewed distributions, or in situations with outliers. Other than the arithmetic mean, there is the geometric mean and the harmonic mean.

## Median

The median is a popular measure of central tendency. It is the 50th percentile of a distribution. To find the median of a number of values, first order them, then find the observation in the middle: the median of 5, 2, 7, 9, and 4 is 5. (Note that if there is an even number of values, one takes the average of the middle two: the median of 4, 6, 8, and 10 is 7.) The median is often more appropriate than the mean in skewed distributions and in situations with outliers.

## Misses

Misses occur when a diagnostic test returns a negative result, but the true state of the subject is positive. For example, if a person has strep throat and the diagnostic test fails to indicate it, then a miss has occurred. The concept is similar to a Type II error in significance testing.

## **Mode**

The mode is a measure of central tendency. It is the most frequent value in a distribution: the mode of 3, 4, 4, 5, 5, 5, 8 is 5. Note that the mode may be very different from the mean and the median.

## **Multiple Regression**

Multiple regression is linear regression in which two or more predictor variables are used to predict the criterion.

## **Negative Association**

There is a negative association between variables X and Y if smaller values of X are associated with larger values of Y and larger values of X are associated with smaller values of Y.

## **Nominal Scales**

A nominal scale is one of four commonly-used levels of measurement. No ordering is implied, and addition/subtraction and multiplication/division would be inappropriate for a variable on a nominal scale. {Female, Male} and {Buddhist, Christian, Hindu, Muslim} have no natural ordering (except alphabetic). Occasionally, numeric values are nominal: for instance, if a variable were coded as Female = 1, Male = 2, the set {1,2} is still nominal.

## **Non-representative**

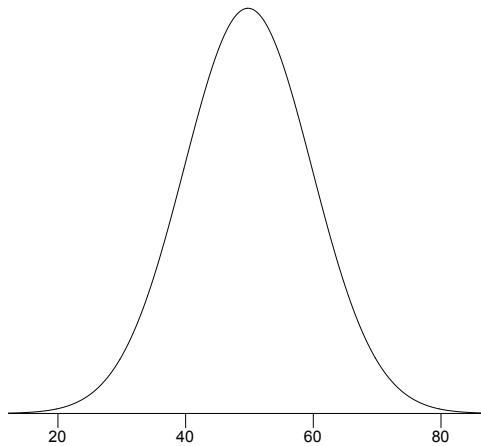
A non-representative sample is a sample that does not accurately reflect the population.

## **Normal Distribution**

One of the most common continuous distributions, a normal distribution is sometimes referred to as a “bell-shaped distribution.” If  $\mu$  is the distribution mean, and  $\sigma$  the standard deviation, then the height (ordinate) of the normal distribution is given by

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

A graph of a normal distribution with a mean of 50 and a standard deviation of 10 is shown below.



If the mean is 0 and the standard deviation is 1, the distribution is referred to as the “standard normal distribution.”

### **Null Hypothesis**

A null hypothesis is a hypothesis tested in significance testing. It is typically the hypothesis that a parameter is zero or that a difference between parameters is zero. For example, the null hypothesis might be that the difference between population means is zero.

Experimenters typically design experiments to allow the null hypothesis to be rejected.

### **Omnibus Null Hypothesis**

The null hypothesis that all population means are equal.

### **One Tailed**

The last step in significance testing involves calculating the probability that a statistic would differ as much or more from the parameter specified in the null hypothesis as does the statistics obtained in the experiment.

A probability computed considering differences in only one direction, such as the statistic is larger than the parameter, is called a one-tailed probability. For example, if a parameter is 0 and the statistic is 12, a one-tailed probability (the positive tail) would be the probability of a statistic being  $\geq$  to 12. Compare with the two-tailed probability which would be the probability of being either  $\leq -12$  or  $\geq 12$ .

### **Ordinal Scales**

One of four commonly-used levels of measurement, an ordinal scale is a set of ordered values. However, there is no set distance between scale values. For instance, for the scale: (Very Poor, Poor, Average, Good, Very Good) is an ordinal scale. You can assign

numerical values to an ordinal scale: rating performance such as 1 for “Very Poor,” 2 for “Poor,” etc, but there is no assurance that the difference between a score of 1 and 2 means the same thing as the difference between a score of 2 and 3.

