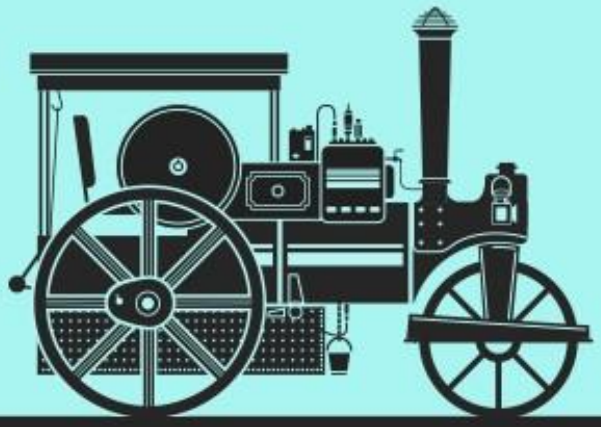


Statistics Using Excel

Succinctly

by Charles Zaiontz

Syncfusion | Technology Resource Portal

Statistics Using Excel Succinctly

By

Charles Zaiontz

Foreword by Daniel Jebaraj

Important licensing information. Please read.

Technical Reviewer: Vittoria Ardino
Copy Editor: Suzanne Kattau
Acquisitions Coordinator: Hillary Bowling, marketing coordinator, Syncfusion, Inc.
Proofreader: Darren West, content producer, Syncfusion, Inc.

Table of Contents

The Story behind the *Succinctly* Series of Books

Daniel Jebaraj, Vice President
Syncfusion, Inc.

Staying on the cutting edge

As many of you may know, Syncfusion is a provider of software components for the Microsoft platform. This puts us in the exciting but challenging position of always being on the cutting edge.

Whenever platforms or tools are shipping out of Microsoft, which seems to be about every other week these days, we have to educate ourselves, quickly.

Information is plentiful but harder to digest

In reality, this translates into a lot of book orders, blog searches, and Twitter scans.

While more information is becoming available on the Internet and more and more books are being published, even on topics that are relatively new, one aspect that continues to inhibit us is the inability to find concise technology overview books.

We are usually faced with two options: read several 500+ page books or scour the web for relevant blog posts and other articles. Just as everyone else who has a job to do and customers to serve, we find this quite frustrating.

The *Succinctly* series

This frustration translated into a deep desire to produce a series of concise technical books that would be targeted at developers working on the Microsoft platform.

We firmly believe, given the background knowledge such developers have, that most topics can be translated into books that are between 50 and 100 pages.

This is exactly what we resolved to accomplish with the *Succinctly* series. Isn't everything wonderful born out of a deep desire to change things for the better?

The best authors, the best content

Each author was carefully chosen from a pool of talented experts who shared our vision. The book you now hold in your hands, and the others available in this series, are a result of the authors' tireless work. You will find original content that is guaranteed to get you up and running in about the time it takes to drink a few cups of coffee.

Free forever

Syncfusion will be working to produce books on several topics. The books will always be free. Any updates we publish will also be free.

Free? What is the catch?

There is no catch here. Syncfusion has a vested interest in this effort.

As a component vendor, our unique claim has always been that we offer deeper and broader frameworks than anyone else on the market. Developer education greatly helps us market and sell against competing vendors who promise to "enable AJAX support with one click" or "turn the moon to cheese!"

Let us know what you think

If you have any topics of interest, thoughts or feedback, please feel free to send them to us at succinctly-series@syncfusion.com.

We sincerely hope you enjoy reading this book and that it helps you better understand the topic of study. Thank you for reading.

Please follow us on Twitter and "Like" us on Facebook to help us spread the word about the *Succinctly* series!

Chapter 1 Introduction

What is Statistics?

Statistics is a field of study that has two principal objectives:

- Collecting, organizing, interpreting, and presenting data (Descriptive Statistics)
- Using mathematical techniques, especially from probability theory, to make inferences and/or predictions based on a sample of experimentally observed data (Inferential Statistics)

Descriptive Statistics involves calculating the mean, median, variance, standard deviation, and other properties of the data, and presenting this information in ways that make the data more meaningful such as histograms, box plots, etc.

Inferential Statistics involves analyzing data and inferring characteristics of a general population based on the same properties of a sample taken from the population. This is what gives the field of statistics its power since, with a relatively small amount of data, we are able to make significant assertions even though such inferences are not 100 percent certain but, rather, are probabilistic in nature.

Why do Statistics in Excel?

There are a number of commonly used, powerful tools for carrying out statistical analyses. The most popular of these are Statistical Package for the Social Sciences (SPSS), Statistical Analysis System (SAS), Stata, Minitab and R. Many people choose to use Excel as their principal analysis tool or as a complement to one of these tools for any of the following reasons:

- It is widely available and so many people already know how to use it
- It is not necessary to incur the cost of yet another tool (as some of the popular tools are quite expensive)
- It is not necessary to learn new methods of manipulating data and drawing graphs
- It provides numerous built-in statistical functions and data analysis tools
- It is much easier to see what is going on since, unlike the more commonly used statistical analysis tools, very little is hidden from the user
- It provides the user with a lot of control and flexibility

This makes Excel an ideal tool for quick analyses and even some serious systematic analyses. However, it has two major shortcomings:

- Many people are not very familiar with the statistical capabilities that are built into Excel
- Excel was not designed to be a comprehensive statistical package and so many commonly used statistical tests are not provided

We will address the first of these shortcomings in this book. We will also give you some techniques for extending the built-in statistical capabilities included in Excel, which partially addresses its second shortcoming.

Where Excel provides the statistical analysis capabilities that you need (and, fortunately, many of the most commonly used tests are included in Excel), it is a great tool to use for the reasons previously mentioned. Where it does not, there are a number of software packages which extend Excel's built-in capabilities. Once such software package is one that I have developed called Real Statistics. You can download it for free here.

Where is Statistics Used?

Statistics plays a central role in research in the social sciences, pure sciences, and medicine. A simplified view of experimental research is as follows:

- You make some observations about the world, and then create a theory consisting of a hypothesis and possible alternative hypotheses that try to explain the observations you have made
- You then test your theory by conducting experiments. Such experiments include collecting data, analyzing the results, and coming to some conclusions about how well your theory holds up
- You iterate this process, observing more about the world and improving your theory

Statistics also plays a major role in decision making for business and government, including marketing, strategic planning, manufacturing, and finance.

Statistics is a discipline which is concerned with the collection and analysis of data based on a probabilistic approach. Theories about a general population are tested on a smaller **sample** and conclusions are made about how well properties of the sample extend to the **population** at large.

What Will This Book Cover?

This book aims to show the reader how to perform statistical analyses using Excel. In particular, it describes:

- Basic statistical tests and analyses
- Methods for carrying out these tests and analyses using Excel
- Numerous examples

We don't have space to cover all of the aspects of statistics in Excel but, as you will see, we will try to give you a good idea of how to do important statistical analyses in Excel.

Chapter 2 Excel Environment

The Basics

For the most part, we will assume that the reader is already familiar with Excel. We don't have the space here to provide a review of Excel. Instead, we will touch on a few of the capabilities that are not as commonly used but that are essential for doing statistical analysis in Excel. We will focus on the most recent versions of Excel in the Windows environment, namely, Excel 2007, 2010, and 2013—although most of what is written will also apply to earlier versions of Excel in Windows as well as Excel for the Macintosh (especially Excel 2011).

User Interface

Excel works with files called **workbooks**. Each workbook contains one or more spreadsheets called **worksheets**. Each worksheet consists of **cells** that are organized in a rectangular grid. The rows of a worksheet are labeled with a number, and the columns are labeled with a letter or series of letters. The first row is labeled 1, the next row is labeled 2, and so on. The first column is labeled A, the next column is labeled B, and so on. The column after Z is labeled AA and then the next columns are labeled AB, AC, and so on. The column after AZ is BA, and then the next columns are labeled BB, BC, and so on. The column after ZZ is labeled AAA, and then the next columns are labeled AAB, AAC, and so on.

Each worksheet in Excel 2007/2010/2013 can contain up to 1,048,578 rows and 16,384 columns (i.e., column A through XFD). Each cell can contain a number, text, truth value or a formula. Cells have both a **value** and an **address** (e.g., the cell with address B5 lays in the fifth row and second column).

A cell **range** consists of all of the cells in a rectangular area on the worksheet (e.g., the range B6:C10 consists of the 10 cells in the rectangle whose opposite corners are the cells B6 and C10. The formula =SUM(B6:C10) has a value which is the sum of the values of all the cells in the range B6:C10). Formulas can use either **absolute addressing**, **relative addressing**, or a combination of both.

The rectangular grid is only one part of the Excel user interface (UI). The entire UI is as depicted in Figure 1:

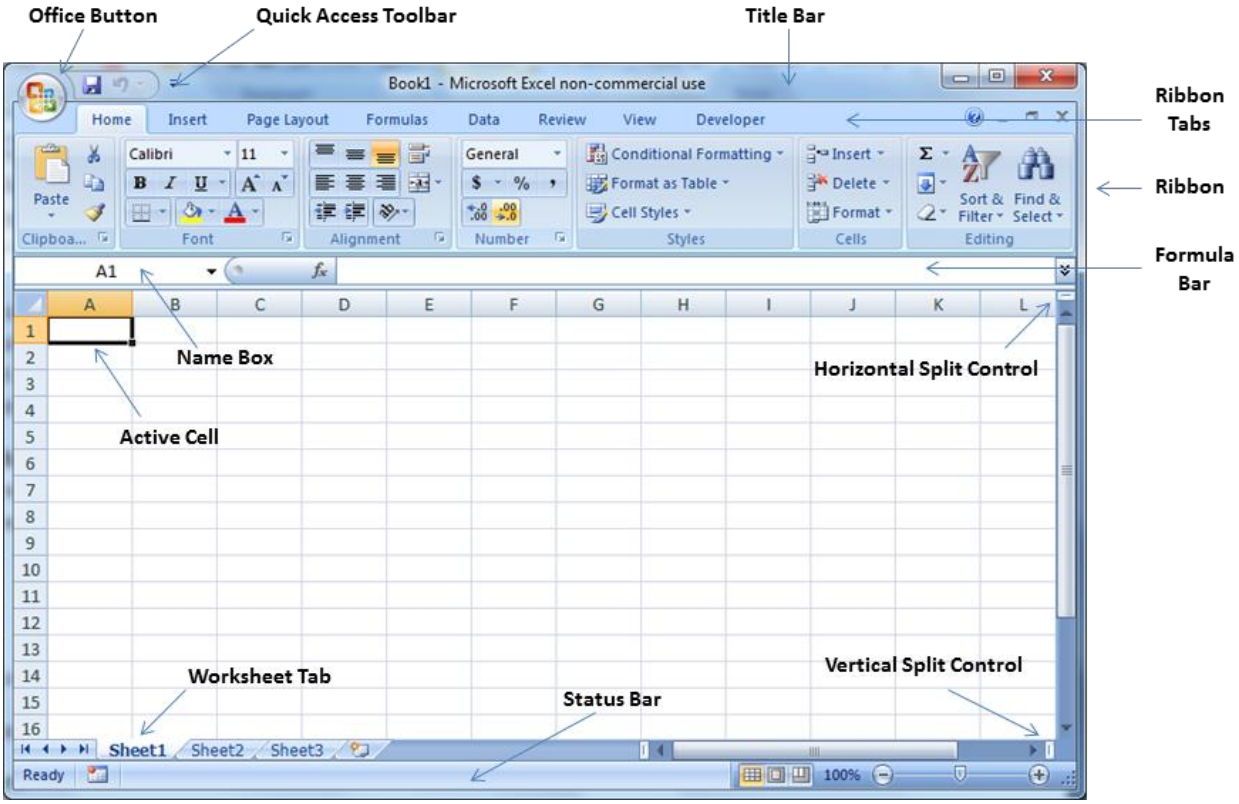


Figure 1: Excel user interface

This is the layout that is used in Excel 2007. The layout used in Excel 2010 and Excel 2013 is almost identical.

The **Ribbon Tabs** are the top-level menu items. In Figure 1, these are **Home, Insert, Page Layout, Formulas**, etc. To access most capabilities in Excel, you click one of these ribbon tabs. For each tab, a different ribbon will be displayed.

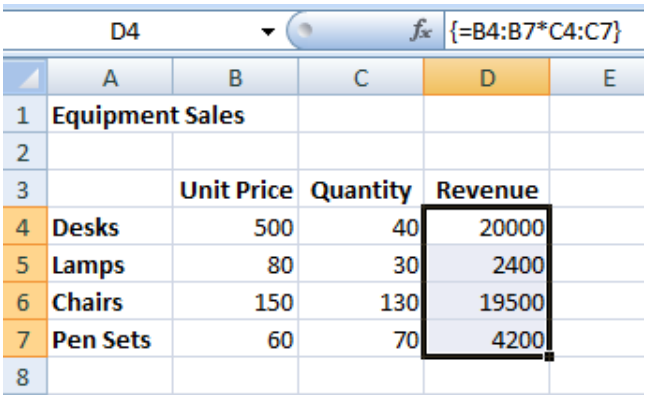
Each **ribbon** consists of a collection of Excel capabilities that are organized into **groups** that correspond to some ribbon tab. For example, the **Home** ribbon (displayed in Figure1) is organized into the **Clipboard, Font, Alignment, Number**, etc. groups. Each group consists of one or more **icons** that correspond to some capabilities in Excel. For example, to center the content of a cell in a worksheet, click that cell and then click the center icon in the Alignment group on the Home ribbon.

We use the following abbreviation for this sequence of steps: **Home > Alignment|Center**. There are also shortcuts for some icons (e.g., to center the contents of a cell, you can click that cell and then enter **Ctrl-E**).

Array Formulas

Most worksheet formulas used in Excel return a single value that is assigned to the cell which contains the formula. Excel also allows you to define a formula that assigns values to a range of cells at the same time. These are called **array formulas** and they will be used quite often in the rest of this book.

Example: Calculate the revenues for each item in the worksheet of Figure 2:



	A	B	C	D	E
1	**Equipment Sales**				
2					
3		**Unit Price**	**Quantity**	**Revenue**	
4	**Desks**	500	40	20000	
5	**Lamps**	80	30	2400	
6	**Chairs**	150	130	19500	
7	**Pen Sets**	60	70	4200	
8					

Figure 2: Array formulas

Given that the revenue for each item is unit price times quantity, we can enter the formula =B4*C4 in cell D4 and then copy this formula into cells D5, D6, and D7 (e.g., by clicking cell D4, pressing **Ctrl-C**, highlighting the range D5:D7, and pressing **Ctrl-V**, or by highlighting the range D4:D7 and pressing **Ctrl-D**).

Another way to do this is via an array formula using the following steps:

1. Highlight the range D4:D7
2. Enter the array formula =B4:B7*C4:C7
3. Press **Ctrl-Shft-Enter**

It is essential to press **Ctrl-Shft-Enter** (step three) and not simply **Enter** for an ordinary formula. Note that the formula that appears in the formula bar is {=B4:B7*C4:C7}. The curly brackets indicate that this is an array formula. If the range B4:B7 is given the name UnitPrice and C4:C7 is given the name Quantity, then the array formula can be entered as =UnitPrice*Quantity (step two).

The array formula appears in all four cells in the range D4:D7. To make changes to the formula, you must edit the entire range, not just one, two, or three of these cells. Similarly, you can't copy or delete a part of the range but, rather, you must copy or delete the entire range. If you attempt to modify a part of the range, you will receive an error message. If you get stuck and see a series of error messages, you just need to press **Esc** to recover.

You can erase a range that contains an array formula by highlighting the entire range and pressing **Delete**. You can write over the array function, replacing it by a value or another formula. The important thing is to use the entire range and not a part of the range.

Note, too, that you can also use array formulas such as {=SUM(B4:B7*C4:C7)}. This returns the value, which is the sum of the revenues of the four types of equipment. Even though this formula returns a single value, and so may be placed in a single cell such as D8, it must be entered as an array formula (since the formula contains an embedded array formula). This means that you need to type =SUM(B4:B7*C4:C7) and then press **Ctrl-Shft-Enter**. If you forget to press **Ctrl-Shft-Enter** and only press **Enter**, you will get an error message.

Array Functions

A few of Excel's built-in functions are array functions where the output of the function is an array (i.e., a range). These functions are managed as previously described for array formulas.

Example: Change the data range in columns A and B of Figure 3 into an equivalent row range:

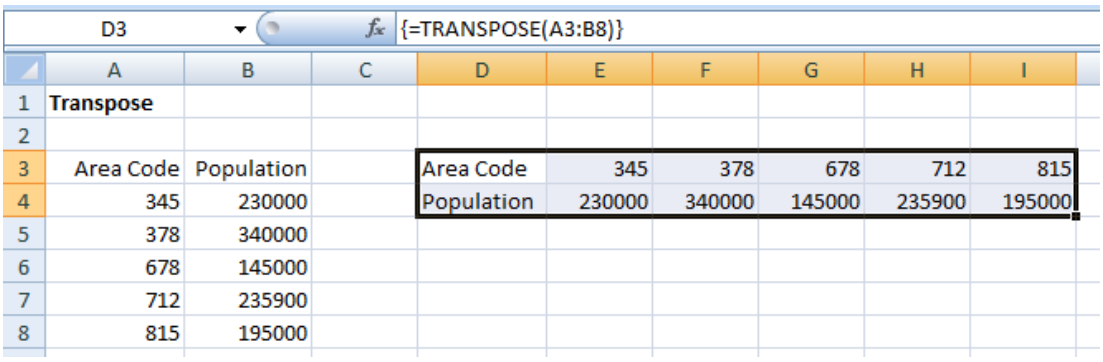


Figure 3: Array function

This can be accomplished by means of Excel's TRANSPOSE array function using the following steps:

1. Highlight the output range D3:I4
2. Enter the array formula =TRANSPOSE(A3:B8)
3. Press **Ctrl-Shft-Enter**

Note that the output range (step one) must be of the right size. In this case, the input range is six rows by two columns, and so the output range must be two rows by six columns. As for array formulas, the formula bar contains the array formula enclosed in curly brackets. Once again, it is important to press **Ctrl-Shft-Enter**.

Data Analysis Tools

Excel provides a number of data analysis tools which are accessible via **Data > Analysis|Data Analysis**.

If this option is not visible, you may need to first install Excel's analysis toolpack. This is done by selecting **Office Button > Excel Options > Add-Ins** in Excel 2007 or **File > Help|Options > Add-Ins** in Excel 2010/2013, and then clicking **Go** at the bottom of the window. Next, you select the **Analysis ToolPak** option in the dialog box that appears, and click **OK**. You will then be able to access the data analysis tools.

