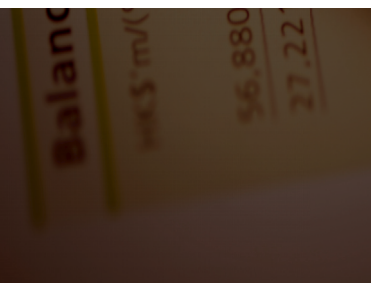
Economics Collection
Philip J. Romero and Jeffrey A. Edwards, *Editors*

Statistics for Economics

Second Edition

Shahdad Naghshpour



BEP BUSINESS EXPERT PRESS

Statistics for Economics

Statistics for Economics

Second Edition

Shahdad Naghshpour

BEP BUSINESS EXPERT PRESS

Statistics for Economics, Second Edition

First published in 2016 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-63157-389-7 (paperback)
ISBN-13: 978-1-63157-390-3 (e-book)

Business Expert Press Economics Collection

Collection ISSN: 2163-761X (print)
Collection ISSN: 2163-7628 (electronic)

Cover and interior design by S4Carlisle Publishing Services Private Ltd., Chennai, India

First edition: 2016

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

To Donna

Abstract

Statistics is the branch of mathematics that deals with real life problems. As such, it is an essential tool for economists. Unfortunately, the way the concept is introduced to students is not compatible with the way economists think and learn. The problem is worsened by the use of mathematical jargon and complex derivations. However, as this book demonstrates, neither is necessary. The book is written in simple English with minimal use of symbols, mostly for the sake of brevity and to make reading literature more meaningful.

All the examples and exercises in the book are constructed within the field of economics, thus eliminating the difficulty of learning statistics with examples from fields that have no relation to business, politics, or policy. Statistics is in fact, no more difficult than economics. Anyone that can comprehend economics can understand and use statistics successfully within this field.

In my opinion, the most important aspect of statistics is its ability to summarize the information imbedded in numerous data into few parameters and capture the essence of data. The ability of capturing the inherent core of data from seemingly random and varying bits of information is unique to statistics. It seems that somehow, statistics is able to find order in chaos.

The second edition incorporates Stata 14.1 and duplicates the answers for all the examples using Stata as well. This will enable the more serious users to be prepared for the next level when more powerful tools are necessary. In most cases no further reading is necessary to perform a more sophisticated method. In the rest of the cases additional subcommands in the form of an option is all that is needed. The book utilizes Microsoft Excel to obtain statistical results as well as to perform additional necessary computations. The spreadsheet is not the software of choice for performing sophisticated statistical analysis. However, it is widely available and almost everyone has some degree of familiarity with it. Using Excel will eliminate the need for students and readers to buy and learn new software, the need for which would itself prove to be another impediment to learning and using statistics.

Keywords

Null and alternative hypotheses, standardization, normal distribution function, statistical inference, test statistics, t distribution function, F distribution function, parameter, mean, standard deviation, interpretation and analysis, coefficient of determination, degrees of freedom, sampling distribution of sample statistics, standard error, unbiased, consistent, efficient, Central Limit Theorem, and margin of error, individual error, average error, mean squared error, analysis of variance (ANOVA).

Contents

Statistics is the Science of Finding Order in Chaos

I wrote this manuscript to share my affection for statistics and to show that comprehending statistics does not require mastery of mathematical jargon or complex formulations and derivations. I do not claim that upon learning the material in this book, you will be considered a statistician or can start a career in statistics; however, I promise you will have a much better understanding of the subject and will be able to apply its methods. I also hope you will gain the wisdom of knowing where the things you have learned will not work and realize that you have to learn new material to handle such cases.

Statistics is the science of life. It does not live outside of real life. Conclusions in statistics are probabilistic in nature compared with deterministic in most branches of mathematics. Every aspect of life benefits from statistics.

Learning statistics is like learning to play a musical instrument or learning a foreign language. Reading and comprehending the material are not sufficient. Simply reading your books and notes is not enough for learning statistics either; you also need to practice. Memorizing the material is also important. It is not sufficient to know the material or where to find it. The same is true about learning foreign languages. Unless you like to walk around with a dictionary or a statistics books under your arm, you must know the material by heart.

The objective of this book is to address the fundamentals of statistical analysis in a simple and easy-to-comprehend way. Instead of covering numerous topics, the book covers interrelated subjects that are necessary for the comprehension of the presented topics. For example, the chapter on distribution functions avoids detailed technical discussions of numerous density functions and instead focuses on comprehension of the material as needed and related to the rest of the book. On the other hand, this

book provides more details on subjects of great use in economics and business such as the mean of several rates, which is hardly mentioned in most statistics books. This knowledge is useful, for example, when trying to average several interest rates. To this end, the geometric and harmonic means are explained in detail, with examples.

The concept of "degrees of freedom" is a good example of the approach taken in the book. Customarily, degrees of freedom are defined as the denominator of sample variance without any explanation or clarification, so the reader must memorize it. In this book, the relationship of the degrees of freedom with the number of parameters that must be estimated is established and the notion is clarified by providing an example, which demonstrates the determination of how many degrees of freedom are lost. Another concept that has received much more attention than is customary is the concept of error. The notion of the error is at the heart of statistical analysis and differentiates statistics from mathematics. The hope is that such attention to detail and explanation of difficult concepts will enable the reader to go beyond mechanical generation of statistical output and instead assist in comprehension of the subject. Most software programs are capable of producing a huge amount of output with few commands, rendering the ability to compute statistical formulas by hand unnecessary. Therefore, the attention is devoted to explanation and comprehension of the subject.

The second edition has augmented the explanations in the first to clarify the subjects even more. The examples are based on economic theory utilizing actual data. The hope is that the use of theory will prove useful in relating the subject to actual empirical applications and help with research.

The only software utilized in the first edition was MS Excel to make the book accessible to more people even if they did not have dedicated statistical software. The second edition also incorporates Stata software for use by more technically oriented readers who have access to sophisticated software. The idea is not to provide a users' guide for Stata. Therefore, the commands and procedures are explained to the extent that they

pertain to the topic at hand, instead of providing a comprehensive list of the software's capabilities. Instead of providing statistical tables Excel, Stata, and the Internet are used to obtain the exact p values used in some of the formulas.

This manuscript has benefited from the tireless and dedicated contributions of Madeline Messick. Her contributions are substantial and without her perseverance the improvements to the second edition would not have been possible; any remaining shortcomings are my responsibility.

Introduction

Economics is a very interesting subject. The scope of the economic domain is vast. Economics deals with market structure, consumer behavior, investment, growth, fiscal policy, monetary policy, the roles of banks, etc. The list can go on for quite some time. It also predicts how economic agents behave in response to changes in economic and noneconomic factors such as price, income, political affiliation, and stability. Economic theory, however, is not specific in its predictions. For example, the theory proves that when the price of a good increases, the quantity supplied increases, provided all the other pertinent factors remain constant, which is also known as *ceteris paribus*.

What the theory does not and cannot state is how much the quantity supplied of a good increases for a given increase in price. The answer to this question seems to be more interesting to most people than the fact that the quantity will increase as a result of an increase in price. While the theory that explains the above relationship is important for economists, for the rest of the population, knowledge of that relationship is worthless if the magnitude of the change is unknown. Assume that for a 10% increase in price, the quantity supplied increases by 1%. This has a different consequence than if the quantity supplied increases by 10%, and totally different consequence than if the quantity supplied increases by 20%. The knowledge of the magnitude of change is as important, if not more important, than the knowledge of the direction of change. In other words, predictions are valuable when they are specific.

Statistics can answer the question of how much the quantity will increase. The science of statistics provides the necessary theories that form the foundation for answering such specific questions, including the necessary conditions to set up the study and collect data and the means to analyze and clarify the meaning of the findings. It also provides the foundation to explain the meaning of the findings using statistical inference.

In order to make an economic decision, it is necessary to know the economic conditions. This is true for all economic agents, from the

smallest to the largest. The smallest economic agent might be an individual with little disposable income, while the largest can be a multinational corporation with thousands of employees or a government.

The first step in making any economic decision is to gain knowledge of the state of the economy. Economic conditions are always in a state of flux. Sometimes it seems that we are not very concerned with mundane economic basics. For example, we may not try to forecast the price of a loaf of bread or a pound of meat. We know the average prices for these items; we consume them on a regular basis and will continue doing so as long as nothing drastic happens. However, if you were to buy a new car you would most likely call around and check some showrooms to learn about available features and prices because we tend not to have up-to-date information on big-ticket items or goods and services that we do not purchase regularly.

The process described above is a kind of sampling, and the information that you obtain is called *sample statistics*, which are used to make an informed decision about the average price of an automobile. When the process is performed according to strict and formal statistical methods, it is called *statistical inference*. This specific sample statistic is called the sample mean. The *mean* is one of numerous statistical measures at the disposal of modern economists.

Another useful measure is the sample median. The *median* is a value that divides the observations into two equal halves, one with values less than the median and the other with values more than median. Statistics explains when each measure should be used and what determines which one is the appropriate measure. For example, the median is the appropriate measure when dealing with home prices and incomes.

Other areas of application of statistics include engineering. For example, to build a bridge, it is important to use the appropriate measure to estimate the necessary capacity of the bridge. In this case neither the mean nor the median would be appropriate, as the bridge must be able to withstand the maximum load. Statistics of extreme outcomes is a main branch of the subject. The estimation would require information about variance and probability as well. In addition to identifying the appropriate tools for the task at hand, statistics also provides methods for obtaining suitable data and procedures for performing analysis to deliver the necessary inference.

One cannot imagine an economic problem that does not depend on statistical analysis. Every year, the Government Printing Office compiles the Economic Report of the President. The majority of the statistics in the report are fact-based information about different aspects of economics; however, many of the statistics are based on some statistical analysis, albeit descriptive statistics. *Descriptive statistics* provide simple yet powerful insight to economic agents and enable them to make more informed decisions.

Another component of statistical analysis is inferential statistics. *Inferential statistics* allow economists and political leaders to test hypotheses about economic conditions. For example, in the presence of inflation, the Federal Reserve Board of Governors may choose to reduce the money supply to cool down the economy and slow down the pace of inflation. The knowledge of how much to reduce the supply of money is not only based on economic theory, but also depends on proper estimation of the current state of the economy as well as the effect of a particular change in the supply of money on the final outcome.

Another widely used application of statistical analysis is in policy decisions. We hear a lot about the erosion of the middle class or that the middle class pays a larger percentage of its income in taxes than do lower and upper classes. How do we know who the middle class is? A set dollar amount of income would be inadequate to define the middle class because of inflation. However, statistical analysis has a much more meaningful and more elegant solution. The concept of interquartile range identifies the middle 50% of the population or income. Formally, the interquartile range is explained as the difference between the first and third quartile; specifically, the difference between the observations above the 75% level of income and those below the 25% represents the 50% of incomes that are in the middle.

Knowledge of statistics can also help us to identify and comprehend daily news. Recently, a report indicated that the chance of accident for teenage drivers increases by 40% when there are passengers in the car that are under 21 years of age. This is a meaningless report. Few teenagers drive alone or have passengers over 21 years of age. Total miles driven by teenagers when there are passengers less than 21 years of age far

exceeds any other types of teenage driving. Other things equal, the more you drive, the higher the probability of an accident. This example indicates that knowledge of statistics is helpful in understanding everyday events and in making sound analysis.

One of the most important aspects of statistics is the establishment of rules that allow the use of a sample to draw inferences about population parameters. Inferential statistics allows us to make decisions about the possibility of an outcome based on its probability, not dissimilar to what we do in real life anyway. If we know that a friend is usually late, we use this information to estimate his approximate arrival time. When we do this, we informally draw conclusions based on our previous experiences with the individual. In statistics the process is formal. We take random samples, and based on statistical theories of sampling distribution and the probabilities of outcomes, we make inferences and predictions about the outcomes. In essence, statistics formalizes the human experience of estimation and prediction and provides theoretical proofs for anticipated outcomes.

This book focuses on a few introductory topics in statistics and provides examples from economics. It takes a different orientation for covering the material than most other books. Chapters 1 and 2 cover descriptive statistics from tabular, graphical, and numeric points of view. A summary table of all the tools introduced in these chapters is provided in Chapter 1 to help you see the big picture of what belongs where. This grouping helps relate topics to each other. Chapter 3 provides some applications of these basic tools in different areas of economics. The purpose of Chapter 3 is to demonstrate that even simple statistics, when used properly, can be very useful and beneficial. Interestingly, some, if not most, of descriptive statistics are either intuitive or commonly utilized in everyday life. However, as the first three chapters demonstrate, it is useful to demonstrate their power using examples from economics.

Chapter 4 introduces some commonly used distribution functions. Most likely these will be new for you. These distribution functions are used as yardsticks to measure different statistics to determine whether they behave as expected or if the statistic should be considered an unusual outcome. When we sample, the resulting sample statistics, such as sample

mean, follow certain distribution functions. These important properties are discussed in Chapter 5, named "Sampling Distribution of Sample Statistics." Chapter 6 formally discusses estimation. Point estimation uses sample statistics directly, while a confidence interval provides a range that covers the population parameter with a desired level of confidence. Finally, Chapter 7 combines materials from Chapters 4 through 6 to perform statistical inference. Statistical inference is a probabilistic statement about the expected outcome of a study. Chapter 8 presents an introduction to regression analysis and Chapter 9 is the conclusion.

CHAPTER 1

Descriptive Statistics

Introduction

A simple fact of life is that most phenomena have a random component. Human beings have a natural height that is different from the natural height of a dog or a tree. However, human beings are not all of the same height. The usually small range is governed by random variation, called error. For example, the range of height for adult human beings is roughly from 52 to 75 inches. This does not mean that 100% of all mankind are in this range. The small portions that are outside this range are considered outliers. Summarizing the height of human beings is very common in statistics. However, in science it is helpful to provide the associated level of confidence in a statement. For example, it is important to state that a particular percentage, say 90%, of women living in the United States have a height between 59 and 69 inches.

One might think that it is important or maybe even necessary to provide a range that covers all cases. However, such a range may prove to be too wide to be of actual use. For example, one might be able to say with 100% certainty that the annual income in the United States is between zero and $100,000,000,000. However, although the lower end is a certainty, the upper end is not as definite. Granted that the chance of anyone making $100,000,000,000 in a year is very low, nevertheless there is no compelling reason against it. Therefore, one has to provide the probability of someone making such a huge income. Since the likelihood is very low, it would be more meaningful to state an income range for a meaningful majority, such as the income range of 95% of people.

For example, it is useful to know that 99% of all people in the United States earned less than $434,682 per individual return in 2012 (Internal Revenue Service 2014[1]), which is the same as saying that the top 1% made

at least that much per return in the same year. According to the same source, the top 10% made more than $125,195 per return. The particular percentage is not important as the choice of the top 1% versus the top 10% (or some other percentages) depends on the task at hand.

For example, the government might want to help the middle class by granting them a tax break to lower their tax burden to an equivalent burden as the upper or lower classes. One way of determining the middle class income of a population is to find the 50% of people whose incomes are in the middle. Another way of stating this is to identify the cutoff income level for the lower 25% of incomes, and the cutoff income level for the upper 25% of incomes. The two cutoffs mark the income range that contains the 50% of incomes in the middle. Computations necessary to determine these and other useful values are the subject of descriptive statistics.

Descriptive statistics provide quick and representative information about a population or a sample, such as that a typical man is 5 ft 10 in, the average high temperature on Fourth of July in Washington, DC, is 85° Fahrenheit,[2] eight out of nine runners in the men's 100-meter dash at the 2012 Olympics finished in less than 10 seconds,[3] etc. These statistics are describing something of interest about the population and condense all the facts into a single parameter. Note the subtle differences in terms such as the "most common," "typical," or "average." Descriptive statistics is the science of summarizing and condensing information in few parameters.

There are many ways of condensing information to create descriptive statistics. Different types of data require different tools. Data can be *qualitative* or *quantitative*. These naming conventions actually refer to the way *variables* are measured and not to an inherent characteristic of a phenomenon. Variables are used for statistical analysis and are measured based on their characteristics. The preferred name for qualitative variables is *categorical* variables, because the word "qualitative" has a value connotation, which is often reflected in the literature.

In many cases, analyzing qualitative and quantitative variables requires different tools, but in some cases the tools are similar for both, if not identical. However, the interpretations of qualitative and quantitative variables

are usually different. Note that a population is not defined as either qualitative or quantitative. Rather, it is the variable of interest in the population that is either qualitative or quantitative. For example, the population may consist of people. If the age of the person is of interest, then the variable is quantitative; but if the gender of the person is of interest, then the variable is qualitative. If the population under study is a firm and the variable is the firm's status as a polluter (i.e., the firm either pollutes or does not pollute), then it is a qualitative variable. However, if the amount of pollution is of interest, then it is a quantitative variable.

Definition 1.1 *Qualitative variables* are nonnumeric. They represent a label for a category of similar items. For example, the status of a firm as a polluter is a qualitative variable.

Definition 1.2 *Quantitative variables* are numerical. The distance each student travels to get to school is a quantitative variable.

Measurement Scales

Variables must be measured in a meaningful way. The following definitions provide brief descriptions of different types of measurement scales. Most of the methods in this text require interval or measurement scales with stronger relational requirements.

Definition 1.3 *Nominal* or *categorical* data are the "count" of the number of times an event occurs. Countries might be grouped according to their policy toward trade and might be classified as open or closed economies. Care must be taken to assure that each case belongs to only one group. An ID number is an example of nominal data. Since the relative size does not matter for nominal data, the customary arithmetic computations and statistical methods do not apply to these numbers.

Definition 1.4 When there are only two nominal types, the data is *dichotomous*. When there is no particular order, a dichotomous variable is called a *discrete dichotomous variable*. Gender is an example of a *discrete* dichotomous variable. Alternatively, when one can place an order on the type of data, as in the case of young and old, then the variable is a *continuous dichotomous variable*.

Definition 1.5 An *ordinal scale* indicates that data is ordered in some way.

Although orders or ranks are represented by numerical values, such values are void of content and cannot be used for typical computations such as averages. The distances between ranks are meaningless. The income of the person who is ranked 20th in a group of ordered income is not one half of the income of someone who is ranked 40th. In an ordinal scale, only the comparisons "greater," "equal," or "less" are meaningful. Ordinal scales are very important in economics, as in the case of utility and indifference curves. It is not necessary to measure the amount of utility (i.e., satisfaction or happiness) one receives from different goods and services; it is sufficient to rank consumer's utility. The customary arithmetic computations (such as adding and multiplication) and statistical methods do not apply to ordinal numbers.

Definition 1.6 A *Likert scale* is a special type of ordinal scale, where the subjects provide the ranking of each variable.

Customarily, an odd number of choices are used in ranking scales to allow the center value to represent the "neutral" case. For example, a subject is asked to rank his or her preference on a scale of 1, very low, to 5, very high. In this case, a choice of 3 would represent a neutral response, indicating no preference.

Definition 1.7 In an *interval scale,* the relative distances of any two sequential values are the same. In the interval scale, the size of the difference between measurements is also important. Each numerical scale is actually measured from an *accepted zero*. This makes use of the type of scale irrelevant as in the case of Celsius and Fahrenheit scales for temperatures. Both scales have an arbitrary zero. Some arithmetic computations such as addition and subtraction are meaningful.

Definition 1.8 A *ratio scale* provides meaningful use of the ratio of measurements in addition to interval size and order of values.

For example, the ratio of sales, gross domestic product (GDP), and output are expressed as a ratio scale. There are numerous other measurement scales, but these have little practical use in economics.

Types of Available Tools

Descriptive statistics provide summaries of information about a population or sample, both of which will be defined shortly. In real life, the amount of information available is vast, and comprehending their intrinsic value is difficult. Descriptive statistics provide some means of condensing massive amounts of information in as few *parameters* as possible.

Definition 1.9 A *population* comprises all possible values of a variable.

Definition 1.10 A *parameter* is a characteristic of a population that is of interest. Parameters are constant and usually unknown.

Examples of parameters include population mean, population variance, and regression coefficients. One of the main purposes of statistics is to obtain information from a sample that can be used to make inferences about population parameters. The estimated parameter value obtained from a sample is called a *statistic*.

Table 1.1 summarizes the descriptive methods for quantitative and qualitative variables. Note that these are only the descriptive statistics and by no means all the methods at our disposal.

Descriptive Statistics for Qualitative Variables

The available descriptive statistics for qualitative variables can be divided into graphical and tabular methods. Each one consists of several customarily used tools. In order to be able to graph the data, it must be tabulated in some fashion; therefore, we will discuss the tabular methods first.

Tabular Methods for Qualitative Variables

The most common tabular methods for qualitative variables are frequency and relative frequency.

Table 1.1 Descriptive statistics

Qualitative Variables	Tabular Methods	Frequency		
		Relative Frequency		
	Graphical Methods	Bar Graphs		
		Pie Charts		
Quantitative Variables	Tabular Methods	Frequency Distribution		
		Relative Frequency		
		Cumulative Distribution		
		Percentiles		
		Quartiles		
		Hinges		
	Graphical Methods	Histograms		
		Ogive		
		Stem and Leaf		
		Dot Plot		
		Scatter Plot		
		Box Plot		
	Numerical Methods	Measures of Location	Mean	Ungrouped Data
				Grouped Data
			Trimmed Mean	
			Median	
			Mode	
		Measures of Dispersion	Range	
			Interquartile Range	
			Variance	Ungrouped Data
				Grouped Data
			Standard Deviation	
			Coefficient of Variation	
		Measures of Association	Covariance	
			Correlation Coefficient	

Frequency Distribution for Qualitative Variables

A frequency distribution shows the frequency of occurrence for nonoverlapping classes.

Example 1.1 In a small town, a small company is responsible for refilling soda dispensers of 30 businesses. The type of business, the average number of cans of soda (in 100 cans), the gender, and race of the owner are presented in Table 1.2. Find the frequencies of the types of businesses of the soda dispensers.

Table 1.2 Some information about soda dispensers

Business Type	Average	Gender	Race
Gas Station	3.8	Male	Black
Drug Store	2.4	Male	Black
Mechanic Shop	3.4	Female	White
Sporting Goods	4	Female	White
Tire Shop	2.8	Female	White
Hardware Store	3.1	Female	White
Drug Store	2.7	Male	White
Mechanic Shop	1.8	Female	Black
Gas Station	2.6	Male	White
Hardware Store	2.8	Male	Black
School	3.7	Female	White
Mechanic Shop	4	Female	White
Hardware Store	2.6	Male	White
Sporting Goods	2.4	Female	Black
Mechanic Shop	2.6	Female	White
Gas Station	3.5	Female	Black
Mechanic Shop	2.1	Male	Black
Drug Store	3.6	Female	Black
Hardware Store	3.4	Male	White
Sporting Goods	3.2	Male	White
Hardware Store	3.5	Male	Black
Mechanic Shop	1.9	Male	White
Mechanic Shop	1.9	Female	White
Hardware Store	1.8	Female	White
Sporting Goods	1.7	Male	White
Drug Store	2.7	Female	White
Hardware Store	2.1	Female	White
Mechanic Shop	2.7	Male	White
Drug Store	3.2	Female	White
Mechanic Shop	3.7	Male	White

Solution 1.1

A frequency distribution for the variable Business Type will clarify the information. For each business, put a line next to its type and cross the lines when the count is five. This will allow a quick reference for determining the frequency.

Gas Stations	///	
Mechanic Shop	𝍸	𝍸
Drug Store	𝍸	
Hardware Store	𝍸	//
Sporting Goods	////	
School	/	
Tire Shop	/	

This is certainly an improvement in making the information more easily understandable, but Table 1.3 makes it even clearer and more condensed. It is easier to determine the locations, how many times each is restocked, as well as finding the most frequent, and the least frequent locations.

Table 1.3 Business Types and frequencies of Business Types

Business Type	Frequency
Gas Stations	3
Mechanic Shop	9
Drug Store	5
Hardware Store	7
Sporting Goods	4
School	1
Tire Shop	1
Total	**30**

The same results can be obtained from Stata using the following command:

```
tabulate businesstype
```

To obtain the results in Stata, first type table 1.2 into the data editor. Then paste the above command into the command window. If the data

is copied from Excel, the column heading "Business Type" in Excel will be changed "businesstype" in Stata due to its naming conventions (Figure 1.1).

Business Type	Freq.	Percent	Cum.
Drug Store	5	16.67	16.67
Gas Station	3	10.00	26.67
Hardware Store	7	23.33	50.00
Mechanic Shop	9	30.00	80.00
School	1	3.33	83.33
Sporting Goods	4	13.33	96.67
Tire Shop	1	3.33	100.00
Total	30	100.00	

Figure 1.1 Stata presentation of frequencies using the tabulate command

Note that Stata sorted the Business Type variable alphabetically. The column "Percent" depicts the relative frequencies and the column "Cum." displays cumulative frequencies, which will be explained shortly.

A table with 30 rows has been reduced to a two-column table with seven rows. If there were 20,000 locations, the resulting table would not be any larger as long as the number of types of business remained the same. Although no one can really understand anything from a table with 20,000 entries, the resulting frequency table would be very clear. This signifies the power of statistics to condense information in as few parameters as possible. The results can be graphed for a more visual presentation. One possible graph is called a *bar graph*. Other graphs, such as *pie charts*, are also available.

The question could have been about the gender or the race of the owner as well. The example could have been about the types of industries in a state, the kinds of automobiles produced at a plant, the kinds of services provided by a firm, or the kinds of goods sold in a store. The method of determining the frequencies would be the same in all such cases.

Relative Frequency for Qualitative Data

Except in rare occasions the magnitudes of the frequencies vary for different populations and samples, making comparison difficult. For better comparison between populations or samples, the relative frequency is used. The *relative frequency* shows the percentage that each class makes up of the total population or sample (Table 1.4). It is obtained by dividing the frequency for each class by the total number in the population or the sample.

Table 1.4 Business Types, frequencies, and relative frequencies of Business Types

Business Type	Frequency	Relative Frequency
Gas Stations	3	0.1
Mechanic Shop	9	0.3
Drug Store	5	0.166
Hardware Store	7	0.233
School	1	0.033
Sporting Goods	4	0.133
Tire Shop	1	0.033
Total	30	0.9998

The sum of the relative frequency is always 1.0. Here, however, the sum is not exactly 1 due to rounding error. Relative frequencies can be displayed in percentages by multiplying the values by 100, a common practice in many software programs, as is the case in Stata as shown in a previous example.

Graphical Methods for Qualitative Variables

The two most commonly used graphical methods for qualitative variables are bar graphs and pie charts. Many other graph types have been introduced with the advent of spreadsheet programs such as Excel and more are available in specialty software such as Stata.

Bar Graphs

A *bar graph* is a graphical representation of the frequency distribution or relative frequency distribution of qualitative data. The names of the

qualitative variables are placed on the X-axis and the frequency is depicted on the Y-axis. A *histogram* and a bar graph are identical except for the fact that the bar graph is used for qualitative variables, while the histogram is used for quantitative variables.

Example 1.2 The following table represents the frequency of the Business Type variable for 30 businesses. Provide a bar graph of the business types where soda dispensers are located.

Business Type	Frequency
Gas Stations	3
Mechanic Shop	9
Drug Store	5
Hardware Store	7
Sporting Goods	4
School	1
Tire Shop	1
Total	**30**

Solution 1.2

The bar graph in Excel is rotated 90 degrees horizontally to the right. In this case, we will use what Excel calls a "column" graph for our bar graph. The sequence of commands to plot a bar graph in Excel is provided for your reference (instructions are similar but may vary slightly for Mac users).

1. Open a new spreadsheet in Excel.
2. Enter in the data from the above table, making sure to leave a space before the line for the total. The data should be captured in cells A1 through B8.
3. Go to "Insert," which is the second tab at the top left on the spreadsheet.
4. Click on "Column" (which looks like a bar graph) and then click on the first chart (top left).
5. Excel will populate a chart similar to the one shown in Figure 1.2.

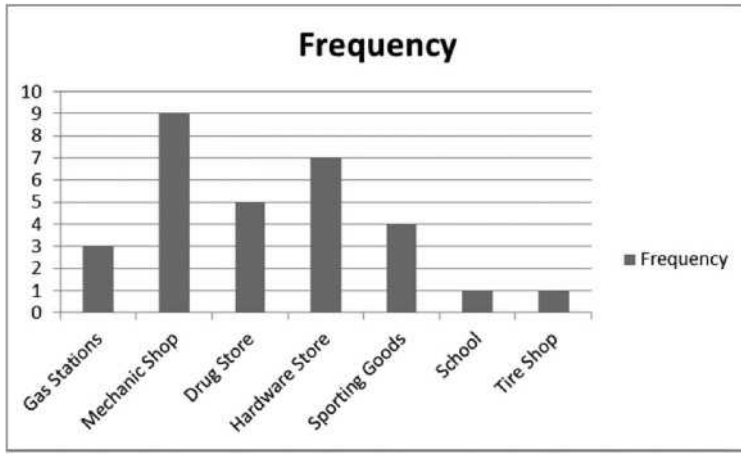


Figure 1.2 Excel display of bar graph of Business Types

In this case, creating a bar graph of the relative frequencies would not provide additional meaningful results. The graph would be identical to the above graph, except the scales on the vertical axis would be the relative frequencies (percentages) and not the actual frequencies. However, the relative frequencies are already known, so further benefit is not gained. Often it is more meaningful to plot the relative frequencies instead of the actual frequencies because you can easily compare relative frequencies, and they are similar to probabilities.

Using the option "plot" for the "tabulate" command in Stata displays a frequency table combined with a horizontally drawn bar chart (Figure 1.3). Use the ungrouped data from Table 1.2.

```
tabulate businesstype, plot
```

Business Type	Freq.	
Drug Store	5	*****
Gas Station	3	***
Hardware Store	7	*******
Mechanic Shop	9	*********
School	1	*
Sporting Goods	4	****
Tire Shop	1	*
Total	30	

Figure 1.3 Stata display of tabulate with bar graph

An alternative to this simple display is to download a subroutine, which is written by Stata users. One such subroutine is written by Cox,[4]

called "catplot" for creating bar graphs for categorical data (Figure 1.4). It is accessed by typing the following into the command line:

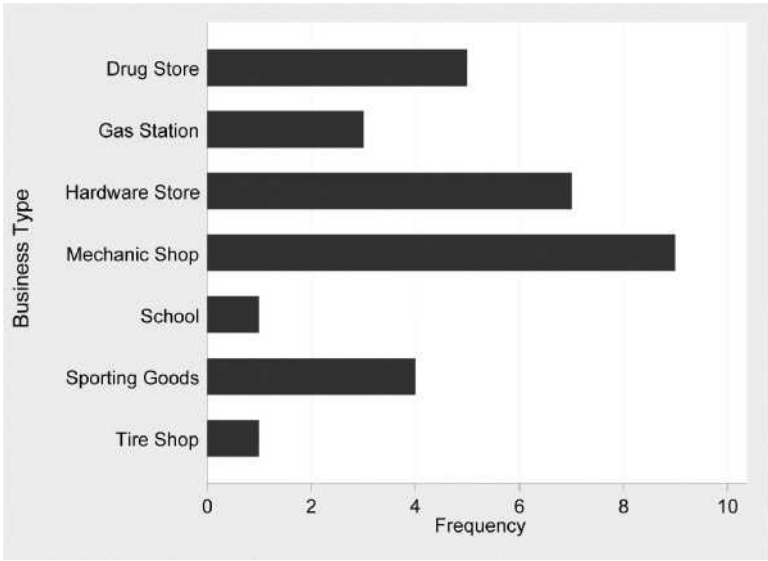


Figure 1.4 Stata display of (horizontal) bar graph of Business Types

```
ssc install catplot
```

Once the software finishes downloading, and again using the data from Table 1.2, type the following command to create the graph:

```
catplot businesstype
```

It is only necessary to have the actual type of the business listed; Stata will calculate the frequencies. Labels and other information can be added to improve the readability and presentation of the graph.

Pie Chart

A *pie chart* is a graphical presentation of frequency distribution and relative frequency. In this regard, the pie chart is similar to the bar graph because one cannot differentiate between the graphs of actual and relative frequencies, except for the scale. Pie charts are more effective when there are few categories or variables; otherwise, the graph becomes cluttered.

A circle is divided into wedges representing each of the categories or variables in a table. If frequencies are charted, their magnitude is placed

under their name. When the pie chart is based on the relative frequencies instead of the frequencies, the scale will be different but not the sizes of the slices on the pie.

Example 1.3 Provide a pie graph for the Business Type variable in Table 1.2.

Solution 1.3

1. Open a new spreadsheet in Excel.
2. Use the data that you entered into Excel in Example 1.2.
3. Go to "Insert," which is the second tab at the top left-hand corner of the spreadsheet. Click on the "Pie."
4. Several options become available. You can select whichever pie shape you wish. Excel will populate a chart similar to the ones below (Figure 1.5).

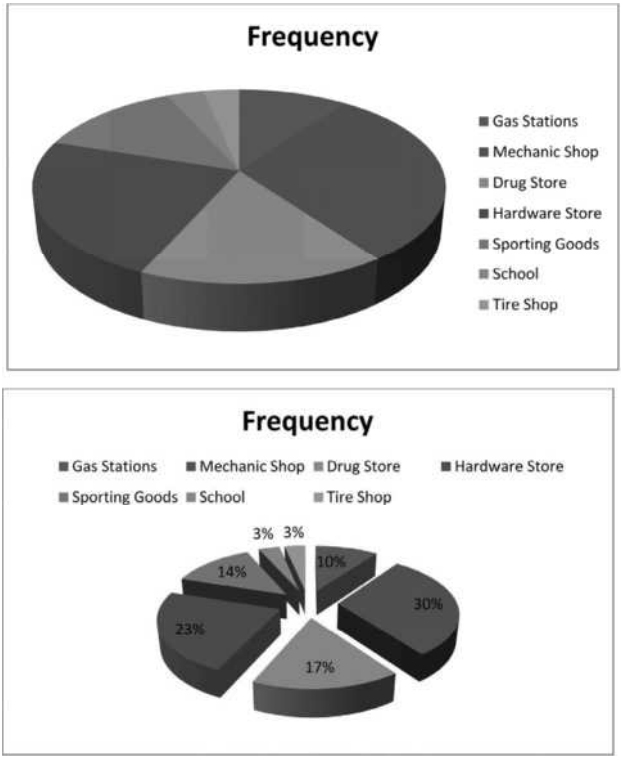


Figure 1.5 Excel display of pie charts of Business Types for soda dispensers

Notice that due to space limitation the legend is placed on the side in the above pie chart.

Type the following in Stata to obtain Figure 1.6:

graph pie, over(businesstype)

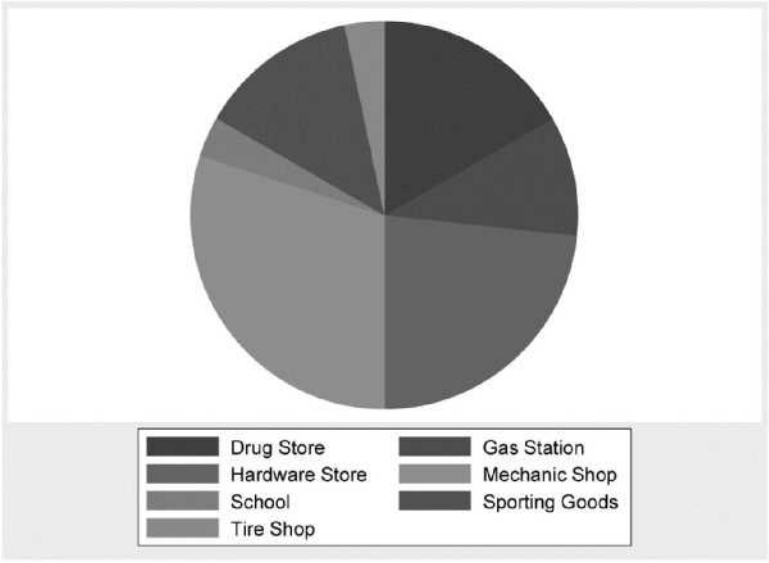


Figure 1.6 Stata display of pie charts of Business Types for soda dispensers

Descriptive Statistics for Quantitative Variables

As depicted in Table 1.1 above, there are more methods available to describe quantitative data. Some are similar to the methods used in qualitative methods, but their interpretations are usually broader.

Tabular Methods for Quantitative Variables

There are three commonly used tabular descriptive statistics for quantitative variables. They are frequency distribution, relative frequency distribution, and cumulative distribution.

Frequency Distribution for Quantitative Variables

A frequency distribution shows the frequency of occurrence for non-overlapping classes. Unlike the qualitative frequency distribution, there are no set and predefined classes or groups. The researcher will determine the size of each class and the number of classes. Such data are called grouped data.

Definition 1.11 *Grouped data* refers to data that are summarized or organized to provide a better and more compact picture of reality. The grouping can be in the form of relative frequency or summarized in cross-tabulation tables or into classes.

Example 1.4 An anthropologist is studying a small community of gold miners in a remote area. The community consists of nine (9) families. The family income is reported in $1,000 of dollars below.

66, 58, 71, 73, 64, 70, 66, 55, and 75

Group the data in a meaningful way.