### **Orthogonal Comparisons**

When comparisons among means provide completely independent information, the comparisons are called “orthogonal.” If an experiment with four groups were conducted, then a comparison of Groups 1 and 2 would be orthogonal to a comparison of Groups 3 and 4 since there is nothing in the comparison of Groups 1 and 2 that provides information about the comparison of Groups 3 and 4.

### **Outer Fence**

In a box plot, the lower outer fence is two steps below the lower hinge whereas the upper inner fence is two steps above the upper hinge.

### **Outlier**

Outliers are atypical, infrequent observations; values that have an extreme deviation from the center of the distribution. There is no universally-agreed on criterion for defining an outlier, and outliers should only be discarded with extreme caution. However, one should always assess the effects of outliers on the statistical conclusions.

### **Outside Values**

A component of a box plot, outside values are more than one step beyond the nearest hinge but not more than two steps. They are beyond an inner fence but not beyond an outer fence.

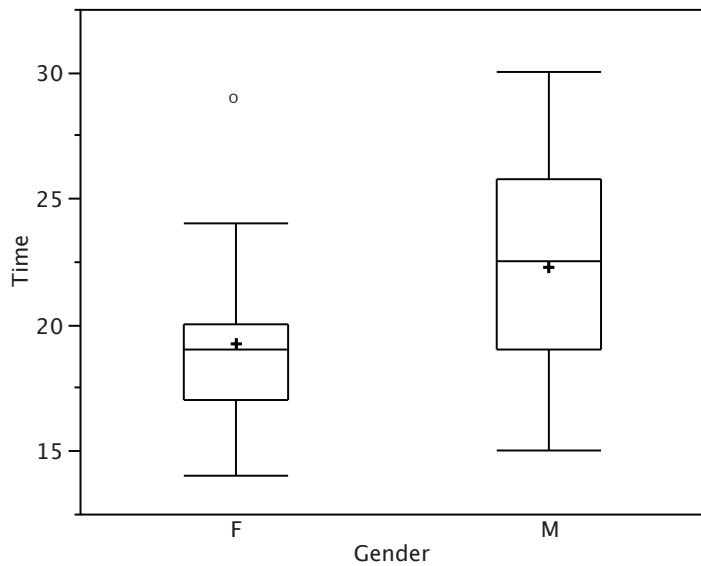
### **Pairwise Comparisons**

Two or more box plots drawn on the same Y-axis. These are often useful in comparing features of distributions. An example portraying the times it took samples of women and men to do a task is shown below.

### **Parallel Box Plots**

Two or more box plots drawn on the same Y-axis. These are often useful in comparing features of distributions. An example portraying the times it took samples of women and

men to do a task is shown below.



### **Parameter**

A value calculated in a population. For example, the mean of the numbers in a population is a parameter. Compare with a statistic, which is a value computed in a sample to estimate a parameter.

### **Partial slope**

The partial slope in multiple regression is the slope of the relationship between the part of the predictor variable that is independent of the other predictor variables and criterion. It is also the regression coefficient for the predictor variable in question.

### **Pearson's r**

Pearson's correlation is a measure of the strength of the linear relationship between two variables. It ranges from -1 for a perfect negative relationship to +1 for a perfect positive relationship. A correlation of 0 means that there is no linear relationship.

### **Percentiles**

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, some statisticians define the 65th percentile as the lowest score that is *greater* than 65% of the scores. Others have defined the 65th percentile as the lowest score that is *greater than or equal* to 65% of the scores. A more sophisticated definition is given below.

The first step is to compute the rank (R) of the percentile in question. This is done using the following formula:

$$R = P/100 \times (N + 1)$$

where  $P$  is the desired percentile and  $N$  is the number of numbers. If  $R$  is an integer, then the  $P$ th percentile is the number with rank  $R$ . When  $R$  is not an integer, we compute the  $P$ th percentile by interpolation as follows:

1. Define  $IR$  as the integer portion of  $R$  (the number to the left of the decimal point).
2. Define  $FR$  as the fractional portion of  $R$ .
3. Find the scores with Rank  $IR$  and with Rank  $IR + 1$ .
4. Interpolate by multiplying the difference between the scores by  $FR$  and add the result to the lower score.