After selecting **Data > Analysis|Data Analysis**, you will be presented with the dialog box in Figure 4:

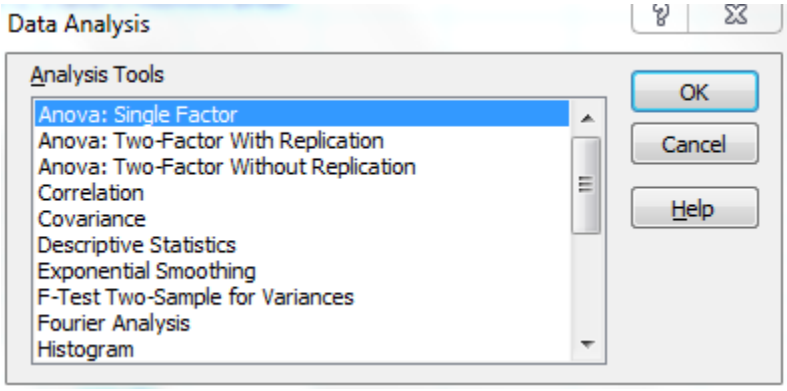


Figure 4: Data analysis dialog box

You can now select any one of the following options useful for statistical analysis:

- ANOVA: Single Factor
- ANOVA: Two-Factor with Repetition
- ANOVA: Two-Factor without Repetition
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test: Two Sample for Variances
- Histogram
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variance
- t-Test: Two-Sample Assuming Unequal Variance
- z-Test: Two-Sample for Means

Each of these options represents a data analysis tool that will be described in this book. Now suppose, for example, that you choose **ANOVA: Single Factor**. You will now be presented with the dialog box shown in Figure 5:

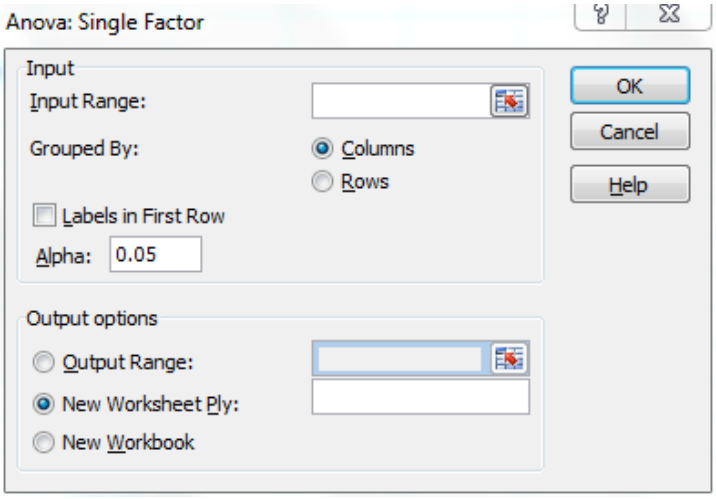


Figure 5: Dialog box for ANOVA: Single Factor

The Input Range consists of the Excel range within which the data elements to be analyzed are stored. Suppose, for example, this data consists of a 4×8 array representing four treats as in Figure 6:

	A	B	C	D	E	F
1		Treat 1	Treat 2	Treat 3	Treat 4	
2	1	34	29	26	10	
3	2	45	11	24	23	
4	3	56	33	23	40	
5	4	21	28	31	36	
6	5	12	44	43	36	
7	6	34	37	29	23	
8	7	17	23	19	21	
9	8	21	54	41	39	
10						

Figure 6: Sample input range

In this case, you insert the range B2:E9 in the **Input Range** field (of the dialog box of Figure 5) and select **Columns**. If you had assigned a name (e.g., Study1) to the range B2:E9, then you could put this name in the **Input Range** field instead of B2:E9.

Alternatively, you could insert B1:E9 in the **Input Range** field and select the **Labels in First Row** check box in the dialog box to indicate that you have included the column headings in the data range. Note that the participant numbers (in column A) are not used.

If the data were arranged where the treatments are listed by row instead of column, then you would select **Rows** and you could select the **Labels in First Column** check box.

The **Alpha** value (which will be described later in the book) is set to 0.05 by default, although you have the option to change this to 0.01 or some other value.

You can now select **New Worksheet Ply** from the **Output** options (and leave the data field blank). In this case, a new worksheet will be created (in the tab prior to the current one) and the ANOVA report will be placed in this worksheet starting at cell A1. You can then copy the results to the current worksheet (or anywhere else you like).

Alternatively, you can select **Output Range** or **New Workbook**, and put the report in some specific output range that you choose or in a new workbook.

Chapter 3 Descriptive Statistics

Basic Terminology

Before exploring **Descriptive Statistics**, we define the following statistical concepts which will be used throughout the rest of the book.

Sample: A subset of the data from the population which we analyze in order to learn about the population. A major objective in the field of statistics is to make inferences about a population based on properties of the sample.

Random sample: A sample in which each member of the population has an equal chance of being included and in which the selection of one member is independent from the selection of all other members.

Random variable: A variable which represents values from a random sample. We will use letters at the end of the alphabet, especially $x, y,$ and z as random variables.

Independent random variable: A random variable that is chosen and then measured or manipulated by the researcher in order to study some observed behavior.

Dependent random variable: A random variable whose value depends on the value of one or more independent variables.

Discrete random variable: A random variable that takes a discrete set of values. Such random variables generally take a finite set of values (e.g., heads or tails, brands of cars, etc.) but they can also include random variables that take a countable set of values (e.g., 0, 1, 2, 3, etc.).

Continuous random variable: A random variable that takes an infinite number of values even in a finite interval (e.g., the height of a building, a temperature, etc.).

Statistic: A quantity which is calculated from a sample and is used to estimate a corresponding characteristic (i.e., **parameter**) about the population from which the sample is drawn.

In the rest of this chapter, we define some statistics which are commonly used to characterize data. In particular, we define metrics of central tendency (e.g., mean and median), variability (e.g., variance and standard deviation), symmetry (i.e., skewness) and peakedness (i.e., kurtosis).

We also provide some important ways of graphically describing data and distributions including histograms and box plots.

Measures of Central Tendency

We consider a random variable x and a data set *S*= $\{x_1, \dots, x_n\}$ of size n which contains possible values of x. The data can represent either the population being studied or a sample drawn from the population.

We seek a single measure (i.e., a **statistic**) which somehow represents the center of the entire data set *S*. The commonly used measures of central tendency are the mean, median, and mode. Besides the normally studied mean (also called the **arithmetic** mean), we also touch briefly on two other types of mean: the **geometric** mean and the **harmonic** mean.

These measures are summarized in Table 1 where R is an Excel range which contains the data elements in the sample or population *S*:

Statistic	Excel 2007	Excel 2010/2013
Arithmetic Mean	AVERAGE(R)	AVERAGE(R)
Median	MEDIAN(R)	MEDIAN(R)
Mode	MODE(R)	MODE.SNGL(R), MODE.MULT(R)
Geometric Mean	GEOMEAN(R)	GEOMEAN(R)
Harmonic Mean	HARMEAN(R)	HARMEAN(R)

Table 1: Measures of central tendancy

Mean

We begin with the **mean**, the most commonly used measure of central tendency. The **mean** (also called the **arithmetic mean**) of the data set *S* is defined as follows:

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

The mean is calculated in Excel using the worksheet function **AVERAGE**.

Example: The mean of *S* = {5, 2, -1, 3, 7, 5, 0, 2} is (2 + 5 – 1 + 3 + 7 + 5 + 0 + 2) / 8 = 2.875. We achieve the same result by using the formula =AVERAGE(C3:C10) in Figure 7:

	A	B	C	D	E	F	G	H	I
1	**Measures of Central Tendency**								
2									
3		5	5	5					
4		2	2	2					
5		-1	-1	-1					
6		3	3	3					
7		7	7	7			50	1.05	5
8		5	5	4			70	1.05	2
9		0	0	0				1.1	-1
10			2	6				1.1	0
11									
12	AVERAGE	3	2.875	3.25		HARMEAN	58.33333	1.074419	#NUM!
13									
14	MEDIAN	3	2.5	3.5		GEOMEAN	59.1608	1.074709	#NUM!
15									
16	MODE	5	5	#N/A					
17	MODE_SNGL	5	5	#N/A					
18									
19	MODE.MULT	5	5	#N/A					
20		5	2	#N/A					

Figure 7: Examples of central tendency in Excel

When the data set *S* is a population, the Greek symbol μ is used for the mean. When *S* is a sample, then the symbol $\bar{x}$ is used.

Median

If you arrange the data in *S* in increasing order, the middle value is the **median**. When *S* has an even number of elements, there are two such values. The average of these two values is the median.

The median is calculated in Excel using the worksheet function **MEDIAN**.

Example: The median of *S* = {5, 2, -1, 3, 7, 5, 0} is 3 since 3 is the middle value (i.e., the 4th of 7 values) in -1, 0, 2, 3, 5, 5, 7. We achieve the same result by using the formula =MEDIAN(B3:B10) in Figure 7.

Note that each of the functions in Figure 7 ignores any non-numeric values including blanks. Thus, the value obtained from =MEDIAN(B3:B10) is the same as that for =MEDIAN(B3:B9).

The median of *S* = {5, 2, -1, 3, 7, 5, 0, 2} is 2.5 since 2.5 is the average of the two middle values 2 and 3 of -1, 0, 2, 2, 3, 5, 5, 7.This is the same result as =MEDIAN(C3:C10) in Figure 7.

Mode

The **mode** of the data set *S* is the value of the data element that occurs most often.

Example: The mode of *S* = {5, 2, -1, 3, 7, 5, 0} is 5 since 5 occurs twice, more than any other data element. This is the result we obtain from the formula =MODE(B3:B10) in Figure 7. When there is only one mode, as in this example, we say that *S* is **unimodal**.

If *S* = {5, 2, -1, 3, 7, 5, 0, 2}, the mode of S consists of both 2 and 5 since they each occur twice, more than any other data element. When there are two modes, as in this case, we say that *S* is **bimodal**.

The mode is calculated in Excel by the formula **MODE**. If range R contains unimodal data, then MODE(R) returns this unique mode. But when R contains data with more than one mode, MODE(R) returns the first of these modes. For our example, this is 5 (since 5 occurs before 2, the other mode, in the data set). Thus, MODE(C3:C10) = 5.

If all of the values occur only once, then MODE returns an error value. This is the case for *S* = {5, 2, -1, 3, 7, 4, 0, 6}. Thus, MODE(D3:D10) returns the error value #N/A.

Excel 2010/2013 provide an array function **MODE.MULT** which is useful for multimodal data by returning a vertical list of modes. When we highlight C19:C20, then enter the array formula =MODE.MULT(C3: C10), and then press **Ctrl-Alt-Enter**, we see that both modes are displayed.

Excel 2010/2013 also provides the function **MODE.SNGL** which is equivalent to MODE.

Geometric mean

The **geometric mean** of the data set *S* is defined as follows:

$$\sqrt[n]{\prod_{i=1}^{n} x_n}$$

This statistic is commonly used to provide a measure of average rate of growth as described in the next example.

The geometric mean is calculated in Excel using the worksheet function **GEOMEAN**.

Example: Suppose the sales of a certain product grow five percent in the first two years and 10 percent in the next two years. What is the average rate of growth over the four years?

If sales in year 1 are \$1, then sales at the end of the four years are (1 + .05)(1 + .05)(1 + .1)(1 + .1) = 1.334. The annual growth rate r is that amount such that $(1 + r)^4$.= 1.334. Thus, r is equal to $.334^{1/4}$.= .0747.

The same annual growth rate of 7.47 percent can be obtained in Excel using the formula GEOMEAN(H7:H10) – 1 = .0747.

Harmonic Mean

The **harmonic mean** of the data set *S* is calculated as follows:

$$n \Big/ \sum_{i=1}^{n} \frac{1}{x_i}$$

This statistic can be used to calculate an average speed as described in the next example.

The harmonic mean is calculated in Excel using the worksheet function **HARMEAN**.

Example: If you go to your destination at 50mph and return at 70mph, what is your average rate of speed?

Assuming the distance to your destination is d, the time it takes to reach your destination is d/50 hours and the time it takes to return is d/70 hours, for a total of d/50 + d/70 hours. Since the distance for the whole trip is 2d, your average speed for the whole trip is as follows:

$$\frac{2d}{\frac{d}{50}+\frac{d}{70}} = \frac{2}{\frac{1}{50}+\frac{1}{70}} = 58.33$$

This is equivalent to the harmonic mean of 50 and 70, and so can be calculated in Excel as HARMEAN(50,70) which is HARMEAN(G7:G8) from Figure 7.

Measures of Variability

Let's now consider a random variable x and a data set *S* = $\{x_1, \ldots, x_n\}$ of size n which contains possible values of x. The data set can represent either the population being studied or a sample drawn from the population.

The mean is the statistic used most often to characterize the center of the data in *S*. We now consider the following commonly used measures of variability of the data around the mean, namely, the **standard deviation**, **variance**, **squared deviation**, and **average absolute deviation**.

In addition, let's also explore three other measures of variability that are not linked to the mean, namely, the **median absolute deviation**, **range**, and **interquartile range**.

Of these statistics, the variance and standard deviation are the most commonly employed.

These measures are summarized in Table 2 where R is an Excel range which contains the data elements in the sample or population S:

Statistic	Excel 2007	Excel 2010/2013	Symbol
Population Variance	VARP(R)	VAR.P(R)	σ^2
Sample Variance	VAR(R)	VAR.S(R)	s^2
Population Standard Deviation	STDEVP(R)	STDEV.P(R)	σ
Sample Standard Deviation	STDEV(R)	STDEV.S(R)	s
Squared Deviation	DEVSQ(R)	DEVSQ(R)	SS
Average Absolute Deviation	AVEDEV(R)	AVEDEV(R)	AAD
Median Absolute Deviation	See below	See below	MAD
Range	MAX(R)-MIN(R)	MAX(R)-MIN(R)	
Interquartile Range	=QUARTILE(R, 3) – QUARTILE(R, 1)	See below	IQR

Table 2: Measures of variability

Variance

The **variance** is a measure of the dispersion of the data around the mean. Where *S* represents a population, the **population variance** (symbol $\boldsymbol{\sigma^2}$) is calculated from the population mean μ as follows:

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

Where *S* represents a sample, the **sample variance** (symbol $\boldsymbol{s^2}$) is calculated from the sample mean $\bar{x}$ as follows:

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

The sample variance is calculated in Excel using the worksheet function **VAR.S**. The population variance is calculated in Excel using the function **VAR.P**. In versions of Excel prior to Excel 2010, these functions are called **VAR** and **VARP**.

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2} represents a population, then the variance = 4.25. This is calculated as follows:

First, the mean = (2+5-1+3+4+5+0+2)/8 = 2.5, and so the squared deviation SS = $(2–2.5)^2$ + $(5–2.5)^2$ + $(-1–2.5)^2$ + $(3–2.5)^2$ + $(4–2.5)^2$ + $(5–2.5)^2$ + $(0–2.5)^2$ + $(2–2.5)^2$ = 34. Thus, the variance = SS/n = 34/8 = 4.25.

If instead *S* represents a sample, then the mean is still 2.5 but the variance = SS/(n–1) = 34/7 = 4.86. These can be calculated in Excel by the formulas VAR.P(B3;B10) and VAR.S(B3:B10) as shown in Figure 8:

	A	B	C	D	E	F
1	**Measures of Variability**					
2						
3		5				
4		2				
5		-1				
6		3				
7		4				
8		5				
9		0				
10		2				
11						
12	AVERAGE	2.5		MEDIAN	2.5	
13				MAD	2	
14	VAR.S	4.857143				
15	VAR.P	4.25		MIN	-1	
16				MAX	5	
17	STDEV.S	2.203893		Range	6	
18	STDEV.P	2.061553				
19					INC	EXC
20	DEVSQ	34		Quartile 1	1.5	0.5
21				Quartile 3	4.25	4.75
22	AVEDEV	1.75		IQR	2.75	4.25

Figure 8: Examples of measures of variability in Excel

Standard Deviation

The **standard deviation** is the square root of the variance. Thus, the population and sample standard deviations are calculated respectively as follows:

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

The sample standard deviation is calculated in Excel using the worksheet function **STDEV.S**. The population variance is calculated in Excel using the function **STDEV.P**. In versions of Excel prior to Excel 2010, these functions are called **STDEV** and **STDEVP**.

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2} is a population, then the standard deviation = the square root of the population variance, that is: $\sqrt{4.25}$ = 2.06. If *S* is a sample, then the sample standard deviation = square root of the sample variance = $\sqrt{4.86}$ = 2.20.

These are the results of the formulas STDEV.P(B3:B10) and STDEV.S(B3:B10) as shown in Figure 8.

Squared Deviation

The **squared deviation** (symbol *SS* for **sum of squares**) is most often used in **ANOVA** and related tests. It is calculated as follows:

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

The squared deviation is calculated in Excel using the worksheet function **DEVSQ**.

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2}, then the squared deviation = 34. This is the same as the result of the formula DEVSQ(B3:B10) as shown in Figure 8.

Average Absolute Deviation

The **average absolute deviation** (**AAD**) of data set *S* is calculated as follows:

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

The average absolute deviation is calculated in Excel using the worksheet function **AVEDEV**.

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2}, then the average absolute deviation = 1.75. This is the same as the result of the formula AVEDEV(B3:B10) as shown in Figure 8.

Median Absolute Deviation

The **median absolute deviation** (**MAD**) of data set *S* is calculated as follows, where $\tilde{x}$ = median of the data elements in *S*:

$$\text{Median}\{|x_i - \tilde{x}|: x_i \in S\}$$

If R is a range which contains the data elements in *S*, then the MAD of *S* can be calculated in Excel by the array formula:

=MEDIAN(ABS(R-MEDIAN(R)))

Even though the value is presented in a single cell, it is essential that you press **Ctrl-Shft-Enter** to obtain the array value, otherwise the result won't come out correctly.

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2}, then the median absolute deviation = 2 since *S* = {-1, 0, 2, 2, 3, 4, 5, 5} and so the median of *S* is (2+3)/2 = 2.5. Thus, MAD = the median of {3.5, 2.5, 0.5, 0.5, 0.5, 1.5, 2.5, 2.5} = {0.5, 0.5, 0.5, 1.5, 2.5, 2.5, 2.5, 3.5}, that is, (1.5+2.5)/2 = 2.

This metric is less affected by extremes in the data (i.e., the **tails**) because the data in the tails have less influence on the calculation of the median than they do on the mean.

Range

The **range** of a data set *S* is a crude measure of variability and consists simply of the difference between the largest and smallest values in *S*.

If R is a range which contains the data elements in *S*, then the range of *S* can be calculated in Excel by the formula:

=MAX(R) – MIN(R)

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2}, then the range = 5 – (-1) = 6

Interquartile Range

The **interquartile range** (**IQR**) of a data set *S* is calculated as the 75th percentile of *S* minus the 25th percentile. The IQR provides a rough approximation of the variability near the center of the data in *S.*

If R is a range which contains the data elements in *S*, then the IQR of *S* can be calculated in Excel by the formula:

=QUARTILE(R, 3) – QUARTILE(R, 1)

Example: If *S* = {2, 5, -1, 3, 4, 5, 0, 2}, then the IQR = 4.25 – 1.5 = 2.75

As we will see shortly, in Excel 2010/2011/2013 there are two ways to calculate the quartiles using QUARTILE.INC, which is equivalent to QUARTILE and QUARTILE.EXC. Using this second approach, we calculate the IQR for *S* to be 4.75 – 0.5 = 4.25.

The variance, standard deviation, average absolute deviation, and median absolute deviation measure both the variability near the center and the variability in the tails of the distribution of the data. The average absolute deviation and median absolute deviation do not give undue weight to the tails. On the other hand, the range only uses the two most extreme points and the IQR only uses the middle portion of the data.

Ranking

Table 3 summarizes the various ranking functions that are available in all versions of Excel for a data set R. We will describe each of these functions in more detail in the rest of the section, plus we will describe additional ranking functions that are only available in Excel 2010/2013.