Solution 1.4

We deliberately chose a small set to demonstrate the point better without boring calculations. In practice, datasets will be much larger, and it would make more sense to condense the data by grouping them into classes. Since only one value is repeated, it does not make sense to build a frequency distribution; no real summary will emerge. If we divide the data into classes, however, we can build the frequency distribution. The range of data is from 55 to 75 (Table 1.5). If the researcher wishes to have five classes, the size of each class would be:

$$\text{Class width } = \frac{Maximum - Minimum}{Number\ of\ classes} = \frac{75-55}{5} = 4$$

Table 1.5 Classes and their frequencies

Classes	Frequency
55–59	2
60–64	1
65–69	2
70–74	3
75	1

The number of classes is arbitrary, and any reasonable number of classes and class widths will work. Avoid extremities and unbalanced classes. An unbalanced class is where the intervals covered by the groups do not match (e.g., 55–57 and 58–68). To avoid decimal points in classes, we added an extra class for values greater than or equal to 75. Other choices for number of classes or class width would be equally valid. The quantitative data groups can include decimal numbers; however, in this case, extra caution is needed to avoid overlapping the classes.

The histogram command in Excel provides the frequency as well as the cumulative frequency. If the option Chart Percentage is selected from the histogram dialog box, the histogram and the ogive will be graphed too. The graph for the cumulative frequencies is called ogive. In carpentry, there is a molding bit for shaping the edge of the wood called Roman ogive. The graphs of the cumulative frequencies usually resemble the finished edge of the Roman ogive molding common on quality furniture.

A list of nine values has been reduced to a two-column table with five rows as shown in Table 1.5. The procedure would be the same for a larger dataset, for example, for the family incomes of the United States with population of over 300,000,000 people. The result can be graphed for more visual presentation. One such graph is called a dot plot (see Example 1.12). Other graphs such as a histogram are also available (see Example 1.9).

Relative Frequency Distribution for Quantitative Variables

The relative frequency for quantitative variables is computed in the same way as those of qualitative variables. The frequency for each class is divided by the total number of observations in the population or sample to obtain the relative frequency.

Example 1.5 Table 1.6 provides the relative frequencies for the family incomes example in the above section.

Solution 1.5

Table 1.6 Relative and cumulative frequencies of family incomes

Class	Frequency	Relative Frequency	Cumulative Frequency
55–59	2	0.222222222	0.222222222
60–64	1	0.111111111	0.333333333
65–69	2	0.222222222	0.555555556
70–74	3	0.333333333	0.888888889
75	1	0.111111111	1

Cumulative Frequency Distribution for Quantitative Variables

In the case of quantitative variables, the classes or values of interest are sequential and have meaningful order, usually from smallest to the largest, as in the previous two examples. This allows us to obtain cumulative frequencies. Cumulative frequencies consist of sums of frequencies up to the value or class of interest. The last value is always one (1) since it represents 100% of observations. See Table 1.6.

Percentiles

A *percentile* is the demarcation value below which the stated percentage of the population or sample lie. For example, 17% of a population or sample lies below the 17th percentile. To obtain a percentile, identify the value that corresponds to the stated percentile. To do this, first sort the data and then find the index (i):

$$i = \frac{p}{100} n$$

where p is the desired percentile and n is either the population or the sample size. When the index is an integer, add 1 to it to get the position of the percentile. If the result is a decimal value, use the next higher integer to get the position of the percentile. Note that the next higher integer is used not the rounded-up value.

For example, the 17th percentile of a dataset containing 84 observations is the 15th observation of the sorted group. We find this by using

the above formula to get the index (i = 0.17 × 84 = 14.28). To find the 17th percentile, we then raise the index, 14.28, to the next higher integer, 15. We conclude that the 15th observation of the sorted data marks the 17th percentile for a sample size of 84.

Example 1.6 A retail store has collected sales data, in thousands of dollars, for 18 weeks. Find the 80th and the 50th percentiles for weekly sales.

66, 58, 71, 73, 64, 70, 66, 55, 75, 65, 57, 71, 72, 63, 71, 65, 55, and 71.

Solution 1.6

Sorting the combined data gives:

55, 55, 57, 58, 63, 64, 65, 65, 66, 66, 70, 71, 71, 71, 71, 72, 73, 75

The 80th percentile is obtained by:

$$i = \frac{80}{100} \times 18 = 14.4$$

Since the result is a real value (i.e., a value with a decimal), use the next higher integer, which is 15 in this example. The number in the 15th position is the 80th percentile. That value is 71. Therefore, 80% of weekly sales are below $71,000.

The 50th percentile is:

$$i = \frac{50}{100} \times 18 = 9$$

Since the index is an integer, use the next higher integer, namely the 10th observation, which is 66. Therefore, 50% of weekly sales are below $66,000.

Quartiles

Quartiles divide the population into four equal portions, each equal to 25% of the population. Like the median and percentiles, the data must be sorted first. The first quartile, Q_1, is the data point such that 25% of the observations are below it. The second quartile, Q_2, is the data point

such that 50% of the observations are below it. The third quartile, Q_3, is the data point such that 75% of the observations are below it.

The first quartile is the same as the 25th percentile. The second quartile is the same as the 50th percentile, as well as the median. The third quartile is the same as the 75th percentile. The quartiles are calculated the same way as the 25th, 50th, and 75th percentiles using the following indices.

Use $i = \frac{25}{100} n$ for the first quartile.

Use $i = \frac{50}{100} n$ for the second quartile.

Use $i = \frac{75}{100} n$ for the third quartile.

If the result of an index is an integer, use the next higher integer to find the location of the quartile. If the result of the index is a real value (i.e., a value with a decimal) the next higher integer will determine the position of the quartile.

Example 1.7 For the weekly sales data of the retail store in Example 1.6, find the first, second, and the third quartiles. The data are repeated for your convenience.

66, 58, 71, 73, 64, 70, 66, 55, 75, 65, 57, 71, 72, 63, 71, 65, 55, and 71.

Solution 1.7

The three quartiles are calculated using the following indexes.

$i = \frac{25}{100} \times 18 = 4.5$ the first quartile is in the 5th position.

$i = \frac{50}{100} \times 18 = 9$ the second quartile is in the 10th position.

$i = \frac{75}{100} \times 18 = 13.5$ the third quartile is in the 14th position

Sort the combined data.

55, 55, 57, 58, 63, 64, 65, 65, 66, 66, 70, 71, 71, 71, 71, 72, 73, 75

Q_1 Q_2 Q_3

Stata provides a detailed output for preset percentiles. See Figure 1.7 for output showing Stata's preset percentiles. Obtaining other percentiles is tedious and not worth the effort in this software. The Stata command is displayed:

summarize *varlist*, detail

Varlist is used as a placeholder to indicate where you should place the name of the variable (or variables) which you wish to summarize. This is a common practice in Stata help files (Figure 1.7).

```
. summarize sales, detail
```

Weekly Sales

	Percentiles	Smallest		
1%	55	55		
5%	55	55		
10%	55	57	Obs	18
25%	63	58	Sum of Wgt.	18
50%	66		Mean	66
		Largest	Std. Dev.	6.34313
75%	71	71		
90%	73	72	Variance	40.23529
95%	75	73	Skewness	-.4952022
99%	75	75	Kurtosis	2.04109

Figure 1.7 Stata display of summary statistics

Hinges

Hinges also divide the data into four equal portions. The hinges, however, use the concept of the median. To obtain hinges, first sort the data then find the median as in examples 1.5 and 1.6. Find the median of the lower half and call it the **lower hinge**. Find the median of the second half and call it the **upper hinge**.

Example 1.8 For the weekly sales data of the retail store in Example 1.6, find the lower and upper hinges.

Solution 1.8

Sort the combined data.

55, 55, 57, 58, **63**, 64, 65, 65, 66, 66**,** 70, 71, 71, **71**, 71, 72, 73, 75

lower hinge Median upper hinge

Graphical Methods for Quantitative Variables

The numbers of available graphical methods for quantitative variables far exceed the number of graphical methods available for qualitative variables. Here, we will address histograms, ogive, stem and leaf, dot plot, scatter plot, and box plot. Box plot uses some of the concepts that are introduced in Chapter 2.

Histogram

A *histogram* is a graphical representation of the frequency distribution or relative frequency distribution of quantitative data. The boundaries of the classes are used for the demarcation of the vertical bars. A histogram and a bar graph are identical except quantitative values are used in the histogram on the X-axis compared with qualitative values for a bar graph.

Example 1.9 The following data represent incomes of gold miners in a small community (this is the same data used in Example 1.4). The corresponding histogram follows (Table 1.7).

Table 1.7 Histogram and related setup in Excel

Classes	Frequency
55–59	2
60–64	1
65–69	2
70–74	3
>75	1

Solution 1.9

1. For many of the exercises in this class, you will need to use the data analysis tools add-in in Excel. If you have not already done so, you will need to access the add-in using either step 2 or 3 below.
2. For PCs, go to "file," "options," and "add-ins." In the "manage" drop-down list, select "Excel Add-ins" and in the dialog box select "Analysis Toolpak."

3. For Macs, you will need to go to http://www.analystsoft.com/en/products/statplusmacle/download.phtml to download StatPlus.
4. If you have problems with the installation, see "Analysis Tookpak" in Microsoft Support for more help.
5. Enter the data from Example 4.1 into a new spreadsheet in Excel.
6. In Column B, you will enter the bin range, which are the demarcation points for each group. In cell B1, type 59, followed in the cells below by 64, 69, and 74.
7. Click on the "Data" tab, then "Data Analysis" in the Analysis group. Select "Histogram."
8. In the dialog box, Column A should be entered as the input range and Column B as the Bin Range.
9. Before clicking "ok," check the box that says "Chart Output."
10. Excel will populate a table and chart similar to the ones below (Figure 1.8).

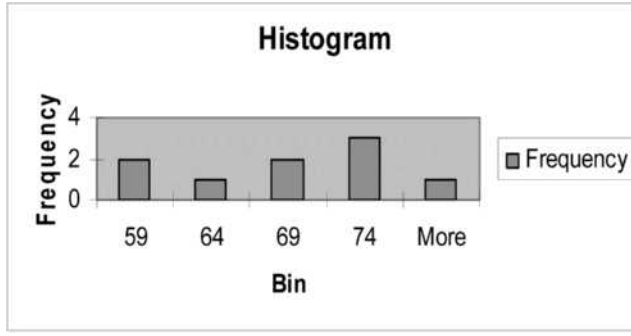


Figure 1.8 Histogram Output from Excel

The above graph can represent the relative frequencies, too. Only the unit of measurement on the Y-axis will differ.

Ogive

In Excel the ogive is obtained from the histogram dialog box by selecting the cumulative percentage option.

Example 1.10 For the nine incomes of the community of gold miners, graph the frequency, relative frequency, and cumulative frequency.

66, 58, 71, 73, 64, 70, 66, 55, and 75.

Solution 1.10

The frequency, relative frequency, and cumulative frequency for these data are given in Table 1.8.

Table 1.8 Frequency, relative frequency, and cumulative frequency

Class	Frequency	Relative Frequency	Cumulative Frequency
55–59	2	0.222	0.222
60–64	1	0.111	0.333
65–69	2	0.222	0.555
70–75	4	0.444	0.999

The ogive gives the cumulative area under the relative frequency histogram (Figure 1.9).

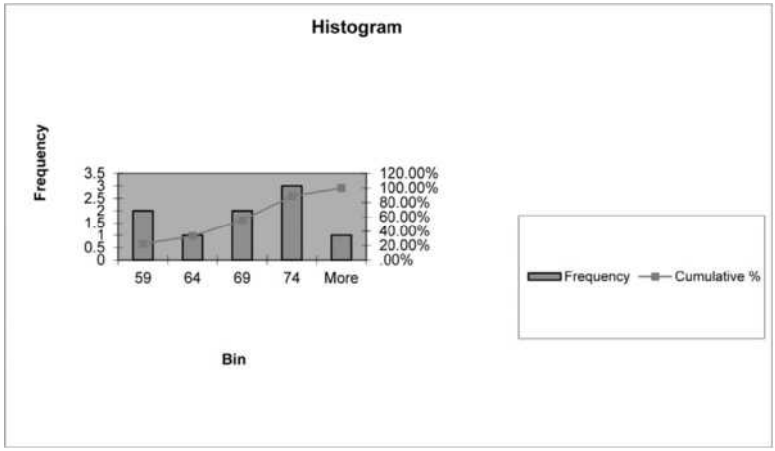


Figure 1.9 Ogive superimposed on a histogram

Stem and Leaf

Stem and leaf is another descriptive way of summarizing information. Tukey introduced the concept of the stem and leaf.[5] Some authors place stem and leaf under exploratory data.[6]

In the stem and leaf, usually the last digit of a value is recorded as the leaf and is placed after the first value which is called a stem. A vertical

line for easy visualization separates the leaves and stems. To create a stem-and-leaf display, place the first digit(s) of each observation to the left of a vertical line. Place the last digit of each observation to the right of the line. The leaf for each observation that shares a stem is placed on the same row (see Figure 1.10).

Example 1.11 Provide a stem-and-leaf graph for the gold miners' data.

66, 58, 71, 73, 64, 70, 66, 55, and 75.

Solution 1.11

```
5 | 8 5
6 | 6  4  6
7 | 1  3  0 5
```

Figure 1.10 Stem-and-leaf graph

Notice that the result resembles a rotated histogram. If the data for each leaf are also sorted, a better summary is obtained, as in Figure 1.11.

```
5 | 5 8
6 | 4  6  6
7 | 0 1  3  5
```

Figure 1.11 Sorted stem-and-leaf graph

If the numbers are too large, the first two or more digits could be placed on the left side. The idea is to select the digits in a manner that makes the summary useful. Stata will display a stem-and-leaf graph of all values when the following command is used: stem *varlist*.

Dot Plot

The dot plot is useful when only one set of data is under consideration. The actual data are placed on the X-axis. For each occurrence of the value, a dot is placed above it. All the dots of equal frequency are at the same height, which has no significant meaning other than reflecting the occurrence of the observation. In the case of multiple occurrences, additional dots are placed above the previous ones. The dots are placed at equal distances for visual as well as representation purpose.

Example 1.12 Provide a dot plot for the gold miners' data (Figure 1.12):

66, 58, 71, 73, 64, 70, 66, 55, and 75.

Solution 1.12

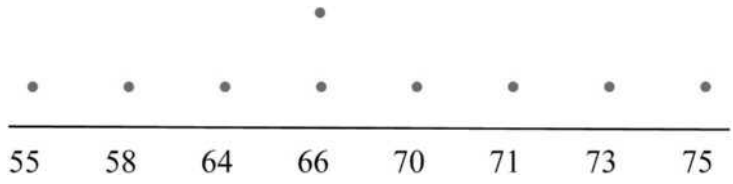


Figure 1.12 Dot plot

A dot plot resembles an exaggerated histogram. Stata uses the command "dotplot" followed by the name(s) of variable(s) to display a dot plot.

Scatter Plot

An observant reader would notice that all the previous examples have been based on only one variable with numerous classifications and categories. In economics and many other branches of science, it is also beneficial to present graphics of two or more variables. Scatter plots are one visual method of presenting several variables graphically, as two or more variables can be combined into one graph.

Example 1.13 Graph a scatter plot of annual income and consumption for the United States (in billions of dollars) for years 1990 through 2014 using the data from Table 1.9.

Solution 1.13

To obtain a scatter plot, type the information from Table 1.9 in an Excel spreadsheet.

1. In cell A1 type "I" for Income followed by the income data from the second column in Table 1.9.
2. In cell B1, type "C" for Consumption followed by consumption data from the third column in Table 1.9.
3. Highlight cells A1 through B22. Go to "Insert," which is the second tab on the top left hand corner of the spreadsheet; choose "Charts."

Table 1.9 Annual income (I) and consumption (C) in the United States, 1990–2014

Year	I	C
1990	4,904.5	3,825.6
1991	5,071.1	3,960.2
1992	5,410.8	4,215.7
1993	5,646.8	4,471.0
1994	5,934.7	4,741.0
1995	6,276.5	4,984.2
1996	6,661.9	5,268.1
1997	7,075.0	5,560.7
1998	7,587.7	5,903.0
1999	7,983.8	6,307.0
2000	8,632.8	6,792.4
2001	8,987.1	7,103.1
2002	9,149.5	7,384.1
2003	9,486.6	7,765.5
2004	10,048.3	8,260.0
2005	10,609.3	8,794.1
2006	11,389.0	9,304.0
2007	11,994.9	9,750.5
2008	12,429.6	10,013.6
2009	12,087.5	9,847.0
2010	12,429.3	10,202.2
2011	13,202.0	10,689.3
2012	13,887.7	11,083.1
2013	14,166.9	11,484.3
2014	14,733.9	11,930.3

Sources: Bureau of Economic Analysis, National Income and Product Account Tables: Table 2.1. Personal Income and Its Disposition and Table 2.3.5. Personal Consumption Expenditures by Major Type of Product.

4. Click on "Scatter," which will reveal several options. Select the option at the top left to obtain a graph similar to Figure 1.13.
5. Examine different options for different effects.

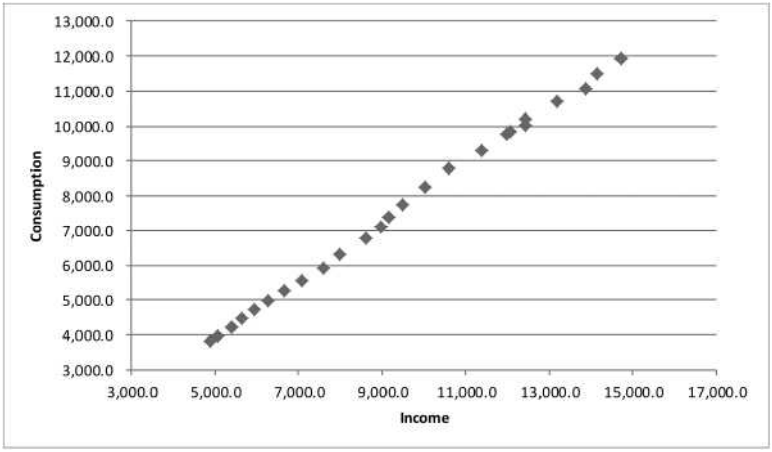


Figure 1.13 Excel display of scatter plot of income–consumption for United States, 1990–2014

Stata provides numerous options for plotting two or more variables in a scatter plot. The simplest form consists of the command "scatter" followed by the names of two variables.

The graph depicted in Figure 1.14 is by no means the extent of possible ways to represent data, for either qualitative or quantitative variables. Many other imaginative ways can be used, some of which are available in popular software such as Excel or dedicated software such as Stata.

Box plot

A *box plot* is a visual representation of several descriptive statistics in a concise manner. Some of the descriptive statistics that are used in a box plot and not explained yet will be explained in Chapter 2. In a box plot, the box visually demonstrates the 25th to 75th percentiles, while a vertical line is used to represent the median (see Figure 1.12 for an example). The graph consists of one box per variable. To form a box plot, draw a box connecting the first and third quartiles, the 25th and the 75th percentiles, respectively. The height of the box is arbitrary. Draw a vertical line at the median to divide the box. Draw lines from the first and third quartiles to data points on

the extreme left and extreme right, but do not extend them to the **outliers.** These lines are called **whiskers**. Whiskers, extend from the edges of the box to the *adjacent* values, capped by an **adjacent line**.[7] The values further away from the box extending past the adjacent lines in either direction are called **outside values** or **outliers**. An outlier is defined as any value that exceeds 1.5 times the interquartile range to the left of the first quartile or to the right of the third quartile.

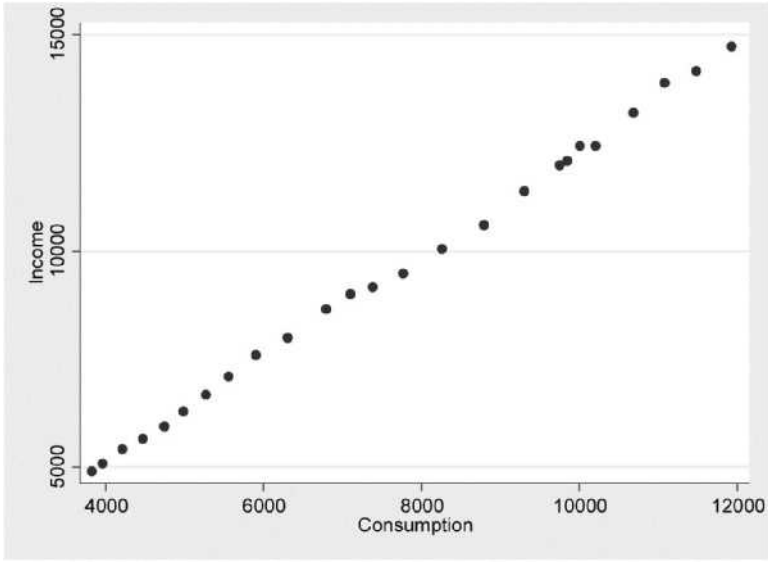


Figure 1.14 Stata display of scatter plot of income–consumption for United States, 1990–2014

Example 1.14 Use the data from Table 1.2 to obtain the box plot of income by the type of business and by gender.

Solution 1.14

The following graph of a box plot is created in Stata (Figure 1.15). Note that there are no outliers for any of the average weekly sales data for any of the Business Types.

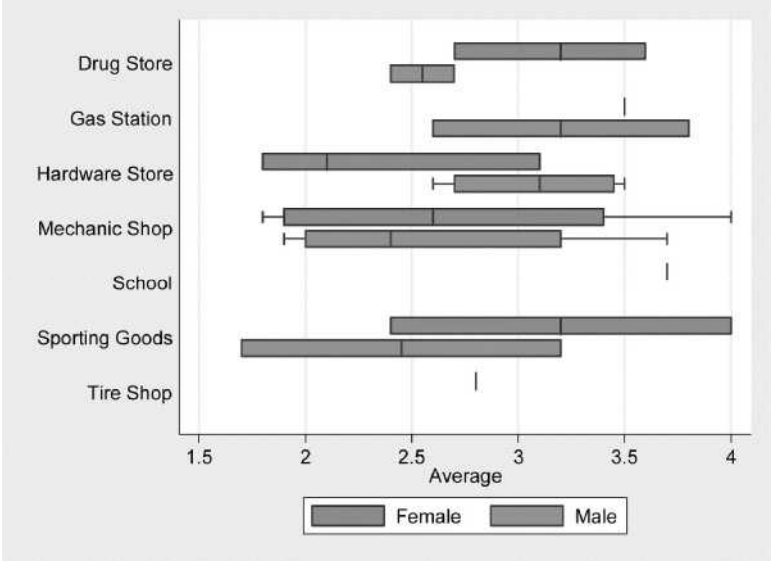


Figure 1.15 Stata display of a box plot of income by location by gender

CHAPTER 2

Numerical Descriptive Statistics for Quantitative Variables

Introduction

One of the purposes of descriptive statistics is to summarize the information in the data for a variable into as few *parameters* as possible. *Measures of central tendency* provide concise summaries of a population. They are also called *measures of location* as in Table 1.1 in Chapter 1. Measures of central tendency are addressed first. However, they are often not enough to provide the full picture of reality, as will be demonstrated when the concept of variance is introduced shortly. The addition of *measures of dispersion*, such as the variance, provides a more complete picture of reality. Measures of dispersion are followed by *measures of association*.

Measures of Central Tendency

This section discusses statistics that represent information about the nucleus of data.

Mean

The arithmetic mean, or simply the mean, is the most commonly used descriptive measure. Other names for the mean are the average, mathematical expectation, and the expected value. This section deals with raw or ungrouped data.

Arithmetic Mean

The *arithmetic mean* is the *representative* or *typical value* that represents a population. The mean is the sum of all the elements in the population divided by the number of the elements.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2.1}$$

The symbol μ, pronounced mu, is used to represent the population mean. The symbol $\sum$, called capital sigma, represents the sum of the values for a variable. The subscript "*i*" represents observations 1 through N. When there is no ambiguity, the index and subscripts are not included:

$$\mu = \frac{\Sigma(X)}{N} \tag{2.2}$$

The population mean is a parameter and provides information about the central tendency of the data. The mean is susceptible to extreme values. Since all values of the population are used in calculation of the mean, a single very large or very small value can have a major impact on the mean. This is not quite as important in the case of a population as it is with samples.

Sample Mean

The *sample mean* is the sum of the sample values divided by the sample size.

$$\hat{\mu} = \bar{X} = \frac{\Sigma(X)}{n} \tag{2.3}$$

Both, $\hat{\mu}$, pronounced mu-hat, and $\bar{X}$, pronounced x-bar, are commonly used in the literature to represent the sample mean. Both are widely accepted. However, $\hat{\mu}$ has several advantages over $\bar{X}$. First it reduces the number of symbols that one has to learn in half. The population parameter is μ and its estimate is $\hat{\mu}$. Second, it eliminates guessing which statistic represents a particular population parameter. Third, it provides a reasonably simple rule to follow. Population parameters are

represented by Greek letters, and sample statistics are represented by the same Greek letter with a hat on it.

Definition 2.1 A *statistic* is a numeric fact or summary obtained from a sample. It is always known, because it is calculated by the researcher, and it is a variable. A statistic is also used to make inferences about the corresponding population parameter.

Example 2.1 An anthropologist is studying communities of gold miners in a remote area. She selects nine (9) families at random from a community. The family income is reported in $1,000 of dollars below. Find the sample mean. These data are hypothetical but plausible.

66, 58, 71, 73, 64, 70, 66, 55, and 75

Solution 2.1

We use these data to show computational detail. A careful reader would remember that the same data were used in Example 1.4 but with a major difference. There, the data were presented as population data. The idea of a small community with nine families is acceptable but a sample of nine is more plausible. We use the same data to avoid wasting time entering new data and we limit data size to avoid tedious computations. The sample mean is as follows:

$$\hat{\mu} = \bar{X} = \frac{66+58+71+73+64+70+66+55+75}{9} = \frac{598}{9} = 66.444$$

The expected income of a family from this sample is $66,444.44. This statistic is an estimate of the population parameter μ.

The procedure for obtaining the population mean is the same. However, there is a major conceptual difference between the sample mean and population mean, as between any *statistic* and the corresponding *parameter*. The former is a *variable*, while the latter is a *constant*. In actual research we seldom, if ever, know population parameters, which necessitates collecting samples and obtaining sample statistics to make inferences about the *unknown* population *parameters*. It is possible to have small population, for example, the population can consist of the two children in a household; but usually they are of little use in economic studies.

If one of few extreme members of the population appears in a sample, especially a small sample, the impact will be detrimental. If the sample mean is erroneous, the estimate of the population mean will be misleading. Irrespective of what values appear in the sample, the sample mean does provide an unbiased estimate of the population mean, provided the sample points are taken at random, a point which will be addressed in more detail in Chapters 5 to 7.

Example 2.2 The closing stock prices for Wal-Mart and Microsoft from May 21 to July 2, 2015 are provided in Table 2.1. We will use these data in many of the examples in this book.

Table 2.1 Closing stock prices for Wal-Mart (WMT) and Microsoft (MSFT) May 21 through July 2, 2015

Date	MSFT	WMT	Date	MSFT	WMT
5/21/2015	47.42	76.11	6/12/2015	45.97	72.43
5/22/2015	46.9	75.86	6/15/2015	45.48	71.93
5/26/2015	46.59	74.9	6/16/2015	45.83	72.35
5/27/2015	47.61	75.19	6/17/2015	45.97	72.73
5/28/2015	47.45	74.84	6/18/2015	46.72	72.98
5/29/2015	46.86	74.27	6/19/2015	46.1	72.74
6/1/2015	47.23	74.73	6/22/2015	46.23	72.79
6/2/2015	46.92	74.53	6/23/2015	45.91	72.57
6/3/2015	46.85	74.89	6/24/2015	45.64	72.38
6/4/2015	46.36	74.15	6/25/2015	45.65	71.86
6/5/2015	46.14	73.06	6/26/2015	45.26	72.12
6/8/2015	45.73	72.61	6/29/2015	44.37	71.42
6/9/2015	45.65	72.47	6/30/2015	44.15	70.93
6/10/2015	46.61	72.93	7/1/2015	44.45	71.88
6/11/2015	46.44	72.94	7/2/2015	44.4	71.86

Find the (sample) average price for Wal-Mart for the period from June 12 to July 2, 2015.

Solution 2.2

$$\hat{\mu} = \bar{X} = \frac{\Sigma(X)}{n} = \frac{1082.97}{15} = \$72.20$$

From the data verify that the mean for the period from May 21 to June 11, 2015 is \$74.23. This indicates that on average the price of the Wal-Mart stock has been falling between May 21 and July 2, 2015.

Example 2.3 Suppose the researcher in Example 2.1 samples incomes from another community; see below. Calculate the sample mean of their incomes.

$$65, 57, 71, 72, 63, 71, 65, 55, 71$$

Solution 2.3

$$\hat{\mu} = \bar{X} = \frac{590}{9} = 65.555\overline{5}$$

The sample mean changed since it is a statistic, which is a variable. This sample mean provides another estimate of the population mean μ. The mean of the combined samples of incomes of miners can be obtained using the individual data from the two groups.

$$\hat{\mu} = \bar{X} = \frac{1188}{18} = 66$$

The same result can be obtained from the previously calculated sample means.

$$\hat{\mu} = \bar{X} = \frac{66.4444 + 65.555}{2} = 66$$

In this case, the sample sizes are equal, so *simple arithmetic* average works. When sample sizes differ, the *weighted average* is used, where the weights would be the corresponding sample sizes.

Trimmed Mean

Trimmed mean is a modification of the mean. It is used when there are unusually high or low observations in the data. Such data are also called

outliers and their exclusions provide more meaningful and representative statistics. The sample data are sorted and a given percentage, say 5%, of the top and the bottom of the data are discarded, and the regular mean is calculated for the remaining data. The trimmed mean is less susceptible to extreme values.

Geometric Mean

The geometric mean is useful when the values change in geometric progression instead of arithmetic progression, as is the case with *growth rates* or *interest rates*. The geometric mean is calculated using the following formula.

$$\text{G.M} = \sqrt[N]{X_1 X_2 \ldots X_N} \tag{2.4}$$

The formula can be expressed using logarithm to avoid taking the *n*th root, as shown in the next section. This was more important before the advent of powerful calculators. The logarithmic formula is a linear sum of its elements.

Example 2.4 Assume that a new company grew at 28% the first year, 15% the second year, and 13% the third year. What is the rate of the growth of company?

Solution 2.4

Note that at the beginning of each year, the following amounts are available.

End of	Growth Rate
Year 1	28% = (100 + 100 × 0.28) = 128
Year 2	15% = (128 + 128 × 0.15) = 147.2
Year 3	13% = (147.2 + 147.2 ×0.13) = 166.336

Therefore, \$100 will grow to \$166.336 over 3 years at the above growth rates for each year. The geometric mean for the growth rate over 3 years is

$$\text{G.M} = \sqrt[3]{1.28 \times 1.15 \times 1.13} = \sqrt[3]{1.66336} = 1.184846$$

The growth rate is (1.184846 − 1) × 100 = 18.4846%. Therefore, $100 will grow to $166.336 over 3 years at this rate as well, which is the case as seen below:

End of	Total at the End of Period			
	Using Actual Rates		Using the Average Rate	
Year 1	28%	(100 + 100 × 0.28) = 128	18. 4846%	(100 + 100 × 0.1848) = 118.48
Year 2	15%	(128 + 128 × 0.15) = 147.2	18.4846 %	(118.48 + 118.48 × 0.1848) = 140.39
Year 3	13%	(147.2 + 147.2 × 0.13) = 166.336	18.4846 %	(140.39 + 140.39 × 0.1849) = 166.336

In the case of the average growth rate, the same result is obtained if the rate is raised to power 3:

$$= 100 \times (1 + 0.184846)^3 = 166.336$$

The expression for calculating the geometric mean in Excel is given below:

= geomean(range)

where the range is any valid Excel range. Make sure the above data are entered as 128, 115, and 113 but not as 28, 15, and 13. In Stata, the following command will display arithmetic, geometric, and harmonic means. In addition, it will display their 95% confidence intervals. The level of confidence can be modified.

ameans *varlist*

Remember that you should enter in your variable name instead of *varlist*. The result of command *ameans* in Stata is displayed in Figure 2.1. In addition to each mean a 95% confidence interval is also provided.

Variable	Type	Obs	Mean	[95% Conf.	Interval]
var1	Arithmetic	3	118.6667	98.43454	138.8988
	Geometric	3	118.4846	100.2009	140.1046
	Harmonic	3	118.3072	101.5974	141.5956

Figure 2.1 Stata display of arithmetic, geometric, and harmonic means

Expressing Geometric Mean in Logarithmic Form

Since the elements of the geometric mean are multiplicative, it can be expressed in natural logarithm form as in Equation (2.5).

$$Ln(G.M.) = \frac{1}{n}\sum_{i=1}^{n} Ln(X_i) = \frac{Ln(X_1) + Ln(X_2) + ... + Ln(X_n)}{n} \quad (2.5)$$

To verify the result using the natural logarithm, use the built-in natural log equation in Excel. In three empty cells, enter the following:

=ln(1.28)	Excel will display 0.24686
=ln(1.15)	Excel will display 0.139762
=ln(1.13)	Excel will display 0.122218

Calculate the arithmetic mean of the resulting values.

$$(0.24686 + 0.139762 + 0.122218)/3 = 0.169613$$

This value is in natural logarithm. Take the antilogarithm by raising the number "e" to the power of 0.169613 to obtain the correct number.

$$= \exp(0.169613) \qquad \text{Excel will display } 1.184846$$

This is the formula for compound interest. To generalize, let P_0 be the initial investment, P_n the amount after n years, and r the interest rate or the rate of growth.

$$P_n = P_0(1 + r)^n \quad (2.6)$$

In the above example, the ending value is known to be $P_n = 166.336$, the beginning value is $P_0 = 100$, and $n = 3$. The (average) rate of growth is

$$166.336 = 100(1 + r)^3$$

$$(1+r) = \sqrt[3]{1.66336} = 1.18484$$

$$r = 1.18484 - 1 = 0.1848$$

The geometric mean is also the proper mean when dealing with the *ratio* of items.

Example 2.5 The ratio of the average income in a country to the price of an average car is 4 in year one and 5 in year two. What is the average ratio of income to the price of a car?

Solution 2.5

Since the average of the *ratios* of prices to incomes is of interest using the arithmetic mean would be incorrect. The correct statistic for the geometric mean is obtained with

$$\sqrt{4\times 5} = \sqrt{20} = 4.472136$$

The average income is 4.47 times the price of average car. The average of the ratio of the average price to average income is given by

$$\sqrt{\frac{1}{4}\times\frac{1}{5}} = \sqrt{0.2\times 0.025} = \sqrt{0.05} = 0.223607$$

The geometric mean of the ratio of income to the price of car is the same as the reciprocal of the geometric mean of the price of car to income.

Harmonic Mean

Monetary policy must be transparent and stated policies must be pursued in order for rational expectations to be formed and to maintain economic stability.[1] Therefore, when the Federal Reserve Bank announces a target interest or exchange rate, it should strive to achieve and maintain that rate. Economic conditions change, so the actual interest or exchange rate fluctuates over time. In order to maintain the desired rate, the government must change the rate several times over the targeted period. To maintain the targeted rate over a period, it is necessary to have the average of the actual rates be equal to the targeted rate. This is achieved by using the *harmonic mean*. The harmonic mean is calculated using the following formula:

$$\text{H.M} = \frac{N}{\sum_{i=1}^{N}\frac{1}{X_1}} = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_N}} \tag{2.7}$$

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of the values.

Example 2.6 A salesman travels to another city to meet a client. To make sure he does not miss the appointment, he drives at 90 miles per hour. After the meeting, he returns more leisurely at 45 miles per hour. What is his average speed?

Solution 2.6

The average speed is not (90 + 45)/2 = 67.5 miles per hour. For simplicity, assume he traveled 90 miles; any other value will work as well.

$$\text{Time while going} = \frac{90}{90} = 1$$

$$\text{Time while returning} = \frac{90}{45} = 2$$

Total travel time = 1 + 2 = 3

$$\text{Average speed} = \frac{distance}{time} = \frac{90+90}{1+2} = 60 \text{ miles per hour}$$

The harmonic mean will give the correct answer where the arithmetic mean failed.

$$\text{H.M} = \frac{2}{\frac{1}{90}+\frac{1}{45}} = \frac{2}{\frac{1+2}{90}} = \frac{2\times 90}{3} = 60 \text{ miles per hour}$$

Using the traveled distance, verify that it would not have taken 3 hours to travel 180 miles as suggested by 67.5 miles per hour obtained by the arithmetic mean. Using the *ameans* command in Stata introduced in Example 2.4 provides the following output.

Variable	Type	Obs	Mean	[95% Conf.	Interval]
var1	Arithmetic	2	67.5	-218.3896	353.3896
	Geometric	2	63.63961	.7784902	5202.378
	Harmonic	2	60	.	.

Figure 2.2 Stata display of arithmetic, geometric, and harmonic Means

The Excel command is

=harmean(range)

where "range" is any valid Excel range holding the data.

Rule 2.1

When n = 2, the geometric mean is equal to the square root of the arithmetic mean times the harmonic mean.

$$\text{G.M.} = \sqrt{(\text{A.H.})(\text{H.M.})} \tag{2.8}$$

Relationship between Arithmetic, Geometric, and Harmonic Means

$$\text{H.M.} \leq \text{G.M.} \leq \text{A.M.} \tag{2.9}$$

The equality sign holds only in the trivial case when all sample values are identical.