### **Per-Comparison Error Rate**

The per-comparison error rate refers to the Type I error rate of any one significance test conducted as part of a series of significance tests. Thus, if 10 significance tests were each conducted at 0.05 significance level, then the per-comparison error rate would be 0.05. Compare with the familywise error rate.

### **Pie Chart**

A graphical representation of data, the pie chart shows relative frequencies of classes of data. It is a circle cut into a number of wedges, one for each class, with the area of each wedge proportional to its relative frequency. Pie charts are only effective for a small number of classes, and are one of the less effective graphical representations.

### **Placebo**

A device used in clinical trials, the placebo is visually indistinguishable from the study medication, but in reality has no medical effect (often, a sugar pill). A group of subjects chosen randomly takes the placebo, the others take one or another type of medication. This is done to prevent confounding the medical and psychological effects of the drug. Even a sugar pill can lead some patients to report improvement and side effects.

### **Planned Comparison**

A comparison that is planned before conducting the experiment or at least before the data are examined. Also called an *a priori* comparison.

### **Platykurtic**

A distribution with short tails relative to a normal distribution is platykurtic. See also “kurtosis.”

### **Point Estimate**

When a parameter is being estimated, the estimate can be either a single number or it can be a range of numbers such as in a confidence interval. When the estimate is a single number, the estimate is called a “point estimate.”

## **Polynomial Regression**

Polynomial regression is a form of multiple regression in which powers of a predictor variable instead of other predictor variables are used. In the following example, the criterion (Y) is predicted by X,  $X^2$  and,  $X^3$ .

$$Y = b_1X + b_2X^2 + b_3X^3 + A$$

## **Population**

A population is the complete set of observations a researcher is interested in. Contrast this with a sample which is a subset of a population. A population can be defined in a manner convenient for a researcher. For example, one could define a population as all girls in fourth grade in Houston, Texas. Or, a different population is the set of all girls in fourth grade in the United States. Inferential statistics are computed from sample data in order to make inferences about the population.

## **Positive Association**

There is a positive association between variables X and Y if smaller values of X are associated with smaller values of Y and larger values of X are associated with larger values of Y.

## **Posterior Probability**

The posterior probability of an event is the probability of the event computed following the collection of new data. One begins with a prior probability of an event and revises it in the light of new data. For example, if 0.01 of a population has schizophrenia then the probability that a person drawn at random would have schizophrenia is 0.01. This is the prior probability. If you then learn that that their score on a personality test suggests the person is schizophrenic, you would adjust your probability accordingly. The adjusted probability is the posterior probability.

## **Power**

In significance testing, power is the probability of rejecting a false null hypothesis.

## **Precision**

A statistic's precision concerns to how close it is expected to be to the parameter it is estimating. Precise statistics are vary less from sample to sample. The precision of a statistic is usually defined in terms of its standard error.

## **Predictor**

A predictor variable is a variable used in regression to predict another variable. It is sometimes referred to as an independent variable if it is manipulated rather than just measured.

## **Prior Probability**

The prior probability of an event is the probability of the event computed before the collection of new data. One begins with a prior probability of an event and revises it in

the light of new data. For example, if 0.01 of a population has schizophrenia then the probability that a person drawn at random would have schizophrenia is 0.01. This is the prior probability. If you then learn that that score on a personality test suggests the person is schizophrenic, you would adjust your probability accordingly. The adjusted probability is the posterior probability.

### **Probability Density**

For a discrete random variable, a probability distribution contains the probability of each possible outcome. However, for a continuous random variable, the probability of any one outcome is zero (if you specify it to enough decimal places). A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable. The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

### **Probability Distribution**

For a discrete random variable, a probability distribution contains the probability of each possible outcome. The sum of all probabilities is always 1.0. See binomial distribution for an example.

### **Probability Value**

In significance testing, the probability value (sometimes called the p value) is the probability of obtaining a statistic as different or more different from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming the null hypothesis is true. The lower the probability value, the stronger the evidence that the null hypothesis is false. Traditionally, the null hypothesis is rejected if the probability value is below 0.05.