Excel Function	Definition	Notes
MAX(R)	The largest value in R (maximum)	=LARGE(R, 1)
MIN(R)	The smallest value in R (minimum)	=SMALL(R, 1)
LARGE(R, n)	The nth largest element in R	LARGE(R, COUNT(R)) = MIN(R)
SMALL(R, n)	The nth smallest element in R	SMALL(R, COUNT(R)) = MAX(R)
RANK(c, R, d)	The rank of element c in R	If d = 0 (or omitted), then the ranking is in decreasing order; otherwise, it is in increasing order
PERCENTILE(R, p)	Element in R at the pth percentile	$0 \le p \le 1$. If PERCENTILE(R,p) = c, then p% of the data

Excel Function	Definition	Notes
		elements are less than c
PERCENTRANK(R, c)	Percentage of elements in R below c	If PERCENTRANK(R,c) = p, then PERCENTILE(R, p) = c
QUARTILE(R, n)	Element in R at the nth quartile	n = 0, 1, 2, 3, 4

Table 3: Ranking functions

MIN, MAX, SMALL, LARGE

MIN(R) = the smallest value in R and **MAX**(R) = the largest value in R.

SMALL(R, n) = the nth smallest value in R and **LARGE**(R, n) = the nth largest value in R.

Here, n can take on any value from 1 to the number of elements in R, that is, COUNT(R).

Example: For range R with data elements {4, 0, -1, 7, 5}:

MIN(R) = -1, MAX(R) = 7.

LARGE(R, 1) = 7, LARGE(R, 2) = 5, LARGE(R, 5) = -1

SMALL(R, 1) = -1, SMALL(R, 2) = 0, LARGE(R, 5) = 7

RANK

All versions of Excel contain the ranking function RANK. For versions of Excel starting with Excel 2010, there is also the equivalent function RANK.EQ.

RANK(c, R, d) = the rank of data element c in R. If d = 0 (or is omitted), then the ranking is in decreasing order, that is, a rank of 1 represents the largest data element in R. If d ≠ 0, then the ranking is in increasing order, and so a rank of 1 represents the smallest element in R.

Example: For range R with data elements {4, 0, -1, 7, 5}:

RANK(7, R) = RANK(7, R, 0) = 1

.

RANK(7, R, 1) = 5

RANK(0, R) = RANK(0, R, 0) = 4

RANK(0, R, 1) = 2

RANK.AVG

Excel's RANK and RANK.EQ functions do not take care of ties. For example, if the range R contains the values {1, 5, 5, 0, 8}, then RANK(5, R) = 2 because 5 is the second-highest ranking element in range R. But 5 is also the third-highest ranking element in the range, and so for many applications, it is useful to consider the ranking to be 2.5, namely, the average of 2 and 3.

Versions of Excel starting with Excel 2010 address this issue by providing a new function **RANK.AVG** which takes the same arguments as RANK but returns the average of equal ranks as previously described.

For versions of Excel prior to Excel 2010, you will need to use the following formula to take care of ties in a similar fashion:

= RANK(c, R) + (COUNTIF(R, c) – 1) / 2

Example: Using the RANK.AVG function, find the ranks of the data in range E17:E23 of Figure 9:

	E	F	G
16		Rank	Reverse
17	1	6	2
18	5	4	4
19	5	4	4
20	0	7	1
21	8	1.5	6.5
22	8	1.5	6.5
23	5	4	4

Figure 9: Average ranking

The result is shown in column F of Figure 9. For example, the average rank of 8 (cell E21 or E22) is 1.5 as shown in cell F21 (or F22) and calculated using the formula =RANK.AVG(E21,E17:E23). If instead you want the ranking in the reverse order (where the lowest value gets rank 1), then the results are shown in column G. This time using the formula =RANK.AVG(E21,E17:E23,1), we see that the rank of 8 is 6.5 as shown in cell G21.

PERCENTILE

For any percentage p (that is, where $0 \le p \le 1$ or equivalently $0\% \le p \le 100\%$), **PERCENTILE**(R, p) = the element at the pth **percentile**. This means that if PERCENTILE(R, p) = c, then p% of the data elements in R are less than c.

If p = k/(n–1) for some integer value k = 0, 1, 2, ... n–1 where n = COUNT(R), then PERCENTILE(R, p) = SMALL(R, k+1) = the k+1th element in R. If p is not a multiple of 1/(n-1), then the PERCENTILE function performs a linear interpolation as described in the following example.

Example: For range R with data elements {4, 0, -1, 7, 5}, R's five data elements divide the range into four intervals of size 25%, that is, 1/(5-1) = .25. Thus:

PERCENTILE(R, 0) = -1 (the smallest element in R)
PERCENTILE(R, .25) = 0 (the second smallest element in R)
PERCENTILE(R, .5) = 4 (the third smallest element in R)
PERCENTILE(R, .75) = 5 (the fourth smallest element in R)
PERCENTILE(R, 1) = 7 (the fifth smallest element in R)

For other values of p, we need to interpolate. For example:

PERCENTILE(R, .8) = 5 + (7 – 5) * (0.8 – 0.75) / 0.25 = 5.4

PERCENTILE(R, .303) = 0 + (4 – 0) * (0.303 – 0.25) / 0.25 = .85

Of course, Excel's PERCENTILE function calculates all these values automatically without you having to figure things out.

PERCENTILE.EXC

Starting with Excel 2010, Microsoft introduced an alternative version of the percentile function named **PERCENTILE.EXC**.

If n = COUNT(R), then for any integer k with 1 ≤ k ≤ n

PERCENTILE.EXC(R, k/(n+1)) = SMALL(R, k), that is, the kth smallest element in range R.

For 0 < p < 1, if p is not a multiple of 1/(n+1), then PERCENTILE.EXC(R, p) is calculated by taking a linear interpolation between the corresponding values in R. For p < 1/(n+1) or p > n/(n+1), no interpolation is possible and so PERCENTILE.EXC(R, p) returns an error.

Excel 2010/2013 also provides the new function **PERCENTILE.INC** which is equivalent to PERCENTILE.

Example: Find the 0 – 100 percentiles in increments of 10% for the data in Figure 10 using both PERCENTILE (or PERCENTILE.INC) and PERCENTILE.EXC:

	A	B	C	D	E	F	G	H	I	J	K	L	M
3	Scores	54	67	34	54	94	55	32	45	87	64	39	60

Figure 10: Data for percentile calculations

The result is shown in Figure 11. For example, the score at the 60th percentile is 58 (cell P10) using the formula =PERCENTILE(B3:M3,O10), while it is 59 (cell S10) using the formula =PERCENTILE.EXC(B3:M3,R10).

	O	P	Q	R	S
1	PERCENTILE			PERCENTILE.EXC	
2					
3	*Percentile*	*Score*		*Percentile*	*Score*
4	0%	32		0%	#NUM!
5	10%	34.5		10%	32.6
6	20%	40.2		20%	37
7	30%	47.7		30%	44.4
8	40%	54		40%	54
9	50%	54.5		50%	54.5
10	60%	58		60%	59
11	70%	62.8		70%	64.3
12	80%	66.4		80%	75
13	90%	85		90%	91.9
14	100%	94		100%	#NUM!

Figure 11: PERCENTILE vs. PERCENTILE.EXC

PERCENTRANK and PERCENTRANK.EXC

PERCENTRANK (or **PERCENTRANK.INC**) and **PERCENTRANK.EXC** are the inverses of PERCENTILE and PERCENTILE.EXC. Thus, PERCENTRANK(R, c) = the value p such that PERCENTILE(R, p) = c. Similarly, PERCENTRANK.EXC(R, c) = the value p such that PERCENTILE.EXC(R, p) = c.

Example: Referring to Figures 10 and 11, we have:

PERCENTRANK(B3:M3,54) = .4

PERCENTRANK.EXC(B3:M3,S12) = .8

QUARTILE and QUARTILE.EXC

For any integer, n = 0, 1, 2, 3 or 4, **QUARTILE**(R, n) = PERCENTILE(R, n/4). If c is not an integer but $0 \le c \le 4$, then QUARTILE(R, c) = QUARTILE(R, INT(c)). Thus:

QUARTILE(R, 0) = PERCENTILE(R, 0) = MIN(R)
QUARTILE(R, 1) = PERCENTILE(R, .25)
QUARTILE(R, 2) = PERCENTILE(R, .5) = MEDIAN(R)
QUARTILE(R, 3) = PERCENTILE(R, .75)
QUARTILE(R, 4) = PERCENTILE(R, 1) = MAX(R)

Example: For range R with data elements {4, 0, -1, 7, 5}:

QUARTILE(R, 0) = PERCENTILE(R, 0) = -1
QUARTILE(R, 1) = PERCENTILE(R, .25) = 0
QUARTILE(R, 2) = PERCENTILE(R, .5) = 4
QUARTILE(R, 3) = PERCENTILE(R, .75) = 5
QUARTILE(R, 4) = PERCENTILE(R, 1) = 7

Starting with Excel 2010, Microsoft introduced an alternative version of the quartile function named **QUARTILE.EXC**. This function is defined so that QUARTILE.EXC(R, n) = PERCENTILE.EXC(R, n/4). Excel 2010/2013 also provides the new function QUARTILE.INC which is equivalent to QUARTILE.

Shape of the Distribution

Symmetry and Skewness

Skewness as a measure of symmetry. If the skewness of *S* is 0, then the distribution of the data in *S* is perfectly symmetric. If the skewness is negative, then the graph is skewed to the left. If the skewness is positive, then the graph is skewed to the right (see Figures 12 and 14 for examples).

Excel calculates the skewness of sample *S* using the formula:

$$\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2))s^3}$$

where $\bar{x}$ is the mean and s is the standard deviation of *S*. To avoid division by zero, this formula requires that n > 2.

When a distribution is symmetric, the mean = median; when the distribution is positively skewed, mean > median, and when the distribution is negatively skewed, mean < median.

Excel provides the **SKEW** function as a way to calculate the skewness of *S.* For example, if R is a range in Excel containing the data elements in *S,* then SKEW(R) = the skewness of *S*.

There is also a population version of the skewness given by the formula:

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

This version has been implemented in Excel 2013 using the function **SKEW.P**.

For versions of Excel prior to Excel 2013, you can use the formula:

SKEW(R)*(n–2)/SQRT(n(n–1))

instead of SKEW.P(R) where R contains the data in $S = \{x_1, \ldots, x_n\}$ and n = COUNT(R).

Kurtosis

Kurtosis as a measure of peakedness or flatness. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

Kurtosis is calculated in Excel as follows:

$$\frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where $\bar{x}$ is the mean and s is the standard deviation of *S*. To avoid division by zero, this formula requires that $n > 3$.

Excel provides the **KURT** function as a way to calculate the kurtosis of *S*. For example, if R is a range in Excel containing the data elements in *S,* then KURT(R) = the kurtosis of *S*.

Graphic Illustrations

We now look at an example of these concepts using the chi-square distribution:

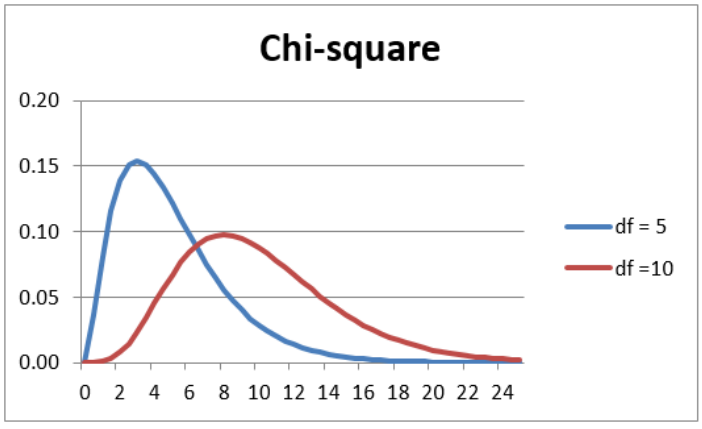


Figure 12: Example of skewness and kurtosis

Figure 12 contains the graph of two chi-square distributions. We will study the chi-square distribution later but, for now, note the following values for the kurtosis and skewness:

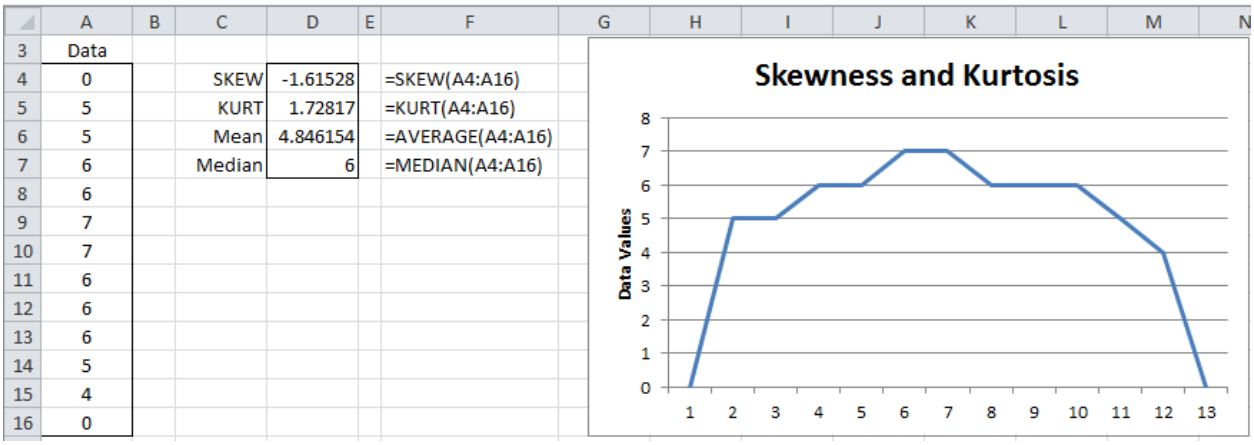
	df = 5	*df = 10*
Kurtosis	5.4	4.2
Skewness	1.264911	0.894427

Figure 13: Comparison of skewness and kurtosis

The red curve (df = 10) is flatter than the blue curve (df = 5) which is reflected in the fact that the kurtosis value of the blue curve is lower.

Both curves are asymmetric and skewed to the right (that is, the fat part of the curve is on the left). This is consistent with the fact that the skewness for both is positive. But the blue curve is more skewed to the right, which is consistent with the fact that the skewness of the blue curve is larger.

Example: Calculate the skewness and kurtosis for the data in range A4:A16 of Figure 14:



	A	B	C	D	E	F
3	Data					
4	0		SKEW	-1.61528		=SKEW(A4:A16)
5	5		KURT	1.72817		=KURT(A4:A16)
6	5		Mean	4.846154		=AVERAGE(A4:A16)
7	6		Median	6		=MEDIAN(A4:A16)
8	6					
9	7					
10	7					
11	6					
12	6					
13	6					
14	5					
15	4					
16	0					

Figure 14: Calculation of skewness and kurtosis in Excel

Descriptive Statistics Data Analysis Tool

Excel provides the **Descriptive Statistics** data analysis tool which produces a summary of the key descriptive statistics for a data set.

Example: Produce a table of the most common descriptive statistics for the scores in column A of Figure 15:

	A	B	C	D
1	**Descriptive Statistics data analysis tool**			
2				
3	Scores		*Scores*	
4	23			
5	38		Mean	30.8182
6	45		Standard Error	4.93193
7	21		Median	23
8	17		Mode	21
9	21		Standard Deviation	16.3574
10	8		Sample Variance	267.564
11	61		Kurtosis	-0.5471
12	21		Skewness	0.62514
13	52		Range	53
14	32		Minimum	8
15			Maximum	61
16			Sum	339
17			Count	11

Figure 15: Output from descriptive statistics data analysis tool

The output from the tool is shown on the right side of Figure 15. To use the tool, select **Data > Analysis|Data Analysis** and select **Descriptive Statistics**. A dialog box appears as in Figure 16:

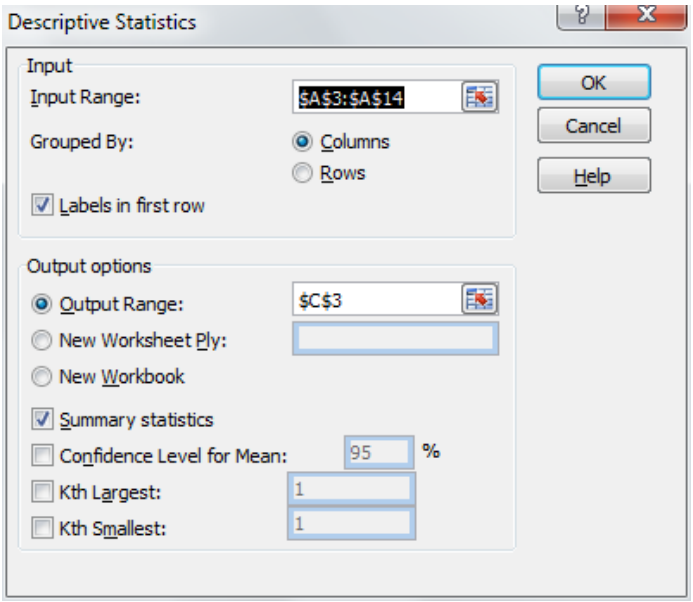


Figure 16: Descriptive statistics dialog box

Now click **Input Range** and highlight the scores in column A (i.e., cells A3:A14). If you include the heading as is done here, check **Labels in first row**. Since we want the output to start in cell C3, click **Output Range** and insert C3 (or click cell C3). Finally, click **Summary statistics** and press **OK**.

Note that, had we checked the **Kth Largest** check box, the output would have also contained the value for LARGE(A4:A14,k) where k is the number we insert in the box to the right of the label **Kth Largest**. Similarly, checking the **Kth Smallest** check box outputs SMALL(A4:A14,k). The **Confidence Interval for Mean** option generates a confidence interval using the t distribution as explained in Chapter 7.

Graphical Representations of Data

Frequency Tables

When you have a lot of data, it can be convenient to put the data in bins, usually of equal size, and then create a graph of the number of data elements in each bin. Excel provides the **FREQUENCY**(R1, R2) array function for doing this, where R1 = the input array and R2 = the bin array.

To use the FREQUENCY array function, enter the data into the worksheet and then enter a bin array. The bin array defines the intervals that make up the bins. For example, if the bin array = 10, 20, 30, then there are four bins, namely, data with values $x \le 10$, data with values x where $10 < x \le 20$, data with values x where $20 < x \le 30$, and, finally, data with values $x > 30$. The FREQUENCY function simply returns an array consisting of the number of data elements in each of the bins.

Example: Create a frequency table for the 22 data elements in the range A4:B14 of Figure 17 based on the bin array D4:D7 (the text "over 80" in cell D8 is not part of the bin array):

	A	B	C	D	E
1	**Frequency**				
2					
3	Scores			Bins	Count
4	34	20		20	5
5	45	34		40	11
6	23	29		60	2
7	22	12		80	3
8	7	72		over 80	1
9	34	23			
10	9	10			
11	66	34			
12	29	23			
13	67	24			
14	44	90			

Figure 17: Example of the FREQUENCY function

To produce the output, highlight the range E4:E8 (i.e., a column range with one more cell than the number of bins) and enter the formula:

=FREQUENCY(A4:B11,D4:D7)

Since this is an array formula, you must press **Ctrl-Shft-Enter**. Excel now inserts frequency values in the highlighted range. Here, E4 contains the number of data elements in the input range with value in the first bin (i.e., data elements whose value is ≤ 20). Similarly, E5 contains the number of data elements in the input range with value in the second bin (i.e., data elements whose value is > 20 and ≤ 40). The final output cell (E8) contains the number of data elements in the input range with value > the value of the final bin (i.e., > 80 for this example).