Mean of Data Summarized as Frequencies

The expanded presentation of arithmetic mean will be informative.

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{X_1}{N} + \frac{X_2}{N} + \cdots + \frac{X_N}{N} = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \cdots + \frac{1}{N}X_N \tag{2.10}$$

The mean is the sum of $1/N$th of each observation. In other words, each observation gets a weight equal to $1/N$th of the total number of observations. Sometimes, each value might have a different weight, say, $f_1, f_2, \ldots f_N$ for each of $X_1, X_2, \ldots, X_N$. In that case the mean is expressed in terms of frequencies of each value.

$$\mu = f_1X_1 + f_2X_2 + \ \ldots f_NX_N = \Sigma fX \tag{2.11}$$

Note that the sum of relative frequencies is $1(\Sigma f = 1)$ and, hence, is not written. In general, the formula for the *frequencies* would be

$$\mu = \frac{f_1X_1 + f_2X_2 + \ldots + f_NX_N}{f_1 + f_2 + \ldots + f_N} \tag{2.12}$$

The formula for the population mean is

$$\mu = \frac{\sum_{i=1}^{N} f_1X_i}{\sum_{i=1}^{N} f_i} \tag{2.13}$$

The formula for the sample mean is similar.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i} \tag{2.14}$$

Weighted Mean

The weighted mean is similar to the mean using relative frequencies, except that the sum of weights need not add up to one.

$$\mu = \frac{w_1 X_1 + w_2 X_2 + \ldots + w_N X_N}{w_1 + w_2 + \ldots + w_N} \tag{2.15}$$

Example 2.7 Refer to Examples 2.1 and 2.3 regarding the incomes of gold miners from two samples.

66, 58, 71, 73, 64, 70, 66, 55, and 75

65, 57, 71, 72, 63, 71, 65, 55, and 71

List the data according to incomes and their frequencies in a table. Use the number of cases, which is the same as frequencies in this example, as weight and calculate the weighted average.

Solution 2.7

Tabulation of the data is exactly the same as in frequency tables, except the number of occurrences of incomes is named weights instead of frequencies. Recall the sample mean for the 18 family incomes is equal to

$$\hat{\mu} = \frac{1188}{18} = 66$$

Observation	Weights
55	2
57	1
58	1
63	1
64	1
65	2
66	2
70	1
71	4
72	1
73	1
75	1
Total	**18**

Adding the observations and dividing by 12 (the number of rows of data) will give an incorrect answer, except when the frequencies are all equal to 1. To obtain the correct mean, each observation must be weighted by the number of times it occurs.

Note that the sum for the population is over the population size (N), while the sum for the sample is for the sample size (n).

$$\hat{\mu} = \frac{55\times 2 + 57\times 1 + 63\times 1 + 64\times 1 + 65\times 2 + 66\times 2 + 70\times 1 + 71\times 4 + 72\times 1 + 73\times 1 + 75\times 1}{18}$$

$$\hat{\mu} = \frac{1188}{18} = 66$$

Therefore, summarizing the data in a frequency distribution does not affect the *mean*. Soon, it will be shown that it does not affect the *variance* either. The following Stata command will calculate the weighted mean for tabulated data. The command can be used to obtain the mean when the weights are actually frequencies. Note that the weights could be values other than frequencies such as the case of calculating the grade point averages where the weights are the credit hours and not the frequencies. The data points are named "observation" and their frequencies are named "weights."

```
mean observation, stdize (observation) stdweight(weights)
```

```
Mean estimation

N. of std strata =      12          Number of obs    =      12
```

	Mean	Std. Err.	[95% Conf.	Interval]
observation	66	0	.	.

Figure 2.3 Stata display of output for weighted average

The following Stata command will also compute the weighted average:

```
sum observation [fw= weights], detail
```

Mean of Grouped Data

Often, the data are available only after being summarized in tables of grouped data making it impossible to obtain statistics using the methods discussed so far. Grouped data provide condensed information about the population without requiring knowledge of statistics. With a slight modification of the previous formulas, one can calculate the appropriate statistics.

The formula for the mean of the population grouped data is

$$\mu = \frac{\sum fM}{\sum f} \tag{2.16}$$

The formula for the mean of the sample grouped data is

$$\hat{\mu} = \frac{\sum fM}{\sum f} \tag{2.17}$$

where "M" is the midpoint of each class. Once again, the only apparent difference between the population and sample formulas is in the number of elements and where they originate; however, the former produces a parameter, while the latter produces a statistic. The mean for grouped data is the same as the weighted mean where the weights are the frequencies. Each value is given a weight equal to its number of occurrences or its frequency.

Example 2.8 Group the data from Example 2.7 into four groups each representing 5-year intervals. Calculate the mean using the grouped data.

Solution 2.8

Classes	Frequency	M
55–59	4	57
60–64	2	62
65–69	4	67
70–75	8	72.5

$$\hat{\mu} = \frac{4\times 57 + 2\times 62 + 4\times 67 + 8\times 72.5}{18}$$
$$= \frac{228 + 124 + 268 + 580}{18} = 66.6667$$

The result has changed slightly. Other groupings will result in different calculated means. The lack of accuracy is due to lack of precise information about the actual magnitude of each observation. The Stata command for weighted average is applicable here as well.

Median

Fifty percent of observations are below the *median*. The median is the value in the middle of sorted data. When there is an even number of observations, use the average of the two numbers in the middle for the median. Customarily the letter *M* is used to designate the median. The median is the same as the 50th *percentile*, as well as the *second quartile*.

Example 2.9 Refer to Example 2.7. Calculate the median for each community separately. Then find the median for the combined data.

Solution 2.9

Sort the incomes for each community; pick the income in the middle. The medians are 66 and 65, respectively, and are shown in boldface numbers.

55, 58, 64, 66, **66**, 70, 71, 73, 75

55, 57, 63, 65, **65**, 71, 71, 71, 72

The median of combined samples cannot be obtained from the separate component sample medians. Therefore, it is necessary to combine the two sets and sort them first.

55, 55, 57, 58, 63, 64, 65, 65, 66, 66, 70, 71, 71, 71, 71, 72, 73, 75

Since there is even number of observations in the combined data, the median is (66 + 66)/2 = 66

Although there is not a specific command in Stata to obtain the median, the command "summarize" provides the 50th percentile, which is the same as the median. To obtain the median in Excel, type the following and include an acceptable range, usually several cells in a column.

=median(range)

Mode

The mode is the most frequent value of a population or a sample. A population, or a sample, may have more than one mode or no mode at all. If all members of the population occur equally frequently, there is either no mode or every element is a mode. If there are two modes, it is called *bimodal*. When the incomes of men and women are measured, there are two, usually distinct, modes, one for men and another for women.

Example 2.10 For the data in Example 2.7, obtain the mode for the incomes of the two communities. Find the mode for the combined data as well.

Solution 2.10

The mode for the first community is 66, and for the second community it is 71. From the frequency distribution table provided in Example 2.7, the mode for the combined data is 71, since there are four families with that income. The mode for the combined population cannot be obtained from the separate component population modes; therefore, it is necessary to combine the two sets of data.

When the data are grouped, the mode is the midpoint of the interval with the greatest frequency. In this case, the range 70 to 75 is the mode. In a *bar graph* or a *histogram*, the tallest bar represents the modal value.

The Stata command *modes varlist* will display the most frequent value for the variable. As usual, "varlist" should be replaced by the specific variable name. Data must be in its raw form and not tabulated as frequency tables for this command to work. It might be necessary to install the command by clicking on the download, after typing "*help modes*" in the command line. Excel offers two functions to obtain the mode: *mode.sngl* and *mode.mult*. The former displays the mode, while the latter displays a vertical array of the more frequently occurring values in the designated range.

Empirical Relationship between the Mean, Median, and Mode

There are some interesting relationships among the three measures of central tendencies.

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}) \tag{2.18}$$

Measures of Dispersion

Measures of dispersion are statistics that indicate how the data are scattered.

Range

The *range* (R) reflects how the data are scattered. It is calculated by subtracting the minimum from the maximum.

$$R = \text{Maximum} - \text{Minimum} \tag{2.19}$$

Example 2.11 For the data in Example 2.7, obtain the range for the combined data.

Solution 2.11

$$R = 75 - 55 = 20$$

Interquartile Range

The *interquartile range* (IQR) is a measure of dispersion that measures the distance between the first and the third quartiles.

$$\text{IQR} = Q_3 - Q_1 \tag{2.20}$$

Example 2.12 For data in Example 2.7, obtain the interquartile range for the combined data.

Solution 2.12

Combine and sort the data.

55, 55, 57, 58, **63,** 64, 65, 65, 66, **66**, 70, 71, 71, **71**, 71, 72, 73, 75

Q_1 Q_2 Q_3

$$\text{IQR} = 71 - 63 = 8$$

The IQR can be used to find the "middle class" of a population or a sample. It gives the range containing the middle 50%.

Variance

The variance is one of the more important parameters of a population and measures of dispersion. The concept of variance is used in many aspects of statistics. To demonstrate the usefulness of the variance, consider the following two hypothetical samples.

44.7778, 44.7778, 44.7778, 44.7778, 44.7778, 44.7778, 44.7778, 44.7778, 44.7778, 44.7778

And

66, 58, 71, 73, 64, 70, 66, 55, and 75

The means for both samples is 44.7778, but the two samples have different spreads. The knowledge of the mean is not sufficient to distinguish the two, so the variance is needed.

The variance is the average of the squared errors. The need for the variance arises from the need to determine and calculate the error, which is an important statistical measure. Seldom, every member of the population has the same value and if they did, statistical analysis would be trivial and irrelevant. Each member of the population will vary from the *expected value* by some magnitude. The deviation may be positive or negative depending on whether the observation is above or below the mean, respectively. This deviation from the mean is often referred to as *individual error*.

Understanding and analyzing the individual errors would be difficult and often meaningless as there are as many individual errors as there are observations. Averaging individual errors would have been a reasonable solution, except for the fact that the average of deviations from mean $(X - \mu)$ is *always zero* because the sum of individual errors $\Sigma (X - \mu)$ is always zero, a useless statistic. To overcome this problem, the deviations are squared to obtain:

$$\sum (X - \mu)^2$$

This value cannot be zero unless all observations are identical; a trivial case. Since larger populations will have larger sum of squared deviations, their average is calculated to enable comparison of different-sized populations. The result is called the *variance*.

Population Variance

The population variance is the sum of the squares of the deviations of values from their mean, divided by population size. Therefore, the variance is the *mean of the squared deviations (MSE)*, thus an average; the distinction is that it is the average of deviations, so it is a measure of dispersion not a measure of central tendency.

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{2.21}$$

The variance is also called *sigma squared* to reflect the fact that it is a squared measure. The variance reflects how much a data point deviates from the expected value, that is, the mean for data. The numerator, the sum of squares of deviations, is abbreviated as *SST*.

Sample Variance

The sample variance is the sum of the squares of the deviations of sample observations from the sample mean divided by the *degrees of freedom*. The concept of degrees of freedom is discussed in Chapter 3, although it will be used in explaining variance concept in the next few pages. Since the sample variance is a statistic, its value will change from one sample to another.

$$\widehat{\sigma^2} = \frac{\sum\left(X-\widehat{\mu_X}\right)^2}{n-1} \tag{2.22}$$

Example 2.13 Refer to the gold miners' income data of Example 2.1. Calculate the sample variances of the incomes for each community.

Solution 2.13

The sample variance for the first community is calculated by first calculating the mean then subtracting the observation, *X*, from the mean (see column 2 below), and finally squaring the result (see column 3). The last step involves summing the values in column 3 and dividing by the number of observations minus 1.

Income	$(X-\hat{\mu})$	$(X-\hat{\mu})^2$
66	–0.444	0.198
58	–8.444	71.309
71	4.556	20.753
73	6.556	42.975
64	–2.444	5.975
70	3.556	12.642
66	–0.444	0.198
55	–11.444	130.975
75	8.556	73.198
Sum	**0.00**	**358.222**

$$\widehat{\sigma_1^2} = \frac{\sum\left(X-\widehat{\mu_1}\right)^2}{n-1} = \frac{358.2222}{8} = 44.7778$$

Similarly, the sample variance for the second community is calculated and is equal to

$$\widehat{\sigma_2^2} = 40.277778$$

Verify that the variance of the combined samples is 40.23529, which is neither the sum of the variances of the two samples, nor the average of their variances. Later, the correct formula to average two variances to obtain the variance of the combined data will be presented in the section titled "Average of Sample Variances." You should also verify that the sum of $(X - \hat{\mu})$ equals zero for both samples, as expected. As shown earlier, the Stata command *summarize varlist, detail* will display the variance for all the variables. The function in Excel is

=var.s(range)

where "range" is a valid range of numbers.

Standard Deviation

The variance, population or sample, is in the square of the unit of the measurement of the observations; for example, if the unit of observation is a yard, the unit of the variance will be in yards-squared. The former is

a measure of length, while the latter is a measure of area. To make the variance comparable to the actual observations, take the square root of the variance.

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \tag{2.23}$$

The σ is called the *standard deviation*, and is pronounced *sigma*. Its counterpart is called *sample standard deviation*, which is denoted by $\hat{\sigma}$, pronounced *sigma hat*.

$$\hat{\sigma} = \sqrt{\frac{\sum \left(X - \widehat{\mu_X}\right)^2}{n-1}} \tag{2.24}$$

The standard deviation represents the *average error* of a population or sample. The standard deviation is a measure of *risk*, too. It reflects how much, on average, a data point deviates from the expected value, that is, the mean of the data.

The standard deviation is the statistical "yardstick" that allows comparison of dissimilar entities. To measure the length of a room, you place a yardstick at the beginning of the room, mark the end of the yardstick, move to the mark and place the yardstick, mark the end of the yardstick, and so on until the entire length is covered. In other words, you divide the length of the room by the length of the yardstick, and the result will be a value in terms of the yard. The divisor provides the unit of measurement. Hence, the unit of measurement of standardized values is the standard deviation.

The Standard Deviation of the Sample Mean

When the value under consideration is the *sample mean*, its distribution is explained by the *sampling distribution of the sample mean*, a topic which is covered in Chapter 5. For the time being, we will simply provide the relationship without background information.

$$\text{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n} \tag{2.25}$$

If the population variance is not known, replace it with the sample variance.

$$\widehat{Var(\hat{\mu})} = \widehat{\sigma^2_{\hat{\mu}}} = \frac{\widehat{\sigma^2}}{n} \tag{2.26}$$

where σ^2 is the population variance and $\widehat{\sigma^2}$ is the sample variance.

Definition 2.2 The square root of the variance of the sample mean is called the *standard deviation of the sample mean*.

Definition 2.3 When the expected value is a *sample statistic* such as sample mean instead of a parameter such as population mean, the resulting *standard deviation* is called the *standard error*.

Note the distinction between the standard deviation of a sample and the standard deviation of sample means. In the case of the former, one sample is taken and its standard deviation is calculated. In the case of the latter, many samples are taken, their means are calculated, and the standard deviation for those means is calculated. The *standard error* is an alternative name for this latter concept. Another important point that causes confusion is the way the standard error is obtained. Although only one sample is taken to estimate the sample standard error, the calculated value represents the standard deviation of a distribution function of the sample means, a concept that is explained in more detail in Chapter 5.

Error

In statistics, error is the amount by which each data point misses the expected value or the average. To avoid using error for two different things, σ, or the standard deviation, is called the error and σ^2, the variance, is referred to as mean squared error (MSE). The term MSE is usually used in situations where part of the variation in the data can be explained by a trend line, treatment effect, block effect, etc. and the remaining unexplained portion is called MSE. The term variance is more commonly used for the population variance, when no portion of it could be explained by other factors.

The *expected value* is the parameter that represents the population. The actual observations deviate from their mean due to random error, which cannot be explained. In statistics, this is called the *error*. The error is the portion of the total variation that cannot be explained. The error is not necessarily a fixed amount. It is the amount not explained by the given tool; change the tool and the error will change.

Some Algebraic Relations for Variance

Two important relationships are used in dealing with variances and are worth reviewing.

1. The variance of constants is zero. Given a constant "C," then

$$\text{Var}(C) = 0 \tag{2.27}$$

2. The variance of a constant multiple of a variable is equal to the square of the constant times the variance of the variable.

$$\text{Var}(CX) = C^2 \text{Var}(X) \tag{2.28}$$

Computational Formula (Shortcut)

The definitional formula for variance may result in lots of rounding during computation, especially for a large population or sample, and cause an erroneous difference. When the estimated mean is an irrational real number, deviations from it will also be irrational real numbers. An irrational real number does not have terminating or repeating decimal places. When this deviation is squared and added up, the small amount of rounding can add up and become substantial. The following formulas correct for this problem by not introducing rounding into the computation until the last stages when values are finally divided by the sample size. The computation for the population variance is

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N} \tag{2.29}$$

The derivation of the computational formula is relatively simple. It is important to point out that the letters *X* and *N*, and so on, are dummy notations and are used to represent a variable. Sometimes other letters, such as Y and M, might be used instead. The concept is the same, only the notation is different. Therefore, the variance can also be written as follows:

$$\sigma^2 = \frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N} \tag{2.30}$$

Another computational formula delays division another step.

$$\sigma^2 = \frac{N\sum X^2 - (\sum X)^2}{N^2} \tag{2.31}$$

The corresponding computational formulas for the sample variance are

$$\widehat{\sigma^2} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} \tag{2.32}$$

This is the more commonly used formula in most texts:

$$\widehat{\sigma^2} = \frac{n\Sigma X^2 - (\Sigma X)^2}{n(n-1)} \tag{2.33}$$

Example 2.14 Use the family income of gold miners from Example 2.1 to calculate the sample variance for family incomes using the computational formula.

Income	X^2
66	4356
58	3364
71	5041
73	5329
64	4096
70	4900
66	4356
55	3025
75	5625
598	40092

Solution 2.14

$$\widehat{\sigma^2} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} = \frac{40092 - \frac{598^2}{9}}{8} = \frac{40092 - 3973.7778}{8}$$
$$= \frac{358.2222}{8} = 44.7777778$$

Or

$$\widehat{\sigma^2} = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)} = \frac{9 \times 40092 - 357604}{9 \times 8} = 44.7777778$$

In this example, the choice of the formula did not make any difference in the accuracy of the results because there was not much of rounding off to begin with.

Average of Several Variances

Sometimes it is important to average several variances. Suppose two or more samples are taken from the same population and estimated (sample) variances are obtained. In order to gain a better estimate of the population variance, all the variances should be averaged. If the sample sizes are the same, a simple average will provide the desired mean. If sample sizes are different, however, the observations, which in this case are the variances, should be weighted. The logical weights are the sample sizes. When estimating variances, sample sizes are replaced by degrees of freedom. This weighted mean of variances is usually called a "pooled" variance. Recall that the formula for the weighted mean is

$$\mu = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_N X_N}{w_1 + w_2 + \cdots + w_N} \tag{2.34}$$

Since we are dealing with pooled variances, and since we use the hat symbol for sample statistics, we will use the commonly used symbol σ^2_{pooled}. The weights $(w_1, w_2, ..., w)$ are degrees of freedom ($n_i - 1$), and the *X*s are the sample variances, $\widehat{\sigma_1^2}, \widehat{\sigma_2^2}, ..., \widehat{\sigma_k^2}$, where $n_1, n_2, ..., n_k$ are sample sizes. The formula for the case of two sample (estimated) variances is

$$\widehat{\sigma^2_{pooled}} = \frac{(n_2 - 1)\widehat{\sigma_1^2} + (n_2 - 1)\widehat{\sigma_2^2}}{(n_1 - 1) + (n_2 - 1)} \tag{2.35}$$

Repeat the pattern for averages of three or more sample variances. The case when samples are from different populations is discussed in Chapter 5.

Example 2.15 Calculate the weighted variance for the variances of the Microsoft stock prices between May 21 to June 11 and June 12 to July 2, 2015 from Example 2.2.

Solution 2.15

We already have the following results:

$$\widehat{\sigma_1^2} = 0.348735 \qquad \widehat{\sigma_2^2} = 0.614498$$

Since the two samples are from the same company and same year, it is likely that the population variance has not changed and these are two estimates of the same variance. Therefore, the two sample variances should be pooled. We also know that the sample size for each sample is 15.

$$\begin{aligned}\widehat{\sigma^2_{pooled}} &= \frac{(n_1 - 1)\widehat{\sigma_1^2} + (n_2 - 1)\widehat{\sigma_2^2}}{n_1 + n_2 - 1} \\ &= \frac{(15-1)(0.348735) + (15-1)(0.614498)}{15+15-2} \\ &= \frac{4.88229 + 8.602969}{28} = \frac{13.48526}{28} = 0.48\end{aligned}$$

In this and similar examples, when the sample sizes are the same, the outcomes of using the weighted average and simple arithmetic average will be the same.

Variance of Data with Frequencies

Earlier, the method of obtaining a mean for data presented in a frequency table was shown. In a similar way, the variance of data from a frequency table can be obtained. The formula for the population variance for the frequency distribution is

$$\sigma^2 = \frac{\sum f(X-\mu)^2}{\sum f} \tag{2.36}$$

The formula for the sample variance for the frequency distribution is

$$\widehat{\sigma^2} = \frac{\sum f(X-\mu)^2}{\sum f - 1} \tag{2.37}$$

where "f" represents the frequency of each variable. The limit of "$\sum$" for a population is "N," while that of a sample is "n."

The computational formula for a population is

$$\sigma^2 = \frac{\sum fX^2 - \dfrac{(\sum fX)^2}{\sum f}}{\sum f} \tag{2.38}$$

The computational formula for a sample is

$$\widehat{\sigma^2} = \frac{\sum fX^2 - \dfrac{(\sum fX)^2}{\sum f}}{\sum f - 1} \tag{2.39}$$

Example 2.16 Use the family income of gold miners from Example 2.7 to calculate the sample variance using the frequency table.

Solution 2.16

The sample variance for the combined 18 family incomes is equal to

$$\widehat{\sigma^2} = \frac{\sum(X-\widehat{\mu_X})^2}{n-1} = 40.23529412$$

Create a frequency table. Note that there are 18 observations and not 12. The calculation of the sample variance for the above data using the computational formula follows:

X	f	Xf	X^2	X^2f
55	2	110	3025	6050
57	1	57	3249	3249
58	1	58	3364	3364
63	1	63	3969	3969
64	1	64	4096	4096

X	f	Xf	X^2	X^2f
65	2	130	4225	8450
66	2	132	4356	8712
70	1	70	4900	4900
71	4	284	5041	20164
72	1	72	5184	5184
73	1	73	5329	5329
75	1	75	5625	5625
Total	18	1188		79092

$$\widehat{\sigma^2} = \frac{79092 - \frac{(1188)^2}{18}}{17} = \frac{79092 - 78408}{17} = \frac{684}{17} = 40.23529$$

The answer is the same as the one obtained using the raw data. Therefore, summarizing the data into frequency distributions does not affect the *variance*. The variance of the grouped data, however, will most likely be different from the actual variance. The following command in Stata will yield the same result.

```
. sum observation [fw= frequency], detail

                        observation
-------------------------------------------------------------
      Percentiles      Smallest
 1%           55            55
 5%           55            57
10%           55            58       Obs                  18
25%           63            63       Sum of Wgt.          18

50%           66                     Mean                 66
                      Largest        Std. Dev.       6.34313
75%           71            71
90%           73            72       Variance       40.23529
95%           75            73       Skewness       -.4952022
99%           75            75       Kurtosis       2.04109
```

Figure 2.4 Stata display of the variance

The variable name is "observation" and the weights are named "frequency." If your variables have different names, make sure to substitute the correct name in the command.

Variance for Grouped Data

The formula for the variance of the grouped data for the population is

$$\sigma^2 = \frac{\sum fM^2 - \frac{\left(\sum fM\right)^2}{\sum f}}{\sum f} \tag{2.40}$$

where M is the midpoint of each class. The formula for the variance of the grouped data for the sample is

$$\widehat{\sigma^2} = \frac{\sum fM^2 - \frac{\left(\sum fM\right)^2}{\sum f}}{\sum f - 1} \tag{2.41}$$

where M is the midpoint of each class.

Measures of Association

Measures of association determine the association between two variables or the degree of association between two variables. The extension to more variables is less common since more advanced methodologies have been developed.

Population Covariance

The covariance is a measure of association between two variables. Covariance associates the deviation of one variable from its own mean to the deviation of another variable from its mean, on average. Let the variables be X and Y and their corresponding means be μ_X and μ_Y. The covariance is defined as

$$\text{Cov}(X,Y) = \frac{\sum(X-\mu_X)(Y-\mu_Y)}{N} \tag{2.42}$$

The covariance is the sum of the cross product of the deviations of the values of X and Y from their means divided by the population size. Sometimes it is written as σ_{XY}, which should not be mistaken as a

standard deviation. This compares to the notation of σ_X^2 for the variance of the population. The covariance of a variable with itself is actually its variance, that is, $\sigma_{X,X} = \sigma_X^2$.

The definitional formula for covariance suffers from the rounding error and also can become very tedious if the means have long significance digits. The computational formula for covariance is

$$\sigma_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{N} \tag{2.43}$$

Sample Covariance

In the sample covariance, the population means are not known and have to be replaced by the sample means. Consequently, the covariance loses a degree of freedom. The theoretical formula for the sample covariance is

$$\widehat{\sigma_{XY}} = \frac{\sum(X - \widehat{\mu_X})(Y - \widehat{\mu_Y})}{n-1} \tag{2.44}$$

The computational formula for the sample covariance is

$$\widehat{\sigma_{XY}} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{n-1} \tag{2.45}$$

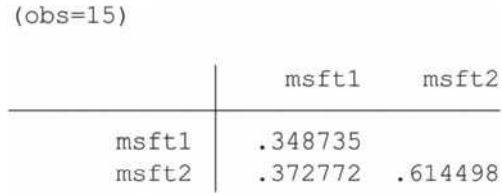
The covariance shows association between two variables. The magnitude of the covariance is a function of the degree of association as well as the units of measurement of the values of the two variables. Changing the unit of measurement, from inch to yard for example, will change the covariance.

Example 2.17 What is the covariance for the closing prices for Microsoft from May 21 to June 11 and June 12 to July 2, 2015, from Example 2.2?

Solution 2.17

Stata calculates the covariance as part of computations for the correlation coefficient. Covariance is obtained as an option with the *correlate* command.

correlate msft1 msft2, covariance



(obs=15)

	msft1	msft2
msft1	.348735	
msft2	.372772	.614498

Figure 2.5 Stata display of covariance

Excel command for covariance is

=covariance.s(range1,range2)

where range is any valid Excel range, such as cells in a column. In this case, you will enter the May 21 to June 11 data as the first range, and the June 12 to July 1 data as the second range, with the two ranges separated by a comma. Verify that in Excel you get 0.372772381 for the covariance.

Correlation Coefficient

The correlation coefficient ρ (rho) uses the measures of association and dispersion to provide a new measure without a unit allowing comparisons of associations between unrelated things measured in different units. The measure of association is the covariance and is placed on the numerator. The measures of dispersion are the standard deviation of *X* and the standard deviation of *Y*, which are placed in the denominator. All three are subject to change when the unit of measurement changes, but the correlation coefficient is immune. The formula for the correlation coefficient is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.46}$$

where σ_{XY} is the covariance, σ_X is the standard deviation of the *X* values, and σ_Y is the standard deviation of the *Y* values. The sample correlation coefficient $\hat{\rho}$ is written as

$$\hat{\rho} = \frac{\widehat{\sigma_{XY}}}{\widehat{\sigma_X}\,\widehat{\sigma_Y}} \tag{2.47}$$

Substitute the formulas for the population covariance and the standard deviations:

$$\rho = \frac{\dfrac{\sum (X-\mu_X)(Y-\mu_Y)}{N}}{\sqrt{\dfrac{\sum (X-\mu_X)^2}{N}}\sqrt{\dfrac{\sum (Y-\mu_Y)}{N}}} \tag{2.48}$$

similarly for the sample correlation coefficient,

$$\hat{\rho} = \frac{\dfrac{\sum (X-\widehat{\mu_X})\sum (Y-\widehat{\mu_Y})}{n-1}}{\sqrt{\dfrac{\sum (X-\widehat{\mu_X})^2}{n-1}}\sqrt{\dfrac{\sum (Y-\widehat{\mu_Y})^2}{n-1}}} \tag{2.49}$$

The corresponding computational forms for the population and the sample are given in equations 2.50 and 2.51, respectively.

$$\rho = \frac{\sum XY - \dfrac{(\sum X)(\sum Y)}{N}}{\sqrt{\sum X^2 - \dfrac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \dfrac{(\sum Y)^2}{N}}} \tag{2.50}$$

$$\rho = \frac{\dfrac{\sum XY - \dfrac{(\sum X)(\sum Y)}{n}}{n-1}}{\sqrt{\dfrac{\sum X - \dfrac{(\sum X)^2}{n}}{n-1}}\sqrt{\dfrac{\sum Y - \dfrac{(\sum Y)^2}{n}}{n-1}}} \tag{2.51}$$

In practice $(n-1)$ are not even written in the formula, as they cancel out. They are shown here for pedagogical reasons. Do not overlook the fact that ρ is a parameter and a constant, while $\hat{\rho}$ is the statistic and a variable. The sample correlation coefficient $\hat{\rho}$ is used to estimate and draw inference about the population correlation coefficient ρ.

Example 2.18 What is the correlation coefficient for the closing prices for Microsoft from May 21 to June 11 and June 12 to July 2, 2015, from Example 2.2?

Solution 2.18

Stata calculates correlation coefficient using the following command:

```
correlate msft1 msft2
```

where msft1 and msft2 are the variable names.

	msft1	msft2
msft1	1.0000	
msft2	0.8053	1.0000

Figure 2.6 Stata display of correlation coefficient

The Excel command for correlation is

```
=correl(range1,range2)
```

where range1 and range2 should be valid Excel ranges, such as cells in a column. It is necessary for the two ranges to be of equal size, that is, have the same numbers of rows. Verify that you get a result of 0.805258913 using Excel.

CHAPTER 3

Some Applications of Descriptive Statistics

Introduction

The descriptive statistics that were covered in Chapters 1 and 2 provide summary statistics and graphical methods to present data in concise and meaningful ways. Although those measures and methods are useful in their own right, they are also used to create more powerful statistical measures, some of which are discussed in this chapter. Later, in Chapter 7, these measures are utilized to provide statistical inference. Statistical inference is the foundation of testing hypotheses in every branch of science.

Coefficient of Variation

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. In other words, CV expresses the standard deviation (the average error) as the percentage of the average of the data. It is a relative measure of dispersion.

$$\text{CV} = \frac{\sigma}{\mu} \tag{3.1}$$

CV is independent of the unit of measurement of the variable. If two populations have the same standard deviation, the one with the smaller mean has relatively more variation. In general, the smaller the CV is, other things equal, the more uniform and compact are the data points. A smaller CV indicates less volatility and risk.

Example 3.1 The manager of the mortgage department in a local bank has gathered the amount of approved second mortgage loans for every 100th customer. Calculate the CV.

5,672	6,578	9,700	12,000	9,000	6,350	4,495	6,900	7,835
8,750	10,000	12,000	6,500	7,200	8,000	18,000	19,000	12,000
4,560	1,500	5,900	5,450	6,500	1,800	1,900	10,500	

Solution 3.1

First calculate the standard deviation, $\hat{\sigma}$, followed by the mean, $\hat{\mu}$. To obtain the CV, divide $\hat{\sigma}$ by $\hat{\mu}$.

$$\widehat{CV} = \frac{\hat{\sigma}}{\hat{\mu}} = \frac{4257.58}{8003.46} = 0.532$$

Assume that two stocks are rated similarly where they have the same characteristics such as objectives and the amount and frequency of dividends. In order to compare the *relative risk* of the two stocks, use the CV. The stock with the lower CV indicates lower variation and hence lower risk. The CV is also useful in comparisons of unrelated data, especially when the units of measurement are different. For example, if the reliability of a gas-powered lawnmower is compared with the reliability of an electric edger, then the machine with the lower CV is more reliable.

The most effective use of the CV is to compare two different experiments by finding the ratios of their respective CVs, as demonstrated in the following example.

Example 3.2 Refer to the stock prices for Wal-Mart and Microsoft from Example 2.2. Determine which one is riskier using all 30 observations.

Solution 3.2

Let's refer to Wal-Mart stock as "1" and to the Microsoft stock as "2." Recall we calculated the means and standard deviations for these stocks.

$$\widehat{\mu_1} = 73.215 \qquad \widehat{\sigma_1} = 1.3721$$

$$\widehat{\mu_2} = 46.096 \qquad \widehat{\sigma_2} = 0.9295$$

$$\text{CV}_1 = \frac{1.3721}{73.215} = 0.01874$$

$$\text{CV}_2 = \frac{0.9295}{46.096} = 0.0202$$

$$\frac{CV_1}{CV_2} = \frac{0.01874}{0.0202} = 0.929$$

Therefore, stock 1, Wal-Mart, is less risky than stock 2, Microsoft, for the sample period.

Z Scores

The *Z* score is a useful and intuitive concept and is used often in statistics. The *Z* score uses two of the more common parameters, the mean and the standard deviation. The problem of accurate and consistent measurement has been a challenge throughout history. The yardsticks differed from one time to another and across different locations and cultures. Different countries and rulers tried to unify the units of measurement. The closest unit to become universally accepted is the meter. One problem with any system of measurement is differences in scale of the items being measured. The following example demonstrates the problem.

County fairs give prizes for the "best" in different categories. For example, the farmer with the biggest produce receives a prize. However, by nature, even the largest peach on record is not a match for any watermelon. Would it be fair to compare the amount of milk a cow produces with that of a goat? In economic terms, how can we compare the output of a small manufacturer with that of a larger one to compare productivity?

In statistics, everything is measured in relative terms. Let's have a peach that weighs 8.4 ounces and a watermelon that weighs 274.9 ounces. Although the watermelon is actually heavier, that might not be the case when other factors are considered. One factor is the average weight of peaches and watermelons. A typical peach is about 6 ounces, whereas a typical watermelon is 22 pounds, or 352 ounces. The peach in this example is somewhat heavier than an average peach, while the watermelon is actually lighter than an average watermelon. Therefore, relatively speaking, the peach is heavier than the watermelon. However, we can do even more meaningful and more precise comparisons.

Dividing the deviation from the mean by the standard deviation takes into account the spread of the data as well. Let's assume that the standard

deviation for peaches is 1.15 ounces. Therefore, the amount this particular peach exceeds its average as measured by its own yardstick is (8.4 – 6) /1.15 = 2.086 standard deviations. Let's assume that the standard deviation for the watermelon is 2.57 pounds or 16 × 2.57 = 41.12 ounces. Therefore, the watermelon is (274.9 – 352)/41.12 = –1.875 standard deviations below its expected weight. This is the essence of what is called a *Z score* and the procedure is known as *standardization*. The Z score is defined as follows:

$$Z = \frac{\text{Observed} - \text{Expected}}{\text{Standard Deviation of the Observed}} \tag{3.2}$$

$$Z = \frac{x - \mu}{\sigma} \tag{3.3}$$

The expected value of an observation is its mean or (μ); its standard deviation is (σ). The distance of an observation (X) from its expected value, $(X - \mu)$, is also called its (individual) *error*. Some of the different aspects of error will be discussed later in this chapter. The Z score is a scaled error. The unit of measurement of a Z score is the standard deviation of the population or its sample estimate. The standardization process can be applied to any data or observation.

Suppose two students are given the task of measuring the error of an observation. The first student finds that the observation deviates from the mean by 41.6666 feet. In fact there are infinite numbers of decimal digits; such numbers are called irrational numbers. Disliking decimal points and especially the irrational ones, he decides to measure the deviation of the data point from the mean in inches and is relieved to find out that it has no decimal point. He presents the error of the observation as 500 inches. The second student, using an electronic measuring tape, subtracts the observation from the mean and reports 0.007891414 miles as the error. While the first error seems large and the second seems small, both are the same; 500 inches is 41.66667 feet or 0.007891414 miles. Z scores provide a unique and comparable measure of error to avoid the confusion that may arise from changes in units of measurement. Every error is reported in the units of its own standard deviation. Since Z scores are reported in terms of the standard deviation, they allow comparison of unrelated data measured in different units.

Z Score for a Sample Mean

If the value under consideration is the sample mean, $\hat{\mu}$, the resulting Z score would be

$$Z = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \tag{3.4}$$

where $\hat{\mu}$ is the sample mean, μ is its expected value, which is also the population mean, and the standard deviation of the sample is $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation of the sample mean is also known as the *standard error*. In order to calculate the Z score for a sample mean, it is necessary to know the population variance. When the population variance is unknown, we must use a t value instead of a Z value. We will discuss t distributions in more detail in Chapter 4. For the time being, we will assume we know the population variance.

Example 3.3 Calculate the Z scores for closing prices of Microsoft stock from May 21 to July 2, 2015. Assume the population's mean and variance are equal to the sample mean and variance, respectively. Treat the sample variance as the population variance.

Solution 3.3

Use Equation (3.4) and the following values from Example 3.2.

$$\widehat{\mu_2} = \mu = 46.096 \qquad \widehat{\sigma_2} = \sigma = 0.9295$$

Note that we are using the notation for population mean and population standard deviation to make sure that the theorem applies (see Table 3.1). The sum of Z scores is provided at the bottom, which is zero. Adding up to zero is a mathematical property of the *individual errors*. See Definition 3.5.