### **Qualitative Variable**

Also known as categorical variables, qualitative variables are variables with no natural sense of ordering. They are therefore measured on a nominal scale. For instance, hair color (Black, Brown, Gray, Red, Yellow) is a qualitative variable, as is name (Adam, Becky, Christina, Dave . . .). Qualitative variables can be coded to appear numeric but their numbers are meaningless, as in male=1, female=2. Variables that are not qualitative are known as quantitative variables.

### **Quantitative Variable**

Variables that are measured on a numeric or quantitative scale. Ordinal, interval and ratio scales are quantitative. A country's population, a person's shoe size, or a car's speed are all quantitative variables. Variables that are not quantitative are known as qualitative variables.

## **Quantile-Quantile Plot**

A quantile-quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line.

## **Random Assignment**

Random assignment occurs when the subjects in an experiment are randomly assigned to conditions. Random assignment prevents systematic confounding of treatment effects with other variables.

## **Random Sampling**

The process of selecting a subset of a population for the purposes of statistical inference. Random sampling means that every member of the population is equally likely to be chosen.

## **Range**

The difference between the maximum and minimum values of a variable or distribution. The range is the simplest measure of variability.

## **Ratio Scale**

One of the four basic levels of measurement, a ratio scale is a numerical scale with a true zero point and in which a given size interval has the same interpretation for the entire scale. Weight is a ratio scale, Therefore, it is meaningful to say that a 200 pound person weighs twice as much as a 100 pound person.

## **Regression**

Regression means “prediction.” The regression of Y on X means the prediction of Y by X.

## **Regression Coefficient**

A regression coefficient is the slope of the regression line in simple regression or the partial slope in multiple regression.

## **Regression Line**

In linear regression, the line of best fit is called the regression line.

## **Relative Frequency**

The proportion of observations falling into a given class. For example, if a bag of 55 M & M's has 11 green M&M's, then the frequency of green M&M's is 11 and the relative frequency is  $11/55 = 0.20$ . Relative frequencies are often used in histograms, pie charts, and bar graphs.

## **Relative Frequency Distribution**

A relative frequency distribution is just like a frequency distribution except that it consists of the proportions of occurrences instead of the numbers of occurrences for each value (or range of values) of a variable.

## **Reliability**

Although there are many ways to conceive of the reliability of a test, the classical way is to define the reliability as the correlation between two parallel forms of the test. When defined this way, the reliability is the ratio of true score variance to test score variance. Chronbach's  $\alpha$  is a common measure of reliability.

## **Repeated Measures Factor**

A within-subjects variable is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects variable in which different groups of subjects are used for each level of the variable. Also called a "repeated measures variable."

## **Repeated Measures Variable**

A within-subjects variable is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects variable in which different groups of subjects are used for each level of the variable. Also called a "repeated measures factor."

## **Representative Sample**

A representative sample is a sample chosen to match the qualities of the population from which it is drawn. With a large sample size, random sampling will approximate a representative sample; stratified random sampling can be used to make a small sample more representative.

## **Robust**

Something is robust if it holds up well in the face of adversity. A measure of central tendency or variability is considered robust if it is not greatly affected by a few extreme scores. A statistical test is considered robust if it works well in spite of moderate violations of the assumptions on which it is based.

## **Sample**

A sample is a subset of a population, often taken for the purpose of statistical inference. Generally, one uses a random sample.