Histogram

A **histogram** is a graphical representation of the output from the FREQUENCY function previously described. You can use Excel's chart tool to graph the output or, alternatively, you can use the Histogram data analysis tool to directly accomplish this.

To use Excel's **Histogram** data analysis tool, you must first establish a bin array (as for the FREQUENCY function previously described) and then select the Histogram data analysis tool. In the dialog box that is displayed, you then specify the input data (**Input Range**) and bin array (**Bin Range**). You have the option to include the labels for these ranges (in which case you click **Labels**).

For the data in Figure 17, the **Input Range** is A4:B14 and the **Bin Range** is D4:D7 (with the Labels check box left unchecked). The output is displayed in Figure 18:

Bin	*Frequency*
20	5
40	11
60	2
80	3
More	1

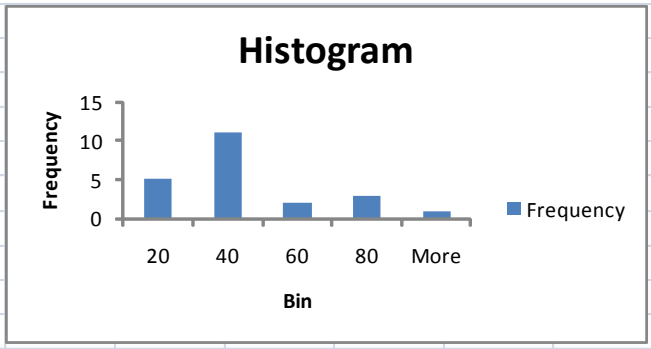


Figure 18: Example of the histogram data analysis

You should exercise caution whenever you create histograms in order to present the data in a clear and accurate way. For most purposes, it is important that the intervals be equal in size (except for an unbounded first and/or last interval). Otherwise, a distorted picture of the data may be presented. To avoid this problem, equally-spaced intervals should generally be used.

Box Plots

Another way to characterize data is via a **box plot**. Specifically, a box plot provides a pictorial representation of the following statistics: maximum, 75th-percentile, median (50th-percentile), 25th-percentile, and minimum.

As we shall see later, box plots are especially useful when comparing samples and testing whether or not data is symmetric.

Example: Create box plots for the three brands shown in range A3:C13 of Figure 19 using Excel's charting capabilities:

	A	B	C	D	E	F	G	H
1	**Boxplot (aka box and whiskers plot)**							
2								
3	Brand A	Brand B	Brand C			Brand A	Brand B	Brand C
4	1020	840	1430		Min	380	300	430
5	1560	940	1750		Q1-Min	142.5	185	342.5
6	560	780	870		Med-Q1	202.5	220	132.5
7	780	650	920		Q3-Med	237.5	120	332.5
8	990	720	1300		Max-Q3	597.5	1025	512.5
9	670	430	890					
10	510	1850	740					
11	490	300	720					
12	380	360	430					
13	880	690	1050					

Figure 19: Box plot data

Select the range containing the data including the headings (A3:C13). Now, create the table in the range E3:H8. The output in column F corresponds to the raw data from column A. Column G corresponds to column B and column H corresponds to column C. In fact, once you construct the formulas for the range F4:F8, you can fill in the rest of the table by highlighting the range F4:H8 and pressing **Ctrl-R**.

The formulas for the cells in the range F4:F8 are as follows:

Cell	Content
F4	=MIN(A4:A13)
F5	=QUARTILE(A4:A13,1)-F4
F6	=MEDIAN(A4:A13)-QUARTILE(A4:A13,1)
F7	=QUARTILE(A4:A13,3)-MEDIAN(A4:A13)
F8	=MAX(A4:A13)-QUARTILE(A4:A13,3)

Table 4: Formulas in the box plot table

Note that, alternatively, you can replace the QUARTILE function in Table 4 by the QUARTILE.EXC function.

Once you have constructed the table, you can create the corresponding box plot as follows:

1. Select the data range E3:H7. Notice that the headings are included on the range but not the last row
2. Select **Insert > Charts|Column > Stacked Column**
3. Select **Design > Data|Switch Row/Column** if necessary so that the x axis represents the brands
4. Select the lowest data series in the chart (i.e., **Min**) and set fill to No fill (and, if necessary, set border color to No Line) to remove the lowest boxes. This is done by right-clicking on any of the three **Min** data series boxes in the chart and selecting **Format Data Series**. In the resulting dialog box, select **Fill|No fill**
5. Repeat the previous steps for the lowest visible data series (i.e., **Q1-Min**). That is, right-click on the Q1-Min data series and select **Format Data Series… > Fill|No fill**. Alternatively, right-click on the Q1-Min data series and press **Ctl-Y**.
6. With the Q1-Min data series still selected, select **Layout > Analysis|Error Bars > More Error Bar Options**. In the resulting dialog box (Vertical Error Bars menu), click **Minus** and **Percentage**, and insert a percentage error of **100%**
7. Click the **Q3-Med** data series (the uppermost one) and select **Layout > Analysis|Error Bars > More Error Bar Options**. In the resulting dialog box (Vertical Error Bars menu), click **Plus** and **Custom**, and then click **Specify Value**. Now specify the range F8:H8, that is, the last row of the table you previously created, in the dialog box that appears (in the **Positive Error Values** field).
8. Remove the legend by selecting **Layout > Labels|Legend > None**

The resulting box plot is:

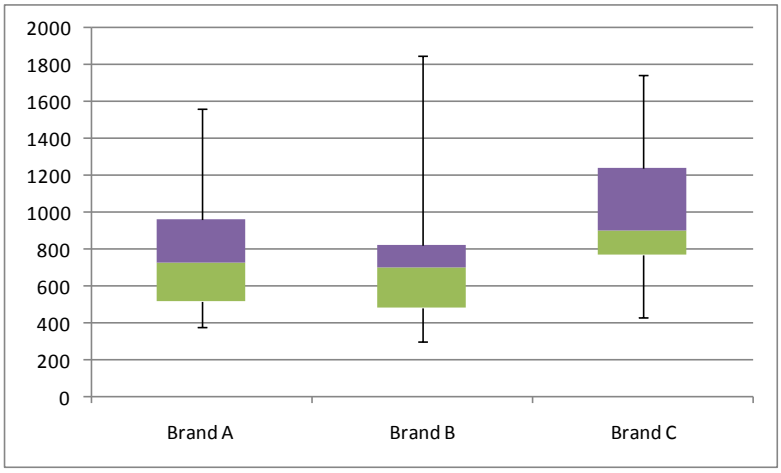


Figure 20: Box plot

For each sample, the box plot consists of a rectangular box with one line extending upward and another line extending downward (usually called **whiskers**). The box itself is divided into two parts. In particular, the meaning of each element in the box plot is described in Table 5:

Element	Meaning
Top of upper whisker	Maximum value of the sample
Top of box	75% percentile of the sample
Line through the box	Median of the sample
Bottom of the box	25% percentile of the sample
Bottom of the lower whisker	Minimum of the sample

Table 5: Box plot elements

From the box plot (see Figure 20), we can see that the scores for Brand C tend to be higher than for the other brands, and those for Brand B tend to be lower. We also see that the distribution of Brand A is pretty symmetric, at least in the range between the 1st and 3rd quartiles, although there is some asymmetry for higher values (or there is potentially an outlier). Brands B and C look less symmetric. Because of the long upper whisker (especially with respect to the box), Brand B may have an outlier.

The approach described in this section works perfectly for non-negative data. When a data set has a negative value, you need to add -MIN(R) to all the data elements where R is the data range containing the data. This will shift the y-axis up by –MIN(R). For example, if R ranges from -10 to 20, then the range described in the chart will range from 0 to 30.

We can also convert the box plot to a horizontal representation of the data (as in Figure 21) by clicking the chart and selecting **Insert > Charts|Bar > Stacked Bar**:

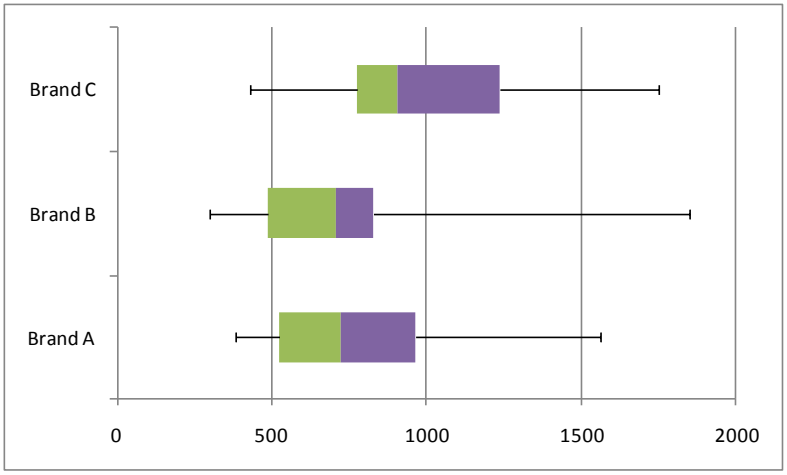


Figure 21: Horizontal box plot

Outliers

One problem that we face when analyzing data is the presence of **outliers** or data elements that are much bigger or much smaller than the other data elements.

For example, the mean of the sample {2, 3, 4, 5, 6} is 4 while the mean of {2, 3, 4, 5, 60} is 14.4. The appearance of the 60 completely distorts the mean in the second sample. Some statistics, such as the median, are more resistant to such outliers. In fact, the median for both samples is 4. For this example, it is obvious that 60 is a potential outlier.

One approach for dealing with outliers is to throw away data that is either too big or too small. Excel provides the TRIMMEAN function for dealing with this issue:

> **TRIMMEAN**(R, p) – calculates the mean of the data in the range R after first throwing away p% of the data, half from the top and half from the bottom. If R contains n data elements and k = the largest whole number ≤ np/2, then the k largest items and the k smallest items are removed before calculating the mean.

Example: Suppose R = {5, 4, 3, 20, 1, 4, 6, 4, 5, 6, 7, 1, 3, 7, 2}. Then TRIMMEAN(R, 0.2) works as follows. Since R has 15 elements, k = INT(15 * .2 / 2) = 1. Thus, the largest element (20) and the smallest element (1) are removed from R to get R′ = {5, 4, 3, 4, 6, 4, 5, 6, 7, 1, 3, 7, 2}. TRIMMEAN now returns the mean of this range, namely, 4.385 instead of the mean of R which is 5.2.

Missing Data

Another problem we face when collecting data is that some data may be missing. For example, when conducting a survey with 10 questions, perhaps some of the people who take the survey don't answer all 10 questions.

The principal way Excel deals with missing data is to ignore the missing data elements. For example, in the various functions in this chapter (AVERAGE, VAR.S, RANK, etc.), any blank or non-numeric cells are simply ignored.

Chapter 4 Distributions

Discrete Distributions

The (**probability**) **frequency function** f, also called the **probability density function** (abbreviated **pdf**), of a discrete random variable x is defined so that for any value t in the domain of the random variable:

$$f(t) = P(x = t)$$

where $P(x = t)$ = the probability that x assumes the value t.

The corresponding (**cumulative**) **distribution function** $F(x)$ is defined by

$$F(t) = \sum_{u \le t} f(u)$$

for any value t in the domain of the random variable x.

For any discrete random variable defined over the domain *S* with frequency function f and distribution function F:

$$0 \le f(t) \le 1 \text{ for all } t \text{ in } S$$

$$\sum_{t \in S} f(t) = 1$$

$$F(t) = P(x \le t)$$

$$P(t_1 < x \le t_2) = F(t_2) - F(t_1)$$

A frequency function can be expressed as a table or a bar chart as described in the following example.

Example: Find the distribution function for the frequency function given in columns A and B in Figure 22. Also, show the graph of the frequency and distribution functions:

	A	B	C	D
1	Frequency/Distribution Functions			
2				
3	x	f(x)	F(x)	
4	1	0.12	0.12	
5	2	0.25	0.37	
6	3	0.08	0.45	
7	4	0.14	0.59	
8	5	0.09	0.68	
9	6	0.18	0.86	
10	7	0.09	0.95	
11	8	0.05	1.00	

Figure 22: Table defining the frequency and distribution functions

Given the frequency function $f(x)$ defined in the range B4:B11 of Figure 22, we can calculate the distribution function $F(x)$ in the range C4:C11 by putting the formula =B4 in cell C4 and the formula =B5+C4 in cell C5, and then copying this formula into cells C6 to C11 (for example, by highlighting the range C5:C11 and pressing **Ctrl-D**).

From the table in Figure 22, we can create the charts in Figure 23:

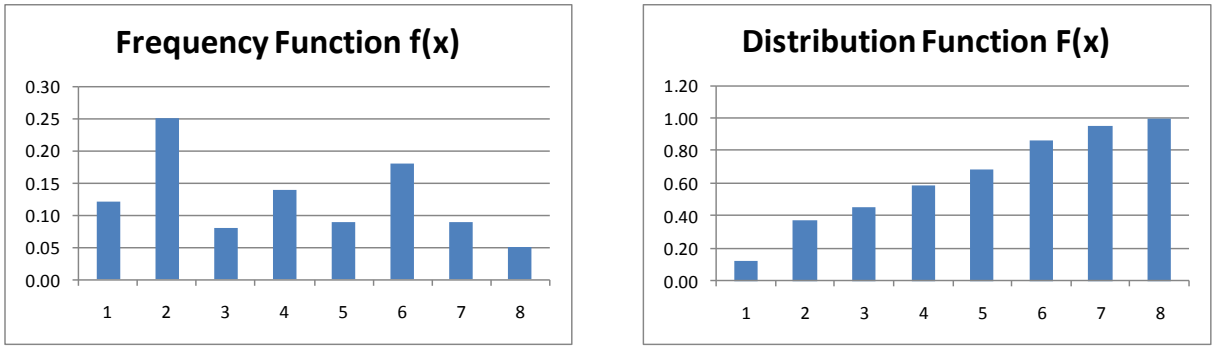


Figure 23: Charts of frequency and distribution functions

Continuous Distributions

While for a discrete random variable x, the probability that x assumes a value between a and b (exclusive) is given by $\sum_{a<x<b} f(x)$, the frequency function f of a continuous random variable can assume an infinite number of values (even in a finite interval), and so we can't simply sum up the values in the ordinary way. For continuous variables, the equivalent formulation is that the probability that x assumes a value between a and b is given by $Area_{a<x<b}\, f(x)$, that is, the area under the graph of $y = f(x)$ bounded by the x-axis and the lines $x = a$ and $x = b$.

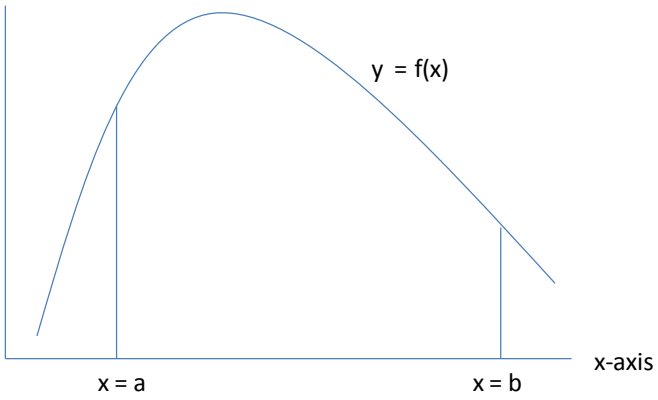


Figure 24: Area under the curve $y = f(x)$

For a continuous random variable x, f is a **frequency function**, also called the **probability density function** (**pdf**), provided:

f is the **frequency function**, more commonly called the **probability density function**, for a particular random variable, provided the area of the region indicated in Figure 24 represents the probability that x assumes a value between a and b, inclusively. f only takes non-negative values and the area between the curve if $y = f(x)$ and the x-axis is 1.

The corresponding (**cumulative**) **distribution function** F is defined by:

$$F(x) = \text{Area}_{t<x} f(t)$$

For any continuous random variable with distribution function F:

$$F(b) = P(x < b)$$

$$F(b) - F(a) = P(a < x < b)$$

Note that the probability that x takes any particular value a is not $f(a)$. In fact, for any specific value a, the probability that x takes the value a is considered to be 0.

Essentially, the area under a curve is a way of summing when dealing with an infinite range of values in a continuum. For those of you familiar with calculus, $\text{Area}_{a<x<b} f(x) = \int_a^b f(x)dx$.

Excel Distribution Functions

The following table provides a summary of the various distribution functions provided by Excel. The entries labelled PDF/CDF specify the probability density function (where the last argument for the function is FALSE), as well as the (left-tailed) cumulative distribution function (where the last argument of the function is TRUE):

Distribution	PDF/CDF	Inverse	Right/Two-tailed	Test
Beta	BETA.DIST	BETA.INV		
Binomial	BINOM.DIST	BINOM.INV		
Chi-square	CHISQ.DIST	CHISQ.INV	CHISQ.DIST.RT CHISQ.INV.RT	CHISQ.TEST
Exponential	EXPON.DIST			
F	F.DIST	F.INV	F.DIST.RT F.INV.RT	F.TEST
Gamma	GAMMA.DIST	GAMMA.INV		
Hyper-geometric	HYPGEOM.DIST			
Lognormal	LOGNORM.DIST	LOGNORM.INV		
Negative Binomial	NEGBINOM.DIST			
Normal	NORM.DIST	NORM.INV		
Poisson	POISSON.DIST			
Standard Normal	NORM.S.DIST	NORM.S.INV		Z.TEST
Student's t	T.DIST	T.INV	T.DIST.RT T.DIST.2T T.INV.2T	T.TEST

Distribution	PDF/CDF	Inverse	Right/Two-tailed	Test
Weibull	WEIBULL.DIST			

Table 6: Excel distribution functions

The functions listed in Table 6 are not available for versions of Excel prior to Excel 2010. For earlier releases of Excel, a similar collection of functions is available as summarized in Table 7:

Table 7: Distribution functions for versions of Excel prior to Excel 2010

Distribution	CDF	PDF	Inverse	Test
Beta	BETADIST		BETAINV	
Binomial	BINOMDIST	BINOMDIST	CRITBINOM	
Chi Square	CHIDIST		CHIINV	CHITEST
Exponential	EXPONDIST	EXPONDIST		
F	FDIST		FINV	FTEST
Gamma	GAMMADIST		GAMMAINV	
Hypergeometric		HYPGEOMDIST		
Lognormal	LOGNORMDIST		LOGINV	
Negative Binomial		NEGBINOMDIST		
Normal	NORMDIST	NORMDIST	NORMINV	
Poisson	POISSON	POISSON		

Distribution	CDF	PDF	Inverse	Test
Standard Normal	NORMSDIST		NORMSINV	ZTEST
Student's T	TDIST		TINV	TTEST
Weibull		WEIBULL		

Chapter 5 Normal Distribution

Normal Distribution

Basic Concepts

The probability density function of the **normal distribution** is defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where e is the constant 2.7183..., and π is the constant 3.1415…

The normal distribution is completely determined by the parameters μ and σ. It turns out that μ is the mean of the normal distribution and σ is the standard deviation.