Theorem 3.1 Chebyshev's Theorem

The proportion of observations falling within K standard deviations of the mean is *at least*

$$\left(1 - \frac{1}{k^2}\right) \tag{3.5}$$

Date	MSFT	*Z* Score
5/21/2015	47.42	1.424079
5/22/2015	46.9	0.864636
5/26/2015	46.59	0.531117
5/27/2015	47.61	1.628496
5/28/2015	47.45	1.456358
5/29/2015	46.86	0.8216
6/1/2015	47.23	1.219668
6/2/2015	46.92	0.886149
6/3/2015	46.85	0.810838
6/4/2015	46.36	0.28367
6/5/2015	46.14	0.046978
6/8/2015	45.73	–0.39412
6/9/2015	45.65	–0.48019
6/10/2015	46.61	0.552635
6/11/2015	46.44	0.369736
6/12/2015	45.97	–0.13592
6/15/2015	45.48	–0.66309
6/16/2015	45.83	–0.28654
6/17/2015	45.97	–0.13592
6/18/2015	46.72	0.67098
6/19/2015	46.1	0.003942
6/22/2015	46.23	0.143807
6/23/2015	45.91	–0.20047
6/24/2015	45.64	–0.49095
6/25/2015	45.65	–0.48019
6/26/2015	45.26	–0.89978
6/29/2015	44.37	–1.8573
6/30/2015	44.15	–2.09398
7/1/2015	44.45	–1.77123
7/2/2015	44.4	–1.82502
Mean	46.09633	0

This is the same as the Z *score* concept. The theorem indicates that we need to find the difference of a value from its mean, that is, $(X - \mu)$. Since the theorem applies to all the values within K standard deviations, that is, $K\sigma$, on either side of the mean, the absolute value is desired. The theorem sets a minimum limit for the $|X - \mu| < K\sigma$. Therefore, Chebyshev's theorem states,

$$P\left(|X - \mu| < K\sigma\right) \geq \left(1 - \frac{1}{K^2}\right) \tag{3.6}$$

Since σ is a nonnegative value, dividing both sides of the inequality $|X - \mu| < K\sigma$ by σ will not change the sign. Therefore,

$$P = \left(\frac{|X - \mu|}{\sigma} < K\right) \geq \left(1 - \frac{1}{K^2}\right)$$

$$P = \left(|Z| < K\right) \geq \left(1 - \frac{1}{K^2}\right) \tag{3.7}$$

As is evident, the Z score is the core of the Chebyshev's theorem. Chebyshev's Theorem is used in the Central Limit Theorem, which will be covered in Chapter 5. The first part of the equation $\{P(|Z| < K)\}$ is the same as the *confidence interval* of a range, which is covered in Chapter 6.

Example 3.4 Determine what percentage of the Microsoft stock prices from May 21 through July 2, 2015, fall within 2 standard deviations of their mean. Verify the correctness of the theorem by counting the prices that are within 2 standard deviations of the mean. Assume the population's mean and variance are the same as the sample mean and variance, respectively. See Example 3.2 for data.

Solution 3.4

The sample mean and sample variance are the same as given in Example 3.2:

$$\mu = \widehat{\mu_2} = 46.096 \qquad \sigma = \widehat{\sigma}_2 = 0.9295$$

Note that we are using the notation for population mean and population standard deviation to make sure that the theorem applies. The theorem requires the knowledge of population mean and standard deviation, and

we are assuming we know them to be the values we obtained from the sample. Insert the information in Equation (3.6):

$$P = \left(|X - \mu| < K\sigma\right) \geq \left(1 - \frac{1}{K^2}\right)$$

$$P(|\,X - 46.096\,| < 2\,(0.9295) \geq \left(1 - \frac{1}{4}\right)$$

$$P(|\,X - 46.096\,| < 1.8590) \geq 0.75$$

Therefore, at least 75% of the 15 observations will be within 2 standard deviations of the mean. Next we calculate the range "within 2 standard deviations."

$$46.096 - 1.8590 = \$44.237$$

$$46.096 + 1.8590 = \$47.955$$

The easiest way to verify the validity of the result is to sort the data. It is evident that only the price on June 30, 2015, is outside the range \$44.237 to \$47.955. Therefore, 29 out of 30 prices (96.67%) are within two standard deviations of the mean, which exceeds the predicted minimal percentage of 75% guaranteed by the theorem.

Correlation Coefficient Is the Average of the Product of Z Scores

The correlation coefficient was introduced in Chapter 2. It measures the degree of association between two variables.

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\dfrac{\sum (X - \mu_X)(Y - \mu_Y)}{N}}{\sigma_X \sigma_Y} = \frac{\sum \left(\dfrac{(X - \mu_X)}{\sigma_X} \times \dfrac{(Y - \mu_Y)}{\sigma_Y} \right)}{N}$$

$$= \frac{\sum Z_X Z_Y}{N}$$

The above derivation depends on the definition of a parameter as a constant. This allows moving parameters, such as standard deviations, into a summation notation. Note that anything that is added and divided by the number of observations is an average number. In this case, the

product of two Z scores ($Z_X Z_Y$) are added and divided by N. Hence, the correlation coefficient is the average of the product of two Z scores.

Standard Error

When data are obtained from a sample, the standard deviation of the estimated sample mean is called a *standard error*. This concept will be addressed in detail when the sampling distribution of the sample mean is discussed in Chapter 5. Since the distributional properties of the *sample standard deviation* are different from that of the *population standard deviation*, we had to assume that the standard deviations obtained from the samples in Examples 3.2 and 3.4 were actually those of the population standard deviation.

$$\text{Standard error} = \sigma_{\hat{\mu}} = \sqrt{\sigma^2_{\hat{\mu}}}$$

Usually, the standard deviation of the sample mean is also *unknown* and has to be estimated, which is represented with a *hat*.

$$\text{Sample standard error} = \widehat{\sigma_{\hat{\mu}}} = \sqrt{\widehat{\sigma^2_{\hat{\mu}}}}$$

When the population variance is not known, the distributional properties of the Z score changes, and thus the resulting equation obtains a *t distribution function*. Distribution functions are explained in Chapter 4. The correct standardization of the sample mean in this case is

$$t = \frac{\hat{\mu} - \mu}{\widehat{\sigma_{\hat{\mu}}}} \tag{3.8}$$

Equation (3.8) has a t distribution instead of a "normal" distribution and is called a t instead of a Z. If the population variance is unknown and has to be estimated by the sample variance, the resulting standardization does not have a normal distribution but instead has a *Student's t distribution*.[1]

Definition 3.1 A *degree of freedom* is the number of elements in a sample that are unconstrained. Degrees of freedom only apply to samples, as

population parameters are constant values. "When computing the variance, if the mean is unknown we lose a degree of freedom."

Example 3.5

This example is designed to demonstrate that the *population mean* governs the outcome of the *sample mean* and how as a result a degree of freedom is lost. Let us have a small population, say of five persons. For example, consider a family with five children ages 3, 5, 7, 8, and 9. Let us take samples without replacement of size 3 from this population. (If those chosen are not returned back into the pool of possible values, the sampling is considered to be without replacement.) Although sampling without replacement affects the probability of an outcome, we are not concerned with that here. This will assure the same child is not included more than once in any sample. There will be 10 different possible samples. The population mean is 32/5 = 6.4, that is, the average age of the children is 6.4 years. The 10 possible samples and their corresponding means are

Sample	Mean
8, 3, 7	6
8, 3, 5	5.333
8, 3, 9	6.667
8, 7, 5	6.667
8, 7, 9	8
8, 5, 9	7.333
3, 7, 5	5
3, 7, 9	6.333
3, 5, 9	5.667
7, 5, 9	7
Total	64

Even though none of the sample means equals the population mean, the population mean has exerted its influence on the sample means. The average of the 10 possible sample means, that is, (64/10) = 6.4 is exactly equal to the population mean. Even if we do not know the population mean, every population has a mean, and that mean will influence all the sample means and the average of all the sample means is always equal to the population mean, a rather simple exercise in algebra, that is, $E(\widehat{\mu}) = \mu$.

In the above example, any 9 of the 10 possible samples can be *chosen freely*. After nine samples are obtained, the 10th, or the last one, is forced to have a mean value such that the average of all the 10 samples means equals the population mean. Let us assume that the fourth possible sample in the above list is the one that is not taken yet, and its mean has not been calculated. Using the sum of the other samples, the mean of this sample has to be (64 – 57.222) = 6.778. Within this last remaining sample, any two numbers can be selected at random, but the last one must be such a number that the average of the three sample units equals 6.778. Two of the three numbers 8, 7, and 5 can be randomly and freely selected; say 5 and 7 are selected. The last number must be 8, since this is the only number that will make the average of this sample equal to 6.778 and the average of all sample means equal to $\mu = 6.4$. In general, $n - 1$ sample points can be selected at random, but the value of the remaining one will be determined automatically by the value of the population. One degree of freedom is lost for every parameter that is unknown and must be estimated by a statistic.

Remember from Chapter 2 that computation of the variance requires the knowledge of the population mean:

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

If the population mean μ is not known, the value of σ^2 cannot be determined. If instead of μ, an estimate is obtained, then the sample mean, or $\hat{\mu}$, is used. The sample variance then is

$$\widehat{\sigma^2} = \frac{\Sigma\left(X-\hat{\mu}\right)^2}{n-1}$$

and it will lose one degree of freedom. The result of the adjustment to the sample variance, that is, dividing by the degrees of freedom, $n - 1$, instead of the sample size, n, is that the sample variance becomes an *unbiased estimate* of the population variance.[1]

Properties of Estimators

Sample statistics are used as estimators of population parameters. Since sample statistics provide a single value, they are also called *point estimates*. It is desirable to be able to compare different point estimates of the same parameters.

Let θ, pronounced *theta*, be the population parameter of interest. Let its estimate be $\hat{\theta}$, pronounced *theta hat*. Like any other point estimates, $\hat{\theta}$ is a sample statistic and a known variable.

Definition 3.2 Unbiasedness If the expected value of a point estimate equals the population parameter, then the estimate is *unbiased*. In symbols

$$E(\hat{\theta}) = \theta \tag{3.9}$$

It can be shown that the sample *mean* $(\hat{\mu})$, *variance* $(\widehat{\sigma^2})$, and *proportion* $(\hat{\pi})$ are all unbiased estimates of their corresponding population parameters.

$$E(\hat{\mu}) = \mu$$

$$E(\widehat{\sigma^2}) = \sigma^2$$

$$E(\hat{\pi}) = \pi$$

Definition 3.3 Efficiency A point estimator is a more *efficient* estimator if it has a smaller variance. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two point estimates of θ and Var $(\hat{\theta}_1)$ < Var $(\hat{\theta}_2)$, then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. For example, the sample mean is more efficient than the sample median in estimating the population mean.

Definition 3.4 Consistency A point estimator is a *consistent* estimator if its variance gets smaller as the sample size increases. The variance of the sample mean is defined as

$$\widehat{\sigma^2} = \frac{\sigma^2}{n}$$

This ratio decreases as the sample size increases. Since the population variance is a parameter, it is a constant; therefore, as the sample size

increases the ratio decreases. Since it is an unbiased estimate of the population mean, it will get closer and closer to the population mean.

Error

Statistics deals with random phenomena. For a set of values X_1, X_2, ..., X_n, there is a representative or expected value (mean). The Greek letter μ is used to represent the expected value. The difference of each value from the expected value, also called the deviation from the mean, represents what is called *individual error*.

Definition 3.5 The *individual error* is the difference between the value of an observation and its expected value. It is also called the *residual*, as it is the residual amount that remains unexplained.

The *expected value* or the *mean* is the best estimate or representative of a population. For example, according to PayScale.com, the median starting salary for economic majors in 2015 was $48,500.[2] If a recent economics graduate is selected at random and his or her actual income is $48,000, then the error associated with this observation is $500. In other words, the observation missed the expected value by $500. The reason for calling the deviation an error is that we do not have any explanation for the deviation other than a random error. Therefore, the *error* is what we cannot explain.

Since observations vary at random, the errors vary at random as well. Furthermore, the portion that cannot be explained depends on the model or procedure used. Sometimes, it is possible to explain part of the variation of observations from their expected value by developing more sophisticated methods.[3] The portions that can be explained by the new procedure are no longer "unexplained" and, thus, not part of the error any more. The remaining unexplained portion is still called an error. Note that unless all the observations in a sample or population are identical, they will deviate from their expected value and, hence, have a random error. A regression model explains part of the individual error using a line, thus improving estimation of the parameter. Regression models are covered in more detail in Chapter 8.

Since there are as many individual errors as there are observations, we need to summarize them into fewer values. A popular and useful

statistic is the average or mean. However, the average of the individual errors is always zero because the sum of all the errors is zero. Recall that individual errors are deviations from their expected values, some of which are negative and the others are positive. Thus, by definition, they cancel each other out and the sum of all deviations from their expected values is *always zero*.

Distributions that are symmetrical have equal numbers of positive and negative individual errors, but this is not a necessary condition for their sum to add to zero. The sum of individual errors for non-symmetric distributions is also zero, in spite of the fact that the counts of negative values are different from the counts of the positive values. This is due to the fact that the expected value or the mean is the same as the *center of gravity* of the data. Imagine data on a line where they are arranged from smallest to largest. Placing a pin at the point of the average will balance the line.

There are several ways to overcome individual errors canceling each other out. One way is to use the absolute value of the individual errors. The average of the absolute values of the individual errors is called *mean absolute error (MAE)*. The mean absolute error is commonly used in time series analysis. One advantage of MAE is that it has the same unit of measurement as the actual observations. Another way to prevent individual errors from canceling each other out is to square them before averaging them. We are already familiar with this concept, which is called the *variance*.

Definition 3.6 The *variance* is the average of the sum of the squared individual errors.

One advantage of the variance over the mean absolute error is that it squares the errors, which gives more power to values that are further away from the expected value. This makes the variance more sensitive in signaling outlying data points, as values farther away from the mean have a larger impact on the variance than those closer to the mean.

The variance's use of the squared values of the individual errors is its shortcoming as well because the unit of measurement for the variance is the square of the unit of measurement of the observations. If the observations are about length in feet, then the variance will be in feet squared,

which is the unit of measurement of an area and not length. Seldom, if ever, the squared values of economic phenomenon have any meaning. If the variable of interest is price measured in dollars, then the unit of measurement of its variance is in dollars-squared, which has no economic meaning. To remedy this problem, it is necessary to take the square root of variance.

Definition 3.7 The *standard deviation* is the square root of the variance.

Definition 3.8 The standard deviation is the *average error*.

How Close Is Close Enough?

By definition, the sum of the residual is zero; if yours fail to be exactly zero, check your formulas and computations. When the formula is correct and there is no computational error, then the problem is due to rounding. Use of computational formulas and less rounding of the data will reduce or eliminate this problem provided the sample size is sufficiently large. If you use five decimal places, the final result can be accurate to about four significant digits. If you have been using five decimal places in your calculations and the sum of the residual is 0.00007, the property has not been violated. It is zero to four decimal places as expected.

Sum of Squares

The sum of squares of deviations of values from their expected value (mean) is a prominent component of statistics. As we saw above, the sum of squares of these individual errors divided by the population size is called population variance. The concept is the same for a sample, except that the divisor is the *degree of freedom* instead of the sample size. Earlier, it was explained that the portion of the phenomenon that cannot be explained is called an error, and that the variance is one way of representing it. When alternative models are used to explain part of the error, it is more meaningful to focus on the numerator alone, at least at first. The numerator of the variance is also called the *total sum of squares (SST)*. Once a set of data is collected, then the total sum of squares

becomes fixed and will not change. The total sum of squares will change only if another sample from the same population is collected at random.

Decomposition of SST is very common in a branch of statistics called *Experimental Design*. In experimental design methodology, SST is decomposed into different components based on the design. These components include treatment SS, block SS, main effect SS, etc. In all cases, there is always a component that remains unexplained, and is referred to as the residual or error SS. By definition, dividing this unexplained remainder by appropriate degrees of freedom would result in the variance of the experiment, customarily known as *mean squared error (MSE)*. As one would expect, the square root of MSE is called *root MSE*, which is the same as the *standard error*. Just as a reminder:

$$\text{MSE} = \frac{\text{Sum}\left(\text{Observed} - \text{Expected}\right)^2}{n-k}$$

Note that the term in the parentheses is the individual error. The (K) in the denominator is the number of parameters in the model, and the entire denominator is the degrees of freedom.

Nonsymmetry of Data

In Chapter 2, several of the relationships between the mean, mode, and median were described. The relationship between these parameters indicates if a graph of the data is symmetric, skewed, or pointed.

Definition 3.9 *Skewness* refers to the extent that a graph of a distribution function deviates from symmetry toward the left or the right.

A distribution function that is not symmetric is either negatively skewed, as in Figure 3.1, or positively skewed, as in Figure 3.2.

Relationship between the Mean, Median, and Mode

The mean, median, and mode of symmetrical distributions are identical. If the distribution is positively skewed, the order of magnitude for these three parameters is

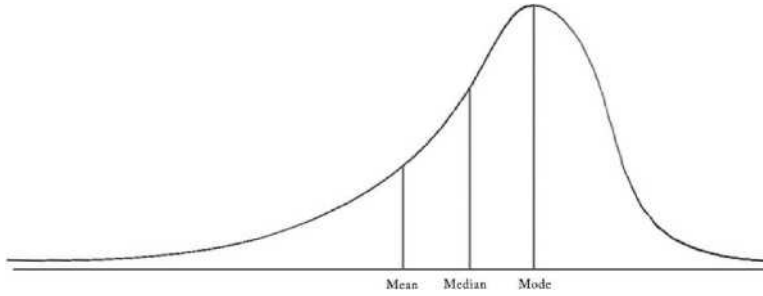


Figure 3.1 Negatively skewed distribution

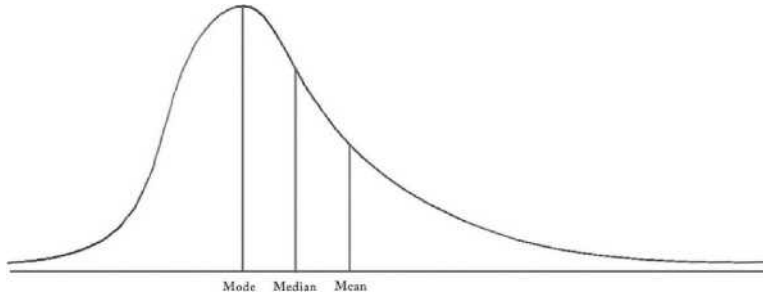


Figure 3.2 Positively skewed distribution

$$\text{Mode} < \text{Median} < \text{Mean} \tag{3.10}$$

If the distribution is negatively skewed, the order of magnitude is

$$\text{Mean} < \text{Median} < \text{Mode} \tag{3.11}$$

Skewness is used to test if the data follows a normal distribution. The normal distribution function is discussed in Chapter 4.

Definition 3.10 The Pearson's coefficient of skewness is defined as

$$S = \frac{3(mean - median)}{standard\ deviation}$$

The range of skewness is $-3 < S < 3$, and for a symmetric distribution $S = 0$. The sign of the Pearson's coefficient of skewness determines whether it is positively or negatively skewed.

Definition 3.11 *Kurtosis* is a measure of pointedness or flatness of a symmetric distribution.

A positive kurtosis indicates the distribution is more pointed than the normal distribution, and a negative value for kurtosis indicates the distribution is flatter than the normal distribution. Kurtosis and skewness are commonly used to test whether a data set follows a normal distribution.

One formulation for computing kurtosis is available that only uses the concepts discussed earlier.

$$K = \frac{Q}{P_{90} - P_{10}} \tag{3.12}$$

where Q is one half of the interquartile range, P_{90} is the 90th percentile and P_{10} is the 10th percentile. A distribution which is more pointed will have a larger kurtosis value and is called *leptokurtic*. A less pointed distribution will have a smaller kurtosis value and is called *platykurtic*. Skewness and kurtosis will be discussed in greater detail in Chapter 4.

CHAPTER 4

Distribution Functions

This chapter covers distribution functions. Section "Probability Distribution Functions" provides an overview of probability functions and introduces the reader to some relevant definitions and concepts. The last section, on continuous distribution functions, introduces the reader to some of the commonly used distribution functions in parametric statistics such as the normal, chi-squared, *t*, and *F* distributions.

Probability Distribution Functions

Probability distributions use charts, equations, or tables to represent the distribution of data. Net worth in the United States can be presented using all the three methods. For example, the U.S. Census provides the following tabular information on the distribution of wealth in the United States:[1]

Table 4.1 Household net worth distribution for persons with baccalaureate degrees

Household Net Worth	Percent
Zero or Negative	13.2
$1–4,999	3.9
$5,000–9,999	3.2
$10,000–24,999	5.5
$25,000–49,999	7.1
$50,000–99,999	10.0
$100,000–249,999	19.4
$250,000–499,999	16.7
$500,000 or over	21.0

Table 4.1 indicates that the net household worth of 10% of people with a baccalaureate degree is between $50,000 and $99,999. The distribution can also be represented as a chart as shown in Figure 4.1.

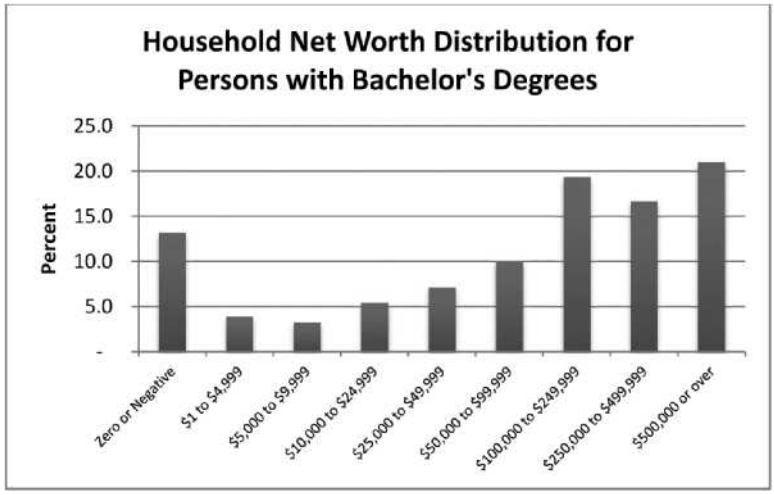


Figure 4.1 Household net worth distribution for persons with bachelor's degrees

However, to determine the effect of a baccalaureate degree on net worth, it would be revealing to compare the distribution of the net worth of those with a bachelor's degree with the net worth of people with a high school diploma. This comparison is shown in Figure 4.2.

It seems to be evident that earning a baccalaureate degree results in higher net worth on average. Although there are people with and without a baccalaureate's degree in all income categories, higher percentages of those with a high school degree dominate lower net worth categories while people with a baccalaureate degree dominate higher categories of net worth.

Probability distribution functions include discrete and continuous functions. Whether a distribution is discrete or continuous depends on whether the variable being modeled is a discrete or continuous variable. This chapter focuses primarily on continuous distribution functions.

Definitions and Concepts

Sometimes it is possible to represent random variables as a function in a way that the function can determine the probability of an outcome of the random variable.

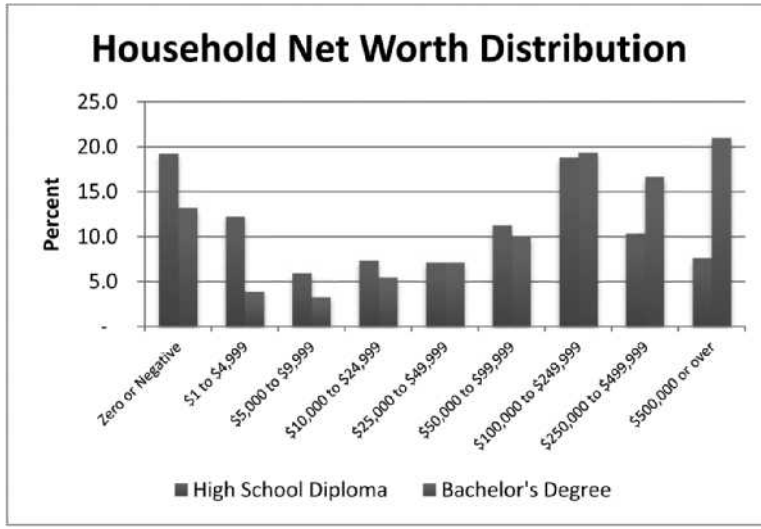


Figure 4.2 Household net worth distribution by educational status

Definition 4.1 The outcome of a *random variable* is determined by chance. The outcome of a random variable might be an "integer number" as in seeing "2" when a die is rolled, or it could be a real number such as the length of time it takes for students to learn this chapter. The subject statistics is used to study the properties of random variables and how they behave.

In its simplest form, the probability distribution consists of values and probabilities. The probability distribution for flipping a coin is

$$f(x) = \begin{cases} \text{Head with probability } 1/2 \\ \text{Tail with probability } 1/2 \end{cases}$$

There are more formal ways to define probability distributions that are beyond the scope of the present text. A distribution function can be presented in the form of a function, a table, or a statement. The subject of probability distributions is vast, but we will focus on the few items needed to continue our discussion. There are two types of random variables, *discrete* and *continuous*.

Definition 4.2 A *discrete random variable* consists of integers (whole numbers) only.

Definition 4.3 A *continuous random variable* can take any value over a range.

Definition 4.4 The *probability density function* of a *discrete* random variable provides the probability for valid values of the discrete random variable. Probability density functions are depicted as $f(x)$.

Definition 4.5 The *probability density function* of a *continuous* random variable provides the probability for valid ranges of the continuous random variable. Probability density functions are depicted as $f(x)$.

Definition 4.6 A *probability distribution* or distribution functions is the cumulative value of a probability density function $f(x)$. In a sense, it is the sum or total of all probabilities up to a particular point.

Probability density functions are also called probability functions, probability mass functions, or frequency functions but universally are shown by $f(x)$. In this text, we may use the term *distribution function* for both discrete and continuous probability density functions as a matter of convenience.

Continuous Distribution Functions

Continuous distribution functions include normal, chi-squared, t, and F distributions. Probably the best known is the normal distribution, as it forms the foundation of many statistical analyses. The normal distribution is sometimes called a bell curve because it takes the shape of a symmetrical bell.

Normal Distribution Functions

The normal distribution function is widely used, especially in the standardized form. If the observations used to calculate Z scores are from a normal distribution, the result is *standardized normal values* or simply *standardized values*. Converting values from different normal distributions with different means and variances to standard normal (with a mean of 0 and standard deviation of 1) allows us to compare them with each other or with the normal table.

Since the normal distribution is a continuous distribution function, and there are infinitely many points on any continuous interval, the probability of any single point is "one out of infinity" or zero. Therefore, for continuous distribution functions, the probability is calculated for an interval instead of a point. Direct computation of such probabilities requires integral calculus. Fortunately, there are tables, both paper-based and electronic, as well as software-generated, that calculate the probability for ranges of values from a normal distribution. All quality statistics software, even spreadsheets such as Excel, are capable of computing the probability for standard normal values. A table of values for the normal distribution can be accessed in Stata by using *findit probtabl* to locate and download a program that once installed will allow you to obtain the normal distribution table by typing the command *ztable*. For tabulated values of chi-squared, *t*, or *F* distributions, use *chitable*, *ttable*, or *ftable*. If you do not wish to use Stata, you can always do an online search for "normal distribution table Aplet" and many results will appear. Aplets are interactive programs that have windows for input, display appropriate probabilities, and provide a graph depicting the probability. Version 14 of Stata can display probabilities and inverse probabilities for most of customary distribution functions. In the command line type *help function* to display a list of available commands.

Properties of Normal Distributions

A normal distribution is depicted in Figure 4.3. The normal distribution curve is unimodal (has only one mode) and symmetric. Consequently, its mean, mode, and median are all the same and fall in the middle of the curve. The tails of the normal curve do not touch the X-axis. The X-axis is actually an asymptote of the functions, which means the curve does not touch the axis even at infinity. One implication is that the probability of an outcome is never zero regardless of how far away that outcome might be from the mean. Nevertheless, the probability of the tail areas becomes very negligible not too far from the center, making it unnecessary to be concerned with infinity. A normal distribution has two parameters, which are its mean and variance. Since this distribution

is commonly used and has numerous applications, it has become known as the normal distribution. The mean and variance of the normal distribution are represented by μ and σ^2, respectively.

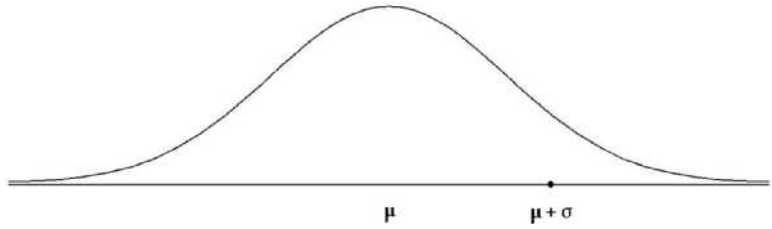


Figure 4.3 Normal distribution with mean = μ and variance = σ^2

The area under the normal curve is equal to 1, as is the area under any distribution density function. Customarily, the distance from the center of a normal distribution is measured by its standard deviation.

When the population variance is not known, using the normal distribution for inference is misleading. The problem is more acute when the sample size is small.

Standardizing Values from a Normal Distribution

Standardizing observations from a normal distribution converts any normal distribution to a normal distribution with a mean equal to 0 and variance equal to 1. For a random variable called X, the notation would be $X \sim N(0,1)$, where the symbol "~" is pronounced "distributed as." Since the square root of "1" is still "1," the variance and the standard deviation are the same and one could as correctly state "standard deviation" instead of "variance." The distribution can also be called *normal standard* $N(0,1)$. The normal standard distribution is used to determine the probability of a statistic in inferential statistics as will be seen in Chapters 6 and 7. A single table of probabilities of values for $N(0,1)$ is sufficient for calculating probabilities for areas under normal distributions with different means and variances. The graph for $N(0,1)$ is exactly the same as the graph in Figure 4.3; however, the one depicted in Figure 4.4 has the added feature that marks the point which is one standard deviation to the right of the center.

Area under a Normal Distribution with Mean 0 and Variance 1

A property of $N(0,1)$ is that inflection points are one standard deviation from the mean. Roughly two-thirds of all observations are within one standard deviation from the mean of $N(0,1)$; the actual percentage, to two decimal place accuracy, is 68.25%. The percentage of values within two standard deviations, to two decimal place accuracy, is 95.45%. The knowledge that the data have a normal distribution improves the probability of observations being within two standard deviations from the 75% indicated by Chebyshev's theorem to over 95% using the normal distribution. In Example 3.3, it was shown that the observed probability was 96.67% instead of 75%. The reason is that the Chebyshev's theorem provides the minimum percentage and that under attainable guidelines the statistics for most real-life outcomes approximate a normal distribution as will be demonstrated in Chapter 5.

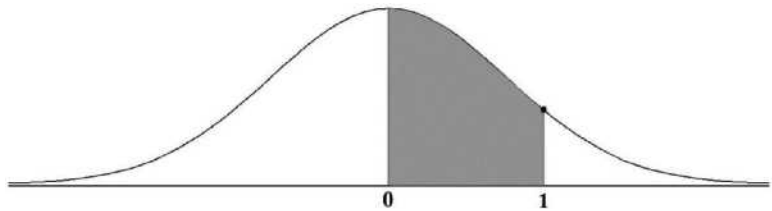


Figure 4.4 The area under the normal distribution between 0 and 1

To calculate the area under a normal distribution, we can use the *probtabl* download from Stata. In this table, the first column represents the tens digit and the first place after the decimal point. The first row represents the second decimal point. As an example, let's say we wanted to find the probability for $Z = 2.45$. On the first column, we would look for 2.4, and on the first row we would look for 0.05 (2.4 + 0.05 = 2.45). At the intersection of the row 2.4 and column 0.05, the probability 0.4929 is presented. The Stata command for obtaining normal probabilities is *normal(Z) and normalden(z)*.

As another example, to obtain the probability of a Z score of 1.0, go to the left margin of the table and identify the row marked 1.0. Then identify the column marked "0." The value at their intersection is 0.3413. This number gives the area under the curve from the midpoint

to one side. To get the area on both sides of the midpoint, we need to double the value. Doubling the probability value provides the probability of an observation falling within one standard deviation on either side of the mean, and is equal to approximately 68% (0.3413 × 2 = 0.6826, or 68%). Computations of other values are similar.

Example 4.1 Find the probabilities for the following Z scores using the normal distribution.

a. $P(-1.52 < Z < 1.49)$
b. $P(-1.69 < Z < -1.58)$

Solution 4.1a

In general, it is best to shade the area for which the probability is needed and perform simple algebra to obtain the results. For example,

$$P(-1.52 < Z < 1.49) = P(-1.52 < Z < 0) + P(0 < Z < 1.49)$$
$$= 0.4357 + 0.4319 = 0.8676$$

See Figure 4.5 for clarification. The above equation represents the probability between –1.52 and 1.49.

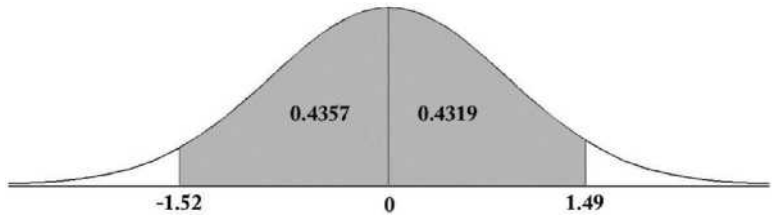


Figure 4.5 Area under the normal distribution curve between –1.52 and 1.49

Solution 4.1b

See Figure 4.6 for finding the shaded area for the second example. Make sure to subtract the number with the smaller absolute value from the larger in order to guarantee that you do not get a negative probability, which would not make any sense.

$$P(-1.69 < Z < -1.58) = P(-1.69 < Z < 0) - P(-1.58 < Z < 0)$$
$$= 0.4545 - 0.4429 = 0.0116$$

Obtaining Probability Values for the Normal Distribution with Excel

The above results can be obtained by using the following command in Excel:

=NORM.DIST(1.49,0,1,1) - 0.5

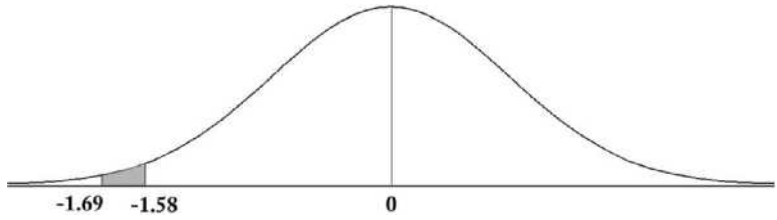


Figure 4.6 Area under the normal distribution between –1.69 and –1.58

which should return the result of 0.43188. The area under the first half of the curve, namely 0.5 must be subtracted since the desired area is from the center to 1.49, while Excel provides the entire area to a point such as 1.49.

When the normal distribution is not standardized, the probability can be obtained directly by using the following command and inserting the mean and standard deviation of the distribution.

=NORM.DIST(X, mean, standard deviation, cumulative)

where "1" represents logical "true" which signals the software to report the cumulative probability to the point 1.49.

Alternatively, point to the arrow on the side of the "Σ AutoSum" on Excel's command ribbon and choose "more functions." From the drop-down window choose "Statistical" in the window labeled "or select a category" and scroll down and select NORM.DIST command (Figure 4.7).

Place the appropriate values in the correct boxes and press OK. The probability is displayed in the cell and the formula is displayed in the function window. The command in Stata is

display normal(1.86) - 0.5

The "display" is necessary to display the output. Without it, the value will be calculated but not displayed. Remember Stata is case sensitive, so type the command as shown.

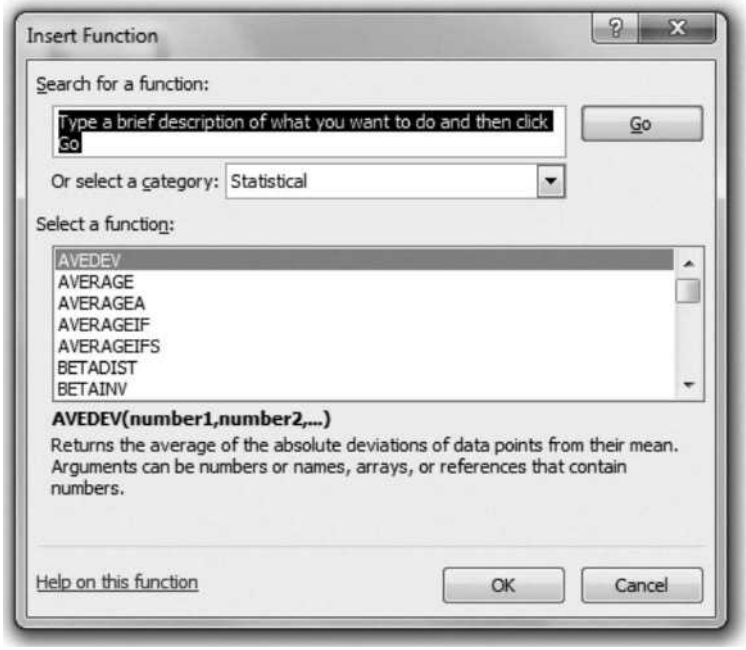


Figure 4.7 Pull-down Window for Excel Functions

Upon selecting the option, a new window opens up as shown in Figure 4.8.