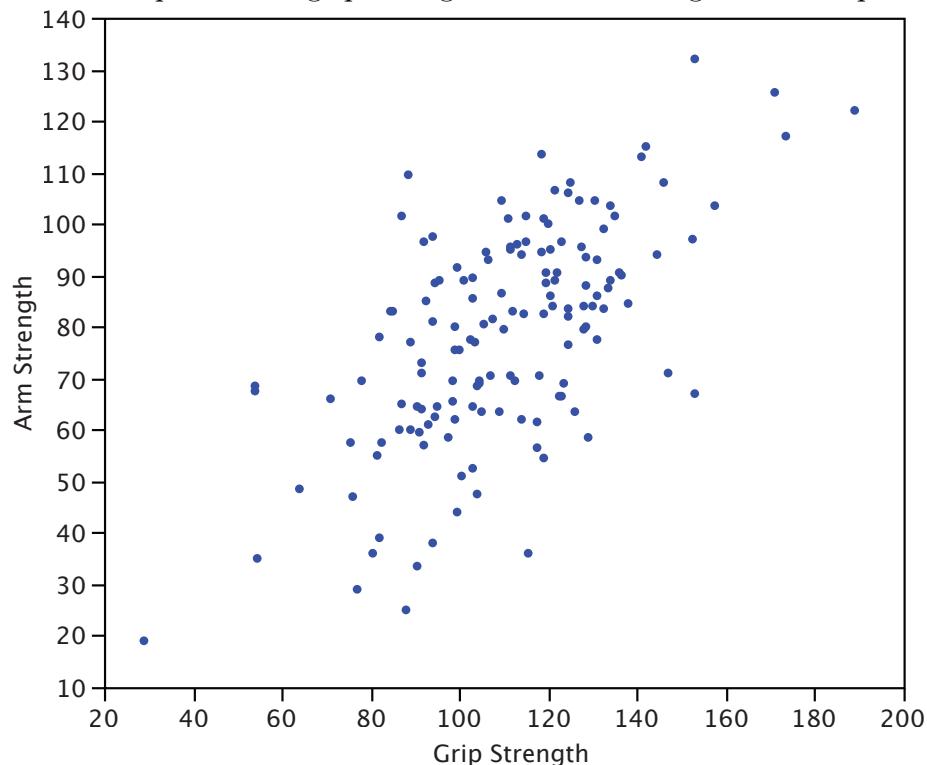
## **Sampling Distribution**

A sampling distribution can be thought of as a relative frequency distribution with a very large number of samples. More precisely, a relative frequency distribution approaches the sampling distribution as the number of samples approaches infinity. When a variable is discrete, the heights of the distribution are probabilities. When a variable is continuous,

the class intervals have no width and the heights of the distribution are probability densities.

### Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables. The scatter plot shown below illustrates the relationship between grip strength and arm strength in a sample of workers.



### Semi-Interquartile Range

The semi interquartile range is the interquartile range divided by 2. It is a robust measure of variability. The Interquartile Range is the (75th percentile – 25th percentile).

### Significance Level

In significance testing, the significance level is the highest value of a probability value for which the null hypothesis is rejected. Common significance levels are 0.05 and 0.01. If the 0.05 level is used, then the null hypothesis is rejected if the probability value is less than or equal to 0.05.

### Significance Testing

A statistical procedure that tests the viability of the null hypothesis. If data (or more extreme data) are very unlikely given that the null hypothesis is true, then the null hypothesis is rejected. If the data or more extreme data are not unlikely, then the null

hypothesis is not rejected. If the null hypothesis is rejected, then the result of the test is said to be significant. A statistically significant effect does not mean the effect is important.

### **Simple effect**

The simple effect of a factor is the effect of that factor at a single level of another factor. For example, in a design with age and gender as factors, the effect of age for females would be one of the simple effects of age.

### **Simple Regression**

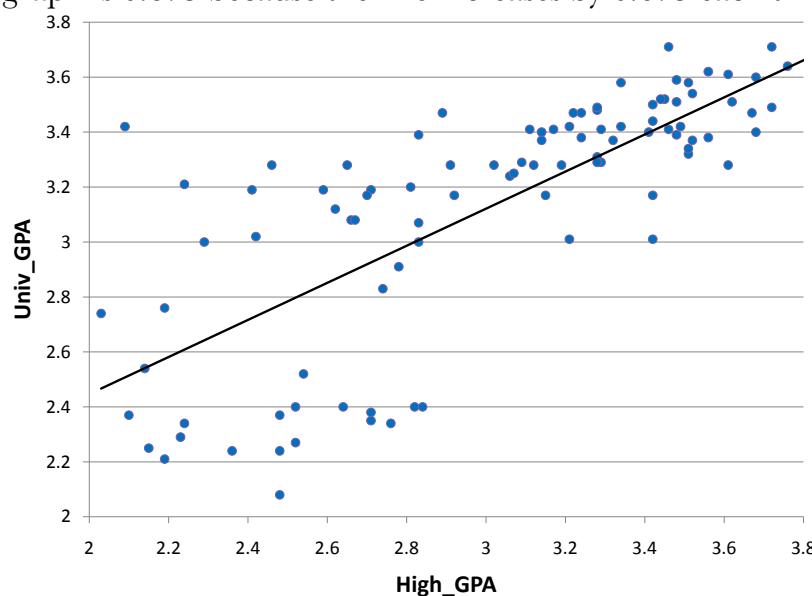
Simple regression is linear regression in which one more predictor variable is used to predict the criterion.