The graph of the normal distribution has the shape of a bell curve as shown in Figure 25:

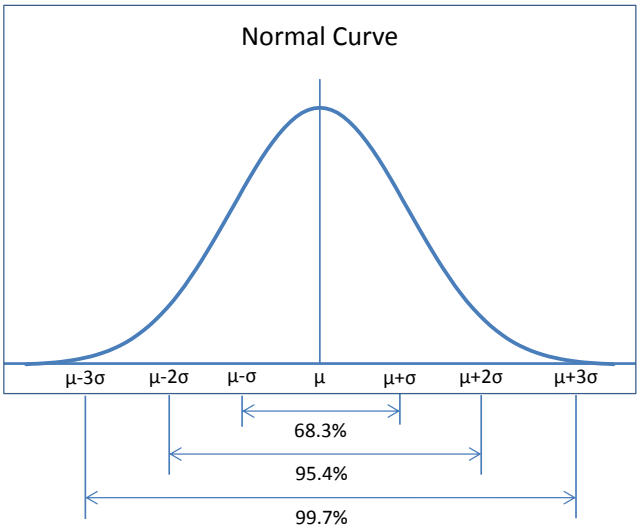


Figure 25: Normal curve

Some characteristics of the normal distribution are as follows:

- Mean = median = mode = μ
- Standard deviation = σ
- Skewness = kurtosis = 0

The function is symmetric about the mean with inflection points (i.e., the points where the curve changes from concave up to concave down, or from concave down to concave up) at $x = \mu \pm \sigma$.

As can be seen from Figure 25, the area under the curve in the interval $\mu - \sigma < x < \mu + \sigma$ is approximately 68.26 percent of the total area under the curve. The area under the curve in the interval $\mu - 2\sigma < x < \mu + 2\sigma$ is approximately 95.44 percent of the total area under the curve, and the area under the curve in the interval $\mu - 3\sigma < x < \mu + 3\sigma$ is approximately 99.74 percent of the area under the curve.

As we shall see, the normal distribution occurs frequently and is very useful in statistics.

Excel Functions

Excel provides the following functions regarding the normal distribution:

NORM.DIST(x, μ, σ, cum) where cum takes the values TRUE and FALSE

NORM.DIST(x, μ, σ, FALSE) = probability density function value at x for the normal distribution with mean μ and standard deviation σ

NORM.DIST(x, μ, σ, TRUE) = cumulative probability distribution function value at x for the normal distribution with mean μ and standard deviation σ

NORM.INV(p, μ, σ) is the inverse of NORM.DIST(x, μ, σ, TRUE), that is,
NORM.INV(p, μ, σ) = the value x such that NORM.DIST(x, μ, σ, TRUE) = p

These functions are not available in versions of Excel prior to Excel 2010. Instead, these versions of Excel provide **NORMDIST** which is equivalent to NORM.DIST and **NORMINV**, which is equivalent to NORM.INV.

Example

Create a graph of the distribution of IQ scores using the Stanford-Binet Intelligence Scale.

This distribution is known to have a normal distribution with mean 100 and standard deviation 16. To create the graph, we first must create a table with the values of the probability density function $f(x)$ for x = 50, 51, …, 150. This table begins as shown on the left side of Figure 26:

	A	B
1	**Normal distribution**	
2		
3	IQ Score	f(x)
4	50	0.0002
5	51	0.0002
6	52	0.0003
7	53	0.0003
8	54	0.0004
9	55	0.0005
10	56	0.0006
11	57	0.0007
12	58	0.0008
13	59	0.0009
14	60	0.0011
15	61	0.0013
16	62	0.0015
17	63	0.0017
18	64	0.0020
19	65	0.0023
20	66	0.0026
21	67	0.0030
22	68	0.0034
23	69	0.0038

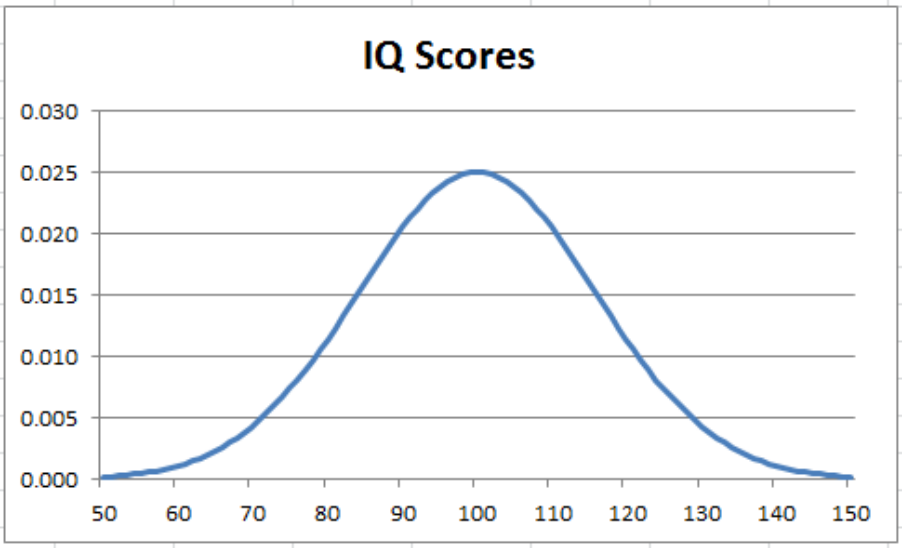


Figure 26: Distribution of IQ scores

We only show the first 20 entries but, in total, there are 101 entries occupying the range A4:B104. The values in column B are calculated using the NORM.DIST function. For example, the formula in cell B4 is:

=NORM.DIST(A4,100,16,FALSE)

We can draw the chart on the right side of Figure 26 by highlighting the range B4:B104 and then selecting **Insert > Charts|Line.** The display of the chart can be improved as described elsewhere in this book.

Standard Normal Distribution

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.

To convert a random variable x with normal distribution with mean μ and standard deviation σ to standard normal form, you use the following linear transformation:

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

The resulting random variable is called a **z-score**. Excel provides the following function for calculating the value of z from x, μ and σ:

STANDARDIZE$(x, \mu, \sigma) = \frac{x-\mu}{\sigma}$

In addition, Excel provides the following functions regarding the standard normal distribution:

NORM.S.DIST(z, cum) = NORM.DIST(z,0,1,cum)

NORM.S.INV(p) = NORM.INV(p,0,1)

These functions are not available in versions of Excel prior to Excel 2010. Instead, these versions of Excel provide **NORMSDIST**(z) which is equivalent to NORM.S.DIST(z, TRUE), and **NORMSINV**(p) which is equivalent to NORM.S.INV(p). For the pdf, the formula equivalent to NORM.S.DIST(z, FALSE) is NORMDIST(z, 0, 1, FALSE).

Lognormal Distribution

A random variable x is **lognormally** distributed provided the natural log of x, $\ln x$, is normally distributed.

Note that the lognormal distribution is not symmetric but is skewed to the right. If you have data that is skewed to the right that fits the lognormal distribution, you may be able to access various tests described elsewhere in this book that require data to be normally distributed.

Excel provides the following two functions:

LOGNORM.DIST(x, μ, σ, cum) = the lognormal cumulative distribution function with mean μ and standard deviation σ at x if cum = TRUE, that is, NORMDIST(LN(x), μ, σ, cum), and the probability density function of the lognormal distribution if cum = FALSE.

LOGNORM.INV(p, μ, σ) = EXP(NORM.INV(p, μ, σ)

These functions are not available in versions of Excel prior to Excel 2010. Instead, these versions of Excel provide LOGNORMDIST(x, μ, σ) which is equivalent to LOGNORM.DIST(x, μ, σ, TRUE), and **LOGINV**(p, μ, σ) which is equivalent to LOGNORM.INV(p, μ, σ).

Sample Distributions

Basic Concepts

Consider a random sample $x_1, x_2, \ldots, x_n$ from a population. As previously described, the mean of the sample (called the **sample mean**) is defined as:

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

$\bar{x}$ can be considered to be a number representing the mean of the actual sample taken but it can also be considered as a random variable representing the mean of any sample of size n from the population.

If $x_1, x_2, \ldots, x_n$ is a random sample from a population with mean μ and standard deviation σ, then it turns out that the mean of the random variable $\bar{x}$ is μ and the standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$.

When the population is normal, we have the following stronger result, namely, that the sample mean $\bar{x}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

Central Limit Theorem

Provided n is large enough (generally about 30 or more), even for samples not coming from a population with a normal distribution, the sample mean $\bar{x}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

This result is called the **Central Limit Theorem** and is one of the most important and useful results in statistics.

Hypothesis Testing

We will now briefly introduce the concept of **hypothesis testing** using the following example.

Example: A company selling electric motors claims that the average life for its motors is at least 120 months. One of its clients wanted to verify this claim by testing a random sample of 48 motors as described in range A3:F10 of Figure 27. Is the company's claim correct?

	A	B	C	D	E	F	G	H	I	J	K
3	157	93	119	151	116	155		sample size	48		=COUNT(A3:F10)
4	131	103	94	111	157	129		hyp mean	120		
5	92	118	104	90	138	127		sample mean	125.9375		=AVERAGE(A3:F10)
6	145	151	133	83	139	148		std dev	23.95955		=STDEV(A3:F10)
7	158	93	135	144	121	118		std err	3.458263		=I6/SQRT(I3)
8	102	153	139	152	85	152		p-value	0.042998		=1-NORM.DIST(I5,I4,I7,TRUE)
9	96	116	121	154	141	96					
10	120	134	134	160	84	153		z test	0.042998		=ZTEST(A3:F10,I4)

Figure 27: One-sample testing of the mean

We first note that the average life of the motors in the sample is 125.9375, which is more than 120 months. But is this a **statistically significant** result or merely due to chance? Keep in mind that, even if the company's claim is false, any given sample may have a mean larger than 120 months.

Now, let's assume that the population of motors is distributed with mean μ and standard deviation σ. To perform the analysis, we first define the **null hypothesis** as follows:

H_0: μ < 120

The complement of the null hypothesis (called the **alternative hypothesis**) is therefore:

H_1: μ ≥ 120

Usually in an analysis, we are actually testing the validity of the alternative hypothesis by testing whether or not to reject the null hypothesis. When performing such analyses, there is some chance that we will reach the wrong conclusion. There are two types of **errors**:

- **Type I** – H_0 is rejected even though it is true (**false positive**)
- **Type II** – H_0 is not rejected even though it is false (**false negative**)

The acceptable level of a Type I error (called the **significance level**) is designated by **alpha** (α) while the acceptable level of a Type II error is designated **beta** (β).

Typically, a significance level of α = .05 is used (although, sometimes other levels such as α = .01 may be employed). This means that we are willing to tolerate up to five percent of type I errors, that is, we are willing to accept the fact that, in 1 out of every 20 samples, we reject the null hypothesis even though it is true.

For our analysis, we use the Central Limit Theorem. Since the sample size is sufficiently large, we can assume that the sample means have a normal distribution, with mean μ and standard deviation $\sigma/\sqrt{n}$, as depicted in Figure 28:

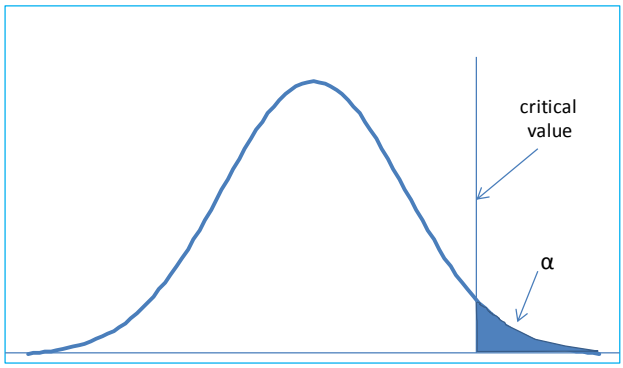


Figure 28: Right-tailed critical region

Keep in mind that this curve represents the distribution of all possible values for the sample mean. We need to perform our test based on the observed sample mean of 125.9375.

The next step in the analysis is to determine the probability that the observed sample mean has a value of 125.9375 or higher, under the assumption that the null hypothesis is true. If this probability is low (i.e., less than α), then we will have grounds for **rejecting** the null hypothesis. Otherwise, we would not reject the null hypothesis (which is also called **retaining** the null hypothesis).

Retaining the null hypothesis is not the same as positing that it is true. It merely means that based on the one observed sample, we don't have sufficient evidence for rejecting it.

The **p-value** (that is, the **probability value**) is the value p of the statistic used to test the null hypothesis. If p < α, then we reject the null hypothesis. For our problem, the p-value is the probability that the observed sample mean has a value of 125.9375 or higher, under the assumption that the null hypothesis is true. This is the same as testing whether the p-value falls in the shaded area of Figure 28 (i.e., the **critical region**) in what is called a **one-tailed test**.

The p-value for our example is calculated in cell I8 of Figure 28. Since p-value = .042998 < .05 = α, we see that the observed value does fall in the critical region, and so we reject the null hypothesis. Therefore, we can show that we have sufficient grounds for accepting the company's claim that their motors last at least 120 months on average.

Note that the population standard deviation σ is unknown. For the analysis shown in Figure 28, we made the assumption that the sample standard deviation is a good estimate of the population standard deviation. For large enough samples, this is a reasonable assumption, although it introduces additional error. In Chapter 6, we show how to perform the same type of analysis using the t distribution without having to make this assumption.

To solve this problem, we could also use Excel's **Z.TEST** function (called **ZTEST** for versions of Excel prior to Excel 2010). See cell I10 in Figure 27.

> **Z.TEST**(R, μ_0, σ) = 1 – NORM.DIST($\bar{x}, \mu_0, \sigma/\sqrt{n}$,TRUE) where $\bar{x}$ = AVERAGE(R) = the sample mean of the data in range R and n = COUNT(R) = sample size. The third parameter is optional. When it is omitted, the value of the sample standard deviation of R is used instead; that is, Z.TEST(R, μ_0) = Z.TEST(R, μ_0, STDEV.S(R)).

We make one final observation. If we had tried to test the null hypothesis H_0: μ = 120, then the alternative hypothesis would become H_0: μ ≠ 120. In this case, we would have two critical regions where we would reject the null hypothesis as shown in Figure 29:

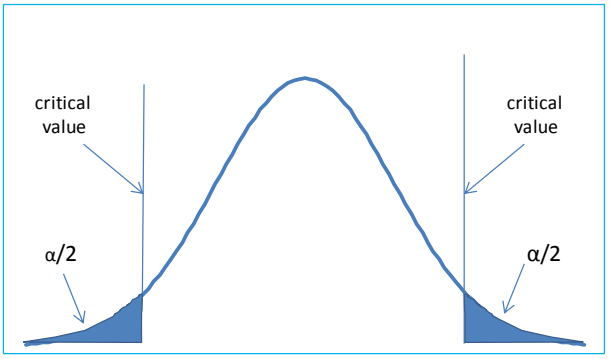


Figure 29: Two-tailed hypothesis testing

This is called a **two-tailed test**, which is more commonly used than the one-tailed test.

Chapter 6 Binomial Distribution

Basic Concepts

Suppose an experiment has the following characteristics:

- The experiment consists of n independent trials, each with two mutually exclusive outcomes (**success** and **failure**)
- For each trial, the probability of success is p (and so the probability of failure is 1 – p)

Each such trial is called a **Bernoulli trial**. Let x be the discrete random variable whose value is the number of successes in n trials. Then the probability distribution function for x is called the **binomial distribution** which is defined as follows:

$$f(x) = C(n,x)p^x(1-p)^{n-x}$$

where $C(n,x) = \frac{n!}{x!(n-x)!}$ and n **factorial** is $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$.

$C(n,x)$ can be calculated in Excel by using the function **COMBIN**(n,x).

Figure 30 shows a graph of the probability density function for the binomial distribution with n = 10 and p = .25:

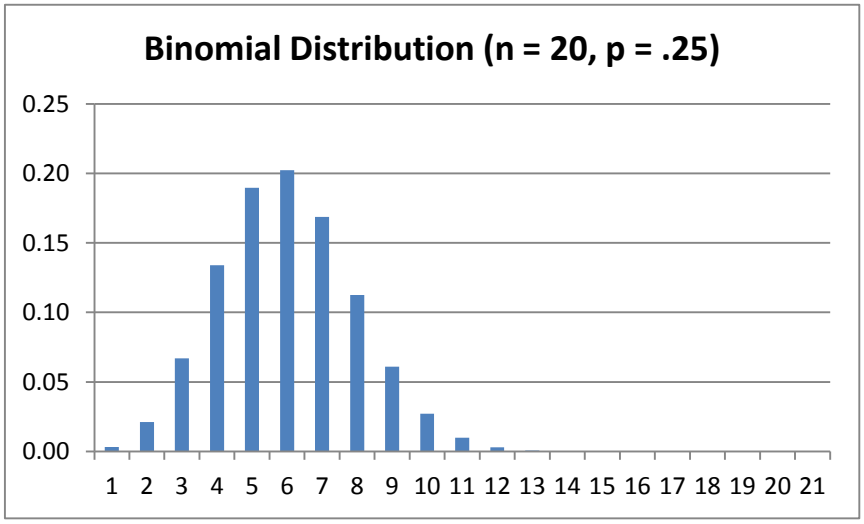


Figure 30: Binomial distribution

The mean of this distribution is np and the variance is np(1 – p).

Excel Functions

Excel provides the following functions regarding the binomial distribution:

BINOM.DIST(x, n, p, cum) where n = the number of trials, p = the probability of success for each trial, and cum takes the value TRUE or FALSE

BINOM.DIST(x, n, p, FALSE) = probability density function value at x for the binomial distribution, that is, the probability that there are x successes in n trials where the probability of success on any trial is p

BINOM.DIST(x, n, p, TRUE) = cumulative probability distribution value at x for the binomial distribution, that is, the probability that there are, at most, x successes in n trials where the probability of success on any trial is p

BINOM.INV(n, p, 1 – α) = the **critical** value; that is, the smallest value of x such that BINOM.DIST(x, n, p, TRUE) ≥ 1 – α

For versions of Excel prior to Excel 2010, the following functions are used instead: **BINOMDIST** which is equivalent to BINOM.DIST, and **CRITBINOM** which is equivalent to BINOM.INV.

Excel 2013 introduces the following new function (where $x \le y \le n$):

BINOM.DIST.RANGE(n, p, x, y) = the probability there are between x and y successes (inclusive) in n trials where the probability of success on any trial is p

Thus,

BINOM.DIST.RANGE(n, p, 0, y) = BINOM.DIST(y, n, p, TRUE)

and for x > 0

BINOM.DIST.RANGE(n, p, x, y) = BINOM.DIST(y, n, p, TRUE)
– BINOM.DIST(x – 1, n, p, TRUE)

The y parameter may be omitted, in which case we have:

BINOM.DIST.RANGE(n, p, x) = BINOM.DIST(x, n, p, FALSE)

Example: What is the probability that, if you throw a die 10 times, it will come up a 6 four times?