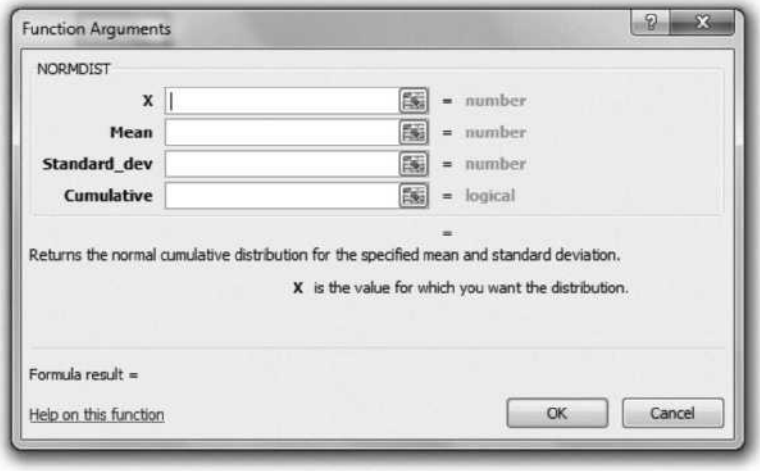


Figure 4.8 Excel Window for inserting the parameters for a normal distributions

Area under a Normal Distribution with any Mean and Variance

To calculate the probability for a normal distribution with any mean and variance, first convert the distribution to the standard normal by standardizing all values. To do so, follow these steps:

1. Rewrite the question using probability notations (see Example 4.1 for some examples).
2. Convert the values into Z scores.
3. Draw a graph and shade the area under investigation.
4. Look up the probability from a table or use a software program to obtain the probability.

When finding the area between two Z values, if the Z values are on different sides of zero, that is, one is positive and the other is negative, find the area between 0 and each Z value for each and *add* the corresponding probabilities. If the Z values are on the same side of 0, that is, either both are negative or both are positive, find the area between 0 and each Z value and *subtract* the smaller probability from the larger one.

Continuous distribution functions such as normal distributions share two characteristics: (1) length and (2) area. A Z score is a length measure. It shows how far a point is from the center (which is also the mean) in terms of standard deviations. In other words, Z scores indicate the number of units a point deviates from the mean. The probability of a particular Z value from the mean is an area. The area between the mean (zero) and a point is the value provided in the standard normal table.

Example 4.2 Assume that the random variable X has a normal distribution with mean 16.9 and standard deviation of 3.01. Find the area associated with the probabilities below using the normal distribution.

a. $P(X < 22.51)$
b. $P(X > 11.3)$
c. $P(13.93 < X < 23.41)$
d. $P(11.43 < X < 15.61)$
e. $P(17.64 < X < 21.45)$

Solution 4.2

The first step is to standardize the values, both alphabetically and numerically.

a. $$P(X<22.51)=P\left(\frac{X-\mu}{\sigma}<\frac{22.51-16.9}{3.01}\right)=P(Z<1.86)$$
$$=P(0<Z<1.86)+0.5=0.4686+0.5=0.9686$$

Note that the area of interest is everything to the left of 1.86. Excel and Stata commands are listed below. Note that Stata requires typing the Z formula.

=NORM.DIST(22.51, 16.9, 3.01, 1) = 0.9688

display normal((22.51-16.9)/3.01) = 0.9688

The *display* command before any valid algebraic equation will display the result.

b. $$P(X>11.3)=P\left(\frac{X-\mu}{\sigma}>\frac{11.3-16.9}{3.01}\right)=P(Z>-1.86)$$
$$=P(-1.86<Z<0)+0.5=0.4689+0.5=0.9689$$

c. $$P(13.93<X<23.41)=P\left(\frac{13.93-16.9}{3.01}<\frac{X-\mu}{\sigma}<\frac{23.41-16.9}{3.01}\right)$$
$$=P(-0.99<Z<2.16)=P(-0.99<Z<0)+P(0<z<2.16)$$
$$=0.3389+0.4846=0.8235$$

d. $$P(11.43<X<15.61)=P\left(\frac{11.43-16.9}{3.01}<\frac{X-\mu}{\sigma}<\frac{15.61-16.9}{3.01}\right)$$
$$=P(-1.82<Z<-0.43)=P(-1.82<Z<0)-P(-0.43<Z<0)$$
$$=0.4656-0.1664=0.2992$$

e. $$P(17.64<X<21.45)=P\left(\frac{17.64-16.9}{3.01}<\frac{X-\mu}{\sigma}<\frac{21.45-16.9}{3.01}\right)$$
$$=P(0.25<Z<1.51)=P(0<Z<1.51)-P(0<Z<0.25)$$
$$=0.4345-0.0987=0.3358$$

Finding the Value that Corresponds to a Given Probability

Example 4.3 Let the random variable X have a normal distribution with a mean of 15 and a standard deviation of 3.

a. What is the cutoff value for the top 1% of this population?
b. Find the interquartile range.

Solution 4.3

a. In this example, the probability of the outcome is given and the value that determines the desired probability is the objective.

$$P(Z > z) = 0.01$$

Search for the probability that corresponds to 0.5 – 0.01 = 0.49 in the body of the normal table. You will need to scan through the body of the table to find the value that most closely matches 0.49. In this case, the closest that we can find is 0.4901. If you then reverse the process that was used before, and add the leftmost row value and top-most column value, you get 2.33. The following Excel and Stata commands can also be used to provide the Z value:

=norm.s.inv(1-0.01)

display invnormal(1-.01)

Next, determine the X value by reversing the computation of the Z score.

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 15}{3} = 2.33$$

$$X = 15 + 3(2.33) = 21.99$$

Therefore, 1% of the population has an X value greater than 21.99. Note that 21.99 is the 99th percentile. The following Excel command accomplishes the same task.

=norm.inv(1-.01,15,3)

b. Note that by definition the first quartile is to the left of the mean.

$$P(Z < z) = 0.25$$

$$P(z < Z < 0) = 0.5 - 0.25 = 0.25$$

Search the body of the table for 0.25. The closest number is 0.2517 that corresponds to the Z score of $z = -0.675$. In the Z score formula solve for the X.

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 15}{3} = -0.68$$

$$X = 15 + 3(-0.675) = 12.975$$

Therefore $X = 12.975$ is the first quartile.

For the third quartile,

$$P(Z < z) = 0.75$$

$$P(0 < Z < z) = 0.75 - 0.5 = 0.25$$

Following the previous procedure $Z = +0.675$,

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 15}{3} = 0.68$$

$$X = 15 + 3(0.675) = 17.025$$

Therefore, the third quartile is 17.025. The interquartile range for the population is 12.975 to 17.025.

Using the Menus in Excel

Note that in Excel the top 1% is entered as 0.99 for probability. Point to the arrow on the side of the "Σ AutoSum" on Excel's ribbon command and choose "more functions." From the drop-down window, choose "Statistical" in the window labeled "or select a category" and scroll down until you get to the NORM.INV command. See Figure 4.7 for more details.

Place the appropriate values in the correct boxes and press OK (Figure 4.9).

The result 21.97904 is displayed, which is slightly different from the result from the table due to rounding. You could have entered the following formula as well.

= NORM.INV(0.99,15,3)

a.

= NORM.INV(0.25,15,3) = 12.97653

= NORM.INV(0.75,15,3) = 17.02347

Again the results are slightly different due to rounding and the degree of precision of the normal table. Interestingly, both the above values and the values for the normal distribution table are obtained from Excel.

In Stata, enter the following command to obtain the Z value associated with a probability:

```
display invnorm(0.25)
```

To obtain the X values, enter the following commands. Note that both syntaxes use the plus (+) sign because Stata assigns the correct signs of negative and positive for values to the left and to the right of the center, respectively.

```
display 15+ invnorm(0.25)*3
display 15+ invnorm(0.75)*3
```

Nonconformity with the Normal Distribution

Nonconformity with the normal distribution could be due to a deviation from symmetry, which is called skewness or due to a deviation in the pointedness of the distribution, which is called kurtosis.

Normality versus Skewness

Normal distributions are essential for statistical analysis. The exact shape of the normal curve depends on the probability density function of the normal distribution. Any deviation from the normal probability density function results in skewness or kurtosis as explained in Chapter 3. It is easy to detect skewness visually because skewed distributions are not symmetric. In Figures 4.9 and 4.10, the graphs for positive and negatively skewed functions are superimposed on that of the normal distribution for comparison.

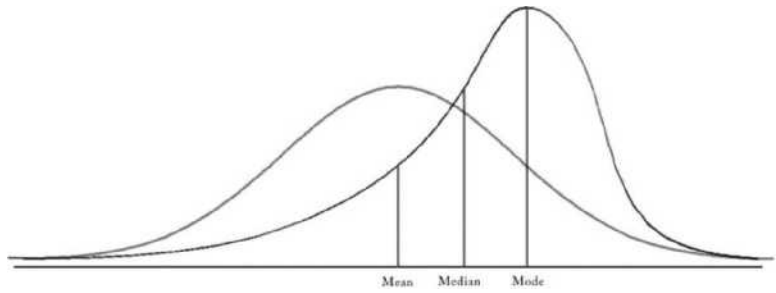


Figure 4.9 Comparison of negative skewness with a normal distribution

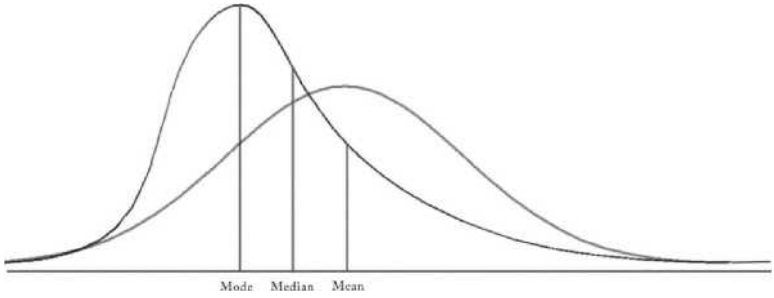


Figure 4.10 Comparison of positive skewness with a normal distribution

The commands for computing skewness in Excel and Stata are

=skew(range)

summarize varlist, detail

The range is a valid range in Excel and *varlist* is the name(s) of the variable(s) in Stata.

Normality versus Kurtosis

Kurtosis measures the degree of flatness or pointedness of a symmetric curve compared with a normal distribution, as explained in Chapter 3. It is beneficial to depict the graph of a flatter curve, called platykurtic, which corresponds to a kurtosis measure with negative value. The more peaked curve, called leptykurtic corresponds to a kurtosis measure with a positive value. Figures 4.11 and 4.12 demonstrate the comparison to the normal curve.

Figure 4.11 Comparison of negative kurtosis (platykurtic or flatter) with normal

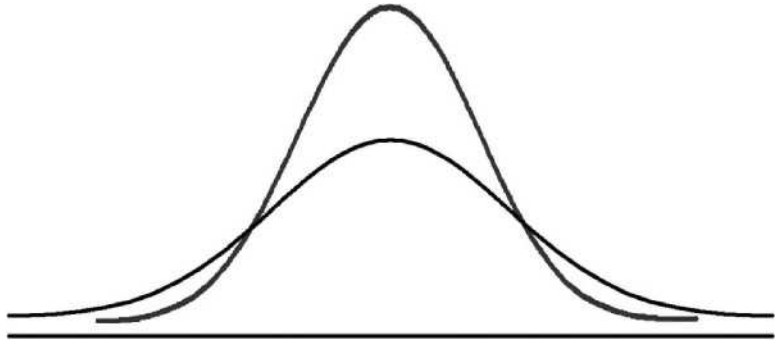


Figure 4.12 Comparison of positive kurtosis (leptokurtic or pointed) with normal

Example 4.4 Calculate the skewness and kurtosis for the Microsoft stock prices from May 21 to July 2, 2015. Use the data from Table 2.1.

Solution 4.4

Use the following two commands, the first in Excel and the second in Stata, to answer the question.

=skew(range)

summarize varlist, detail

The range is a valid range in Excel and *varlist* is the name(s) of variable(s) in Stata. The results of *skew* and *kurtosis* commands from Excel may differ from the Stata *summarize* command as the two programs use different methods of computing the results. With the Microsoft stock price data, Excel gives a result of –0.4967 and Stata gives –0.4715. The output from Stata was presented earlier.

Chi-squared (χ^2) Distribution Functions

The chi-squared distribution is one of the distribution functions that are derived from normal distribution function.

Theorem 4.1 Let the random variable X have a normal distribution with mean and positive variance:

$$X \sim N\left(\mu, \sigma^2\right)$$

Then, the random variable

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

will have a chi-squared (χ^2) distribution with one (1) degree of freedom.

$$V \sim \chi^2(1)$$

Note that $Z = \frac{X-\mu}{\sigma}$; hence, Z^2 is a χ^2 (1). Therefore, the standard normal values can be squared to obtain probabilities for the chi-square distribution with 1 degree of freedom.

$$P(|Z| < 1.96) = 0.95$$

$$P(Z^2 < (1.96)^2) = P(Z^2 = 3.842) = 0.95$$

This is identical to the chi-squared value with one degree of freedom. In confidence intervals and tests of hypothesis, one can use either a normal distribution or a chi-squared distribution. Each one is beneficial in different settings. In Stata, the user-defined command *chitable* displays a table of chi-squared values (first you must install *probtabl*; for more information, see the discussion under normal distribution functions earlier in the chapter). Probabilities are listed at the top row and the degrees of freedom are on the left margin. The Stata command for obtaining chi-squared probabilities are *chi2(df, x)* and *chi2dn(df, x)* for cumulative and probability density, respectively, where *df* stands for degrees of freedom and the *x* inside the parentheses is the desired chi-squared value.

The Excel command is

=Chisq.dist(x, deg_freedom, cumulative)

where x is the desired value, deg_freedom is the degrees of freedom, and cumulative is either 0 or 1. To obtain the probability for the right or left hand area, add ".r" and ".l" before the left parenthesis, respectively.

Theorem 4.2 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution $N(\mu, \sigma^2)$. Recall that

$$\hat{\mu} = \frac{\Sigma X}{n} \text{ and } \widehat{\sigma^2} = \frac{\Sigma\left(X - \hat{\mu}\right)^2}{n-1}$$

It can be proven that

i. $\hat{\mu}$ and $\widehat{\sigma^2}$ are independent

ii. $\dfrac{(n-1)\widehat{\sigma^2}}{\sigma^2}$ is distributed as a chi-squared with $(n-1)$ degrees of freedom

$$\frac{(n-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-1)$$

t Distribution Functions

This is another distribution function, which is derived from the normal distribution. This distribution corrects for loss of degrees of freedom in empirical studies. In addition, it is also related to the chi-squared distribution.

Theorem 4.3

Let X be a random variable that is $N(0,1)$, and let U be a random variable that is χ^2 (r). Assume Z and U are independent. Then

$$t = \frac{X}{\sqrt{\dfrac{U}{r}}}$$

has a t distribution with r degrees of freedom.

Theorem 4.4 Let X be a random variable that is $N(\mu, \sigma^2)$. Use the customary "hat" notation to represent sample mean and sample variance. Then the following relationship has a t distribution with $(n-1)$ degrees of freedom.

$$t = \frac{\frac{(\hat{\mu} - \mu)}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{n\hat{\sigma}^2}{(n-1)\sigma^2}}}$$

The Stata command for obtaining *t* probabilities *are t(df,t)* and *tden(df,t)* for cumulative and probability density, respectively, where *df* stands for degrees of freedom and the *t* inside the parenthesis is the desired *t value.* The command *ttable* will display the table of *t* values. This command will work only if you have installed *probtabl* in the previous section (using the command *findit probtabl*). The Excel commands are *t.dist*, *t.dist.2t*, and *t.dist.rt*, for left-tailed, two-tailed, and right-tailed probabilities. The inverse is obtained by *t.inv* and *t.inv.2t.*

F Distribution Functions

The final continuous distribution function that is considered is derived from the ratios of two chi-squared distributions; hence, it too belongs to the family of normal distributions.

Theorem 4.5

Let U and V be independent chi-squared variables with r_1 and r_2 degrees of freedom, respectively. Then

$$F = \frac{U/r_1}{V/r_2}$$

has an F distribution with r_1, r_2 degrees of freedom.

$$\frac{U/r_1}{V/r_2} \sim F_{r1,\, r2}$$

The relation between F and Z is such that

$$F = e^{2Z}$$

The *F*-statistic consists of the ratio of two variables, each with a chi-squared, χ^2, distribution divided by their corresponding degrees of freedom.

The Stata command for obtaining *F* probabilities is *ftable* (*probtabl* must be installed first using *findit probtabl*). The Stata command for obtaining F probabilities are *f(df1, df2,x)* and *fdn(df1, df2, x)* for cumulative and probability density, respectively, where *df* stands for degrees of freedom and the *x* inside the parenthesis is the desired *F* value. The commands in Excel are f.dist and f.dist.rt for left-tailed and right-tailed probabilities, respectively. The syntaxes are

=f.dist(x, deg_freedom1, deg_freedom2, cumulative)

=f.dist.rt(x, deg_freedom1, deg_freedom2)

The following command returns the inverse of the left hand tail value for a particular probability:

=f.inv(probability, deg_freedom1, deg_freedom2)

CHAPTER 5

Sampling Distribution of Sample Statistics

Sampling

Samples, which are subsets of a population, can be collected in a variety of ways. As will become evident in this chapter, random sampling is important for establishing the necessary foundation for statistical analysis. However, sampling techniques are not limited to random sampling. Sampling theory establishes the customary properties of statistics for each non-random sample. Although each sampling technique has its advantages and disadvantages, the present text will not focus on various sampling techniques. After a brief discussion about sampling, the attention will be focused on the properties and advantages of random sampling. Theories that are necessary for performing statistical inference and are related to sampling are discussed in this chapter.

In Chapter 2, a statistic was defined as

> ***A statistic*** *is a numeric fact or summary obtained from a sample. It is always known, because it is calculated by the researcher, and it is a variable. Usually, a statistic is used to draw inferences about the corresponding population parameter.*

The only way to know the parameters of a population is to conduct a census. A census is a survey of everyone or everything in a population. Censuses are expensive, and contrary to common belief, they are not always more accurate than a sample or necessarily correct. On average, it takes more than 2 years to release the census results in the United States. During this time, data change; for example, new babies are born, some people die, and others move. Sometimes, in order to expedite the release of census information, a sample is obtained from the census data.

Sometimes, taking a census is not an option. This is not only due to the time and money involved, but also because the census itself might be destructive. For example, in order to find out the average life of light bulbs, they must be turned on and left until they burnout. Barring mistakes, this would provide the average life of the light bulbs, but then there will not be any light bulbs left. Similarly, determining whether oranges were not destroyed by frost requires cutting them open. There are lots of other reasons where it is unrealistic, if not impossible, to conduct a census to obtain information about a population and its parameters.

Surveying a sample can address some of the problems associated with conducting a census. However, collecting sample data is neither inexpensive nor effortless. Sampling textbooks devote substantial effort to explain how to obtain random samples from a population. One simple, but not necessarily pragmatic or efficient way is to assign ID numbers to all members of the population and then pull the desired number of sample units by using random drawings of the ID numbers.

In this instance, our interest in sampling is very limited. We are interested in obtaining an estimate from a relatively small portion of a population to obtain insight about the population parameter. The knowledge of parameters allows meaningful analysis about the nature of the characteristics of interest in the population and is vital for making decisions about the population of the study.

As indicated earlier, summary values obtained from a sample are called *statistics*. Since statistics are random variables, different samples result in slightly different outcomes. It is possible to have many samples and thus many sample statistics, such as the sample *mean*. It turns out that the sample means have certain properties that are useful as they allow us to conduct *statistical inference*.

Definition 5.1 *Statistical inference* is the method of using sample statistics to draw conclusions about a population parameter.

Statistical inference requires the population of interest to be defined clearly and exclusively. Furthermore, the sample must be *random*. Theories that explain how statistics are used to make inferences about population parameters will be explained shortly.

The method of using information from a sample to make an inference about a population is called *inductive* statistics. In inductive statistics, we observe specifics to draw an inference about the general population. This chapter introduces the necessary theories for inductive inference, whereas Chapters 6 and 7 provide specific methods for making inferences under different situations. The alternative procedure of *deductive* inference starts from the general and makes assertions about the specific. For example, if a firm has 500 employees and 300 of them are men, then the probability of choosing a male worker at random is 300/500 = 0.60 or 60%.

Statistical inference makes probabilistic statements about the expected outcome. It is essential to realize that since random events occur probabilistically, there is no "certain" or "definite" outcome value. Therefore, it is essential to provide the probability associated with the expected outcome.

Sample Size

Before we discuss the role of randomness and the usefulness or effectiveness of a sample, it is important to understand how other factors influence the effectiveness of the sample statistics in providing *reliable* inferences about a population parameter. Even if a sample is chosen at random, two other factors attribute to the reliability of the sample statistics. They are the *variance* of the population and the *sample size*.

In order to provide the perfect inference about a population parameter for a population with identical members, the necessary sample size is one (1). For example, if the output of a firm is always the same, say 500 units per day, then choosing any single given day at random would be sufficient to determine the firm's output. Note that in the previous example, there was no need to sample at random, although one can argue that any day that is chosen is a random day. However, if the output changes every day due to random factors such as sickness, mistakes during production, or breakdown of equipment, then the sample size must increase.

It is important to understand the possible difference in output among the days of the week and the month, if applicable. For example, Mondays and Fridays might have lower output. By Friday, workers might be tired

and not very productive. Machinery might need cleaning and break more often toward the end of the week. On Mondays, workers might be sluggish and cannot perform up to their potential. Workers might be preoccupied toward the end of the month or early in the month when they are running out of money, or when their bills are due.

These are just some of the issues that might affect the outcome, hence the sample size. Therefore, there should be a direct relationship between the sample size chosen and the variance of the population, and larger samples must be taken from populations with larger variance. Since statistical inference is probabilistic, to obtain higher levels of confidence, we should take larger samples. It can be shown that the required sample size for estimating a mean is given by

$$n = \frac{Z^2_{\alpha/2}\sigma^2}{E^2}$$

where $Z^2_{\alpha/2}$ is the square of the Z score for the desired level of significance, σ^2 is the variance of the population, and E is the tolerance level of error. The reason for dividing the level of significance by 2 is that the probabilities of extreme values are evenly spread on both the low end and the high end of the distribution.

Definition 5.2 The *reliability* of a sample mean $(\hat{\mu})$ is equal to the probability that the deviation of the sample mean from the population mean is within the *tolerable level of error* (E).

$$\text{Reliability} = P(-E \leq \hat{\mu} - \mu \leq E)$$

Example 5.1 Assume that a population's standard deviation for a particular output level is 29. Also, assume that we desire to limit our error to 5%, which makes the level of significance 95%. Let the tolerable level of error to be 4. Determine the required sample size.

Solution 5.1

First obtain a Z score. In Figure 5.1, the area in the middle is 95% of the area under the curve. Therefore, the area at the two tails adds up to 5%, or 0.05 probability.

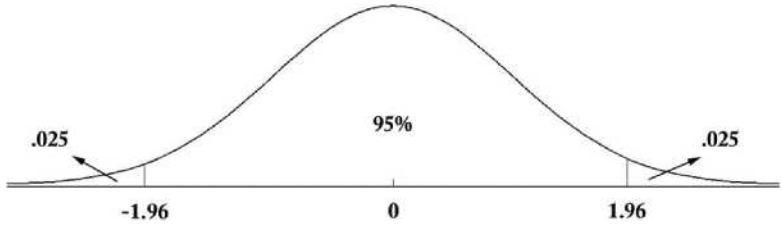


Figure 5.1 Graph of a normal distribution

Since most tables are designed to calculate the area between zero and a demarcation point on the right-hand side, it is necessary to obtain that area first. Half of the level of significance, or 0.025, is on the right hand, which makes the area between zero and the demarcation point equal to 0.5 – 0.025= 0.475, which according to the table corresponds to a Z score of 1.96. The commands in Stata and Excel are

=NORMINV(1-0.025,0,1)

display invnormal(.975)

Note that the answer from Excel is 1.959964. The necessary sample size is given by

$$n = \frac{(1.96^2)(29^2)}{4^2} = \frac{(3.84)(841)}{16} = 201.9241$$

Therefore, the minimum sample size should be 202. Since fractional samples are not possible, we must always use the next higher integer to assure the minimum desired level of accuracy.

Example 5.2 A 95% reliability for the sample mean is given by the shaded area in Figure 5.2.

$$\text{Reliability} = P(\mu - E \le \hat{\mu} \le \mu + E) = .95$$

Shortly, we will see that the tolerable level of error is equal to

$$E = 1.96\frac{\sigma}{\sqrt{n}}$$

An astute student will notice that in order to estimate the sample size, it is necessary to know the variance. It is also imperative to understand that in order to calculate the variance, one needs the mean, which, apparently, is

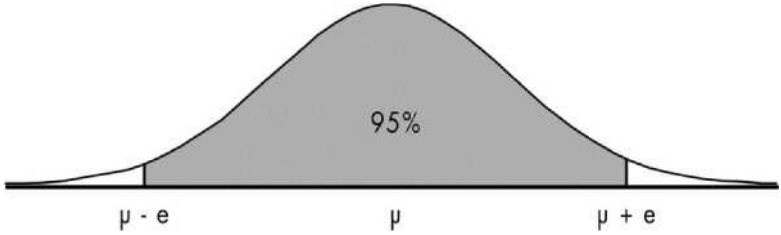


Figure 5.2 The range for 95% reliability of the sample mean

not available; otherwise we would not have to estimate it. Sometimes one might have enough evidence to believe that the variance of a population has not changed, while its mean has shifted. For example, everybody in a country is heavier, but the spread of the weights among people have remained the same. In situations that the known variance is also believed to have changed, the only solution is to take a pre-sample to have a rough idea about the mean and the variance of the population and then use the sample estimates as a starting point to determine a more dependable sample size.

Sampling Distribution of Statistics

As stated earlier, sample statistics are random variables and change from sample to sample. This means that the actual observed statistics are only one outcome of all the possible outcomes. A *sampling distribution* of any statistic explains how the statistic differs from one sample to another. The most commonly used statistics are the sample mean, proportion, and variance. Therefore, we will study their sampling distributions in a systematic way. We begin with the sampling distribution for one sample mean and distinguish between the cases when the population variance is known and when it is unknown. Next, we introduce two sample means and address the cases of known and unknown variances. However, before embarking on this mission, it is necessary to discuss the Law of Large Numbers and the Central Limit Theorem, which are the foundations of inferential statistics.

Theorem 5.1 Law of Large Numbers

For a sequence of independent and identically distributed random variables each with mean (μ) and variance (σ^2), the probability that the

difference between the sample mean and the population mean is greater than an arbitrary small number will approach zero as the sample size approaches infinity.

The theorem indicates that as the number of observations increases, the average of their means approaches the population's mean. Since sample means are statistics and random, their values vary. The theory neither requires nor implies that any of the observations will be equal to the population mean. The law of large numbers is essential for the central limit theorem.

Theorem 5.2 Central Limit Theorem

Assume a sequence of random variables X_1, X_2, … X_n are independent and are identically distributed, each with mean (μ) and variance (σ^2), then the distribution function of the Z scores of the sample means will converge to standard normal as the sample size increases.

The distribution function of the sample means, not the actual observations, will be standard normal. The theorem depends on sampling with replacement. In practice, however, sampled units are not returned to the population to be included in the next draw of a sample unit. Exclusion of a member of the population from successive draws for a sample changes the probability of the remaining population members, thus invalidating the outcome. When a population is substantially large, the change in probability of excluding one member is very small. For example, if the unit of observation is people, then excluding one person in the United States from a population after being drawn would affect the probability by about 1 in 330,000,000. When a population is finite, a correction factor is added to the formula to reduce the problem, which will be explained and demonstrated shortly.

Sampling Distribution of a One Sample Mean

In this section, we will consider separately the case where the population variance is known, and that where the population variance is unknown.

Population Variance Is Known

The sample mean is a statistic. Assume we know the variance of a population from which the sample mean is obtained. Let $(\hat{\mu})$ be the mean of a random sample of size *n* from a distribution with a finite mean (μ) and a finite and *known* positive variance (σ^2). Using the Central Limit Theorem, the following is true about the sample mean $(\hat{\mu})$:

1. The distribution function for $(\hat{\mu})$ can be approximated by a *normal* distribution
2. $E(\hat{\mu}) = \mu$
3. $\sigma^2_{\hat{\mu}} = \dfrac{\sigma^2}{n}$

Therefore, we can use the standard normal table values for comparison of the standardized values of the sample mean. The knowledge of population variance σ^2 is essential for calculation of $\sigma^2_{\hat{\mu}}$, Property 2, above, states that the expected value of the sample mean $\hat{\mu}$ will be equal to the population parameter μ. In other words, the average of all such sample statistics will equal the actual value of the population parameter. As the sample size increases, the sample variance of the estimate $\widehat{\sigma^2_{\mu}}$ decreases. Therefore, as the sample size increases, the sample statistic (estimate) gets closer and closer to the population parameter μ.

The distribution function of the sample mean for samples of size 30 will be close to the normal distribution even when the variance is estimated rather than known. For random variables from a population that is symmetric, unimodal, and of the continuous type, a sample of size 4 or 5 might result in a very close approximation of the normal distribution. If the population is approximately normal, then the sample mean would have a normal distribution when sample size is as little as 2 or 3.

Example 5.3 Assume that the variance for the daily production of a good is 2,800 pounds. Find the sampling distribution of the sample mean $(\hat{\mu})$ for a sample of size 67.

Solution 5.3

1. The sampling distribution of the sample mean $(\hat{\mu})$ is normal
2. $E(\hat{\mu}) = \mu$
3. $\sigma^2_{\hat{\mu}} = \dfrac{2800}{67} = 41.79$

As we see, there is very little computation involved. Nevertheless, the theoretical application is enormous. The magnitude of the variance of the sample mean (41.79) is much smaller than the variance of the population (2,800).

Example 5.4 Assume that the variance for daily production of a good is 2,800 pounds. What is the probability that in a sample of 67 randomly selected days the output is 15 pounds or more below average?

Solution 5.4

Note that no sample mean is obtained and there is no need to know the population mean. We are interested in the deviation from the population average.

$$P\left[\left(\hat{\mu}-\mu\right) < -15\right] = P\left[\frac{\hat{\mu}-\mu}{\sqrt{\frac{\sigma^2}{n}}} < \frac{-15}{\sqrt{\frac{2800}{67}}}\right] = P\left[Z < \frac{-15}{\sqrt{41.79}}\right]$$

$$= P\left[Z < \frac{-15}{6.47}\right] = P\left[Z < -2.32\right]$$

$$= 0.5 - P\left[0 < Z < 2.32\right] = 0.5 - 0.4898 = 0.0102$$

Therefore, for this production process the chance of a sample mean being more than 15 pounds below the target is 102 in 10,000. Accordingly, potentially over 100 customers in every 10,000 or roughly 1 in 100 might receive a "lighter" product and would ask for a refund or replacement.

Population Variance Is Unknown

Let $(\hat{\mu})$ be the mean of a random sample of size *n* from a distribution with a finite mean and a finite and *unknown* positive variance (σ^2). According to the Central Limit Theorem,

1. The distribution of $(\hat{\mu})$ can be approximated by a *t* distribution function
2. $E(\hat{\mu}) = \mu$
3. $\widehat{\sigma^2_{\hat{\mu}}} = \dfrac{\widehat{\sigma^2}}{n}$

Therefore, we can use the *t* table values, which can be obtained from Stata through the *ttable* command, for comparison of the standardized values of the sample mean. As the sample size increases, the distinction between the normal distribution and the *t* distribution vanishes. The results can be proven for sample proportions as well. In each case, the mean and variance of the estimator will be different.

When the population size is finite, which is relatively small compared to the sample size, it is necessary to add a *correction factor* to the formulas. When the ratio of sample size to population size, $\dfrac{n}{N}$, is greater than 5%, you should use a correction factor with the variance. The correction factor is $\dfrac{N-n}{N-1}$. For finite populations, the variance of the sample mean becomes

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n}\frac{N-n}{N-1} \qquad \text{known variance}$$

$$\widehat{\sigma^2_{\hat{\mu}}} = \frac{\widehat{\sigma^2}}{n}\frac{N-n}{N-1} \qquad \text{unknown variance}$$

Summary

Distribution function for a one sample mean

	Distribution	Mean	Variance
Population Variance Is Known	Normal	μ	$\dfrac{\sigma^2}{n}$
Population Variance Is Unknown	*t*	μ	$\dfrac{\widehat{\sigma^2}}{n}$

Sampling Distribution of a One Sample Proportion

Let $(\hat{\pi})$ be a proportion from a random sample of size n from a distribution with a finite proportion (π) and a finite positive variance (σ^2). When both $n\ \pi \geq 5$ and $n\ (1-\pi) \geq 5,$ then the following theorem is correct based on the Central Limit Theorem:

1. The distribution of $(\hat{\pi})$ can be approximated by a *normal* distribution function
2. $E(\hat{\pi}) = \pi$
3. $\sigma_{\hat{\pi}}^2 = \dfrac{\hat{\pi}(1-\hat{\pi})}{n}$

Note that in order to obtain the variance of the sample proportion, we must estimate the population proportion using the sample proportion. Thus, $\sigma_{\hat{\pi}}^2 = \widehat{\sigma_{\hat{\pi}}^2}$. Therefore, standard normal table values can be used to obtain occurrence probabilities. The symbol (π) is used to represent the population proportion and has nothing to do with the mathematical constant (π) = 3.141593.

An added advantage of using the sampling distribution of the sample proportion is that you can use the normal approximation to estimate the probability of outcomes for the *binomial distribution function* without direct computation or the use of a binomial distribution table. When both $n\ \pi \geq 5$ and $n\ (1-\pi) \geq 5,$ it is reasonable to approximate a binomial distribution using a normal distribution.

Sampling Distribution of Two Sample Means

The extension from the distribution function of a single sample mean to two means is simple and follows naturally. However, it is necessary to introduce the appropriate theories.

Theorem 5.3 The Expected Value of the Sum of Random Variables

Let $Y = X_1 + X_2 + \ldots X_n$, where Xs are independent random variables. The expected value of Y is equal to the sum of the expected values of Xs.

$$E(Y) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

Theorem 5.3 requires that the Xs be independent.

Sampling Distribution of the Difference of Two Means

When conducting inferences about two population parameters, there are two sample statistics, one from each population. Often, in order to conduct an inference, the relationship between the parameters, and hence the corresponding statistics, has to be modified and written as either the difference of the parameters or the ratio of the parameters. This requires knowledge of the distribution function for the *difference* of two sample statistics or the distribution function for the *ratio* of two sample statistics. In this section, the sampling distribution of the difference of two sample means is discussed, followed by the distribution function for the difference of two sample proportions. Later the distribution function for the ratio of two variances will be addressed.

The Two Sample Variances are Known and Unequal

Let $(\widehat{\mu_1})$ and $(\widehat{\mu_2})$ be the means of two random samples of sizes n_1 and n_2 from distributions with finite means μ_1 and μ_2 and finite positive *known* and *unequal* variances of σ_1^2 and σ_2^2. According to the Central Limit Theorem and Theorem 5.3,

1. The distribution of $(\widehat{\mu_1} - \widehat{\mu_2})$ can be approximated by a *normal* distribution function
2. $E(\widehat{\mu_1} - \widehat{\mu_2}) = \mu_1 - \mu_2$
3. $\text{Var}(\widehat{\mu_1} - \widehat{\mu_2}) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_1}$

Note that the variances of the two samples are added together, while the means are subtracted. Therefore, we can use the normal table values for comparison of the standardized values of the differences of sample means.

The Two Sample Variances are Known and Equal

Let $(\widehat{\mu_1})$ and $(\widehat{\mu_2})$ be the means of two random samples of sizes n_1 and n_2 from distributions with finite means μ_1 and μ_2 and finite positive

known and *equal* variances σ_1^2 and σ_2^2. Let n_1 and n_2 be the respective sample sizes. According to the Central Limit Theorem and Theorem 5.3,

1. The distribution of $(\widehat{\mu_1} - \widehat{\mu_2})$ can be approximated by a *normal* distribution function
2. $E(\widehat{\mu_1} - \widehat{\mu_2}) = \mu_1 - \mu_2$
3. $\text{Var}(\widehat{\mu_1} - \widehat{\mu_2}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$

Since $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The Two Sample Variances are Unknown and Unequal

Let $(\widehat{\mu_1})$ and $(\widehat{\mu_2})$ be the means of two random samples of sizes n_1 and n_2 from distributions with finite means μ_1 and μ_2 and finite positive *unknown* and *unequal* variances σ_1^2 and σ_2^2, respectively. According to the Central Limit Theorem and Theorem 5.3,

1. The distribution of $(\widehat{\mu_1} - \widehat{\mu_2})$ can be approximated by a *t* distribution function
2. $E(\widehat{\mu_1} - \widehat{\mu_2}) = \mu_1 - \mu_2$
3. $\text{Var}(\widehat{\mu_1} - \widehat{\mu_2}) = \frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}$

The Two Sample Variances are Unknown and Equal

Let $(\widehat{\mu_1})$ and $(\widehat{\mu_2})$ be the means of two random samples of sizes n_1 and n_2 from distributions with finite means μ_1 and μ_2 and finite positive *unknown* and *equal* variances σ_1^2 and σ_2^2, respectively. According to the Central Limit Theorem and Theorem 5.3,

1. The distribution of $(\widehat{\mu_1} - \widehat{\mu_2})$ can be approximated by a *t* distribution function
2. $E(\widehat{\mu_1} - \widehat{\mu_2}) = \mu_1 - \mu_2$

3. $\text{Var}(\widehat{\mu}_1 - \widehat{\mu}_2) = \widehat{\sigma^2_{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_1}\right)$

where

$$\widehat{\sigma^2_{Pooled}} = \frac{(n_1 - 1)\widehat{\sigma_1^2} + (n_2 - 1)\widehat{\sigma_2^2}}{n_1 + n_2 - 2}$$

The concept of pooled variances was discussed in Chapter 2 in the section named "Average of Several Variances."