### **Skew**

A distribution is skewed if one tail extends out further than the other. A distribution has a positive skew (is skewed to the right) if the tail to the right is longer. It has a negative skew (skewed to the left) if the tail to the left is longer.

### **Slope**

The slope of a line is the change in Y for each change of one unit of X. It is sometimes defined as “rise over run” which is the same thing. The slope of the black line in the graph is 0.675 because the line increases by 0.675 each time X increases by 1.0.



### **Squared Deviation**

A squared deviation is the difference between two values, squared. The number that minimizes the sum of squared deviations for a variable is its mean.

### **Standard Deviation**

The standard deviation is a widely used measure of variability. It is computed by taking the square root of the variance. An important attribute of the standard deviation as a

measure of variability is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score.

### **Standard Error**

The standard error of a statistic is the standard deviation of the sampling distribution of that statistic. For example, the standard error of the mean is the standard deviation of the sampling distribution of the mean. Standard errors play a critical role in constructing confidence intervals and in significance testing.

### **Standard Error of Measurement**

In test theory, the standard error of measurement is the standard deviation of observed test scores for a given true score. It is usually estimated with the following formula in which  $s_{test}$  is the standard deviation of the test scores and  $r_{test,test}$  is the reliability of the test.

$$S_{measurement} = S_{test} \sqrt{1 - r_{test,test}}$$

### **Standard Error of the Estimate**

The standard error of the estimate is the standard deviation of the error of prediction in linear regression. It is a measure of the accuracy of prediction.

In the population is is calculated with the following formula:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

In a sample, it is estimated with:

$$s_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N - 2}}$$

### **Standard Error of the Mean**

he standard error of the mean is the standard deviation of the sampling distribution of the mean. The formula for the standard error of the mean in a population is:

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

where  $\sigma$  is the standard deviation and  $N$  is the sample size. When computed in a sample, the estimate of the standard error of the mean is:

$$s_M = \frac{s}{\sqrt{N}}$$

### **Standard Normal Distribution**

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

### **Standard Normal Deviate**

The number of standard deviations a score is from the mean of its population. The term “normal deviate” should only be used in reference to normal distributions. The transformation from a raw score  $X$  to a  $z$  score can be done using the following formula:

$$z = (X - \mu)/\sigma$$

Transforming a variable in this way is called “standardizing” the variable. It should be kept in mind that if  $X$  is not normally distributed then the transformed variable will not be normally distributed either.

### **Standardize**

A variable is standardized if it has a mean of 0 and a standard deviation of 1. The transformation from a raw score  $X$  to a standard score can be done using the following formula:

$$X_{\text{standardized}} = (X - \mu)/\sigma$$

where  $\mu$  is the mean and  $\sigma$  is the standard deviation. Transforming a variable in this way is called “standardizing” the variable. It should be kept in mind that if  $X$  is not normally distributed then the transformed variable will not be normally distributed either.

### **Statistics**

1. What you are studying right now, also known as statistical analysis, or statistical inference. It is a field of study concerned with summarizing data, interpreting data, and making decisions based on data.
2. A quantity calculated in a sample to estimate a value in a population is called a “statistic.”

### **Stem and Leaf Display**

A quasi-graphical representation of numerical data. Generally, all but the final digit of each value is a stem, the final digit is the leaf. The stems are placed in a vertical list, with each matched leaf on one side. Stem and leaf displays can be very useful for visualizing small data sets with no more than two significant digits. An example is shown below. In

this example, you multiply the stems by 10 and add the value of the leaf to obtain the numeric value. Thus the maximum number of touchdown passes is  $3 \times 10 + 7 = 37$ .

3 2337
2 001112223889
1 2244456888899
0 69

### **Step**

One of the components of a box plot, the step is 1.5 times the difference between the upper hinge and the lower hinge. See also: H-spread.

### **Stratified Random Sampling**

In stratified random sampling, the population is divided into a number of subgroups (or strata). Random samples are then taken from each subgroup with sample sizes proportional to the size of the subgroup in the population. For instance, if a population contained equal numbers of men and women, and the variable of interest is suspected to vary by gender, one might conduct stratified random sampling to insure a representative sample.