We can model this problem using the binomial distribution with n = 10 and p = 1/6 as follows:

$$f(4) = C(10,4)\left(\frac{1}{6}\right)^4\left(1-\frac{1}{6}\right)^{10-4} = 0.054266$$

Alternatively, the problem can be solved using the Excel formula:

BINOM.DIST(4, 10, 1/6, FALSE) = 0.054266

The probability that the die will come up a 6 at least four times can be calculated by the Excel formula:

1 – BINOM.DIST(3, 10, 1/6, TRUE) = 0.06728

Hypothesis Testing

Example: Suppose you have a die and you suspect that it is biased towards the number 3. So you run an experiment in which you throw the die 10 times and count that the number 3 comes up four times. Determine whether the die is biased.

The population random variable x = the number of times 3 occurs in 10 trials has a binomial distribution. Let π be the population parameter corresponding to the probability of success on any trial. We define the following null and alternative hypotheses:

H_0: $\pi \le 1/6$; that is, the die is not biased towards the number 3

H_1: $\pi > 1/6$

Now, setting α = .05, we have:

$P(x \le 4)$ = BINOM.DIST(4, 10, 1/6, TRUE) = 0.984538 > 0.95 = 1 – α.

And so we reject the null hypothesis with 95 percent **level of confidence**.

Example: We suspect that a coin is biased towards heads. When we toss the coin nine times, how many heads need to come up before we are confident that the coin is biased towards heads?

We use the following null and alternative hypotheses:

H_0: $\pi \le .5$

H_1: $\pi > .5$

Using a confidence level of 95 percent (i.e., α = .05), we calculate:

BINOM.INV(n, p, 1 – α) = BINOM.INV(9, .5, .95) = 7

And so 7 is the critical value. If seven or more heads come up, then we are 95 percent confident that the coin is biased towards heads and so we can reject the null hypothesis.

Note that BINOM.DIST(6, 9, .5, TRUE) = .9102 < .95 while BINOM.DIST(7, 9, .5, TRUE) = .9804 ≥ .95.

Chapter 7 Student's t Distribution

Basic Concepts

The one sample hypothesis test using the normal distribution as described in Chapter 5 is fine when one knows the standard deviation of the population distribution and the population is either normally distributed or the sample is sufficiently large that the Central Limit Theorem applies.

The problem with this approach is that the standard deviation of the population is generally not known. One way of addressing this is to use the standard deviation *s* of the sample instead of the standard deviation σ of the population, employing the t distribution.

The (**Student's**) **t distribution** with k **degrees of freedom** has probability density function given by the formula:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-(k+1)/2}$$

where $\Gamma(y)$ is the gamma function.

The overall shape of the probability density function of the t distribution resembles the bell shape of a normally distribution with mean 0 and variance 1, except that it is a bit lower and wider. As the number of degrees of freedom grows, the t distribution approaches the standard normal distribution and, in fact, the approximation is quite close for k ≥ 30:

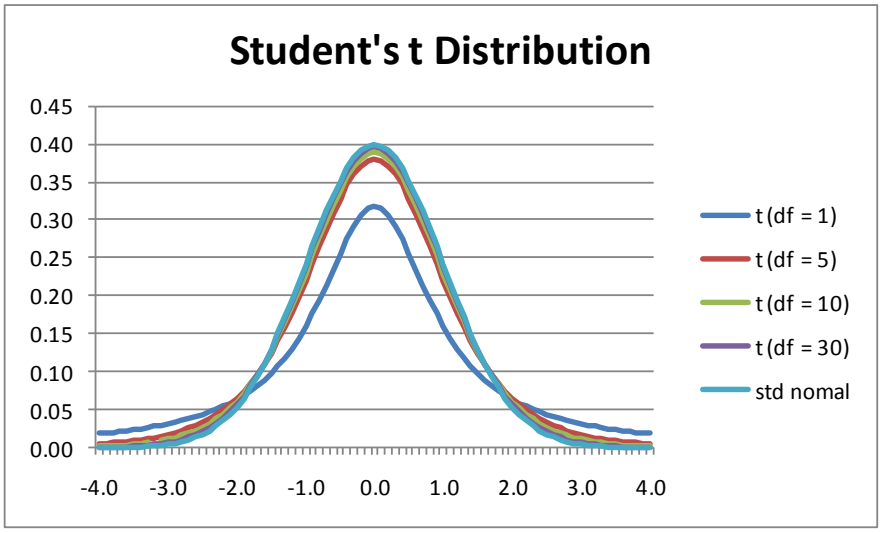


Figure 31: Chart of the t distribution by degrees of freedom

If x has a normal distribution with mean μ and standard deviation σ, then for samples of size n the random variable:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a t distribution with n – 1 degrees of freedom. The same is true even when x doesn't have a normal distribution, provided that n is sufficiently large (by the Central Limit Theorem).

This test statistic is the same as $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ from the Central Limit Theorem with the population standard deviation σ replaced by the sample standard deviation *s*. What makes this useful is that usually the standard deviation of the population is unknown while the standard deviation of the sample is known.

Excel Functions

Excel provides the following functions regarding the t distribution:

T.DIST(x, df, cum) = the probability density function value at x for the t distribution with df degrees of freedom when cum = FALSE and the corresponding cumulative distribution function at x when cum = TRUE

T.INV(p, df) = the value x such that T.DIST(x, df, TRUE) = p, that is, inverse of T.DIST(x, df, TRUE)

T.DIST.RT(x, df) = the right tail at x of the t distribution with df degrees of freedom

T.DIST.2T(x, df) = the sum of the right tail of the t distribution with df degrees of freedom at x plus the left tail at -x, where x ≥ 0 (the function yields an error value when x < 0

T.INV.2T(p, df) = the value xsuch that T.DIST.2T(x, df) = p, that is, inverse of T.DIST.2T(x, df)

These functions were not available prior to Excel 2010. The following table lists how these functions are used as well as their equivalents using the functions **TDIST** and **TINV**, which were used prior to Excel 2010:

Tail	Distribution	Inverse
Right tail	T.DIST.RT(x, df) = TDIST(x, df, 1) if x ≥ 0 = 1-TDIST(-x, df, 1) if x < 0	T.INV(1-α, df) = TINV(2*α, df)

Tail	Distribution	Inverse
Left tail	T.DIST(x, df, TRUE) = 1-TDIST(x, df, 1) if x ≥ 0 = TDIST(-x, df, 1) if x < 0	T.INV(α, df) = -TINV(2*α, df)
Two-tail	T.DIST.2T(x, df) = TDIST(x, df, 2)	T.INV.2T(α, df) = TINV(α, df)

Table 8: Excel t distribution functions

Hypothesis Testing

One-Sample Testing

As explained previously, the t distribution provides a good way to perform one sample test on the mean when the population variance is not known, provided the population is normal or the sample is sufficiently large so that the Central Limit Theorem applies.

It turns out that the t distribution provides good results even when the population is not normal, and even when the sample is small, provided the sample data is reasonably symmetrically distributed about the sample mean. This can be determined by graphing the data. The following are indications of symmetry:

- The box plot is relatively symmetrical; that is, the median is in the center of the box and the whiskers extend equally in each direction
- The histogram looks symmetrical
- The mean is approximately equal to the median
- The coefficient of skewness is relatively small

Example: A weight reduction program claims to be effective in treating obesity. To test this claim, 12 people were put on the program and the number of pounds of weight gained/lossed was recorded for each person after two years. This is shown in columns A and B of Figure 32. Can we conclude that the program is effective?

	A	B	C	D	E	F	G
1	**One sample t test**						
2							
3	Subject	Weight Loss		count	12		=COUNT(B4:B15)
4	1	23		mean	5.5		=AVERAGE(B4:B15)
5	2	18		std dev	11.42167		=STDEV(B4:B15)
6	3	-5		std err	3.297152		=E5/SQRT(E3)
7	4	7					
8	5	1		hyp mean	0		
9	6	-8		α	0.05		
10	7	12		tails	2		
11	8	-8		df	11		=E3-1
12	9	20		t stat	1.668106		=(E4-E8)/E6
13	10	13		p value	0.123479		=T.DIST.2T(E12,E11)
14	11	-2		t crit	2.200985		=T.INV.2T(E9,E11)
15	12	-5		sig	no		=IF(E13<E9,"yes","no")

Figure 32: One sample t test

A negative value in column B indicates that the subject gained weight. We judge the program to be effective if there is some weight loss at the 95 percent significance level. We choose to conduct a two-tailed test since there is risk that the program might actual result in weight gain rather than loss. Thus, our null hypothesis is:

H_0: $\mu = 0$; that is, the program is not effective.

From the box plot in Figure 33, we see that the data is quite symmetric and so we use the t test even though the sample is small:

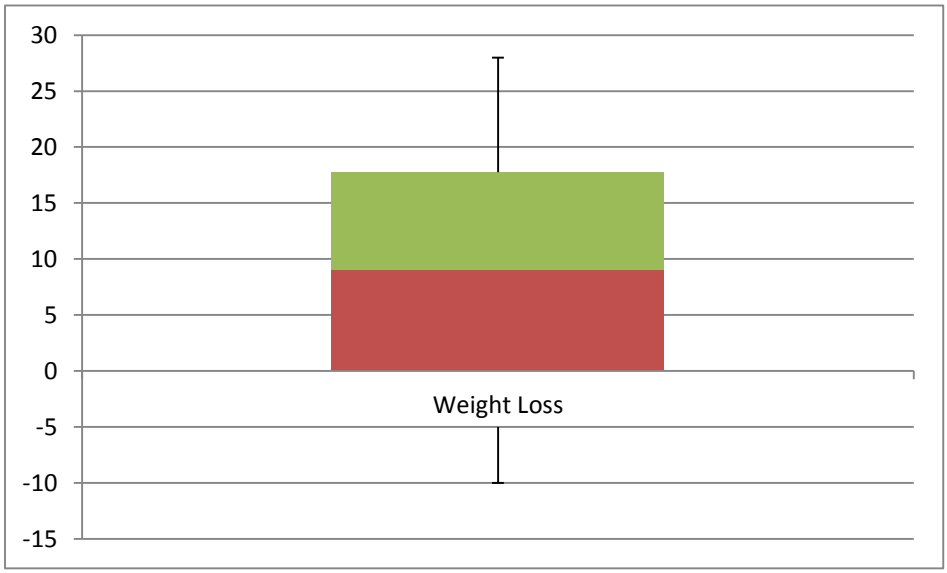


Figure 33: Box plot of the sample data

Column E of Figure 33 contains all the formulas required to carry out the t test. Since Excel only displays the values of these formulas, we show each of the formulas (in text format) in column G so that you can see how the calculations are performed.

Thus, where R is B4:B15, we see that n = COUNT(R) = 12, $\bar{x}$ = AVERAGE(R) = 5.5, s = STDEV(R) = 11.42, and standard error = $s/\sqrt{n}$ = 3.30. From this we see that:

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{5.5-0}{3.3} = 1.668 \text{ with } df = 11 \text{ degrees of freedom.}$$

Since p-value = T.DIST.2T(t, df) = T.DIST.2T(1.668, 11) = .123 > .05 = α, the null hypothesis is not rejected. This means there is an 12.3 percent probability of achieving a value for t this high, assuming that the null hypothesis is true and, since 12.3 percent > 5 percent, we can't reject the null hypothesis.

The same conclusion is reached since:

t-crit = T.INV.2T(α, df) = T.INV.2T(.05, 11) = 2.20 > 1.668 = t-obs.

Here we used T.DIST.2T and T.INV.2T since we are employing a two-tailed test.

Testing Two Independent Samples

Assuming equal population variances

Our goal is to determine whether or not the means of two populations are equal given two independent samples, one from each population. We start by assuming that population variances are equal even if unknown.

Suppose that $\bar{x}$ and $\bar{y}$ are the sample means of two independent samples of size n_x and n_y respectively. Also suppose that s_x and s_y are the sample standard deviations of two sets of data. If x and y are normal or if n_x and n_y are sufficiently large for the Central Limit Theorem to hold, then the statistic:

$$t = \frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{s\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}$$

has a t distribution with $n_x + n_y - 2$ degrees of freedom where s^2 is the pooled variance defined by:

$$s^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{(n_x-1)+(n_y-1)}$$

We can use the statistic t to test the null hypothesis that the two population means are equal or equivalently that $\mu_x - \mu_y = 0$. It turns out that, even when x and y are not normal and n_x and n_y are not particularly large, as long as the two samples are reasonably symmetric, then this test yields pretty good results.

Example: A food company wants to determine whether or not their new formula for peanut butter is significantly tastier than their old formula. They chose two random samples of 10 people and ask the people in the first sample to taste the peanut butter using the existing formula, and they ask the people in the second sample to taste the peanut butter with the new formula. They then ask all 20 people to fill out a questionnaire rating the tastiness of the peanut butter they tasted.

Based on the data in Figure 34, determine whether there is a significant difference between the two types of peanut butter.

As we can see from Figure 34, the mean score for the new formula is 15 while the mean score for the old formula is 11. But is this difference just due to random effects or is there a statistically significant difference?

We also note that the variances for the two samples are 13.3 and 18.8. It turns out that these two values are close enough to satisfy the equal variance assumption.

	A	B	C
3	New	Old	
4	13	12	
5	17	8	
6	19	6	
7	11	16	
8	20	12	
9	15	14	
10	18	10	
11	9	18	
12	12	4	
13	16	11	
14			
15	10	10	size
16	15	11.1	mean
17	13.33333	18.76667	variance

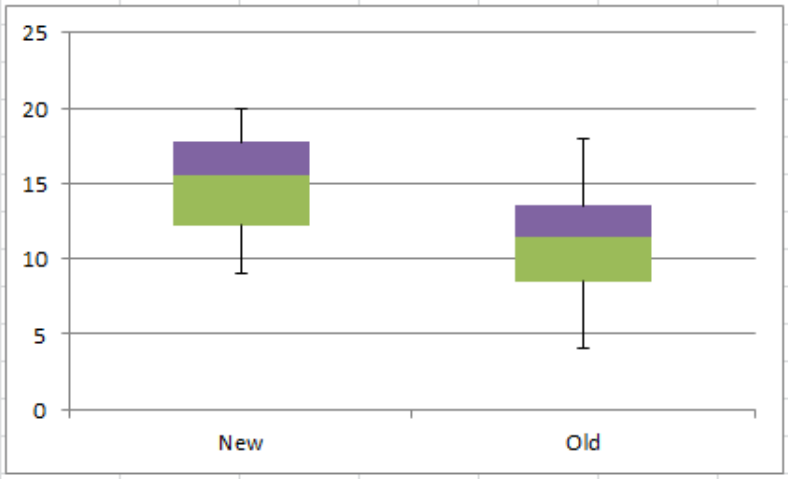


Figure 34: Sample data

Finally, if we draw the box plots for the two data sets, we see that both are relatively symmetric (that is, for each sample the colored areas in each box are approximately equal in size, and the upper and lower whiskers are fairly equal in length).

Thus, the assumptions for the t test are met. To carry out the test, we use Excel's **t-Test: Two Sample Assuming Equal Variances** data analysis tool which, as usual, can be accessed by selecting **Data > Analysis|Data Analysis**. The result is shown in Figure 35:

t-Test: Two-Sample Assuming Equal Variances

	New	*Old*
Mean	15	11.1
Variance	13.33333	18.76667
Observations	10	10
Pooled Variance	16.05	
Hypothesized Mean Difference	0	
Df	18	
t Stat	2.176768	
P(T<=t) one-tail	0.021526	
t Critical one-tail	1.734064	
P(T<=t) two-tail	0.043053	
t Critical two-tail	2.100922	

Figure 35: Two-sample t test assuming equal variance

Assuming that we are performing a two-tailed test, we see that the p-value = 0.043 which is less than the usual alpha value of .05. Thus, we can conclude that there is a significant difference between the mean scores for the two peanut butter formulations.

Assuming unequal population variances

In the previous section, we assumed that the two population variances were equal (or at least not very unequal). When the assumption of equal population variances is not met (or when we don't have enough evidence to draw this conclusion), we can modify the approach used in the previous section by employing a slightly different version of the t test.

Suppose that $\bar{x}$ and $\bar{y}$ are the sample means of two independent samples of size n_x and n_y respectively. Also suppose that s_x and s_y are the sample standard deviations of the two sets of data. If x and y are normally distributed or if n_x and n_y are sufficiently large for the Central Limit Theorem to hold, then the statistic:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

has a t distribution with the following degrees of freedom.

$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

Example: Repeat the previous analysis using the data in Figure 36:

	A	B	C
3	New	Old	
4	20	12	
5	32	8	
6	2	6	
7	25	16	
8	5	12	
9	18	14	
10	21	10	
11	7	18	
12	28	4	
13	31	11	
14			
15	10	10	size
16	18.9	11.1	mean
17	118.3222	18.76667	variance

Figure 36: Revised sample data

This time, we see that there is a big difference between the variances (118.3 vs. 18.8).

Generally, even if one variance is up to four times the other, the equal variance assumption will give good results. This rule of thumb is clearly violated in this example, and so we need to use the t test with unequal population variances, namely, Excel's **t-Test: Two Sample Assuming Unequal Variances** data analysis tool.

The output is shown in Figure 37:

t-Test: Two-Sample Assuming Unequal Variances

	New	*Old*
Mean	18.9	11.1
Variance	118.3222	18.76667
Observations	10	10
Hypothesized Mean Difference	0	
df	12	
t Stat	2.106655	
P(T<=t) one-tail	0.028434	

t Critical one-tail	1.782288
P(T<=t) two-tail	0.056869
t Critical two-tail	2.178813

Figure 37: Two-sample t test assuming unequal variance

Note that the degrees of freedom have been reduced from 18 in Example 1 to 12 in this example (actually 11.78 although Excel rounds this up to 12). Since for the two-tail test, p-value = 0.056869 > .05 = α, we cannot reject the null hypothesis that there is no significant difference between the two formulations.

If we had used the equal variance test, we would have gotten the results shown in Figure 6. This would have led us to a different conclusion since, for the two-tail test, p-value = 0.049441 < .05 = α, and therefore we would have concluded that there was a significant difference between the two formulations.

t-Test: Two-Sample Assuming Equal Variances

	New	*Old*
Mean	18.9	11.1
Variance	118.3222	18.76667
Observations	10	10
Pooled Variance	68.54444	
Hypothesized Mean Difference	0	
df	18	
t Stat	2.106655	
P(T<=t) one-tail	0.024721	
t Critical one-tail	1.734064	
P(T<=t) two-tail	0.049441	
t Critical two-tail	2.100922	

Figure 38: Same problem assuming equal variance

Testing Paired Samples

In **paired sample** hypothesis testing, a sample from the population is chosen and two measurements for each element in the sample are taken. Each set of measurements is considered a sample. Unlike the hypothesis testing studied so far, the two samples are not independent of one another.

For example, if you want to determine whether drinking a glass of wine or drinking a glass of beer has the same or different impact on memory, one approach is to take a sample of, say, 40 people. You then have half of them drink a glass of wine and the other half drink a glass of beer. You then give each of the 40 people a memory test and compare the results. This is the approach with independent samples.

Another approach is to take a sample of 20 people and have each person drink a glass of wine and take a memory test. You then have the same people drink a glass of beer and again take a memory test. Finally, you compare the results. This is the approach used with paired samples.