Summary of the Sampling Distribution of Sample Means

Do not let these seemingly different and possibly difficult formulas confuse you. They are similar. The most common case is case 3 (The Two Sample Variances are Unknown and Unequal). The first three cases can use this formula without any problem. The last case takes advantage of the fact that there are two estimates of the single unknown variance of the population instead of one. Logic dictates that it would be better to average the two estimates using their respective sample sizes as weights as explained in Chapter 2. Table 5.1 provides a summary of the various sample statistics, their distribution functions, and their parameters for one and two sample means.

Table 5.1 Summary of sampling distribution for sample mean statistics

Sample Statistic	Population Variance(s)	Distribution Function	Mean	Variance of the Sample Statistic
$\widehat{\mu}$	Known	Normal	μ	$\frac{\sigma^2}{n}$
$\widehat{\mu}$	Unknown	*t*	μ	$\frac{\widehat{\sigma^2}}{n}$
$\widehat{\mu}_1 - \widehat{\mu}_2$	Known and Unequal	Normal	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}$
$\widehat{\mu}_1 - \widehat{\mu}_2$	Known and Equal	Normal	$\mu_1 - \mu_2$	$\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$
$\widehat{\mu}_1 - \widehat{\mu}_2$	Unknown and Unequal	*t*	$\mu_1 - \mu_2$	$\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}$
$\widehat{\mu}_1 - \widehat{\mu}_2$	Known and Equal	*t*	$\mu_1 - \mu_2$	$\widehat{\sigma^2_{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Sampling Distribution of the Difference of Two Proportions

Let $\widehat{\pi}_1$ and $\widehat{\pi}_2$ be proportions of interest in two random samples of sizes n_1 and n_2 from distributions with finite proportions of π_1 and π_2 and finite positive variances of σ_1^2 and σ_2^2. According to the Central Limit Theorem,

1. The distribution of $(\widehat{\pi}_1 - \widehat{\pi}_2)$ can be approximated by a *normal distribution* function
2. $E(\widehat{\pi}_1 - \widehat{\pi}_2) = \pi_1 - \pi_2$
3. $\text{Var}(\widehat{\pi}_1 - \widehat{\pi}_2) = \dfrac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \dfrac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}$

Therefore, we can use the normal table values for comparison of the standardized values of sample proportions. In practice, $\pi_1 - \pi_2$ are not known; otherwise, it would have been an exercise in futility. Therefore, their estimates, $\widehat{\pi}_1$ and $\widehat{\pi}_2$, respectively, are used in calculating the variance of $(\widehat{\pi}_1 - \widehat{\pi}_2)$. The distribution function, expected value (mean), and variance for one and two sample proportions are given in Table 5.2.

Table 5.2 Summary of the sampling distributions of sample proportions

Sample Statistic	Population Variance(s)	Distributio n Function	Mean	Variance of the Sample Statistic
$\hat{\pi}$	NA	Normal	π	$\dfrac{\hat{\pi}(1-\hat{\pi})}{n}$
$\widehat{\pi}_1 - \widehat{\pi}_2$	NA	Normal	$\pi_1 - \pi_2$	$\dfrac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \dfrac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}$

Sampling Distribution of the Sample Variance

Theorem 5.4 Let random variable X have a normal distribution with mean μ and variance σ^2, then the random variable

$$V = \left(\frac{X-\mu}{\sigma}\right)^2 = Z^2$$

has a chi-squared distribution with one (1) degree of freedom, which is shown as $\chi^2(1)$ as was shown in Chapter 4. Z is the same Z score as discussed in previous chapters, which consists of individual error (deviation from the mean) divided by average error (standard deviation). Chi-squared distributions are cumulative. Therefore, when n chi-squared distribution functions are added up, the result is another chi-squared distribution with n degrees of freedom.

Theorem 5.5 Let random variables $X_1, X_2, \ldots X_n$ have normal distributions each with mean μ and variance σ^2, then the following sum

$$\sum\left(\frac{X-\mu}{\sigma}\right)^2$$

has a chi-squared distribution with n degrees of freedom. From this relation, we can build confidence intervals for one and two variances, and conduct tests of hypothesis for one and two variances.

Sampling Distribution of Two Samples Variances

When conducting inferences about two population parameters, there are two sample statistics, one from each population. Often, in order to conduct an inference, the relationship between the parameters, and hence the corresponding statistics, has to be modified and written as either the difference of the parameters or the ratio of the parameters. This requires knowledge of the distribution function of the difference of two sample statistics or the distribution function of the ratio of two sample statistics. In this section, the sampling distribution of the ratio of two sample variances is discussed.

Theorem 5.6

Let random variables $X_1, X_2, \ldots X_m$ have normal distributions with mean μ_1 and variance σ_1^2, and the random variables $Y_1, Y_2, \ldots, Y_n$ have normal distributions each with mean μ_2 and variance σ_2^2, then the random variable

$$F = \frac{\sum\left(\frac{X-\mu_1}{\sigma}\right)^2}{\sum\left(\frac{Y-\mu_2}{\sigma}\right)^2}$$

has an F distribution with m and n degrees of freedom.

Note that in Theorem 5.6, we could have expressed the numerator and denominator in terms of the corresponding chi-square distributions as stated in Theorem 5.5, which in turn is built upon Theorem 5.4.

In practice, when population variances are not known, they are substituted by their respective sample variances.

$$F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}$$

Customarily, the larger sample variance is placed in the numerator to make comparison with the F-table easier. Tabulated values of F are greater than or equal to 1.

The most common use of F distributions at this level is the test of hypothesis of equality of two variances. This test also provides a way of determining whether or not to pool variances when testing for the equality of two means with unknown population variances. To do this, first test the equality of the variances. If the hypothesis of no difference is rejected, then the variances are different and are not pooled.

Another use of the F distribution is in testing three or more means. When testing a hypothesis that involves more than two means, t distributions cannot be used. Tests of hypothesis and the use of t and F distributions are discussed more in Chapter 6.

Efficiency Comparison between the Mean and the Median

Let $\widehat{\mu}$ be the sample mean and $\widehat{M}$ be the sample median. The expected value of both the sample mean and sample median is equal to population mean. That is, both provide unbiased estimates of the population mean. However, the sample mean is more efficient than the sample median

in estimating the population mean as shown below. Remember from Chapter 1, an estimator of a parameter is more efficient than another estimator if it has smaller variance. The variance of the sample mean $(\widehat{\mu})$ is

$$\widehat{\sigma^2_{\hat{\mu}}} = \frac{\sigma^2}{n}$$

It can be shown that the variance of the median is

$$\text{Variance (Sample Median)} = \frac{\pi\sigma^2}{2n}$$

where $\pi = 3.141593\ldots$

$$\frac{\text{var}(\widehat{\mu})}{\text{var}(\widehat{M})} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} = \frac{2}{3.14159} = 0.64$$

Therefore, $\widehat{\mu}$ is more efficient than the median in estimating the population mean. The variance of the sample median from a sample of size 100 is about the same as the variance of the sample mean from a sample of size 64. Therefore, the sample mean from a sample of size 64 is as efficient as a sample of size 100 using the sample median in estimating the population mean.

It is worthwhile to note the following discrepancy, which is caused by having different orientations or starting points:

$$\text{Var}(\widehat{M}) = 1.57\ \text{Var}(\widehat{\mu})$$

$$\text{Var}(\widehat{\mu}) = 0.6366\ \text{Var}(\widehat{M})$$

In other words, to obtain the same efficiency in estimating the population mean using a sample mean with a sample of size 100, it is necessary to obtain a sample of size 157 when using sample median.

Recall that when extreme observations exist, the sample median is preferred to the sample mean because it is not influenced by extreme values. For example, in real estate, it is of interest to know the price of a typical house. The industry reports the prices of homes sold each month. Usually, only a small fraction of existing homes is sold in any

given month. This causes large fluctuations in the average prices of homes sold, which constitutes the sample. The industry reports the median instead of the average price for the listings to avoid large fluctuations, which can send the wrong signal and cause uncertainty or panic in the market. Since the sample median is less efficient than the sample mean in estimating the population mean, larger samples are needed in order to obtain a realistic estimate of home values. Therefore, it is advisable to use median prices over several months.

CHAPTER 6

Point and Interval Estimation

Estimation *versus* Inference

There are two applications of statistics: descriptive and inferential. Descriptive statistics summarize data in forms of tables, graphs, or computed values and are used for estimation. We can use descriptive statistics to describe population data or sample data. Inferential statistics is used to draw conclusions about a population parameter using sample statistics—they are used to make decisions. Obtaining sample statistics for inferential statistics is the same as obtaining them for descriptive statistics. Whether or not a statistic is considered to be descriptive or inferential depends on its use. The statistics obtained from a sample are called *estimates*, to emphasize the fact that they are estimates for their respective parameters. Estimation is important as we have demonstrated in previous chapters.

Discussions up to this point have pertained to descriptive statistics. When sample estimates are used to test claims about a population parameter and to indicate how far the estimates are from parameters, we are in the domain of inferential statistics. Some statisticians believe that the primary objective of statistics is to make *inferences* about population parameters using sample statistics. Using sample statistics to make deductions about the population parameters is called *statistical* inference. Statistical inference can be based on *point estimations* or *confidence intervals*, both of which will be covered shortly. They are closely related, and in some cases, they are interchangeable.

Point Estimation

A point estimate is the statistic obtained from a sample. The reason for the name is because the estimate consists of a single value. Examples of point estimates include the sample mean ($\hat{\mu}$), sample proportion ($\hat{\pi}$), sample variance ($\widehat{\sigma^2}$), and sample median $\widehat{M}$. These statistics are used to estimate the population mean (μ), population proportion (π), population variance (σ^2), and population median (M), respectively. The sample median can be used to estimate both the population median and the population mean as seen in Chapter 5. Any single-valued estimate obtained from a sample is a point estimate. Good estimates are close to their corresponding population parameter, have low variation, and converge to their respective parameters as the sample size increases. Proper sampling provides accurate estimates of the unknown population parameter. As discussed in Chapter 3, a suitable estimate is *unbiased*, *consistent*, and *efficient*.

Although point estimates are useful in providing descriptive information about a population, their usefulness is limited because it is not possible to ascertain how far they are from the targeted parameter. In order to provide levels of confidence and a probability for the *margin of error*, one needs to know the distribution function of the sample statistics.

Once the *sampling distribution* of the sample statistic is known, the probability of observing a certain sample statistic can be calculated with the aid of the corresponding table.

Example 6.1 Calculate point estimates of the mean, median, variance, standard deviation, and coefficient of variation for the stock prices of Microsoft for May 21 to July 2, 2015.

Solution 6.1

The necessary computations as well as the procedure for obtaining the results from Excel and Stata were provided starting with Example 2.2 in Chapter 2 (Figure 6.1).

```
. summarize, detail

                            MSFT
-------------------------------------------------------------
      Percentiles      Smallest
 1%         44.15         44.15
 5%         44.37         44.37
10%        44.425          44.4       Obs                  30
25%         45.65         44.45       Sum of Wgt.          30

50%         46.12                     Mean            46.09633
                        Largest       Std. Dev.       .9294879
75%         46.85         47.23
90%        47.325         47.42       Variance        .8639477
95%         47.45         47.45       Skewness       -.4715205
99%         47.61         47.61       Kurtosis        2.61937
```

Figure 6.1 Stata command and output for detailed summary statistics

$$\hat{\mu} = 46.096, \widehat{\sigma^2} = 0.864, \widehat{M} = 46.12, \hat{\sigma} = 0.9295, CV = 0.020$$

Interval Estimation

Statistics deals with random phenomena. Nothing remains constant in life. Methods of production change; processes are modified; machines get out of calibration; and new techniques are applied. In all these cases, statistics are used to determine what remains constant and what changes. In descriptive statistics, sample statistics are used to estimate the population parameters.

Interval estimation, such as the use of confidence intervals, augments point estimates by providing a margin of error for the point estimate. The *margin of error* is a range that is added to or subtracted from the point estimate of the population parameter. Confidence intervals are based on the point estimate of the parameter and the *distribution function of the point estimate*. It is affected by the desired level of certainty, the variance of the data, and the sample size.

Calculating Confidence Intervals

Interval estimation is a simple notion and is defined as

$$\text{Point estimate} \pm \text{Margin of error} \tag{6.1}$$

Definition 6.1 The *margin of error* is obtained by multiplying the Z score corresponding to a desired level of certainty and the standard deviation of the estimate of a parameter.

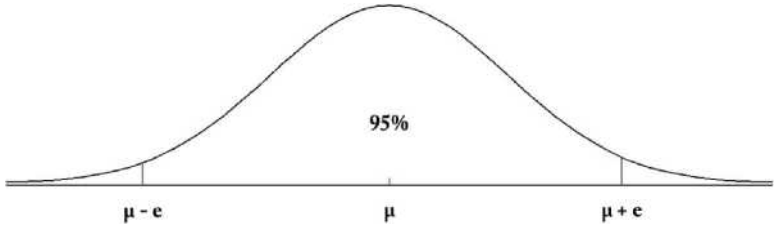


Figure 6.2 Margin of error on the normal distribution

Figure 6.2 depicts the margin of error of 95% certainty. This means that we are 95% certain that the confidence interval covers the true population parameter. The Z score for a normal probability of 0.475 (half of 0.95) is 1.96. Thus, the margin of error in this case is 1.96 times the standard deviation of the estimate. When the estimate of the population mean is the sample mean, instead of a single observation, the resulting standard deviation is called the *standard error*. On the basis of the Central Limit Theorem, when the statistic is the sample mean or sample proportion, the margin of error is obtained by

$$\text{Margin of Error } = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.2}$$

It is important to realize that the margin of error formula in Equation (6.2) depends on the knowledge of the population variance. When the population variance is unknown and the sample variance has to be used, then the formula must be adjusted by replacing the population standard deviation with the sample standard deviation, and consequently the Z value must be replaced by the t value as in Equation (6.3)

$$\text{Margin of Error } = t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \tag{6.3}$$

Example 6.2 Calculate the margin of error for the Microsoft stock price for the period May 21 to July 2, 2015, using an 83% level of confidence.

Solution 6.2

First obtain the Z value that corresponds to half of the 0.83 level of confidence.

$$P(0 < Z < X) = \left(\frac{0.83}{2}\right) = 0.415$$

Using any of the previously shown methods, obtain the Z score of 1.375, which corresponds to the probability of 0.415. Note that we are using the sample variance as if it were the *actual* population variance. According to Example 6.1, the variance of all 30 observations is 0.864.

Margin of Error

$$= (1.375)\sqrt{\frac{0.864}{30}} = (1.375)\sqrt{0.0288} = (1.375)(0.1697) = 0.233$$

This would have been the correct margin had we known population variance. Since the population variance is unknown, we must use the t value instead of Z value. Unfortunately, t values for an alpha of $(1 - 0.83)/2 = 0.085$ are not readily available from conventional t tables. However, Microsoft Excel provides the necessary number by using the following command:

=t.inv(0.085, 29)

where the first number, $(1 - 0.83)/2 = 0.085$, is half of the desired confidence level, and 29 is equal to $n - 1$. Excel displays –1.407. The negative sign indicates the point is to the left of the center. Therefore, the probability between –1.407 and 1.407 equals to 0.83. In order to get the positive-signed value instead use 0.085 + 0.83 = 0.915. Alternatively 1 – 0.085 = 0.915 can be used.

=t.inv(0.915, 29)

The above formula in Excel gives the positive result of 1.407. Therefore,

Margin of Error

$$= (1.407)\sqrt{\frac{0.864}{30}} = (1.407)\sqrt{0.0288} = (1.407)(0.1697) = 0.238768$$

This is the correct value of the margin of error because it is using the t value, as required when the population variance is unknown. The consequence of not having normal distribution is the widening of the margin of error from ± 0.233 to ± 0.239.

The concept of margin of error applies to sample statistics but not to population parameters. Population parameters are constant values and do not have a margin of error. Sample statistics, which are random variables used for estimating their respective population parameters, have a margin of error in estimating the parameter of interest. As seen in Equation (6.2), the margin of error is *directly* related to the square root of variance of the population and the level of confidence, as indicated by the Z score, and *inversely* related to the square root of the sample size.

The probability between –Z and +Z from a standard normal, that is, a normal distribution with mean of zero (0) and variance of one (1), is shown by $(1-\alpha)\%$. This is because the sum of the areas outside of the above range is equal to α. Chapter 7 provides more explanation for the naming of these areas and more detail on the meaning of the term α. Customarily, normal distribution tables are calculated for half of the area, because of symmetry. Thus, the Z value, which corresponds to one half of the α, is shown as $Z_{\alpha/2}$.

Interval Estimation for One Population Mean

Definition 6.2 The $(1-\alpha)\%$ *confidence interval for the mean of one population* (μ) when the variance is known is given by

$$\hat{\mu} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{6.4}$$

Definition 6.3 The $(1-\alpha)\%$ confidence interval for the mean of one population (μ) when the variance is unknown is given by

$$\hat{\mu} \pm t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$$

To obtain the confidence interval, it is necessary to calculate the following two values:

$$\text{Lower Bound (LB)}: \hat{\mu} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \hat{\mu} - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \tag{6.5}$$

$$\text{Upper Bound (UB)}: \hat{\mu} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \hat{\mu} + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \tag{6.6}$$

Example 6.3 Provide a 95% confidence interval for Microsoft stock price for the period of May 21 to July 2, 2015.

Solution 6.3

Since the population variance is *unknown* and the sample size is *small*, we need to use the *t distribution*. The *t* value for the 95% confidence interval is ± 2.04523. The 95% probability refers to the center of the graph while customarily the formula uses the notation referring to the two end points, and hence the use of α, and its half in the formula. Therefore, 1 − 0.95 = 0.05 and half of that is 0.025 for each end. To obtain the correct *t* value, look for the 2.5% probability with 29 degrees of freedom in the *t* table or use the following Excel command:

=t.inv(.025,29) = -2.04523

The value reported by Excel is negative because it is designed to report the lower-end critical value. To obtain a positive value, use 0.95 + 0.025 = 0.975 in the built-in command.

For this example, we will use the average Microsoft stock price of 46.096, the *t* value calculated directly above, and 0.1697, which is the standard error.

Lower Bound = 46.096 – 2.04523 (0.1697) = 45.75

Upper Bound = 46.096 + 2.04523 (0. 1697) = 46.44

A 95% confidence interval for the mean Microsoft stock price is given by the range $45.75 to $46.44, compared with the smaller range $45.86 to $46.33 obtained using *Z*, which requires the knowledge of the population variance and normality. Estimating a parameter adds uncertainty and error; thus, *t* values are larger than *Z* values to account for the fact that the population variance is unknown and must be estimated (Figure 6.3). Accordingly, the corresponding confidence interval

is wider than the one calculated using a Z table when we assume to know the population variance.

The corresponding command in Stata is

means *varlist*

where *varlist* is the name of the variable. Recall that this command displays three commonly used means (see Example 2.4 and Figure 2.1).

```
. means msft
```

Variable	Type	Obs	Mean	[95% Conf.	Interval]
msft	Arithmetic	30	46.09633	45.74926	46.44341
	Geometric	30	46.08721	45.73976	46.43731
	Harmonic	30	46.07803	45.73018	46.43123

Figure 6.3 Stata command and output for confidence interval for Microsoft stock prices

Stata also provides the more powerful command *ci*. Use both the Microsoft and the Walmart data from Example 2.2 (Figure 6.4).

```
. ci msft wmt
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
msft	30	46.09633	.1697005	45.74926	46.44341
wmt	30	73.215	.2505143	72.70264	73.72736

Figure 6.4 Stata command and output using CI command

To compare the time periods from the first half of the dataset to the second half, you will first have to create a new column named "period." Assign the value of 1 for rows from May 21 through June 11 and the value 2 for period June 12 through July 2, 2015. To do this, first type *gen period=1*. Then go to the data editor and change the number for period to 2 for the dates from June 12 through July 2. You can do this by clicking in the cell and typing the number. Alternatively, enter the data in Excel before copying them into Stata.

Although this is a sorted data by definition, it is necessary to sort the data in Stata to be able to create confidence intervals. Adding the option *by (period) total* will provide the confidence interval for the first, second, and combined periods (Figure 6.5). The last block is the same as the one demonstrated above.

```
. sort period

. ci msft wmt, by(period) total

-> period = 1

    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
        msft |        15    46.71733    .1524762        46.3903    47.04436
         wmt |        15      74.232    .3020757       73.58411    74.87989

-> period = 2

    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
        msft |        15    45.47533    .2024019       45.04122    45.90944
         wmt |        15      72.198    .1448784       71.88727    72.50873

> -
-> Total

    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
        msft |        30    46.09633    .1697005       45.74926    46.44341
         wmt |        30      73.215    .2505143       72.70264    73.72736
```

Figure 6.5 Stata command and output using CI command using an option

Definition 6.4 The $(1-\alpha)\%$ confidence interval for the *proportion of one population* (π) is given by equation:

$$\hat{\pi} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \tag{6.7}$$

Example 6.4 Calculate a 95% confidence interval for the proportion of stock prices of Microsoft that are higher than \$46.90. Use the sample between May 21 and June 11, 2015 provided in Example 3.2.

Solution 6.4

Upon verification, there are 6 stock prices over \$46.90 among the 15 observations in this period.

$$\hat{\pi} = \frac{6}{15} = 0.4$$

The Z value corresponding to 95% confidence is 1.96.

$$\text{LB} = 0.4 - 1.96\left(\sqrt{\frac{0.4\times 0.6}{15}}\right) = 0.4 - 1.96\times 0.1265$$
$$= 0.4 - 0.2479 = 0.1521$$

$$\text{UB} = 0.4 + 1.96\left(\sqrt{\frac{0.4\times0.6}{15}}\right) = 0.4 + 1.96\times0.1265$$
$$= 0.4 + 0.2479 = 0.6479$$

The range 0.1521 to 0.6479 covers the true population proportion of Microsoft stock prices that are $46.90 or higher.

Since the population variance is not known and the sample size is smaller than 30, we should have used the *t* distribution instead of the normal distribution. In practice, the sample size is much larger when proportions are used. The results using the *t* values are given by

$$\text{LB} = 0.4 - 2.145\left(\sqrt{\frac{0.4\times0.6}{15}}\right) = 0.4 - 2.145\times0.1265$$
$$= 0.4 - 0.2713 = 0.1287$$

$$\text{UB} = 0.4 + 2.145\left(\sqrt{\frac{0.4\times0.6}{15}}\right) = 0.4 + 2.145\times0.1265$$
$$= 0.4 + 0.2713 = 0.6713$$

The range 0.1287 to 0.6713 covers the true population proportion of Microsoft stock prices that are $46.90 or higher. Notice that the range became wider when a *t* distribution value is used.

The Stata option of *binomial* provides the confidence interval for proportions (Figure 6.6). For the above example, create a new column where any price higher than $46.90 receives the value of 1; call it "higher" (try using *gen higher=(msft>46.90)* to create the variable). The results are slightly different due to difference in methods.

```
. ci higher, binomial

                                                    — Binomial Exact —
    Variable |       Obs        Mean    Std. Err.   [95% Conf. Interval]
-------------+--------------------------------------------------------
      higher |        15          .4   .1264911    .1633643    .6771302
```

Figure 6.6 Stata command and output using binomial distribution

Note that the confidence interval is narrower than the one using *Z* or *t*. In fact even this interval covers more than 95% of cases.[1] The *agresti* option provides an even narrower interval (Figure 6.7).