### **Studentized Range Distribution**

The studentized range distribution is used to test the difference between the largest and smallest means. It is similar to the t distribution which is used when there are only two means.

### **Sturgis' Rule**

One method of determining the number of classes for a histogram, Sturgis' rule is to take  $1 + \log_2(N)$  classes, rounded to the nearest integer.

### **Sum of Squares Error**

In linear regression, the sum of squares error is the sum of squared errors of prediction. In analysis of variance, it is the sum of squared deviations from cell means for between-subjects factors and the Subjects x Treatment interaction for within-subject factors.

### **Symmetric Distribution**

In a symmetric distribution, the upper and lower halves of the distribution are mirror images of each other. In a symmetric distribution, the mean is equal to the median.

### **t distribution**

The t distribution is the distribution of a value sampled from a normal distribution divided by an estimate of the distribution's standard deviation. In practice, the value is typically a statistic such as the mean or the difference between means and the standard

deviation is an estimate of the standard error of the statistic. The t distribution is leptokurtic.

### **t test**

Most commonly, a significance test of the difference between means based on the t distribution. Other applications include (a) testing the significance of the difference between a sample mean and a hypothesized value of the mean and (b) testing a specific contrast among means.

### **Third Variable Problem**

A type of confounding in which a third variable leads to a mistaken causal relationship between two others. For instance, cities with a greater number of churches have a higher crime rate. However, more churches do not lead to more crime, but instead the third variable, population, leads to both more churches and more crime.

### **Touchdown Pass**

In American football, a touchdown pass occurs when a completed pass results in a touchdown. The pass may be to a player in the end zone or to a player who subsequently runs into the end zone. A touchdown is worth 6 points and allows for a chance at one (and by some rules two) additional point(s).

### **Trimean**

The trimean is a robust measure of central tendency; it is a weighted average of the 25th, 50th, and 75th percentiles. Specifically it is computed as follows:

$$\text{Trimean} = 0.25 \times \text{25th} + 0.5 \times \text{50th} + 0.25 \times \text{75th}.$$

### **Trimmed Mean**

The trimmed mean is a robust measure of central tendency generally falling between the mean and the median. As in the computation of the median, all observations are ordered. Next, the highest and lowest alpha percent of the data are removed, where alpha ranges from 0 to 50. Finally, the mean of the remaining observations is taken. The trimmed mean has advantages over both the mean and median, but is analytically more intractable.

### **True Score**

A person's true score on a test is the mean score they would get if they took the test over and over again assuming no practice effects. In practice, the true score is not known but it is an important theoretical concept.

### **Tukey HSD Test**

The "Honestly Significantly Different" (HSD) test developed by the statistician John Tukey to test all pairwise comparisons among means. The test is based on the "studentized range distribution."

## **Two Tailed**

The last step in significance testing involves calculating the probability that a statistic would differ as much or more from the parameter specified in the null hypothesis as does the statistics obtained in the experiment.

A probability computed considering differences in both direction (statistic either larger or smaller than the parameter) is called two-tailed probability. For example, if a parameter is 0 and the statistic is 12, a two-tailed probability would be the probability of being either  $\leq -12$  or  $\geq 12$ . Compare with the one-tailed probability which would be the probability of a statistic being  $\geq$  to 12 if that were the direction specified in advance.

## **Type I Error**

In significance testing, the error of rejecting a true null hypothesis.

## **Type II Error**

In significance testing, the failure to reject a false null hypothesis.

## **Unbiased**

A sample is said to be unbiased when every individual has an equal chance of being chosen from the population.

An estimator is unbiased if it does not systematically overestimate or underestimate the parameter it is estimating. In other words, it is unbiased if the mean of the sampling distribution of the statistic is the parameter it is estimating. The sample mean is an unbiased estimate of the population mean.

## **Unplanned Comparison**

When the comparison among means is decided on after viewing the data, the comparison is called an “unplanned comparison” or a *post-hoc* comparison. Different statistical tests are required for unplanned comparisons than for planned comparisons.

## **Upper Hinge**

The upper hinge is one of the components of a box plot; it is the 75th percentile.