The advantage of this second approach is that the sample can be smaller. Also, since the sampled subjects are the same for beer and wine, there is less of a chance that some external factor (or **confounding variable**) will influence the result. The problem with this approach is that it is possible that the results of the second memory test will be lower simply because the person has imbibed more alcohol. This can be corrected by sufficiently separating the tests, for example, by conducting the test with beer a day after the test with wine.

It is also possible that the order in which people take the tests influences the result (for example, the subjects learn something on the first test that helps them on the second test or perhaps taking the test the second time introduces a degree of boredom that lowers the score). One way to address these **order effects** is to have half the people drink wine on Day 1 and beer on Day 2, while for the other half, the order is reversed (this is called **counterbalancing**).

Obviously, not all experiments can use the paired sample design. For example, if you are testing differences between men and women, then independent samples will be necessary.

Example: A clinic provides a program to help their clients lose weight and asks a consumer agency to investigate the effectiveness of the program. The agency takes a sample of 15 people, weighs each person in the sample before the program begins, and then weighs each person three months later to produce the results in the table in Figure 39. Determine whether the program is effective.

	A	B	C	D
1	**Two sample t test with paired samples**			
2				
3	Person	Before	After	Difference
4	1	210	197	13
5	2	205	195	10
6	3	193	191	2
7	4	182	174	8
8	5	259	236	23
9	6	239	226	13
10	7	164	157	7
11	8	197	196	1
12	9	222	201	21
13	10	211	196	15
14	11	187	181	6
15	12	175	164	11
16	13	186	181	5
17	14	243	229	14
18	15	246	231	15
19				
20	mean			10.933333
21	std dev			6.3298236

Figure 39: Sample data

The program is effective if there is a significant reduction in the weight of the people in the program. Let x = the reduction in weight three months after the program starts. The null hypothesis is:

H_0: $\mu = 0$; that is, any differences in weight is due to chance (two-tailed test).

Essentially, the test to use is the one sample test as previously described, but on the difference in the weight before the program minus the weight after the program. Since all the values in column D of Figure 39 are positive, we see that all the subjects in the sample lost weight. But is the reduction significant?

We could create the analysis (as in Figure 32) using column D of Figure 39 as the single sample. Instead, we will use Excel's **t-Test: Paired Two Sample for Means** data analysis tool. We begin by selecting **Data > Analysis|Data Analysis**.

When the dialog box shown in Figure 40 appears, you need to fill in the values as shown in the figure and click **OK**:

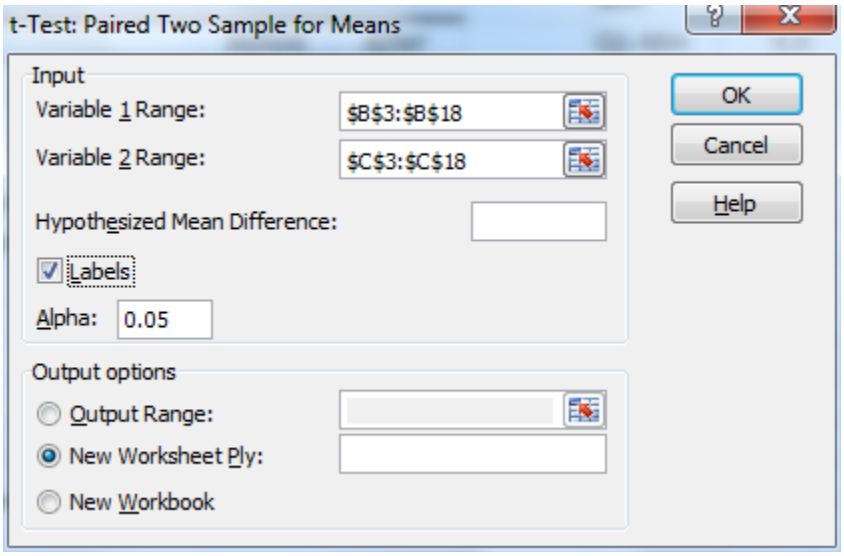


Figure 40: T-Test: Paired two-sample for means dialog box

The output of the analysis is displayed in Figure 41:

t-Test: Paired Two Sample for Means

	Before	*After*
Mean	207.9333	197
Variance	815.781	595
Observations	15	15
Pearson Correlation	0.98372	

Hypothesized Mean Difference	0
df	14
t Stat	6.6897
P(T<=t) one-tail	5.14E-06
t Critical one-tail	1.76131
P(T<=t) two-tail	1.03E-05
t Critical two-tail	2.144787

Figure 41: Paired sample t-test analysis

Since the p-value for the two-tailed test is 1.03E-05, we can conclude that there is a significant difference between the mean weights. Since the mean afterward is lower than the mean weight beforehand, we can conclude that there is a significant reduction in weight. Thus, it is fair to say that the program is effective (at least for the three-month period).

Using the T.TEST Function

In addition to the approaches shown in the last few sections, Excel provides the T.TEST function to perform the three different types of two sample t tests.

T.TEST(R1, R2, *tails, type*) = the t-test value for the difference between the means of two samples R1 and R2 where *tails* = 1 (one-tailed) or 2 (two-tailed) and *type* takes the values:

1. The samples have paired values from the same population
2. The samples are from populations with the same variance
3. The samples are from populations with different variances

These three types correspond to the Excel data analysis tools:

1. t-Test: Paired Two Sample for Mean
2. t-Test: Two-Sample Assuming Equal Variance
3. t-Test: Two-Sample Assuming Unequal Variance

The T.TEST function is not available for versions of Excel prior to Excel 2010. For earlier versions of Excel, the TTEST function is available, which provides identical functionality.

For example, we can repeat the analysis from Figure 35 using the T.TEST function. As we can see from Figure 42, the p-value for the two-tailed test for two independent samples from populations with the same variances is 0.043053 (cell E40). This is the same as the result previously found:

	E	F	G
39	0.043456		=T.TEST(A4:A13,B4:B13,2,3)
40	0.043053		=T.TEST(A4:A13,B4:B13,2,2)

Figure 42: T.TEST for two independent samples

Figure 42 also shows the value of the test if we didn't assume equal variances (cell 39). As you can see, there is very little difference between the two results.

Confidence Intervals

Until now, we have made what are called **point estimates of the mean** or differences between means. We now define the **confidence interval** which provides a range of possible values instead of just a single measurement. We can illustrate how this works using the t distribution:

The 1 – α confidence interval for the population mean is:

$$observed\ mean\ \pm t_{crit} \cdot std\ error$$

Example: Calculate the 95 percent confidence interval for the one sample t test example (see Figure 32):

$$mean_{obs} \pm t_{crit} \cdot std\ err = 5.5\ \pm 2.201\ \cdot 3.297 = 5.5 \pm\ 7.257$$

This yields a 95 percent confidence interval of (-1.757, 12.757) for the population mean. Since the interval contains the hypothetical mean of 0, we are justified once again in our decision not to reject the null hypothesis.

Versions of Excel starting with Excel 2010 provide the following function to calculate the confidence interval for the t distribution:

> **CONFIDENCE.T**(α, s, n) = k such that ($\bar{x}$ – k, $\bar{x}$ + k) is the 1 – α confidence interval for the population mean; that is, CONFIDENCE.T(α, s, n) = t_{crit} · std error, where n = sample size and s = sample standard deviation.

Thus, for the last example, we calculate CONFIDENCE.T(.05,11.422,12) = 7.257 which yields a 95 percent confidence interval of $5.5 \pm\ 7.257$, as before.

Chapter 8 Chi-square and F Distribution

Chi-square Distribution

Basic Concepts

The chi-square distribution with k degrees of freedom has the probability density function:

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}$$

where Γ is the gamma function and k does not have to be an integer and can be any positive real number. Note that chi-square is commonly written as χ^2.

The following are the graphs of the probability density function with degrees of freedom df= 5 and 10. As df grows larger, the fat part of the curve shifts to the right and becomes more like the graph of a normal distribution:

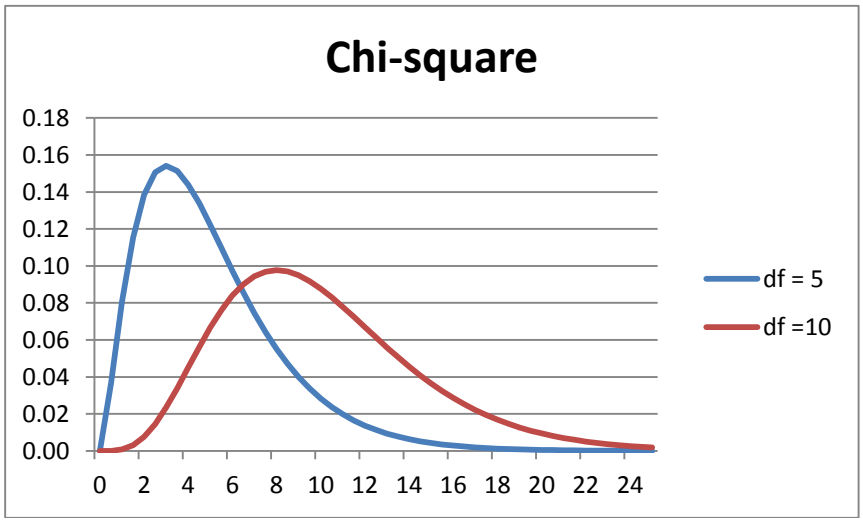


Figure 43: Chart of the chi-square distribution

The key relationship between the chi-square and normal distribution is that, if x has a standard normal distribution, then x^2 has a chi-square distribution with one degree of freedom.

Excel Functions

Excel provides the following functions regarding the chi-square distribution:

CHISQ.DIST(x, df, cum) = the probability density function value at x for the chi-square distribution with df degrees of freedom when cum = FALSE and the corresponding cumulative distribution function at x when cum = TRUE

CHISQ.INV(p, df) = the value x such that CHISQ.DIST(x, df, TRUE) = p, that is, inverse of CHISQ.DIST(x, df, TRUE)

CHISQ.DIST.RT(x, df) = the right-tail at x of the chi-square distribution with df degrees of freedom

CHISQ.INV.RT(p, df) = the value x such that CHISQ.DIST.RT(x, df) = p, that is, inverse of CHISQ.DIST.RT(x, df)

These functions were not available prior to Excel 2010. The following table lists how these functions are used, as well as their equivalents using the functions **CHIDIST** and **CHIINV** which were used prior to Excel 2010:

Tail	Distribution	Inverse
Right-tail	CHISQ.DIST.RT(x, df) = CHIDIST(x, df)	CHISQ.INV.RT(α, df) = CHIINV(α, df)
Left-tail	CHISQ.DIST(x, df, TRUE) =1-CHIDIST(x, df)	CHISQ.INV(α, df) = CHIINV(1-α, df)

Table 9: Formulas for the chi-square distribution in Excel

Contingency Tables

The chi-square distribution is commonly used to determine whether or not two sets of data are independent of each other. Such data are organized in what are called **contingency tables** as described in the following example.

Example: A survey is conducted of 175 young adults whose parents are classified either as wealthy, middle class, or poor to determine their highest level of schooling (e.g., graduated from university, graduated from high school, or neither). The results are summarized in the contingency table in Figure 44 (Observed Values). Based on the data collected, is a person's level of schooling independent of their parents' wealth?

	A	B	C	D	E
3	**Observed Values**				
4					
5		Univ	High Sch	None	Total
6	Wealthy	20	15	10	45
7	Middle Class	40	25	20	85
8	Poor	8	14	23	45
9	Total	68	54	53	175

Figure 44: Contingency Table (observed values)

We set the null hypothesis to be:

H_0: Highest level of schooling attained is independent of parents' wealth.

To use the chi-square test, we need to calculate the expected values that correspond to the observed values in the table in Figure 44. To accomplish this, we use the fact that if two events are independent, then the probability that both events occur is equal to the product of the probabilities that each will occur.

We also assume that the proportions for the sample are good estimates of the probabilities of the expected values.

We now show how to construct the table of expected values (see Figure 45). We know from Figure 44 that 45 of the 175 people in the sample are from wealthy families. So the probability that someone in the sample is from a wealthy family is 45/175 = 25.7 percent. Similarly, the probability that someone in the sample graduated from university is 68/175 = 38.9 percent. But based on the null hypothesis, being from a wealthy family is independent of graduating from university and so, the expected probability of both is simply the product of the two events or 25.7 percent · 38.9 percent = 10.0 percent. Thus, based on the null hypothesis, we expect that 10.0 percent of 175 = 17.5 people are from a wealthy family and have graduated from university:

	G	H	I	J	K
3	**Expected Values**				
4					
5		Univ	High Sch	None	Total
6	Wealthy	17.48571	13.88571	13.62857	45
7	Middle Class	33.02857	26.22857	25.74286	85
8	Poor	17.48571	13.88571	13.62857	45
9	Total	68	54	53	175

Figure 45: Expected values

Thus, for example, cell H6 can be calculated by the formula =H$9*$K6/K9. If you highlight the range H6:J8 and press **Ctrl-D** and then **Ctrl-R** you will fill in the rest of the table.

Alternatively, you can fill in the table of expected values by highlighting the range H6:J8 and entering the array formula:

=MMULT(K6:K8,H9:J9)/K9

Independence Testing

The **Pearson's chi-square test statistic** is defined as:

$$\sum_{i=1}^{k} \frac{(obs_i - exp_i)^2}{exp_i}$$

For sufficiently large samples, the Pearson's chi-square test statistic has approximately a chi-square distribution with (row count – 1) (column count – 1) degrees of freedom.

The test requires that the following conditions be met:

- Random sample: Data must come from a random sampling of the population
- Independence: The observations must be independent of each other. This means chi-square cannot be used to test correlated data (e.g., matched pairs)
- Cell size: All of the expected frequency values in the contingency table are at least 5, although with large tables, occasional values lower than 5 will probably still yield good results. With a 2×2 contingency table, it is better to have even larger frequency values.

In Figure 46, we show how to use the chi-square test to carry out the analysis:

	A	B	C	D
27	χ2	15.81809		=SUM((B6:D8-H6:J8)^2/H6:J8)
28	df	4		=(COUNTA(B5:D5)-1)*(COUNTA(A6:A8)-1)
29	p-value	0.003273		=CHISQ.DIST.RT(B27,B28)
30	α	0.05		
31	χ2-crit	9.487729		=CHISQ.INV.RT(B30,B28)

Figure 46: Chi-square test

Since p-value = 0.003273 < .05 = α, we reject the null hypothesis and conclude that a person's level of schooling is not independent of their parents' wealth.

Excel provides the following function which carries out the chi-square test for independence for contingency tables:

CHISQ.TEST(R1, R2) = CHISQ.DIST.RT(x, df) where R1 = the array of observed data, R2 = the array of expected value, x is the chi-square statistic, and df = (row count – 1) (column count – 1).

For our example, CHISQ.TEST(B6:D8,H6:J8) = 0.003273.

The CHISQ.TEST function is only available in versions of Excel starting with Excel 2010. For previous versions of Excel, the equivalent function **CHITEST** can be used.

F Distribution

Basic Concepts

If x_1 and x_2 have chi-square distributions with df_1 and df_2 degrees of freedom respectively, then the quotient F = x_1 / x_2 has **F distribution** with df_1, df_2 degrees of freedom.

The F distribution is commonly used to compare two variances. It is useful in ANOVA, regression, and other commonly used tests. In particular, if we draw two independent samples of size n_1 and n_2 respectively from two different normal populations with the same variance, then the quotient of the two sample variances has F distribution with $n_1 - 1, n_2 - 1$ degrees of freedom.

Excel Functions

Excel provides the following functions regarding the F distribution:

F.DIST(x, df, cum) = the probability density function value at x for the F distribution with df degrees of freedom when cum = FALSE and the corresponding cumulative distribution function at x when cum = TRUE

F.INV(p, df) = the value x such that F.DIST(x, df, TRUE) = p, that is, inverse of F.DIST(x, df, TRUE)

F.DIST.RT(x, df) = the right-tail at x of the F distribution with df degrees of freedom

F.INV.RT(p, df) = the value x such that F.DIST.RT(x, df) = p, that is, inverse of F.DIST.RT(x, df)

These functions were not available prior to Excel 2010. The following table lists how these functions are used as well as their equivalents using the functions **FDIST** and **FINV** that were available prior to Excel 2010:

Tail	Distribution	Inverse

Tail	Distribution	Inverse
Right-tail	F.DIST.RT(x, df_1, df_2) = FDIST(x, df_1, df_2)	F.INV.RT(α, df_1, df_2) = FINV(α, df_1, df_2)
Left-tail	F.DIST(x, df_1, df_2, TRUE) = 1 – FDIST(x, df_1, df_2)	F.INV(α, df_1, df_2) = FINV(1-α, df_1, df_2)

Table 10: F distribution formulas in Excel

Hypothesis Testing

As we previously observed, if we draw two independent samples of size n_1 and n_2 respectively from two different normal populations with the same variance, then the quotient of the two sample variances has F distribution with $n_1 - 1, n_2 - 1$ degrees of freedom.

This fact can be used to test whether or not the variances of two populations are equal. In order to deal exclusively with the right-tail of the distribution, when taking ratios of sample variances we should put the larger variance in the numerator of the quotient.

In order to use this test, the following must hold:

- Both populations are normally distributed
- Both samples are drawn independently from each other.
- Within each sample, the observations are sampled randomly and independently of each other

Example

A company is comparing two methods for producing high-precision bolts and wants to choose the method with the least variability. It has taken a sample of the lengths of the bolts using both methods as given in the table on the left side of Figure 47:

	A	B	C	D	E	F	G	H
3	Method 1	Method 2		F-Test Two-Sample for Variances				
4	104.07	103.71						
5	103.17	104.11			*Method 1*	*Method 2*		
6	103.62	103.25		Mean	103.9317	104.3733		
7	103.71	105.57		Variance	0.713888	0.508638		
8	103.89	104.12		Observations	12	15		
9	104.58	104.73		df	11	14		
10	103.11	104.94		F	1.403528			=E7/F7
11	105.01	103.51		P(F<=f) one-tail	0.271364			=F.DIST.RT(E10,E9,F9)
12	104.29	103.89		F Critical one-tail	2.565497			=F.INV.RT(0.05,E9,F9)
13	105.33	103.47						
14	102.28	104.83						
15	104.12	105.02						
16		105.31						
17		104.45						
18		104.69						

Figure 47: F test to compare two variances

To carry out the test, we use Excel's **F-Test Two Sample For Variances** data analysis tool which, as usual, can be accessed by selecting **Data > Analysis|Data Analysis**. The result is shown on the right side of Figure 47.

Since p-value = .271364 > .05 = alpha, we cannot reject the null hypothesis. So we must conclude that there is no significant difference between the two variances.

Excel Test Function

Excel provides the following test statistic function:

> **F.TEST**(R1, R2) = two-tailed F-test comparing the variances of the samples in ranges R1 and R2 = the two-tailed probability that the variance of the data in ranges R1 and R2 are not significantly different.

Note that, unlike the F-Test data analysis tool or F.DIST, F.DIST.RT, F.INV and F.INV.RT functions, which are one-tailed, the F.TEST function is two-tailed. For the previous example:

> F.TEST(A4:A18,B4:B18) = .5427

which is double the value obtained for the one-tailed p-value in cell E11 of Figure 47.

Chapter 9 Analysis of Variance

Introduction

Essentially, **analysis of variance** (ANOVA) is an extension of two-sample hypothesis testing for comparing means (when variances are unknown) to more than two samples.