```
. ci higher, binomial agresti

                                                    — Agresti-Coull —
    Variable |        Obs        Mean    Std. Err.   [95% Conf. Interval]
-------------+--------------------------------------------------------------
      higher |         15          .4    .1264911    .1975013    .6432753
```

Figure 6.7 Stata command and output using binomial distribution modified

Definition 6.5 The $(1-\alpha)\%$ confidence interval for the *difference of means of two populations* $(\mu_1 - \mu_2)$ when the population *variances are known and unequal* is given by Equation (6.8).

$$\widehat{\mu_1} - \widehat{\mu_2} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.8}$$

Definition 6.6 The $(1-\alpha)\%$ confidence interval for the *difference of means of two populations* $(\mu_1 - \mu_2)$ when the population *variances are known and equal* is given by Equation (6.9).

$$\left(\widehat{\mu_1} - \widehat{\mu_2}\right) \pm Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{6.9}$$

Definition 6.7 The $(1-\alpha)\%$ confidence interval for the *difference of means of two populations* $(\mu_1 - \mu_2)$ when the population *variances are unknown and unequal* is given by Equation (6.10).

$$\left(\widehat{\mu_1} - \widehat{\mu_2}\right) \pm t_{\alpha/2}\sqrt{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}} \tag{6.10}$$

Example 6.5 Obtain a 95% confidence interval for the difference in mean Microsoft stock prices between May 21 to June 11 and June 12 to July 2 of 2015.

Solution 6.5

Use the data from Example 6.3 where you created the *period* variable. The necessary formula for when the population variance is unknown is given in Equation (6.10). Assume the variances of the two samples are not equal in order to be able to compare the outcome with those from the next example. To improve the estimate, average the two variances using the formula for pooled variance. Avoid rounding the values and use computational formulas to avoid computational error.

$$\widehat{\mu}_1 = 46.72 \qquad \widehat{\sigma_1^2} = 0.348735$$

$$\widehat{\mu}_2 = 45.48 \qquad \widehat{\sigma_2^2} = 0.614498$$

$$= \text{t.inv}(.025,28) = -2.04841$$

Recall that we need to use both positive and negative t values.

$$\begin{aligned}
\text{LB} &= (\widehat{\mu}_1 - \widehat{\mu}_2) - t_{\alpha/2}\sqrt{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}} \\
&= (46.72 - 45.48) - 2.04841\sqrt{\frac{0.348735}{15} + \frac{0.614498}{15}} \\
&= 1.24 - 2.04841\sqrt{0.0232 + 0.0409} = 1.24 - 2.04841\sqrt{0.064216} \\
&= 1.24 - 2.04841(0.253408) = 1.24 - 0.519083 = 0.72
\end{aligned}$$

$$\begin{aligned}
\text{UB} &= (\widehat{\mu}_1 - \widehat{\mu}_2) + t_{\alpha/2}\sqrt{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}} \\
&= (46.72 - 45.48) + 2.04841\sqrt{\frac{0.348735}{15} + \frac{0.614498}{15}} \\
&= 1.24 + 2.04841\sqrt{0.0232 + 0.0409} = 1.24 + 2.04841\sqrt{0.064216} \\
&= 1.24 + 2.04841(0.253408) = 1.24 + 0.519083 = 1.76
\end{aligned}$$

The range \$0.72 to \$1.76 covers the difference of the means of the two periods of stock prices for Microsoft with 95% probability. Note that the range does not cover zero, which indicates that the averages of the stock prices for the two periods are not the same. The price has been falling steadily. The average price of the first 15 observations is higher than the average price of the last 15 as is evident in Figure 6.8.

The easiest way to obtain confidence interval for differences of means in Stata is through the *ttest* command that provides test of hypotheses about the equality of two means, which is covered in Chapter 7.

Definition 6.8 The $(1-\alpha)\%$ confidence interval for the *difference of means of two populations* $(\mu_1 - \mu_2)$ when the population *variances are unknown and equal* is given by Equation (6.11).

$$(\widehat{\mu}_1 - \widehat{\mu}_2) \pm t_{\alpha/2}\sqrt{\sigma^2_{Pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{6.11}$$

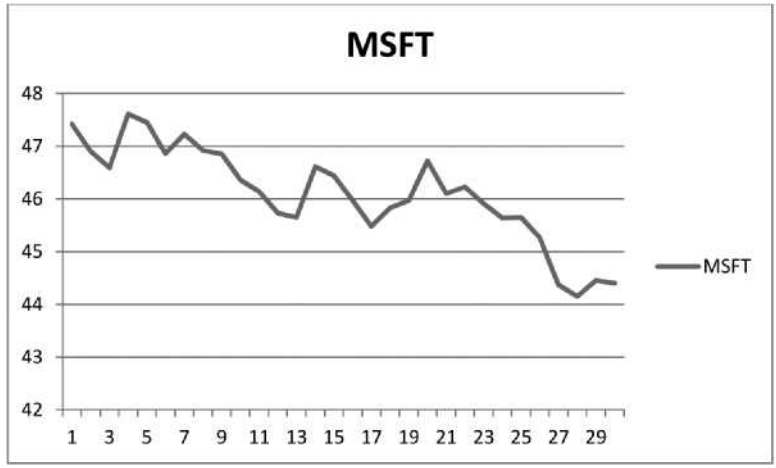


Figure 6.8 Microsoft stock prices, May 21 to July 2, 2015

where

$$\sigma^2_{Pooled} = \frac{(n_1 - 1)\widehat{\sigma_2^2} + (n_2 - 1)\widehat{\sigma_1^2}}{n_1 + n_2 - 2}$$

Example 6.6 Obtain a 95% confidence interval for the difference in Microsoft stock prices between May 21 to June 11 and June 12 to July 2, 2015.

Solution 6.6

Since the data belong to the same company and are so close in time period, it is reasonable to assume that there is one variance for the company's stock prices and the two sample statistics are two estimates of the same population variance. Therefore, it is necessary to find their weighted average and then use Equation (6.11). We already have the following results:

$$\widehat{\mu_1} = 46.71 \qquad \widehat{\sigma_1^2} = 0.348735$$

$$\widehat{\mu_2} = 45.48 \qquad \widehat{\sigma_2^2} = 0.614498$$

$$= \text{t.inv}(.025,28) = -2.04841$$

$$\sigma^2_{Pooled} = \frac{(n_1 - 1)\widehat{\sigma_1^2} + (n_2 - 1)\widehat{\sigma_2^2}}{n_1 + n_2 - 2}$$

$$= \frac{(15-1)(0.348735)+(15-1)(0.614498)}{15+15-2}$$

$$= \frac{4.88229+8.60297}{28} = 0.48162$$

$$\text{LB} = (\widehat{\mu}_1 - \widehat{\mu}_2) - t_{\alpha/2}\sqrt{\widehat{\sigma^2_{Pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= (46.72-45.48) - 2.04841\sqrt{0.48162\left(\frac{1}{15}+\frac{1}{15}\right)}$$

$$= 1.24 - 2.04841\sqrt{0.48162(0.0667+0.0667)}$$

$$= 1.24 - 2.04841\sqrt{0.64216}$$

$$= 1.24 - 2.04841(0.253409) = 1.24 - 0.51908 = 0.72$$

$$\text{UB} = (\widehat{\mu}_1 - \widehat{\mu}_2) - t_{\alpha/2}\sqrt{\widehat{\sigma^2_{Pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= (46.72-45.48) + 2.04841\sqrt{0.48162\left(\frac{1}{15}+\frac{1}{15}\right)}$$

$$= 1.24 + 2.04841\sqrt{0.48162(0.0667+0.0667)}$$

$$= 1.24 + 2.04841\sqrt{0.64216}$$

$$= 1.24 + 2.04841(0.253409) = 1.24 + 0.51908 = 1.76$$

The range \$0.72 to \$1.76 covers the difference of the means of the two periods of stock prices for Microsoft with 95% probability.

The reason these results are exactly the same as the result for the previous case, where we did not assume the equality of the variances, is that the two sample sizes are equal. When weights for variances are equal, the results of the arithmetic and weighted average are always identical. When samples have different sizes, the results will be different.

Definition 6.9 The $(1-\alpha)\%$ confidence interval for the *difference of two population proportions* $(\pi_1 - \pi_2)$ is given by:

$$(\widehat{\pi}_1 - \widehat{\pi}_2) \pm Z_{\alpha/2}\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1}+\frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}} \qquad (6.12)$$

Example 6.7 Calculate a 95% confidence interval for the difference of the proportion of stock prices of Microsoft that are more than or equal

to \$45.90 for periods May 21 to June 11 and June 12 to July 2, 2015 provided in Example 3.2.

Solution 6.7

Upon inspection, it becomes evident that 13 of the first 15 observations and 6 of the second group are greater than or equal to \$45.90. Their respective sample proportions are

$$\widehat{\pi}_1 = \frac{13}{15} = 0.867 \qquad \widehat{\pi}_2 = \frac{6}{15} = 0.4$$

The Z value corresponding to 95% confidence is 1.96. Insert these values in Equation (6.12) to obtain the results.

$$\begin{aligned} \text{LB} &= (0.867 - 0.4) - 1.96\sqrt{\frac{0.867(1-0.867)}{15} + \frac{0.4(1-0.4)}{15}} \\ &= 0.467 - (1.96)\sqrt{0.007687 + 0.016} = 0.467 - (1.96)(0.153906) \\ &= 0.467 - 0.302 = 0.165 \end{aligned}$$

$$\begin{aligned} \text{UB} &= (0.867 - 0.4) + 1.96\sqrt{\frac{0.867(1-0.867)}{15} + \frac{0.4(1-0.4)}{15}} \\ &= 0.467 + (1.96)\sqrt{0.007687 + 0.016} = 0.467 - (1.96)(0.153906) \\ &= 0.467 + 0.302 = 0.769 \end{aligned}$$

The range 0.165 to 0.769 covers the difference of the proportion of stock prices for Microsoft that is greater than or equal to \$45.90 in the two periods May 21 to June 11 and June 12 to July 2, 2015.

Definition 6.10 The $(1-\alpha)\%$ confidence interval for *one population variance* (σ^2) is given by

$$\frac{(n-1)\widehat{\sigma^2}}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\widehat{\sigma^2}}{\chi^2_{1-\alpha/2}} \tag{6.13}$$

Note that the chi-squared distribution is not symmetric; therefore, we cannot use the $\pm$ signs to form the confidence interval. Also, it is important to note that the term $\chi^2_{\alpha/2}$ refers to the right side of the distribution and, hence, it is larger than $\chi^2_{1-\alpha/2}$, which refers to the left side of the distribution. Dividing the same numerator by a larger number provides a smaller

result, hence the lower bound, whereas dividing the same numerator by a smaller value gives a larger result, hence the upper bound.

Example 6.8 Find the 95% confidence interval for the variance of stock prices for Microsoft. Use the same sample for May 21 to June 11, 2015 as in Example 6.5.

Solution 6.8

To obtain the boundaries of the 95% confidence interval, it is necessary to find the demarcations for 2.5% and 97.5% probabilities, which correspond to the lower 2.5% and the upper 2.5% (=1 − 0.975). The sample variance for the 15 observations and the chi-squared values for 0.025 and 0.975 (from the chi-squared table) with 14 degrees of freedom are

$$\widehat{\sigma^2} = 0.348735$$

$$\chi^{14}_{0.025} = 5.629\,.$$

$$\chi^{14}_{0.975} = 26.119$$

Use Equation (6.13) to build the confidence interval,

$$\text{LB} = \frac{(n-1)\widehat{\sigma^2}}{\chi^2_{\alpha/2}} = \frac{(15-1)0.348735}{26.119} = 0.1869$$

$$\text{UB} = \frac{(n-1)\widehat{\sigma^2}}{\chi^2_{1-\alpha/2}} = \frac{(15-1)0.348735}{5.629} = 0.8674$$

The range 0.1869 to 0.8674 covers the population variance of Microsoft stock prices with 95% confidence.

Definition 6.11 The $(1-\alpha)\%$ confidence interval for the *ratio of two population variances* $\dfrac{\sigma_1^2}{\sigma_2^2}$ is given by Equation (6.14).

$$\frac{\widehat{\sigma_1^2} \Big/ \widehat{\sigma_2^2}}{F_{1-\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\widehat{\sigma_1^2} \Big/ \widehat{\sigma_2^2}}{F_{\alpha/2}} \tag{6.14}$$

Since the F distribution is not symmetric, we cannot use the $\pm$ signs to form the confidence interval. The term $F_{\alpha/2}$ refers to the right side of the

distribution and, hence, it is larger than $F_{1-\alpha/2}$, which refers to the left side of the distribution. Dividing the same numerator by a larger number provides a smaller result, hence the lower bound, whereas dividing the same numerator by a smaller value gives a larger result, hence the upper bound.

Example 6.9 Find the confidence interval for the ratio of the variances for the two periods May 21 to June 11and June 12 to July 2, 2015, for Microsoft stock prices.

Solution 6.9

Let's mark the data from May 21 to June 11 with the subscript "1" and from June 12 to July 2 with the subscript "2." The variances and the *F* values for 0.025 and 0.975 with 14 and 14 degrees of freedom are

$$\sigma_1^2 = 0.348735$$

$$\sigma_2^2 = 0.614498$$

$$F_{0.025}^{14.14} = 2.978588$$

$$F_{0.975}^{14.14} = 0.339061$$

$$\text{LB} = \frac{\frac{0.614498}{0.348735}}{2.978588} = \frac{1.76207652}{2.978588} = 0.59158$$

$$\text{UB} = \frac{\frac{0.614498}{0.348735}}{0.339061} = \frac{1.76207652}{0.339061} = 5.19693$$

The range 0.59158 to 5.19693 covers the ratio of variances for the two periods for Microsoft stock prices with 95% confidence.

The Stata command for testing the equality of two variances provides the confidence interval for the ratio of two variances and is presented in Chapter 7.

Determining the Sample Size

In Chapter 4, we showed the necessary sample size for estimating the single mean of a population. The sample size in that case was obtained

by algebraic manipulation of the margin of error in Equation (6.4), which is repeated below for your reference.

$$\hat{\mu} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Setting the margin of error equal to a desired margin of error, E, and solving for n results are given in the following formula.

$$n = \frac{\sigma^2 \left(Z_{\alpha/2} \right)^2}{E^2} \tag{6.15}$$

Example 6.10 What size sample is needed to be within \$0.10 of the actual price with 95% confidence if the variance is 0.372182?

Solution 6.10

$$n = \frac{0.372182^2 \times 1.96^2}{0.1^2} = 53.21$$

Therefore, the necessary sample size is 54. Note that the variance in this formula is the population variance. When the population variance is unknown, use the sample variance instead, but remember to use the t value instead of the Z value.

Similar algebraic manipulations are applied to obtain sample sizes for cases with unknown variances involving either the one or two population means. We will only show the formula for one population proportions for reference.

$$n = \frac{\hat{\pi}(1-\hat{\pi}) \left(Z_{\alpha/2} \right)^2}{E^2} \tag{6.16}$$

Example 6.11 What size sample is needed to be within 5% of the population proportion with 95% confidence when the sample proportion is 0.4?

Solution 6.11

$$n = \frac{(0.4 \times 0.6) 1.96^2}{0.05^2} = 368.79$$

Therefore, the necessary sample size is n = 369.

Inference with Confidence Intervals

The primary objective of statistics is to make inferences about population parameters using sample statistics. Using sample statistics to make deductions about population parameters is called statistical inference. Statistical inference can be based on point estimation, confidence intervals, or a test of hypothesis. These are closely related and in some aspects they are interchangeable. The inference can be based on the estimation theory or decision theory. A test of hypothesis is a tool for decision theory. Estimation theory consists of point estimation and interval estimation. This section deals with *confidence interval estimation*.

Population parameters are unknown and constants. Sample statistics, which are random by nature, are used to provide estimates of population parameters. If sampling is random, then the sample statistic is a good estimate of the corresponding population parameter. A good sample statistic has certain desirable properties, as discussed in Chapter 3. These statistics are called point estimates because they provide a single value as the estimate of the population parameter. If the estimator is "good," then it should be close to the unknown true value of the population parameter. The single estimate does not indicate proximity to the true parameter or probability of being close to the true parameter. Confidence intervals give both an idea of the actual value of the population parameter and a probability, or a level of confidence, that the interval includes the population parameter (Tables 6.1 and 6.2). Confidence intervals can be used for decision theory as well. The level of significance is $(1-\alpha)\%$. Fail to reject the null hypothesis whenever the confidence interval covers the hypothesized parameter.

Table 6.1 Summary of confidence intervals for one population parameter

Parameter	Statistics	Distribution	Variance	Confidence Interval
μ	$\hat{\mu}$	Normal	Known	$\hat{\mu} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
			Unknown	$\hat{\mu} \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$
			Known	$\hat{\mu} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Parameter	Statistics	Distribution	Variance	Confidence Interval
		Unknown	Unknown	$\hat{\mu} \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$
π	$\hat{\pi}$	Normal or Unknown	Always Unknown	$\hat{\pi} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
σ^2	$\widehat{\sigma^2}$	Normal	Always Unknown	$\frac{(n-1)\sigma^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)\sigma^2}{\chi^2_{1-\alpha/2}}$

Table 6.2 Confidence intervals for two samples

Parameter	Statistic	Status of Variances	Confidence Interval
$\mu_1-\mu_2$	$\widehat{\mu_1}-\widehat{\mu_2}$	Known and Unequal	$\widehat{\mu_1}-\widehat{\mu_2} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_1}}$
		Known and Equal	$(\widehat{\mu_1}-\widehat{\mu_2}) \pm Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n_2}+\frac{1}{n_2}\right)}$
		Unknown and Unequal	$(\widehat{\mu_1}-\widehat{\mu_2}) \pm t_{\alpha/2}\sqrt{\frac{\widehat{\sigma_1^2}}{n_1}+\frac{\widehat{\sigma_2^2}}{n_1}}$
		Unknown and Equal	$(\widehat{\mu_1}-\widehat{\mu_2}) \pm t_{\alpha/2}\sqrt{\widehat{\sigma^2_{Pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
$\pi_1-\pi_2$	$\widehat{\pi_1}-\widehat{\pi_2}$	Always Unknown	$(\widehat{\pi_1}-\widehat{\pi_2}) \pm Z_{\alpha/2}\sqrt{\frac{\widehat{\pi_1}(1-\widehat{\pi_1})}{n_1}+\frac{\widehat{\pi_2}(1-\widehat{\pi_2})}{n_2}}$
$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}$	Always Unknown	$\frac{\widehat{\sigma_1^2}/\widehat{\sigma_2^2}}{F_{1-\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\widehat{\sigma_1^2}/\widehat{\sigma_2^2}}{F_{\alpha/2}}$

CHAPTER 7

Statistical Inference with Test of Hypothesis

Hypothesis

A hypothesis is a claim about a population parameter. Hypothesis testing allows researchers to draw inferences about population parameters using sample statistics. The process of statistical inference involves a decision about the plausibility of a hypothesis.

Definition 7.1 The *null hypothesis* reflects the status quo, how things have been, or are currently. A null hypothesis can represent a historical fact or deliberate setting of a mechanical device. The null hypothesis is depicted by the symbol $\mathbf{H_0}$, pronounced h-sub-zero.

Definition 7.2 The *alternative hypothesis* is the researcher's claim that there exists some difference or change from the null hypothesis. Although *hypothesis* and *claim* are used interchangeably in conversation, in statistics the word claim is used for the alternative hypothesis. The alternative hypothesis is depicted by the symbol $\mathbf{H_1}$, pronounced *h-sub-one*.

Statistical Inference

Utilizing the distributional properties of sample statistics, the likelihood of an occurrence of an outcome is calculated given that the null hypothesis is true. If the outcome is unlikely the null hypothesis is refuted. When the observed statistic is reasonably close to the hypothesized value nothing unexpected has occurred and any (minor) difference is attributed to *random error*. If the observed statistic is substantially different from the hypothesized value, either the null hypothesis is true and an unusual

event with very low probability has occurred, or the null hypothesis is false since the more likely outcomes would appear in a random sample. *Statistical inference* consists of accepting outcomes with high probability and rejecting outcomes with low probability.

Definition 7.3 A *statistical hypothesis* is an assertion about the distribution of one or more random variables.

Definition 7.4 When a hypothesis completely specifies the distribution, it is called a *simple* statistical hypothesis; otherwise, it is called a *composite* statistical hypothesis. In this text, we deal with the simple hypothesis exclusively. The format for a simple hypothesis is

$$H_0\text{: A parameter} = \text{A constant} \tag{7.1}$$

Null Hypothesis

The null hypothesis reflects the *status quo*. It is about how things have been or are currently. For example, the average life of a car is 7 years. The null hypothesis can be a statement about the nature of something; the height of an average man is 5 ft 10 in. The null hypothesis might be the deliberate setting of equipment, such as that a soda-dispensing machine deposits 12 ounces of liquid in a can. It is important to realize that the researcher is not making any claims; these are common knowledge for which there is a consensus. A hypothesis can be about any parameter of a distribution function such as 54% of adults are Democrats; or the variance for weekly sales is 50. The following represents the above (null) hypotheses. Notice that the stated null hypotheses are *simple hypotheses*.

Single Mean	Single Proportion
H_0: $\mu = 7$	H_0: $\pi = 0.54$
H_0: μ = 5'10"	**Single Variance**
H_0: $\mu = 12$	H_0: $\sigma^2 = 50$

In hypothesis testing, the expected value of the outcome of an experiment is the hypothesized value. The hypothesized value reflects the status quo and will prevail until, through a process of inference, we

gather enough evidence to reject it. The observed statistic is not a hypothesis. The null hypothesis for a variance, such as the weekly sales null above, must be nonnegative.

A hypothesis is tested only when there is a justification to question the status quo. For example, your soda can is finished too quickly and you are still thirsty. There is a possibility that the can was under-filled, a justification for testing to see whether the cans actually have 12 ounces of soda, on average. It is tempting to state that the manufacturer is making the "claim" that the can contains 12 oz.; however, their statement is of an assertion or a promise and not a claim. As we will see shortly, the alternative hypothesis is the claim of the researcher, which is also known as the research question.

Null Hypothesis for the Equality of Two Parameters

A hypothesis can be used to test the equality of two parameters as well. Comparable things should be equal. For example, average productivity of a man and a woman would be assumed to be the same until proven otherwise. Let θ_1 and θ_2 (pronounced theta) be two parameters from two populations. We want to test whether they are the same. The null hypothesis should *not* be written as

$$H_0: \theta_1 = \theta_2 \tag{7.2}$$

This hypothesis is setting one parameter equal to the other, which makes it a composite hypothesis. Using algebra, the hypothesis can be modified to convert it to a simple hypothesis. There are two possible modifications. Equation (7.2) can be written in the following two forms.

$$H_0: \theta_1 - \theta_2 = 0 \tag{7.3}$$

$$H_0: \frac{\theta_1}{\theta_2} = 1 \tag{7.4}$$

Chapter 5 provided distributional properties for comparing two parameters. A brief summary is presented here. The distribution function for the *difference of two means* of random variables, each with a normal distribution, is also normal. Therefore, the normal distribution

should be used for testing the equality of two means: $H_0: \mu_1 - \mu_2 = 0$. When the variance is unknown and sample size is small, a *t* distribution should be used.

The distribution function for the *difference of two proportions* of random variables, each with a normal distribution, is also normal. Therefore, the normal distribution should be used for testing the equality of two proportions: $H_0: \pi_1 = \pi_2$. This hypothesis too should be modified to resemble a simple hypothesis as $H_0: \pi_1 - \pi_2 = 0$. The distribution function for the ratios of two variances of random variables, each with a chi-squared distribution, is an *F* distribution. Therefore, the *F* distribution should be used for testing the equality of two variances: $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$.

Alternative Hypothesis

The alternative hypothesis is the *claim* a researcher has against the null hypothesis, designated by $\mathbf{H_1}$. It is the research question or the main purpose of the research. A test, or alternative, hypothesis is used when the plausibility of the null hypothesis is in doubt.

How to Determine the Alternative Hypothesis

The claim of the research, that is, the research question, determines the alternative hypothesis. Every alternative hypothesis is a claim that the null hypothesis has changed. When the claim is that the value of the parameter in the null hypothesis has declined, the appropriate sign is the "less than" sign (<). Focus on the meaning and not the wording. When the claim is that the value of the parameter in the null hypothesis has increased, then the appropriate sign is the "greater than" sign (>). These two alternative hypotheses are known as *one-tailed hypotheses*. When the claim is not specific or is indeterminate, meaning we are unsure if the value of the parameter has decreased or increased, then the appropriate sign is the "not equal sign" (≠). This alternative is known as a *two-tailed hypothesis*. None of the three alternative cases include the equal sign (=), because the equal sign is used in the null hypothesis.

The formation of the null and the alternative hypotheses are the main problem of the novice. Remember the following:

- A simple null hypothesis is always of the form:

  H_0: A parameter = A constant

- An alternative hypothesis can take on a number of forms:

  The claim might be that the parameter is greater than (>), less than (<), or not equal to (≠) to a constant, which reflects the claim that the parameter has increased (>), decreased (<), or it simply has changed (≠). Only one of the alternatives listed below is used in a given test.

  H_1: A parameter < A constant

  H_1: A parameter > A constant

  H_1: A parameter ≠ A constant

Examples of an Alternative Hypothesis for a Single Mean

Possible claims against the above examples of null hypothesis are used for demonstration. A consumer advocacy group claims that car manufacturers are cutting corners to improve profitability and making inferior cars that do not last as long. The claim is that the average life of a car is less than the historically established 7 years.

$$H_0: \mu = 7$$

$$H_1: \mu < 7$$

Men are getting taller because of better nutrition and more exercise.

$$H_0: \mu = 5'10''$$

$$H_0: \mu > 5'10''$$

The quality manager claims the soda dispensing machine is out of calibration. The expected calibrated value is 12.

$$H_0: \mu = 12$$

$$H_1: \mu \neq 12$$

A political science researcher believes that, due to economic globalization and political turmoil around the world, the percent of Democrats has declined. She has information from a previous study that the percentage of Democrats was 54%.

$$H_0: \pi = 0.54$$

$$H_1: \pi < 0.54$$

Increased promotional advertising by a firm and its competitors has increased the variance of weekly sales.

$$H_0: \sigma^2 = 50$$

$$H_1: \sigma^2 > 50$$

The alternative hypothesis is a claim against the status quo. If there is no claim, there is no alternative hypothesis and, hence, no need for a test. The nature of the claim determines the sign of the alternative. The sign of the alternative hypothesis depends on the claim and nothing else. Table 7.1 provides a summary of null and alternative hypotheses for testing one mean, proportion, and variance as well as the hypotheses for testing the equality of two means, proportions, and variances.

Table 7.1 Summary of null and alternative hypotheses

Case	Null Hypothesis	Alternative Hypothesis	Comments
Single Mean	μ = a constant	μ > a constant μ < a constant $\mu \neq$ a constant	The claim determines the sign of the alternative hypothesis
Single Proportion	π = a constant	π > a constant π < a constant $\pi \neq$ a constant	The claim determines the sign of the alternative hypothesis
Single Variance	σ^2 = a constant	σ^2 > a constant σ^2 < a constant $\sigma^2 \neq$ a constant	The claim determines the sign of the alternative hypothesis
Two Means	$\mu_1 - \mu_2$ = a constant	$\mu_1 - \mu_2$ > a constant $\mu_1 - \mu_2$ < a constant $\mu_1 - \mu_2 \neq$ a constant	Use to test the equality of two means

(*Continued*)

Case	Null Hypothesis	Alternative Hypothesis	Comments
Two Proportions	$\pi_1 - \pi_2 =$ a constant	$\pi_1 - \pi_2 >$ a constant $\pi_1 - \pi_2 <$ a constant $\pi_1 - \pi_2 \neq$ a constant	Use to test the equality of two proportions
Two Variances	$\frac{\sigma_1^2}{\sigma_2^2} =$ a constant	$\frac{\sigma_1^2}{\sigma_2^2} >$ a constant	Use to test the equality of variances. Usually, no other alternative is tested.

Test Statistics

In Chapter 5, we saw that when the statistic is a sample mean $(\widehat{\mu})$, a sample proportion $(\widehat{\pi})$, a difference of two sample means $(\widehat{\mu}_1 - \widehat{\mu}_2)$, or a difference of two sample proportions $(\widehat{\pi}_1 - \widehat{\pi}_2)$, the Central Limit Theorem asserts that each of these sample statistics has a normal distribution. The sample variance $(\widehat{\sigma^2})$ has a chi-squared distribution, whereas the ratio of two sample variances $\left(\frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}\right)$ has an F distribution. Usually, cases involving two statistics test their equality. The test statistic for a single mean (μ), two means $(\mu_1 - \mu_2)$, single proportion (π), and two proportions $(\pi_1 - \pi_2)$ is provided by

$$\text{Test Statistic} = \frac{\text{Observed} - \text{Expected}}{\text{Standard Deviation of Observed}} \tag{7.5a}$$

In the context of test of hypothesis regarding a parameter it is better to express Equation (7.5) as

$$\textit{Test}\ \text{Statistic} = \frac{\textit{Estimated} - \textit{Hypothesized}}{\text{Standard Error}} \tag{7.5b}$$

where *observed* is the sample statistic, *expected* is the value of the null hypothesis, and standard deviation of the observed value is the standard error. The Central Limit Theorem provides the distribution function and the standard error. The sampling distribution of sample statistics covered in Chapter 5 provides a summary of parameters, statistics, and sampling variances of one and two populations. Whether the correct statistic for this hypothesis is a Z-test or t-test depends on whether the

population variance is known and the sample size. Use Z-test when the population variance is known or when it is unknown and the sample size is large. Z and t statistics are used to test hypotheses about one mean, one proportion, two means, or two proportions. In the case of two means or two proportions, the hypotheses must be modified to resemble a simple hypothesis.

The test statistic for a single variance (σ^2) is given by

$$\chi^2 = \frac{(n-1)\widehat{\sigma^2}}{\sigma_0^2} \tag{7.6}$$

Subscript zero represents the hypothesized null value, which is a constant.

The test statistic for equality of two variances $\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$ is given by

$$F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}} \tag{7.7}$$

Table 7.2 summarizes the relevant materials from Chapters 5, 6, and 7.

Table 7.2 Test statistics for testing hypotheses

Case	Null Hypothesis	Variance	Test Statistics
Single Mean	μ	Known	$Z = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
		Unknown	$t = \frac{\hat{\mu} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$
Single Proportion	π	Unknown	$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
Single Variance	σ^2	Unknown	$\chi = \frac{(n-1)\widehat{\sigma^2}}{\sigma_0^2}$

Case	Null Hypothesis	Variance	Test Statistics
Two Means.	$\mu_1 - \mu_2$	Known and Unequal	$Z = \frac{\left(\widehat{\mu_1} - \widehat{\mu_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
		Known and Equal	$Z = \frac{(\widehat{\mu_1} - \widehat{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
		Unknown and Unequal	$t = \frac{(\widehat{\mu_1} - \widehat{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}}}$
		Unknown and Equal	$t = \frac{(\widehat{\mu_1} - \widehat{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\widehat{\sigma_{\text{Pooled}}^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
Two Proportions	$\pi_1 - \pi_2$	Unknown	$Z = \frac{(\widehat{\pi_1} - \widehat{\pi_2}) - (\pi_1 - \pi_2)}{\sqrt{\frac{\widehat{\pi_1}(1-\widehat{\pi_1})}{n_1} + \frac{\widehat{\pi_2}(1-\widehat{\pi_2})}{n_2}}}$
Two Variances	$\frac{\sigma_1^2}{\sigma_2^2}$	Unknown	$F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}$

All the null hypotheses are set equal to a constant. In the case of equality of two means and two proportions, the constant is *zero*. In the case of equality of two variances, the constant is *one*. Subscript zero represents the hypothesized null value, which is a constant.

Probability of Event Occurrence

Any event that has a probability of occurrence will occur sometime. Some events have a higher probability of occurrence than others, so they will occur more often. The essence of statistical inference is that events with high probability of occurrence are assumed to occur while events with a low probability of occurrence are assumed not to occur. Every event has a *complement* event, which completes and exhausts all possible outcomes of an event. The complement of seeing a "head" in a coin toss is "not seeing a head" or "seeing a tail." The complement of having an

economic recession is "not having an economic recession," which is not necessarily "having an economic boom." The sum of the probabilities of an event and its complement is *one* (1). Unless the probability of an event is exactly 50% for one of the two possible outcomes, the event or its complement must have a higher probability of occurrence than the other.

The probability of having an accident while driving through an intersection when the traffic light is green is lower than when the light is red. Therefore, it is assumed that accidents do not occur when crossing an intersection while the traffic light is green. Note that there is still a chance of having an accident when crossing during a green light. It is also possible to go through a red light without having an accident. When an event has low probability we "assume" it will not occur. This example has a special twist to it. For every car involved in an accident while crossing an intersection when the light is green, there is another car that ran the red light. The probability of an accident for the latter is high while the probability for the former is low. Of the few people who run a red light most end up in an accident, while the majority of people who drive through a green light do not have an accident. Although the number of people who have an accident going through a green light is equal to the number of people that have an accident running a red light, nevertheless, the corresponding probabilities are very different. The null and alternative hypotheses are formed in the following way:

H_0: Driving through a red traffic light does *not* cause an accident
H_1: Driving through a red traffic light does cause an accident

Types of Error

The process of testing a hypothesis is similar to convicting a criminal. *The null hypothesis* is a conjecture to the effect that everybody is assumed to be innocent unless proven otherwise. If there is any reason to doubt this innocence, a *claim* is made against the *null hypothesis*, which is called an alternative hypothesis. Within this domain the evidence is collected, which is the same as taking a *sample*. The type of crime is decided, as indicated by charges of misdemeanor, felony, etc., which is similar to a

test statistic. Finally, based on the evidence a judgment is rendered, either innocent or guilty, which is the *inference*. If the prosecutor fails to provide evidence of guilt, it does not mean the accused is innocent. The degree or the probability that the person was innocent (but was convicted) is the probability of *type I error* or the *p value.*

The null hypothesis of innocence is rejected if the probability of being innocent is low in light of the evidence. Otherwise, we *fail to reject* the null hypothesis. It is possible that the null hypothesis is false, a sample statistic with a low probability was observed, and we erroneously fail to reject the null hypothesis (i.e., the person was guilty but found not guilty). This kind of error is known as *type II error*. Note that the jury's verdict is either "guilty" or "not guilty" and never "innocent." Similarly, in inferential statistics we either "reject the null hypothesis" or "fail to reject the null hypothesis" but never "accept the null hypothesis" or "accept the alternative hypothesis" (Table 7.3).

Definition 7.5 *Type I Error* occurs when the null hypothesis is true but it is rejected.

Definition 7.6 *Type II error* occurs when the null hypothesis is false but is not rejected.

Table 7.3 Summary of types of error in inference

	H_0 is Rejected	H_0 is *Not* Rejected
H_0 is True	Type I Error	No Error
H_0 is False	No Error	Type II Error

It is not possible to commit a type I error if the null hypothesis is not rejected. It is not possible to commit a type II error if the null hypothesis is rejected. There is also a *type III error*, which is defined below.

Definition 7.7 *Type III error* is rejecting a null hypothesis in favor of an alternative hypothesis with the wrong sign.

As an example of type III error, let's return to an example from above about the life of cars. If you recall, we had the following null and alternative hypotheses:

$$H_0: \mu = 7$$

$$H1: \mu < 7$$

A type III error would have occurred if a researcher rejected the null hypothesis that cars last 7 years in favor of the alternative hypothesis that they last fewer than 7 years, when in fact they are actually lasting longer than 7 years. Because the alternative hypothesis claimed that cars were lasting fewer than 7 years, the null cannot be rejected in favor of the alternative hypothesis.

Statistical Inference with the Method of p Value

There are two approaches for ascertaining an inference, the method of p value and the method of critical region. The two approaches are similar. In both, the observed values of the sample statistics, such as the sample mean or proportion, are standardized as Z or t statistics. When the null hypothesis is true, the observed statistic should be *close* to the hypothesized value, which means the corresponding Z or t statistic should be *close to zero*. This is because the numerator of Z or t statistics is the difference between the sample statistic and the hypothesized population parameter. Z or t becomes larger as the calculated statistic increases, which indicates the observed statistic is distinct from the hypothesized parameter value. Consequently, the area under the corresponding distribution function between the center and the Z or t value, which represents the probability, becomes larger and at the same time the area under curve further away from Z or t value becomes smaller. This probability, corresponding to the *area under the tail* section, reflects the probability of observing a *more extreme value* than the observed statistic. The smaller this probability is, the less likely the null hypothesis is correct. This probability is actually the probability of committing a type I error if the null hypothesis is rejected.

Definition 7.8 The value representing the probability of the area under the tail end of the distribution is called the *p value*. It is also called the *Observed Significance Level (OSL)*.

Rule 7.1 Reject the null hypothesis when the p value is *small enough.*

For a given sample size, reducing type I error increases type II error. In order to reduce both types of error, it is necessary to increase the sample size.

When the null hypothesis is true, the significance level indicates the probability or likelihood that the observed results could have happened by chance. When the null hypothesis is true, the observed results should have high probability. Consequently, when the p value is "large," there is no reason to doubt the null hypothesis. However, if the observed outcomes happen to have low probability, it casts doubt about the plausibility of the null hypothesis. When the outcome contradicts the null hypotheses, it implies that the null hypothesis is less likely to be true by virtue of observing the outcome obtained from the sample. In other words, the p value is the probability of "seeing what you saw," which is reflected in the other common name for p value, OSL.

Statistical Inference with the Method of Critical Region

An alternative approach to the p value decision rule is to calculate a *critical value* from an appropriate distribution function based on a preselected level of type I error, customarily 1%, 5%, or 10%, and compare the test statistic with it. The Z or t corresponding to the selected type I error is obtained from a table or software program, as shown in Chapter 6. For example, the critical value for a two-tailed test at 5% is ± 1.96. Reject the null hypothesis when the calculated statistic is more extreme than ± 1.96 (i.e., when the critical value is either greater than +1.96 or less than −1.96).

Rule 7.2 Reject the null hypothesis when the test statistic is more extreme than the critical value.

Either method will yield the same conclusion for a given level of type I error. The method of *p value* is preferred because it gives the exact probability of making a type I error. In contrast, in the method of critical region, the probability of type I error is never exact, except by rare chance. Another advantage of the *p value* is that it allows the researcher to make a better informed decision.

Steps for a Test of Hypothesis

1. Determine the Scope of the Test
2. State the Null Hypothesis
3. Determine the Alternative Hypothesis
4. Determine a Suitable Test Statistic
5. Calculate the Test Statistic
6. Provide Inference

Statistical Inference Using Confidence Intervals

Confidence intervals are obtained using a point estimate, a margin of error, and the probability of type I error. Therefore, they have all the necessary components for drawing inferences. For example, a 95% confidence interval indicates that the range has 95% probability of covering the true parameter, which also means the probability of type I error is 5%.

Rule 7.3 *Reject* the null hypothesis when the confidence interval *does not* cover the hypothesized value. *Fail to reject* when the confidence interval *does* cover the hypothesized value.

A test of hypothesis using confidence intervals is identical to a two-tailed test. The approach is based on the critical region method.

Example 7.1 Determine whether the closing price of Microsoft stock exceeds $46.35. Use data from May 21 to June 11, 2015.

Solution 7.1

Based on the statement in the problem, the alternative hypothesis is

$$H_0: \mu = 46.35 \qquad H_1: \mu > 46.35$$

From previous examples, we have the following statistics.

$$\hat{\mu} = 46.71733 \qquad \widehat{\sigma^2} = 0.348735$$

Since the population variance is unknown, we need to use

$$t = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} = \frac{46.71733 - 46.35}{\sqrt{\frac{0.348735}{15}}} = \frac{0.36733}{\sqrt{0.023249}} = \frac{0.36733}{0.152476} = 2.409$$

Using the following Excel command, we obtain the exact *p value* for the right-hand tail.

$$= \text{t.dist.rt}(2.409,14) = 0.015$$

Figure 7.1 depicts the Stata command and output for performing the same *t* test.

```
. ttest msft1==46.35

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   msft1 |      15    46.71733    .1524762    .5905379    46.3903    47.04436
------------------------------------------------------------------------------
    mean = mean(msft1)                                            t =   2.4091
Ho: mean = 46.35                                 degrees of freedom =       14

  Ha: mean < 46.35               Ha: mean != 46.35                 Ha: mean > 46.35
 Pr(T < t) = 0.9848         Pr(|T| > |t|) = 0.0303          Pr(T > t) = 0.0152
```

Figure 7.1 Stata command and output for testing a mean

The probability of obtaining an average stock price of $46.71, if the true population average is $46.35, is 0.015. This is a low probability. It indicates that if the null hypothesis were to be rejected 100 times, only 1.5 times would the decision be incorrect and a type I error committed. Therefore, we reject the null hypothesis in favor of the alternative hypothesis. To avoid a fraction for the number of cases with type I error, the statement could be reworded as 15 in 1,000 rejections. Since the OSL is low enough, the null hypothesis is rejected.

Stata displays all three possible alternatives, including left-tailed, two-tailed, and right-tailed hypotheses, to choose from. The results for each alternative hypothesis can be seen along the bottom of the output.

Example 7.2 Test the claim that more than 50% of the stock prices for Microsoft close higher than $46.75. Use the sample from May 21 to June 11, 2015.

Solution 7.2 Sorting the data makes it easier to obtain the portion of sample prices over $46.75. Upon inspection, there are 8 closing prices higher than $46.75 during the sample period. Dividing the number of days (8) that the stock closed over $46.75 by the number of days in the

sample (15) gives the result that on 53% of the days the stock closed over $46.75.

$$\hat{\pi} = \frac{8}{15} = 0.53$$

While on the face of it this number appears to be higher than 50%, we want to know the likelihood that the result was due to chance. Based on the statement in the problem, the alternative hypothesis is

$$H_0 : \pi = 0.50 \qquad\qquad H_1 : \pi > 0.50$$

From Table 7.2, the correct formula is

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.53 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{15}}} = \frac{0.03}{\sqrt{0.01667}} = \frac{0.03}{0.13} = 0.23$$

The probability of the region more extreme than $Z = 0.23$ is given by

$$P(Z > 0.23) = 0.5 - P(0 < Z < 0.23) = 0.5 - 0.0910 = 0.4090$$

Since the probability of type I error would be high (at 0.409) if the null hypothesis is rejected, the decision should be to "fail to reject" the null hypothesis. We cannot confidently say that the result of 53% was not due to chance. The Stata output is presented in Figure 7.2 (*use gen higher = (msft1 > 46.75)* to create the "higher" variable).

```
. ttest higher==.5

One-sample t test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  higher |      15   .5333333   .1333333   .5163978   .2473618    .8193049
---------+--------------------------------------------------------------------
    mean = mean(higher)                                     t =   0.2500
Ho: mean = .5                                 degrees of freedom =       14

    Ha: mean < .5               Ha: mean != .5                 Ha: mean > .5
 Pr(T < t) = 0.5969         Pr(|T| > |t|) = 0.8062          Pr(T > t) = 0.4031
```

Figure 7.2 Stata command and output for testing a proportion

Example 7.3 Test the claim that the variance for the stock prices of Microsoft is greater than 0.30. Use the sample from May 21 to June 11, 2015.

Solution 7.3 Based on the statement in the problem, the alternative hypothesis is

$$H_0 : \sigma^2 = 0.30 \qquad H_1 : \sigma^2 > 0.30$$

From Table 7.2, the appropriate formula is

$$\chi^2 = \frac{(n-1)\widehat{\sigma^2}}{\sigma^2} = \frac{(15-1)(0.348735)}{0.3} = \frac{4.88229}{0.3} = 16.2743$$

Using the following Excel command, we obtain the *p value* for the right-hand side probability. The sign of the alternative hypothesis determines which tail should be used. An alternative hypothesis of "greater than" will use the right-hand tail, while "less than" will use the left-hand tail.

= chisq.dist.rt(16.2743,14) = 0.2969

The probability of committing a type I error is almost 30%. Since the probability of type I error is too high, we fail to reject the null hypothesis that the variance is greater than 0.30. Stata uses the standard deviation in its computation. First find the square root of the null hypothesis, in this case 0.5477, and then use it for comparison. The sign "= =" in Stata is a logical command and means "if equal to." Check that the probability associated with a chi-squared statistic of 16.2743 matches the result of direct computation above (the probability for a right-tailed test can be found at the bottom right-hand corner of the table in Figure 7.3).

```
. display .3^.5
.54772256

. sdtest msft1==.54772256

One-sample test of variance
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
msft1	15	46.71733	.1524762	.5905379	46.3903	47.04436

```
    sd = sd(msft1)                                 c = chi2 =  16.2743
Ho: sd = .54772                 degrees of freedom =       14

   Ha: sd < .54772             Ha: sd != .54772                 Ha: sd > .54772
 Pr(C < c) = 0.7031         2*Pr(C > c) = 0.5938            Pr(C > c) = 0.2969
```

Figure 7.3 Stata command and output for testing a variance

Example 7.4 Are the means for Microsoft stock prices for periods May 21 to June 11 and June 12 to July 2, 2015 the same?

Solution 7.4

The objective is to determine whether $\mu_1 = \mu_2$. Since this format is not of the form "a parameter equal to a constant," we rewrite the hypothesis as

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0$$

Since no particular directional claim has been made, the test is a two-tailed test. The following information is available.

$$\widehat{\mu_1} = 46.72 \qquad \widehat{\sigma_1^2} = 0.348735$$

$$\widehat{\mu_2} = 45.48 \qquad \widehat{\sigma_2^2} = 0.614498$$

Since we do not know whether the variances are equal, we will test their equality first, that is, $\sigma_1^2 = \sigma_2^2$. Since this is in the form of "a parameter equal to another parameter," it has to be modified to resemble "a parameter = a constant format."

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

It is customary to express the alternative hypothesis for a ratio as "greater than one" rather than less than one. To assure the ratio of the sample variances is actually greater than one, always place the sample variance that is larger in the numerator.

From Table 7.2, the appropriate formula to test the equality of two variances is

$$F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}} = \frac{0.614498}{0.348735} = 1.762$$

The *p* value for this statistic is obtained from the following Excel command:

=f.dist.rt(1.762,14,14) = 0.1505

The Stata test for equality of two variances is displayed in Figure 7.4.

```
. sdtest msft1==msft2

Variance ratio test

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   msft1 |      15    46.71733    .1524762    .5905379    46.3903    47.04436
   msft2 |      15    45.47533    .2024019    .7838992   45.04122    45.90944
---------+--------------------------------------------------------------------
combined |      30    46.09633    .1697005    .9294879   45.74926    46.44341
------------------------------------------------------------------------------
    ratio = sd(msft1) / sd(msft2)                                 f =   0.5675
Ho: ratio = 1                                     degrees of freedom =   14, 14

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
 Pr(F < f) = 0.1505         2*Pr(F < f) = 0.3010         Pr(F > f) = 0.8495
```

Figure 7.4 Stata command and output for testing two variances

Note that the reported probabilities are cumulative for all values up to the calculated statistic. Since the probability of type I error is not low enough, we fail to reject the null hypothesis that the two variances are equal. Therefore, from Table 7.2 the following test statistic is used for testing the equality of the mean prices for the two periods when the variances are equal and should be pooled.

$$t = \frac{(\widehat{\mu_1} - \widehat{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2_{Pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(45.47533 - 46.71733) - (0)}{\sqrt{0.1916\left(\frac{1}{15} + \frac{1}{15}\right)}}$$

$$= \frac{-1.242}{\sqrt{0.0255}} = \frac{-1.242}{0.1598} = -7.77$$

Since the alternative hypothesis is two-tailed, we use the following Excel command to obtain the exact *p value*:

= t.dist.2t(7.77,28) = 1.83012E-08

The Stata command is displayed below. Note the "Paired *t* test" at the top left-hand corner, which indicates the use of pooled variances. Since the test is two-tailed, the relevant results are listed under H_a: mean(diff)! = 0 in the center of the last line (Figure 7.5).

Since the *p value* is low enough, we reject the null hypothesis that the average prices of Microsoft stock are the same for the periods of May 21 to June 11 and June 12 to July 2, 2015.

```
. ttest msft1== msft2

Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
msft1	15	46.71733	.1524762	.5905379	46.3903	47.04436
msft2	15	45.47533	.2024019	.7838992	45.04122	45.90944
diff	15	1.242	.120468	.4665707	.9836213	1.500378

```
    mean(diff) = mean(msft1 - msft2)                          t =  10.3098
Ho: mean(diff) = 0                         degrees of freedom =       14

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Figure 7.5 Stata command and output for testing two means

CHAPTER 8

An Introduction to Regression Analysis

Until this point, with the exception of the covariance and correlation coefficient, the focus has been on a single variable. Most economic concepts are complex and studying them requires considering other economic, social, cultural, and political factors. Even the simplest economic concepts such as quantity demanded have at least two variables, a quantity and a price. To estimate or forecast income, it is necessary to decide whether the orientation of the study is macro- or microeconomics. At the microeconomic level, income depends on acquired human capital, natural ability, talent, work ethic, exertion, years of experience, and seniority, to name a few. Some seemingly unrelated factors such as race and gender also affect one's income. At the macroeconomic level, the determinants of income, which in this case should be referred to as the national income, are functions of national productivity, resources, population size, overall level of education, economic cycles, seasonal cycles, and other factors.

There are many powerful tools in statistical analysis that permit simultaneously analyzing the factors that determine a phenomenon such as income. One such tool is regression analysis. Regression analysis gives a general idea of how one or more variables are associated with another variable. For example, if we did a regression analysis of income using the variables mentioned in the above paragraph, we would find that, on average, an additional year of education leads to higher income. Regression also helps us know if the association between education and income is purely due to chance and if that association is positive or negative. For education, we would expect to see a positive association between education and income.

In addition, regression provides an indication of how strong the association is between variables. As we saw in Chapter 4, those with bachelor's

degrees have a higher net worth than those with only a high school diploma. Regression results provide specific details as to how much more, on average, each additional year of schooling is worth.

Regression methodology acknowledges that most real life occurrences are subject to random error. Per capita income is the expected income of a person selected at random after all the contributing factors are considered. However, the income is not necessarily the same for all people with identical contributing factors such as those listed above, for example, education and seniority. In addition to natural deviations in outcomes, sometimes some factors are not accounted for either because they are not known, no valid measurement is available, or the data are not collected. Regression methodology minimizes these errors to obtain a linear model that better approximates the reality.