## **Upper Adjacent Value**

One of the components of a box plot, the higher adjacent value is the largest value in the data below the 75th percentile.

## **Variability**

Variability refers to the extent to which values differ from one another. That is, how much they vary. Variability can also be thought of as how spread out a distribution is. The standard deviation and the semi-interquartile range are measures of variability.

## **Variable**

Something that can take on different values. For example, different subjects in an experiment weigh different amounts. Therefore “weight” is a variable in the experiment.

Or, subjects may be given different doses of a drug. This would make “dosage” a variable. Variables can be dependent or independent, qualitative or quantitative, and continuous or discrete.

### **Variance**

The variance is a widely used measure of variability. It is defined as the mean squared deviation of scores from the mean. The formula for variance computed in an entire population is:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

where  $\sigma^2$  represents the variance,  $\mu$  is the mean, and  $N$  is the number of scores.

When computed in a sample in order to estimate the variance in the population, the formula is:

$$s^2 = \frac{\sum(X - M)^2}{N - 1}$$

where  $s^2$  is the estimate of variance,  $M$  is the sample mean, and  $N$  is the number of scores in the sample.

### **Variance Sum Law**

The variance sum law is an expression for the variance of the sum of two variables. If the variables are independent and therefore Pearson's  $r = 0$ , the following formula represents the variance of the sum and difference of the variables  $X$  and  $Y$ :

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Note that you add the variances for both  $X + Y$  and  $X - Y$ .

If  $X$  and  $Y$  are correlated, then the following formula (which the former is a special case) should be used:

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

where  $\rho$  is the population value of the correlation. In a sample  $r$  is used as an estimate of  $\rho$ .

### **Within-Subjects Design**

An experimental design in which the independent variable is a within-subjects variable.

### **Within-Subjects Factor**

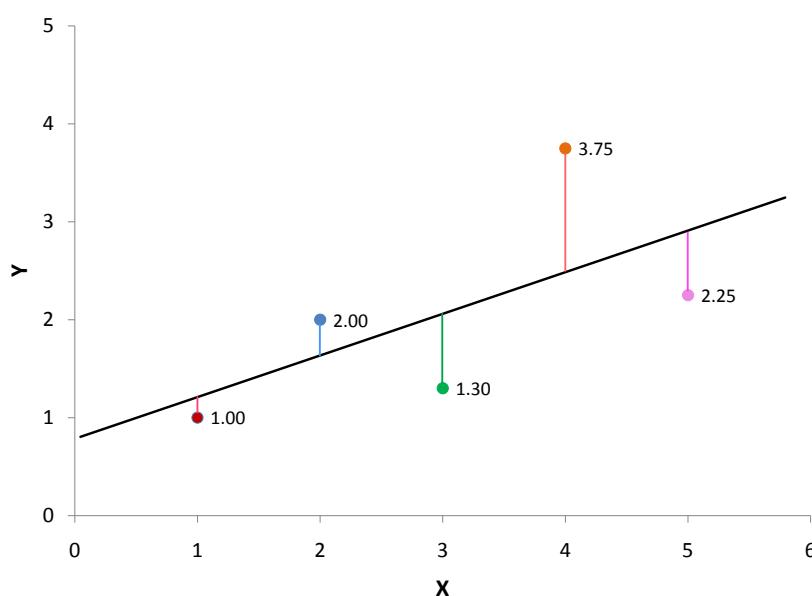
A within-subjects factor is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects factor in which different groups of subjects are used for each level of the variable.

### **Within-Subjects Variable**

A within-subjects variable is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects variable in which different groups of subjects are used for each level of the variable.

### **Y Intercept**

The Y-intercept of a line is the value of Y at the point that the line intercepts the Y axis. It is the value of Y when X equals 0. The Y intercept of the black line shown in the graph is 0.785.



### **z score**

The number of standard deviations a score is from the mean of its population. The term “standard score” is usually used for normal populations; the terms “z score” and “normal deviate” should only be used in reference to normal distributions. The transformation from a raw score  $X$  to a z score can be done using the following formula:

$$z = (X - \mu)/\sigma$$

Transforming a variable in this way is called “standardizing” the variable. It should be kept in mind that if  $X$  is not normally distributed then the transformed variable will not be normally distributed either.