Recall that a basic definition of error in statistics is whatever that cannot be explained. A correctly specified regression model explains part of the unexplained error using explanatory variables such as those listed in the first paragraph. Note that errors are deviations of observations from the *expected value*. In descriptive statistics, the expected value is simply the mean. In regression analysis, the *regression line* is the expected value. Consequently, models with more explanatory power will have smaller errors and a better fit to the data. In statistics it makes more sense to focus on averages rather than individual outcomes, which are subject to variance. Averaging individual errors removes the random error, which by definition has an expected value of zero. To overcome this problem individual errors are squared, as discussed in Chapter 2 when the concept of variance was introduced. Analogously, regression methodology minimizes the squares of individual error, which explain the customary name of *least squares* for the method. A more detailed explanation of this process is available in *Regression for Economics*.[1]

Explanatory Variables

Explanatory variables are factors that are not affected by the model but influence the dependent or response variable. The dependent variable or response variable is the phenomenon of interest that is claimed to be affected by explanatory variable(s). As the number of explanatory variables

increases, the explanatory power of the regression model should increase. This statement is valid only if the added variables are contributing factors and not correlated with each other. The most reliable estimates are obtained by including the correct explanatory variables, also called the *independent variables*. Independent variables to be included in the model are identified by researchers through the use of theories in economic and related fields of study as well as experience. Although there is no intellectual reason that justifies or even explains lower incomes of females or minorities, studies have established discernable differences in income attributable to gender and race. Variables that affect the response variable but are not based on theory are known as *control variables*. In economics, the control variables are accounted for under *ceteris paribus*, which means "other things equal."

The simple regression model consists of one dependent variable and one independent variable plus an error term.

$$Income = \beta_0 + \beta_1\, Education + \varepsilon \tag{8.1}$$

where *Income* is the dependent variable, *Education* is the independent variable, β_0 is the intercept, β_1 is the slope, and ε is the error term.

The independent variables are exogenous to the model and are not to be explained in any manner. The model in Equation (8.1) does not provide any input on why or how people decide on their level of education, while income is explained, in a linear fashion, by levels of education. The latter is the response variable, while the former is the determining factor. The Greek letters β_0 and β_1 are the parameters of the model. They are also called the intercept and the slope, respectively. The interpretation of β_1 is that for every unit change in education, income will change by the magnitude of β_1 and in the direction of its sign. The intercept β_0 provides an estimate of income when education level is zero. Finally, ε, the error term, accounts for everything else that affects income, other than education, plus the random error.

This simplistic model explains income with one variable. The hypothesized claim for the slope of the regression line, β_1, is that it is expected to be positive. This claim is based on theory and common sense. One expects higher income with more education, since education is a kind of investment called human capital. An educated person is more knowledgeable

and hence, more productive, and thus, deserves higher income per unit of time than someone with less education, other things equal.

An advantage of regression analysis is that in addition to estimating the magnitude of the effect of each explanatory variable on the response variable, it also provides a test of hypothesis about its statistical significance. A typical inference about a regression model consists of two different tests. The one that tests the overall significance of the model is based on the F test. In this case, the amount of the variation in the dependent variable that is explained by the model is compared with the amount that still remains unexplained. The portion that is explained by the model is called *mean squared model* (MSM). Recall that the sum of squares of the portions not explained, divided by the appropriate degrees of freedom, is the same as the variance, which in the jargon of regression analysis is called *mean squared error* (MSE).

Like any other variance, mean squared regression also has a chi-squared distribution. Theorem 4.5 from Chapter 4 states that the ratio of two chi-squared distribution functions follows an F distribution. Therefore, a test of the relative magnitude of the portion of the variation in the dependent variable that is explained by the model to the portion that is not is done using an F statistic. The null and alternative hypotheses for the model are

$$H_0: \text{Model is } \textit{not} \text{ good} \qquad H_1: \text{Model is good}$$

Inference is the same: reject the null hypothesis when p value is small enough; fail to reject it otherwise.

Once the null hypothesis that the model is not good is rejected, the slopes of individual variables are tested for significance using t statistics. The customary null hypothesis is that the slope of the independent variable is zero. A slope equal to zero indicates that the corresponding variable does not have any explanatory power.

$$H_0: \beta_{Education} = 0 \qquad H_1: \beta_{Education} > 0$$

The appropriate test statistic for this hypothesis is a t statistic. Testing procedure here too is the same as usual. Reject the null hypothesis if the corresponding p value is low enough. All software designed to perform statistical analysis can perform regression analysis with a relatively easy set

of commands and/or procedures. In fact, many of the commercially available software programs are menu-driven similar to a typical application software. Most have reasonably good help features that will show the necessary steps or commands. Even Excel, a spreadsheet software program, has a menu-driven procedure to perform regression analysis, to provide test statistics for testing the model and slopes, and to provide estimates of slopes and the explanatory power of the model. For more detail, refer to *Regression for Economics*.[2]

Example 8.1 Test the hypothesis that education increases income.

Solution 8.1

As pointed out earlier, there are numerous measures of income, from per capita personal income to national income; the former represents a microeconomics aspect of income, while the latter is a macroeconomics perspective. For this example, we use the national income that is obtained from http://www.bea.gov/histdata/Releases/Regional/2010/PI/state/preliminary_March-23-2011/SA1-3.csv. You will need to use the first line of data, Personal Income, for the years 1970 to 2010. The data on education, which are in 1,000s, are obtained from Table A-1 at http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html (look at the right of the page for a link to the data in Excel). A copy of retrieved data is provided in Table 8.1.

First, we regress income on the total number of people in the United States with education, regardless of the level of education; the variable is named "Education." There is not enough space to discuss and explain all the numbers that are generated; for a full explanation, see *Regression for Economics*.[2] Figures 8.1A and 8.1B display Excel and Stata outputs, respectively. We will focus on the row named "Education." Data for education are in the thousands of people and income is in billions of dollars. The coefficient for the independent variable is 0.13647217353. This indicates that for every one unit increase in the number of educated people, that is, 1,000 additional people, national income increases by $0.13647217353 billion or $136,472,173.53. Equivalently, for one more educated person, national income increases by $136,472.17.

Table 8.1 Data on education and income, 1970–2010, United States

Year	Population (in Thousands)[1]	Years of School Completed (in Thousands)[1]						Income (in Billions)[2]
		Elementary		High school		College		
		0–4 yr	5–8 yr	1–3 yr	4 yr	1–3 yr	4 yr or More	
1970	109,310	5,747	24,519	18,682	37,134	11,164	12,062	864.6
1971	110,627	5,574	24,029	18,601	38,029	11,782	12,612	932.1
1972	111,133	5,124	22,503	18,855	39,171	12,117	13,364	1,023.6
1973	112,866	5,100	21,838	18,420	40,448	12,831	14,228	1,138.5
1974	115,005	5,106	21,200	18,274	41,460	13,665	15,300	1,249.3
1975	116,897	4,912	20,633	18,237	42,353	14,518	16,244	1,366.9
1976	118,848	4,601	19,912	18,204	43,157	15,477	17,496	1,498.1
1977	120,870	4,509	19,567	18,318	43,602	16,247	18,627	1,654.2
1978	123,019	4,445	19,309	18,175	44,381	17,379	19,332	1,859.5
1979	125,295	4,324	18,504	17,579	45,915	18,393	20,579	2,077.9
1980	130,409	4,390	18,426	18,086	47,934	19,379	22,193	2,316.8
1981	132,899	4,358	17,868	18,041	49,915	20,042	22,674	2,595.9
1982	135,526	4,119	17,232	18,006	51,426	20,692	24,050	2,778.8
1983	138,020	4,119	16,714	17,681	52,060	21,531	25,915	2,969.7

1984	140,794	3,884	16,258	17,433	54,073	22,281	26,862	3,281.3
1985	143,524	3,873	16,020	17,553	54,866	23,405	27,808	3,515.9
1986	146,606	3,894	15,672	17,484	56,338	24,729	28,489	3,725.1
1987	149,144	3,640	15,301	17,417	57,669	25,479	29,637	3,955.3
1988	151,635	3,714	14,550	17,847	58,940	25,799	30,787	4,275.3
1989	154,155	3,861	14,061	17,719	59,336	26,614	32,565	4,618.2
1990	156,538	3,833	13,758	17,461	60,119	28,075	33,291	4,904.5
1991	158,694	3,803	13,046	17,379	61,272	29,170	34,026	5,071.1
1992	160,827	3,449	11,989	17,672	57,860	35,520	34,337	5,410.8
1993	162,826	3,380	11,747	17,067	57,589	37,451	35,590	5,646.8
1994	164,512	3,156	11,359	16,925	56,515	40,014	36,544	5,934.7
1995	166,438	3,074	10,873	16,566	56,450	41,249	38,226	6,276.5
1996	168,323	3,027	10,595	17,102	56,559	41,372	39,668	6,661.9
1997	170,581	2,840	10,472	17,211	57,586	41,774	40,697	7,075.0
1998	172,211	2,834	9,948	16,776	58,174	42,506	41,973	7,587.7
1999	173,754	2,742	9,655	15,674	57,935	43,176	43,803	7,983.8
2000	175,230	2,742	9,438	15,674	58,086	44,445	44,845	8,632.8
2001	180,389	2,810	9,518	16,279	58,272	46,281	47,228	8,987.1
2002	182,142	2,902	9,668	16,378	58,456	46,042	48,696	9,149.5
2003	185,183	2,915	9,361	16,323	59,292	46,910	50,383	9,486.6

(*Continued*)

Table 8.1 (*Continued*)

Year	Population (in Thousands)[1]	Years of School Completed (in Thousands)[1]						Income (in Billions)[2]
		Elementary		High school		College		
		0~4 yr	5~8 yr	1~3 yr	4 yr	1~3 yr	4 yr or More	
2004	186,876	2,858	8,888	15,999	59,811	47,571	51,749	10,048.3
2005	189,367	2,983	8,935	16,099	60,893	48,076	52,381	10,609.3
2006	191,884	2,951	8,791	16,154	60,898	49,371	53,720	11,389.0
2007	194,318	2,830	8,462	16,451	61,490	49,243	55,842	11,994.9
2008	196,305	2,599	8,226	15,516	61,183	50,994	57,787	12,429.6
2009	198,285	2,785	8,043	15,587	61,626	51,670	58,574	12,087.5
2010	199,928	2,615	7,836	15,260	62,456	51,920	59,840	12,429.3
2011	201,543	2,589	7,688	14,763	61,911	53,249	61,343	13,202.0
2012	204,579	2,484	7,800	14,993	62,113	53,900	63,291	13,887.7
2013	206,899	2,344	7,578	14,595	61,704	55,173	65,506	14,166.9
2014	209,287	2,525	7,388	14,545	62,240	55,709	66,879	14,733.9

[1]Table A-1, Educational Attainment; 1947, and 1952 to 2002 March Current Population Survey, 2003 to 2014 Annual Social and Economic Supplement to the Current Population Survey (noninstitutionalized population, excluding members of the Armed Forces living in barracks); 1960 Census of Population, 1950 Census of Population, and 1940 Census of Population (resident population). http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html

[2]Table 2.1. Personal Income and Its Disposition; Bureau of Economic Analysis; Last Revised on: June 24, 2015; http://www.bea.gov/iTable/index_regional.cfm

Summary Output						
Regression Statistics						
Multiple *R*	0.979161895					
R Square	0.958758016					
Adjusted *R* Square	0.9577989					
Standard Error	882.8512227					
Observations	45					
ANOVA						
	df	SS	MS	*F*	*Significance F*	
Regression	1	7.79E+08	7.79E+08	999.6269	2.1E–31	
Residual	43	33515330	779426.3			
Total	44	8.13E+08				
	Coefficients	Standard Error	*t* Stat	*p* value	Lower 95%	Upper 95%
Intercept	–15364.55351	697.7345	–22.0206	4.84E–25	–16771.7	–13957.4
Education	0.13647217353	0.004316	31.61688	2.1E–31	0.127767	0.145177

Figure 8.1A Excel regression of income on total education, 1970–2010, United States

```
. regress income population
```

Source	SS	df	MS	Number of obs =	45
				F(1, 43) =	999.58
Model	779136355	1	779136355	Prob > F =	0.0000
Residual	33517099.3	43	779467.424	R-squared =	0.9588
				Adj R-squared =	0.9578
Total	812653454	44	18469396.7	Root MSE =	882.87

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
population	.1364723	.0043165	31.62	0.000	.1277671	.1451774
_cons	-15364.46	697.7529	-22.02	0.000	-16771.61	-13957.31

Figure 8.1B Stata regression of income on total education, 1970–2010, United States

To run a regression in Excel, use the following:

Data| Data Analysis| Regression

In the "Input Y Range:" enter the column coordinates of the dependent variable. In the "Input X Range:" enter the coordinates of the cells containing the independent variable (while there can only be one dependent variable, there can be as many independent variables as necessary). The range must be rectangular in shape and the number of rows must be the same as those assigned to the dependent variable. Click on "labels" if the first row of data contains variable names and press return for results.

The Stata command for regression is of the form as given below:

regress *varname varlist*

where *varname* is the dependent variable *income* and *varlist* can consist of several independent variables separated by space. Do not include any punctuation marks.

Next, let us regress income on the number of people with 4 or more years of college (Figures 8.2A and 8.2B).

Summary Output						
Regression Statistics						
Multiple *R*	0.994259501					
R Square	0.988551956					
Adjusted *R* Square	0.988285723					
Standard Error	465.139678					
Observations	45					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	803347508.7	8.03E+08	3713.1	2.23E–43	
Residual	43	9303261.563	216354.9			
Total	44	812650770.2				
	Coefficients	Standard Error	*t* Stat	*p* value	Lower 95%	Upper 95%
Intercept	–3299.379358	172.1136018	–19.1698	1.09E–22	–3646.48	–2952.28
≤4	0.26226168419	0.00430394	60.93521	2.23E–43	0.253582	0.270941

Figure 8.2A Excel regression of income on 4 or more years of college education, 1970–2010, United States

```
. regress Income Fouryears

      Source |       SS           df       MS      Number of obs   =        45
-------------+----------------------------------   F(1, 43)        =   3712.57
       Model |  803348840         1  803348840     Prob > F        =    0.0000
    Residual |  9304614.66       43  216386.387    R-squared       =    0.9886
-------------+----------------------------------   Adj R-squared   =    0.9883
       Total |  812653454        44  18469396.7    Root MSE        =    465.17

------------------------------------------------------------------------------
      Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Fouryears |   .2622619   .0043043    60.93   0.000     .2535815    .2709423
      _cons |  -3299.281   172.1261   -19.17   0.000    -3646.406   -2952.155
------------------------------------------------------------------------------

.
```

Figure 8.2B Stata regression of income on 4 or more years of college education, 1970–2010, United States

The coefficient for the variable representing people with 4 or more years of college is 0.26226168419. Therefore, every additional person with 4 or more years of college education increases the national income by $262,261.68. Higher levels of education increases national income.

One shortcoming of these examples is that they do not consider other factors that affect income, as discussed earlier. The simple regression analysis is easily extended to include all the variables that a researcher deems necessary. The main determining factor for including a variable in a regression model is the theory in the discipline in which the research is conducted. Literature on the economic impact of education suggests a model given in Equation (8.2).

$$Income = \beta_0 + \beta_1\, Education + \beta_2\, Experience + \beta_3\, Race + \beta_4\, Gender + \beta_5\, Determination + \varepsilon \quad (8.2)$$

The list of variables need not be exhaustive. In general, the explanatory power of a model, as measured by R^2 (pronounced r-squared), increases as the number of variables increases. However, the possibility of including irrelevant variables also increases. Because numerous factors can inflate R^2, you should choose variables based on theory and not because of the value of R^2.[2] One reason is that variables in social sciences are often somewhat correlated. For example, education is not really independent of race or gender. Although there is no biological reason for education to be influenced by race or gender, the reality of the United States is that it is.

Similarly, there is no reason to include race and gender as control variables in the model explaining the determinants of income. The second reason for limiting the number of variables is that usually a few important variables are sufficient to provide reasonable estimates or forecasts of the dependent variable. Parsimony is a desired feature of models; if two models are performing about the same, the one with fewer variables is preferred. However, neither the desire for parsimony nor to increase R^2 should be the determining factor for selecting the exogenous variables. It can be shown that a model with the correct variables will result in unbiased, consistent, and efficient estimation of parameters.

In Equation (8.2), the variable named "determination" merits additional comments. There is no doubt that the amount of effort that a person exerts affects his or her income. Thus, the hypothesized sign of β_5 is positive. However, there is no acceptable way of measuring one's resolve. It is fairly easy to identify those that slack off or those that exert themselves, but neither can be measured. More importantly, any arbitrary ranking or measurement of the "determination" of a person is inaccurate and incomplete in the sense that it cannot be compared because it is a not a cardinal measure; a common problem in economics. For example, there is no cardinal measure of utility, which is a very important economic concept. A detailed discussion of how we deal with the inability to measure utility with cardinal measures is beyond the scope of the present text. In brief, there are two possible options. The first one is to accept that it is not a measurable phenomenon and not worry about including it in the model. The consequence is that the error term is enlarged, and there will be more variation in the dependent variable that remains unexplained than if we could have measured "determination" and used it in the model. This exclusion has serious consequences and is usually covered under *misspecification* of the model. The second method is to use a proxy variable that could represent the desired variable, albeit not precisely or accurately. A proxy variable is a variable that is highly correlated with a variable that cannot be observed or measured, such as determination in this example, but is not correlated with other independent variables in the model. One such variable is the difference of one's income from the average income in the previous job.

A generic presentation of a model with an unknown number of variables is of the form as given below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + X_2 + \ldots + \beta_K X_K + \varepsilon \qquad (8.3)$$

Each beta represents the contribution of the corresponding factor to an explanation of the dependent variable, keeping all the other factors constant.

Regression analysis is a powerful and useful tool used in many areas of science but has a special place in economics. As one might expect, there are many issues that pertain to economic research that are not necessarily applicable to other areas of science. We have already seen two such issues. One is the fact that the independent variables are often somewhat related to each other in economics. This is due to the fact that many, if not all, economic factors are subject to the same economic and social realities. The same applies to individual firms and people. In other fields, it is easier to ensure exogenous variables are independent from each other, which is a requirement of regression analysis.[2] The second issue is the role of factors that are social in nature and reflect the social/cultural structure of a country. For example, the fact is that race and gender influence one's income. Consequently, a special branch of science has been created called *econometrics*. Econometrics is the application of statistical methods, including regression, to economics.

CHAPTER 9

Conclusion

The present text is a brief introduction to statistics. The main focuses have been on explanation, application, interpretation, and a sense of appreciation for statistics. The hope is that the reader has become interested in statistics and will pursue the topic further. In fact, the main reason for including the regression chapter, Chapter 8, is to show additional possibilities that go beyond a single-variable analysis. It demonstrates that we can explore the influence of one or more variables on a variable of interest. Within the subject of regressions, one can explore theoretical and empirical aspects of cross section and time series data. Regression analysis has been augmented to utilize data that are qualitative in nature. The qualitative data can be used as dependent variables or independent variables. Many economics decisions can be represented as qualitative dependent variables, for example, the decision to buy a good or not to buy it, obtain a college degree, take a vacation (i.e., consume leisure), or to save, to name a few. Qualitative variables can also be independent variables, such as race, gender, political persuasion, or nationality.

The present text groups related topics and focuses on the interrelationship of different topics; a good example is Table 1.1. The table divides descriptive statistics into two major categories of qualitative variables and quantitative variables. The scope of methods for quantitative variables is much broader than those for the qualitative variables because the methods used for qualitative variables are also applicable for quantitative variables, but the reverse is not true necessarily. Within each category, the analytical methods are broken down to tabular methods and graphical methods. Recall that these are all descriptive methods and their purposes is to provide insight to the nature of data and to condense massive amounts of information into as few parameters as possible. Descriptive statistics are estimates.

Although graphical and tabular methods are very helpful in providing a visual description of data, the analytical power of statistics is more evident in the numerical methods that apply to quantitative variables. It is customary to distinguish among three different classifications of quantitative variables: measures of central tendency, measures of dispersion, and measures of association. Each of these measures provides different analytical perspectives and allows researchers to differentiate among different types of data where certain aspects might be similar while the nature of data is very different, for example, as in the case of two populations with the same means but different variances.

The knowledge about parameters provides an insight into the nature of data. Massive databases are overwhelming and chaotic collection of numbers. In spite of the fact that the human brain is extremely good at finding order in events when the order is not easy to detect, the data are too large, or the relationships are too complex, it needs statistics to comprehend what is going on. A good example to clarify the above point is the saying "to miss the forest for the trees." Statistics provides a way of summarizing the evidence.

Inferential statistics is used to test and to determine the likelihood of the outcome. The human brain is susceptible to finding patterns and association even when there is none. All superstitions are based on sequences of events that are random but erroneously are considered patterns or causal relationships. It is necessary to use statistical analysis to identify real patterns and avoid superstition.

Statistics provides tools to determine, with an appropriate level of probability, the outcome of a certain phenomenon or how to explain one or more variables using one or more other variables. It removes conjecture out of estimation by providing necessary and sufficient conditions to obtain unbiased, consistent, and efficient estimators. It replaces opinion-based conclusions with a probabilistic inference, which would result in the same conclusion for a given level of significance.

In Chapter 3, we put the few descriptive tools that were introduced in Chapter 2 into use by showing the applications of Z scores and the coefficient of variation. That chapter also provided additional tools to improve the analytical power of statistics. The concept of error is one of the major contributions of statistics to science. This notion allows us to divide

variations in a phenomenon, which is ever present in all real life situations, into two components: one that can be explained and one that cannot be explained. One object of statistical analysis is to reduce the magnitude of the part that is unexplained using theory and improved model selection. For example, the mean explains part of the variation in data and leaves part unexplained. In Chapter 8, we saw a glimpse of regression analysis where part of the previously "unexplainable" error is explained by appropriate independent variables that are identified using theories from economics and other disciplines. There are numerous modifications to the simple regression analysis that allows further reduction of the unexplained portion of variations.

A contributing factor is the discovery of a host of distribution functions. These mathematical relationships have certain known properties that are used as benchmarks in inferential statistics. The most important one among distribution functions is the normal distribution. Although many of natural events resemble the normal distribution function, many do not. Nevertheless, the use of theorems such as the Chebyshev's Theorem and the Central Limit Theorem allows us to use the properties of the normal distribution in dealing with some of the statistics obtained from real data that either have a complicated distribution function or do not even have a known distribution function. For example, the distribution function of the quantity demanded of a good is usually unknown. However, the average quantity demanded of several samples has a normal distribution. The link between the above theorems and statistics is the main subject of sampling distribution of sample statistics. We devoted Chapter 5 to this topic exclusively. Sample statistics are compared with corresponding distributions to make inferences.

The next step for most economists is to learn regression analysis using cross section data, followed by time series, and finally using panel data. Almost all economic programs require at least one course in econometrics, which is the application of linear models such as regression analysis to economic issues. More serious students that pursue graduate work in economics are required to learn and sometimes prove the applicable theorems used in econometrics; however, a purely pragmatic approach of learning the methods is utilized by many curriculums.

Theoretical requirements for analyzing cross section data are different from those of time series data. When cross section and time series data are combined the data set is called panel data. In panel data analysis, the problems that cause difficulty in the regressions using cross section or time series data are utilized to provide a better analysis. For example, the existence of correlation among units over time and the presence of correlations among independent variables are incorporated into the analysis rather than excluded or avoided. Probably the best example of this point is the analysis based on *seemingly unrelated data*. In this methodology, the fact that similar firms are subject to similar economic conditions, and thus, respond in similar manners in certain areas is the foundation of the methodology. Another recent development is *Spatial Econometrics* where spatial information is incorporated in the form of weights assigned to economic events. For example, it is reasonable to expect "neighboring" counties to act more similar than distant counties. There are numerous ways of defining neighbors, such as distance and shared borders. Finally, the hope is that this manuscript has been able to answer some of the questions readers had and also sparked an interest in this fascinating subject.

Glossary

Bar graph is a graphical representation of the frequency distribution or relative frequency distribution of qualitative data.

Binomial distribution function is a probability distribution representing a dichotomous variable with a constant probability of occurrence.

Box plot is a graphical representation of several basic descriptive statistics in a concise manner.

Categorical variable contains qualitative values.

Center of gravity of the data is the same as the expected value, or mean.

Central limit theorem states that in repeated random samples from a population, the sample means will have a normal distribution function; the expected value of the sample mean is equal to the true value of the population mean, and the variance of the sample mean is equal to population variance divided by the sample size.

Ceteris paribus is Latin for "other things being equal."

The **coefficient of variation** is the ratio of the standard deviation to the mean.

Chi-square represents the distribution function of a variance.

Claim is a testable hypothesis.

Confidence interval provides an interval that contains the population parameter with a desired level of confidence.

Consistent estimator is an estimator whose variance becomes smaller as sample size increases.

Continuous dichotomous variables exist when one can place an order on the type of data such as being young *versus* old.

Continuous random variable can assume any real value. It represents all the values over a range.

Correction factor is used when the sample size is small or the sample is more than 5% of the population.

Cross-sectional analysis is a study of a snapshot of regions at a given time.

Cumulative frequencies consist of sums of frequencies up to the value or class of interest.

Deductive statistics starts from general information to make inferences about specifics.

Degree of freedom is the number of elements that can be chosen freely in a sample.

Dependent variable is the variable of interest that is explained by statistical analysis. Other names such as endogenous variable, *Y*-variable, response variable, or output are used as well.

Descriptive statistics provide summary estimates of data.

Dichotomous variables, exist when there are only two variables measured in nominal scale, can take only one of the two values (i.e., yes or no, off or on, 0 or 1), also called dummy variables in econometrics.

Discrete dichotomous is a dichotomous variable that can take only integer values.

Discrete random variable is a variable with integer values.

Dot plot is a histogram consisting of points depicting frequencies.

Dummy variable is a dichotomous variable that can take only on a value of 0 or 1. Dummy variables are used to indicate the presence or absence of a characteristic, such as female. It is a special case of nominal data.

Econometrics is the application of statistics to economics.

Efficient estimator is an estimator with the smallest variance compared with other estimator(s).

Error is the difference between an observed value and its expected value. Error is the portion of variation that cannot be explained.

Errors in measurement refer to incorrectly measuring or recording the values of sample points.

Expected value is the theoretically expected outcome, such as the arithmetic mean.

Experimental design is a statistics method where the experiment is controlled for different variables to ensure desired levels of confidence for the estimates.

***F* statistic** is used to test hypotheses about ratios of variances.

Frequency distribution shows the frequency of occurrence for nonoverlapping classes.

Grouped data are summarized or organized to provide a better and more compact picture of reality.

Harmonic mean is the average of rates. It is the reciprocal of the arithmetic mean of the reciprocal of the values.

A **histogram** is a graphical representation of the frequency distribution or relative frequency distribution of quantitative data.

Independent variable is a variable that is used to explain the response or dependent variable. It also has other names such as exogenous variable, *X*-variable, regressor, input, factor, or predictor variable.

Individual error is the difference between an observed value and its expected value.

Inductive statistics observes specifics to make inferences about the general population.

Inferential statistics is a methodology that allows making decisions based on the outcome of statistics from a sample.

An **interval scale** includes relative distances of any two sequential values, such as a Fahrenheit scale.

Kurtosis is a measure of pointedness or flatness of a symmetric distribution.

A **Likert scale** is an ordinal scale, where the subjects provide the ranking of each variable.

The **lower hinge** is the 25th percentile of a box plot.

Mean is the arithmetic average. It represents the center of gravity of data.

Mean absolute error (MAE) is the average of the absolute values of individual errors.

Mean squared error is the same as variance.

Measurement scales indicate the type of variable such as ordinal, cardinal, or interval scales.

Measures of association determine the association between two variables or the degree of association between two variables. They consist of covariance and correlation coefficients.

Measures of central tendency, such as the mean and median, provide concise meaningful summaries of the central properties of a population.

Measures of dispersion reflect how data are scattered. The most important dispersion measures are the variance and the standard deviation.

Median is a value that divides observations into two equal groups. It is the midpoint between a group of numbers ranked in order.

Mode is the most frequent value of a population.

Nominal or **categorical** data are the "count" of the number of times an event occurs.

Normal distribution is a common distribution function that reflects many randomly occurring events in life.

Null hypothesis reflects the status quo or how things have been or are currently.

Observed significance level is another name for the p-value, which is the probability of seeing what you saw.

Observed value is the value of a sample point.

Ogive is a graph for cumulative frequencies.

Ordinal scale indicates that data are ordered in some way, but the numbering has no value other than representing rank.

***P* value** represents the probability of type I error for inference about a coefficient.

A **parameter** is a characteristic of a population that is of interest; it is constant and usually unknown.

A **percentile** is the demarcation value below which the stated percentage of the population or sample lie.

A **pie chart** is a graphical presentation of frequency distribution and relative frequency.

Point estimate is a statistic that consists of a single value, such as the mean or variance.

Probability is the likelihood of occurrence of an event, expressed in the form of a ratio or a percentage.

Probability distribution determines the probability of the outcomes of a random variable.

Probability density distribution for a discrete random variable determines the probability of an occurrence of a *discrete* random variable and is represented as $f(x)$.

Probability distribution for a continuous random variable determines the probability of occurrence of a *continuous* random variable and is represented as $f(x)$.

Qualitative variables are non-numeric and represent a label for a category.

Quantitative variables are numerical and countable values.

Quartiles divide the population into four equal portions, each equal to 25% of the population.

Random variables are selected in a random fashion and by chance.

A **ratio scale** is an interval scale, with the additional criteria that it has an absolute zero value.

Real numbers consist of all rational and irrational numbers. They include all possible values on a line.

Relative frequency shows the percentage of each class to the total population or sample.

Relative variability is the comparison of variability using the coefficient of variation.

Reliability of a sample mean $(\widehat{\mu})$ is equal to the probability that the deviation of the sample mean, from the population mean, is within the tolerable level of error (e).

Root mean squared error is the square root of the mean square error and is the same as the standard error.

Sample standard deviation is the average error of the sample. This is the standard deviation obtained from a sample, but it is not the same as standard error.

Sample statistics are random values obtained from a sample. They estimate the corresponding population parameters and are used to make inferences about them.

Sample variance is an estimate of the population variance. It is the sum of the squares of the deviations of values from the sample mean divided by the degrees of freedom.

Sampling is the process of obtaining a sample from a population.

Sampling distribution explains the governing rules and characteristics of sample statistics.

Scatter plot is a graph customarily used in presenting data from a regression analysis model.

Simple hypothesis gives an exact value for the unknown parameter of the assumed distribution function based on established rules.

Skewness is a measure of deviation from symmetry in a distribution.

Standard deviation is the square root of variance and represents the average error of a population or sample.

Standard error is the standard deviation of the estimated sample statistics.

Standardization is the conversion of the value of an observation into its Z score equivalent.

A **statistic** is a numerical value calculated from a sample that is variable and known.

Statistical hypothesis is an assertion about distribution of one or more random variables.

Statistical inference is the process of drawing conclusions based on evidence obtained from a sample. All statistical inferences are probabilistic.

Stem and leaf is a graphical way of summarizing data and is a type of a descriptive statistic.

Stochastic means probabilistic. Stochastic variables are random.

***t* distribution** is a modification of the normal distribution function for small samples or when the variance is not known.

A **testable hypothesis** is a claim against established norms and beliefs.

Time series analysis is a special branch of statistics that deals with time series data.

Tolerable level of error is the amount of error that the researcher is willing to accept.

Tolerance level is a measure for detecting multicolinearity. It is the reciprocal of Variance Inflation Factor (VIF). A tolerance value less than 0.1 is an indicative of the presence of multicollinearity.

Total sum of square represents the total variation in the dependent variable.

Trimmed mean is a modification of the mean, where outliers are discarded.

Type I error is rejecting the null hypothesis when it is true.

Type II error is failure to reject the null hypothesis when it is false.

Type III error is rejecting a null hypothesis in favor of an alternative hypothesis with the wrong sign.

Typical refers to the average value.

Unbiased refers to an estimate whose expected value is equal to the corresponding population parameter.

The **upper hinge** is the 75th percentile of a box plot.

Validity is the lack of measurement error.

Variance is the sum of the squares of the deviations of values from their mean, divided by population size. It is the average of the squared individual errors.

Weighted mean is similar to the mean except the weights for the observations are not equal and instead represent their contribution to the total. Calculation of GPA, where the grade received is weighted by the number of credit hours of the class, is an example of a weighted mean.

***Z* score** is a statistic that is based on standard units. It is standardized for the purpose of comparing variables that are measured in different units.

Endnotes

Chapter 1

1. Internal Revenue Service. 2014. "Statistics of Income." http://www.irs.gov/pub/irs-soi/14taxstatscard.pdf
2. Wunderground. 2015. Average Temperatures for Washington, DC. Accessed July 6, 2015.
3. International Olympic Committee. London 2012 Men's 100 Meter race results. Accessed July 6, 2015. http://www.olympic.org/olympic-results/london-2012/athletics/100m-m
4. Cox, Nicholas J. 2004. "Speaking Stata: Graphing Categorical and Compositional Data." *The Stata Journal* 4(2), 190–215.
5. Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley
6. Anderson, David, R., Dennis J. Sweeney, and Thomas, A. Williams. 2011. *Statistics for Business and Economics*. Mason, OH: South-Western College Publisher.
7. Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley

Chapter 2

1. Naghshpour, S. 2014. *Monetary Policy within the IS-LM Framework*. New York, NY: Business Expert Press.

Chapter 3

1. Gosset, William S. 1908. "Probable Error of a Correlation Coefficient." *Biometrika* 6, 302–310.
2. PayScale.com. 2015. Economist Salary (United States). http://www.payscale.com/research/US/Job=Economist/Salary.
3. Naghshpour, Shahdad. 2016. *Regression for Economics*. New York, NY: Business Expert Press.

Chapter 4

1. US Census Bureau, Survey of Income and Program Participation, 2008 Panel, Wave 10.

Chapter 6

1. Agresti, Alan and Brent A. Coull. 1998. "Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions." *The American Statistician* 52(2), 119-126.

Chapter 8

1. Naghshpour, Shahdad. 2016. *Regression for Economics*. New York, NY: Business Expert Press.
2. Naghshpour, 2016, *Regression*.

Index

OTHER TITLES FROM THE ECONOMICS COLLECTION

Philip Romero, The University of Oregon and
Jeffrey Edwards, North Carolina A&T State University, Editors

- *U.S. Politics and the American Macroeconomy* by Gerald T. Fox
- *Seeing the Future: How to Build Basic Forecasting Models* by Tam Bang Vu
- *Emerging and Frontier Markets: The New Frontline for Global Trade* by Marcus Goncalves and José Alves
- *Doing Business in Emerging Markets: Roadmap for Success* by Marcus Goncalves, José Alves, and Rajabahadur V. Arcot
- *Comparing Emerging and Advanced Markets: Current Trends and Challenges* by Marcus Goncalves and Harry Xia
- *What Hedge Funds Really Do: An Introduction to Portfolio Management* by Philip J. Romero and Tucker Balch
- *Learning Basic Macroeconomics: A Policy Perspective from Different Schools of Thought* by Hal W. Snarr
- *Basel III Liquidity Regulation and Its Implications* by Mark Petersen and Janine Mukuddem-Petersen
- *Macroeconomics: Integrating Theory, Policy and Practice for a New Era* by David G. Tuerck
- *The Basics of Foreign Exchange Markets: A Monetary Systems Approach* by William D. Gerdes
- Advanced Economies and Emerging Markets: Prospects for Globalization by Marcus Goncalves, et al

Statistics for Economics

Shahdad Naghshpour

Statistics is the branch of mathematics that deals with real life problems. The book is written in simple English with minimal use of symbols, mostly for the sake of brevity and to make reading literature more meaningful.

All the examples and exercises in the book are constructed within the field of economics, thus eliminating the difficulty of learning statistics with examples from fields that have no relation to business, politics, or policy. Statistics is in fact, not more difficult than economics.

The second edition incorporates Stata 14.1 and duplicates the answers for all the examples using Stata as well. This will enable the more serious users to be prepared for the next level when more powerful tools are necessary. The book utilizes Microsoft Excel to obtain statistical results as well as to perform additional necessary computations. The spreadsheet is not the software of choice for performing sophisticated statistical analysis. However, it is widely available and almost everyone has some degree of familiarity with it. Using Excel will eliminate the need for students and readers to buy and learn new software, the need for which would itself prove to be another impediment to learning and using statistics.

Shahdad Naghshpour is a professor of International Development Doctoral Program at the University of Southern Mississippi. He has published in journals such as Peace Economics, Peace Science and Public Policy, Journal of Economics and Finance, Review of Regional Studies, International Journal of Trade and Global Markets, International Journal of Monetary Economics and Finance, Politics and Policy, and International Journal of Economics among others. He has authored eight books as well. Dr. Naghshpour has revived several awards for research and teaching as well as numerous grants. He is a member of Montclair Who's Who in Collegiate Faculty. Dr. Naghshpour is on the editorial boards of five peer-reviewed journals and currently is the Vice President of the Academy of Economics and Finance. He has been a consultant to numerous state agencies and private companies.

ECONOMICS COLLECTION

Philip J. Romero and Jeffrey A. Edwards, *Editors*

ISBN 978-1-63157-389-7

BUSINESS EXPERT PRESS