We start with the one-factor case. We will define the concept of factor later but, for now, we simply view this type of analysis as an extension of the two independent sample t test with equal population variances.

One-way ANOVA

One-way ANOVA Example

A food company wants to determine whether or not any of their three new formulas for peanut butter is significantly tastier than their old formula. They create four random samples of 10 people each (one for each type of peanut butter) and ask the people in each sample to taste the peanut butter for that sample. They then ask all 40 people to fill out a questionnaire rating the tastiness of the peanut butter they tried.

Based on the data in Figure 48, determine whether there is a significant difference between the four types of peanut butter:

	A	B	C	D
3	Old	New 1	New 2	New 3
4	12	11	19	10
5	8	9	20	17
6	6	16	15	20
7	16	15	14	11
8	12	18	20	5
9	14	19	17	9
10	10	5	18	16
11	18	9	15	13
12	4	12	16	4
13	11	17	12	14

Figure 48: Sample data

The null hypothesis for this example is that any difference between the four types of peanut butter is due to chance, that is:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Basic Concepts

Before we proceed with the analysis for this example, we should review a few basic concepts.

Suppose we have k samples, which we will call **groups** (or **treatments**). These are the columns in our analysis (corresponding to the four types of peanut butter in the example). We will use the index j for these. Each group consists of a sample of size n_j. The sample elements are the rows in the analysis. We will use the index i for these.

Suppose the jth group sample is $\{x_{1j}, \dots, x_{n_j j}\}$ and so the total sample consists of all the elements $\{x_{ij}: 1 \le i \le n_j, 1 \le j \le k\}$. We will use the abbreviation $\bar{x}_j$ for the mean of the jth group sample (called the **group mean**) and $\bar{x}$ for the mean of the total sample (called the **total** or **grand mean**).

Let the **sum of squares** for the jth group be $SS_j = \sum_i (x_{ij} - \bar{x}_j)^2$. We now define the following terms:

$$SS_T = \sum_j \sum_i (x_{ij} - \bar{x})^2$$

$$SS_B = \sum_j n_j (\bar{x}_j - \bar{x})^2$$

$$SS_W = \sum_j SS_j = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$$

SS_T is the sum of squares for the **total** sample, that is, the sum of the squared deviations from the grand mean. SS_W is the sum of squares **within** the groups, that is, the sum of the squared means across all groups. SS_B is the sum of the squares **between** group sample means, that is, the weighted sum of the squared deviations of the group means from the grand mean.

We also define the following degrees of freedom, where $n = \sum_{j=1}^{k} n_j$:

$$df_T = n - 1 \qquad df_B = k - 1 \qquad df_W = \sum_{j=1}^{k} (n_j - 1) = n - k$$

Finally, we define the **mean square**, as $MS = SS/df$, and so:

$$MS_T = SS_T \,/\, df_T \qquad MS_B = SS_B \,/\, df_B \qquad MS_W = SS_W \,/\, df_W$$

We summarize these terms in the following table:

	$\boldsymbol{df}$	$\boldsymbol{SS}$	$\boldsymbol{MS}$
T	$n-1$	$\sum_i \sum_i \left(x_{ij} - \bar{x}\right)^2$	SS_T/df_T
B	$k-1$	$\sum_i n_j\left(\bar{x}_j - \bar{x}\right)^2$	SS_B/df_B
W	$n-k$	$\sum_i \sum_i \left(x_{ij} - \bar{x}_j\right)^2$	SS_W/df_W

Table 11: Summary of ANOVA terms

Clearly MS_T is the variance for the total sample, MS_W is the sum of the group sample variances, and MS_B is the variance for the "between sample", that is, the variance of $\{n_1\bar{x}_1, \ldots, n_k\bar{x}_k\}$.

It is also not hard to show that:

$$SS_T = SS_W + SS_B$$

$$df_T = df_W + df_B$$

It turns out that, if the null hypothesis is true, then MS_W and MS_B are both measures of the same error. Thus, the null hypothesis becomes equivalent to the hypothesis that the population versions of these statistics are equal, that is:

$$\sigma_B = \sigma_W$$

We can, therefore, use the F-test described in Chapter 8 to determine whether or not to reject the null hypothesis. This means that if the x_{ij} are independently and normally distributed, and all the μ_j are equal (null hypothesis), and all the σ_j^2 are equal (homogeneity of variances), then the test statistic:

$$F = \frac{MS_B}{MS_W}$$

has an F distribution with df_B, df_W degrees of freedom.

Analysis

To carry out the analysis for the example, we will use the **ANOVA: Single Factor** data analysis tool. To access this tool, as usual, press **Data > Analysis|Data Analysis** and fill in the dialog box that appears as in Figure 49:

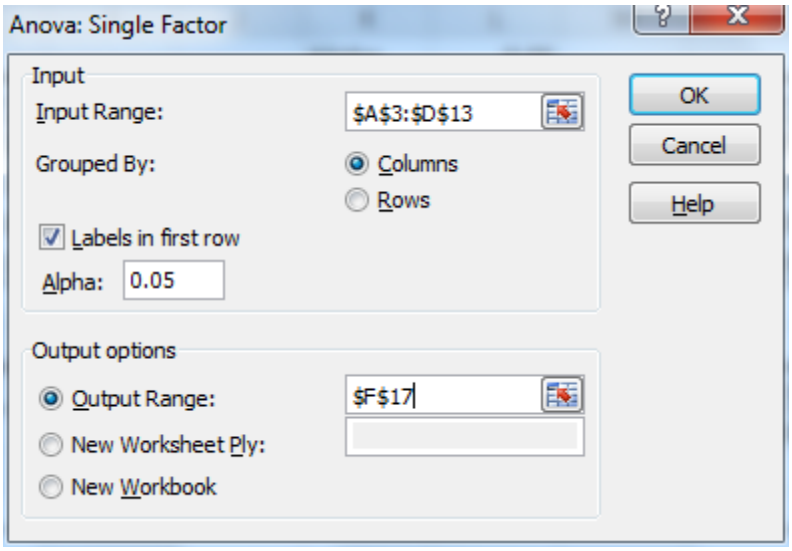


Figure 49: Dialog box for ANOVA: Single-factor data analysis tool

The output is as shown in Figure 50:

	F	G	H	I	J	K	L
17	Anova: Single Factor						
18							
19	SUMMARY						
20	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
21	Old	10	111	11.1	18.76667		
22	New 1	10	131	13.1	21.21111		
23	New 2	10	166	16.6	7.155556		
24	New 3	10	119	11.9	26.32222		
25							
26							
27	ANOVA						
28	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
29	Between Groups	176.675	3	58.89167	3.206928	0.034463	2.866266
30	Within Groups	661.1	36	18.36389			
31							
32	Total	837.775	39				

Figure 50: ANOVA: Single-factor data analysis

All the fields in Figure 50 are calculated as previously described. We can see that the test statistics is F = 3.206928. Since p-value = F.DIST.RT(3.206928, 3, 36) = 0.034463 < .05 = α, we reject the null hypothesis that there is no significant difference between the evaluations of the four different types of peanut butter.

Follow-up Analysis

Although we now know that there is a significant difference between the four types of peanut butter, we still don't know where the differences lay.

It appears from Figure 50 that New 2 has a higher rating than the other types of peanut butter. We can use a t test to determine whether or not there is a significant difference between the rating for New 2 and New 1, the next highest-rated peanut butter. The result is shown in Figure 51:

t-Test: Two-Sample Assuming Unequal Variances

	New 1	*New 2*
Mean	13.1	16.6
Variance	21.21111	7.155556
Observations	10	10
Hypothesized Mean Difference	0	
Df	14	
t Stat	-2.07809	
P(T<=t) one-tail	0.028289	
t Critical one-tail	1.76131	
P(T<=t) two-tail	0.056577	
t Critical two-tail	2.144787	

Figure 51: Follow-up analysis: New 1 vs. New 2

From Figure 51, we see that there is no significant difference between the two types of peanut butter based on a two-tailed test (0.056577 > .05). If we instead compare the New 2 with the old formula, we get the results shown in Figure 52:

t-Test: Two-Sample Assuming Unequal Variances

	Old	*New 2*
Mean	11.1	16.6
Variance	18.76667	7.155556
Observations	10	10
Hypothesized Mean Difference	0	

Df	15
t Stat	-3.41607
P(T<=t) one-tail	0.001915
t Critical one-tail	1.75305
P(T<=t) two-tail	0.003829
t Critical two-tail	2.13145

Figure 52: Follow-up analysis: Old vs. New 2

This time, we see that there is a significant difference between the two types of peanut butter using a two-tailed t-test (.003829 < .05).

The problem with this approach is that doing multiple tests incurs higher amounts of what is called **experimentwise error**. Remember that when we use a significance level of α = .05, we accept that five percent of the time we will get a type I error. If we perform three such tests, then we will essentially increase our overall type I error to $1 - (1 - .05)^3 = .14$. This means that 14 percent of the time, we will have a type I error which is higher than we would like.

The general approach for addressing this issue is to either reduce α using **Bonferroni's correction** (for example, for three tests we use α/3 = .05/3 = .0167) or use a different type of test (for example, **Tukey's HSD** or **REGWQ**) which is beyond the scope of this book.

Levene's Test

As previously mentioned, the ANOVA test requires that the group variances be equal. There is a lot of leeway here and, even when the variance of one group is four times another, the analysis will be pretty good. Another option for testing homogeneity of group variances is by using **Levene's test.**

For Levene's test, the residuals e_{ij} of the group means from the cell means are calculated as follows:

$$e_{ij} = x_{ij} - \bar{x}_j$$

An ANOVA is then conducted on the absolute value of the residuals. If the group variances are equal, then the average size of the residual should be the same across all groups.

There are three versions of the test: using the mean (as described), using the median, or using the trimmed mean.

Example: Use Levene's test (residuals from the median) to determine whether the four groups in the ANOVA example have significantly different population variances.

We begin by calculating the medians for each group (range B14:E14 of Figure 53). For example, cell B14 contains the formula =MEDIAN(B4:B13).

Next, we create the table (in range B19:E29) with the absolute residuals from the median. We do this by entering the formula =ABS(B4-B$14) in cell B20, highlighting the range B20:E29 and then pressing **Ctrl-R** followed by **Ctrl-D**.

Finally, we use the **ANOVA: Single Factor** data analysis tool as described earlier in this chapter on the Input Range B19:E29. The output is shown on the right side of Figure 53:

	A	B	C	D	E	F	G	H	I	J	K	L	M
3		Old	New 1	New 2	New 3		Anova: Single Factor						
4		12	11	19	10								
5		8	9	20	17		SUMMARY						
6		6	16	15	20		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
7		16	15	14	11		Old	10	33	3.3	6.84444		
8		12	18	20	5		New 1	10	39	3.9	4.48889		
9		14	19	17	9		New 2	10	22	2.2	1.78889		
10		10	5	18	16		New 3	10	41	4.1	7.65556		
11		18	9	15	13								
12		4	12	16	4								
13		11	17	12	14		ANOVA						
14	med	11.5	13.5	16.5	12		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
15	var	18.7667	21.2111	7.15556	26.3222		Between Groups	21.875	3	7.29167	1.40374	0.25752	2.86627
16							Within Groups	187	36	5.19444			
17		Absolute value of residuals											
18							Total	208.875	39				
19		Old	New 1	New 2	New 3								
20		0.5	2.5	2.5	2								
21		3.5	4.5	3.5	5								
22		5.5	2.5	1.5	8								
23		4.5	1.5	2.5	1								
24		0.5	4.5	3.5	7								
25		2.5	5.5	0.5	3								
26		1.5	8.5	1.5	4								
27		6.5	4.5	1.5	1								
28		7.5	1.5	0.5	8								
29		0.5	3.5	4.5	2								

Figure 53: Levene's test

Since p-value = .25752 > .05 = α, we cannot reject the null hypothesis. So we can conclude that there is no significant difference between the four group variances. Thus, the ANOVA test previously conducted satisfies the homogenity of variances assumption.

Factorial ANOVA

Example

A new fertilizer has been developed to increase the yield on crops. The makers of the fertilizer want to better understand which of the three formulations (or blends) of this fertilizer are most effective for wheat, corn, soy beans, and rice (the crops). They test each of the three blends on five samples of each of the four types of crops. The crop yields for the five samples of each of the 12 combinations are as shown in Figure 54.

Determine whether or not there is a difference between crop yields based on these three types of fertilizer.

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

Figure 54: Sample data for Two-Factor ANOVA

Before we proceed with the analysis for this example, we should review a few basic concepts.

Basic Concepts

We now extend the one-factor ANOVA model previously described to a model with more than one factor.

A **factor** is an independent variable. A k factor ANOVA addresses k factors. The model for the last example contains two factors: crops and blends.

A **level** is some aspect of a factor; these are what we called groups or treatments in the one factor analysis previously described. The blend factor contains three levels and the crop factor contains four levels.

Let's look at the Two-Factor model in more detail.

Suppose we have two factors, A and B, where factor A has r levels and factor B has c levels. We organize the levels for factor A as rows and the levels for factor B as columns. We use the index i for the rows (i.e., factor A) and the index j for the columns (i.e., factor B).

This results in an $r \times c$ table whose entries are $\{X_{ij}: 1 \le i \le r, 1 \le j \le c\}$ where X_{ij} is a sample for level i of factor A and level j of factor B. Here, $X_{ij} = \{x_{ijk}: 1 \le k \le n_{ij}\}$. We further assume that the n_{ij} are all equal of size m and use k as an index to the sample entries.

We use terms such as $\bar{x}_i$ (or $\bar{x}_{i.}$) as an abbreviation for the mean of $\{x_{ijk}: 1 \le j \le r, 1 \le k \le m\}$. We use terms such as $\bar{x}_j$ (or $\bar{x}_{.j}$) as an abbreviation for the mean of $\{x_{ijk}: 1 \le i \le c, 1 \le k \le m\}$.

We also expand the definitions of sum of squares, mean square, and degrees of freedom for the one-factor model described earlier to the Two-Factor model as shown in Table 12:

SS	***Df***	***MS***
$SS_T = \sum_k \sum_j \sum_i \left(x_{ijk} - \bar{x}\right)^2$	$df_T = n - 1$	$MS_T = SS_T / df_T$
$SS_A = mc \sum_i (\bar{x}_i - \bar{x})^2$	$df_A = r - 1$	$MS_A = SS_A / df_A$
$SS_B = mr \sum_j \left(\bar{x}_j - \bar{x}\right)^2$	$df_B = c - 1$	$MS_B = SS_B / df_B$
$SS_{AB} = m \sum_j \sum_i \left(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}\right)^2$	$df_{AB} = (r-1)(c-1)$	$MS_{AB} = SS_{AB} / df_{AB}$
$SS_W = \sum_k \sum_j \sum_i \left(x_{ijk} - \bar{x}_{ij}\right)^2$	$df_W = n - rc$	$MS_W = SS_W / df_W$

Table 12: Two-Factor ANOVA terminology

Note that, in addition to the A and B factors, the model also contains the intersection between these factors (labeled AB in the table) which consists of the m elements in each of the $r \times c$ combinations of factor levels.

It is not hard to show that:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$df_T = df_A + df_B + df_{AB} + df_E$$

It turns out that, if the null hypothesis is true, then MS_W and MS_B are both measures of the same error. Thus, the null hypothesis is equivalent to the hypothesis that the population versions of these statistics are equal, that is:

$$\sigma_B = \sigma_W$$

We can, therefore, use the F-test described in Chapter 8 to determine whether or not to reject the null hypothesis, that is, if the x_{ij} are independently and normally distributed and all μ_j are equal (the null hypothesis), and all the σ_j^2 are equal (homogeneity of variances), then the test statistic:

$$F = \frac{MS_B}{MS_W}$$

has an F distribution with df_B, df_W degrees of freedom.

Analysis

To carry out the analysis for the last example, we will use the **ANOVA: Two Factor with Replication** data analysis tool. To access this tool, as usual, select **Data > Analysis|Data Analysis** and fill in the dialog box that appears as in Figure 55:

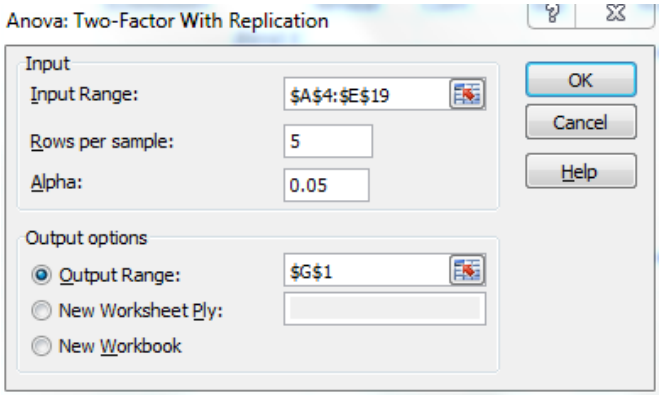


Figure 55: Dialog box for ANOVA: Two-Factor with Replication analysis tool

After pressing **OK**, the output shown in Figures 56 and 57 appear:

ANOVA: Two-Factor With Replication

SUMMARY	Wheat	Corn	Soy	Rice	Total
Blend X					
Count	5	5	5	5	20
Sum	659	677	879	706	2921
Average	131.8	135.4	175.8	141.2	146.05
Variance	844.2	707.8	278.7	354.2	782.3658
Blend Y					
Count	5	5	5	5	20
Sum	716	798	701	827	3042

Average	143.2	159.6	140.2	165.4	152.1
Variance	498.7	978.3	165.7	217.3	511.0421

Blend Z					
Count	5	5	5	5	20
Sum	822	868	932	862	3484
Average	164.4	173.6	186.4	172.4	174.2
Variance	443.3	428.8	212.3	175.8	330.6947

Total				
Count	15	15	15	15
Sum	2197	2343	2512	2395
Average	146.4667	156.2	167.4667	159.6667
Variance	705.8381	871.0286	605.981	404.9524

Figure 56: Two-Factor ANOVA descriptive statistics

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	8782.9	2	4391.45	9.933347	0.000245	3.190727
Columns	3411.65	3	1137.217	2.572355	0.064944	2.798061
Interaction	6225.9	6	1037.65	2.347138	0.045555	2.294601
Within	21220.4	48	442.0917			
Total	39640.85	59				

Figure 57: Two-Factor ANOVA

We can now draw some conclusions from the ANOVA table in Figure 57. Since the p-value (crops) = .0649 > .05 = α, we can't reject the Factor B null hypothesis. So we can conclude (with 95 percent confidence) that there are no significant differences between the effectiveness of the fertilizer for the different crops.

Since the p-value (blends) = .00025 < .05 = α, we reject the Factor A null hypothesis and conclude that there is a statistical difference between the blends.

We also see that the p-value (interactions) = .0456 < .05 = α, and so we conclude that there are significant differences in the interaction between crop and blend. We can look more carefully at the interactions by plotting the mean interactions between the levels of the two factors (see Figure 58). Lines that are roughly parallel indicate the lack of interaction, while lines that are not roughly parallel indicate interaction.

From the first chart, we can see that Brand X has quite a different pattern from the other brands (especially regarding soy). Although less dramatic, Brand Y is also different from Brand Z (especially since the line for Brand Y is trending up towards soy but trending down towards Rice, exactly the opposite of Brand Z):

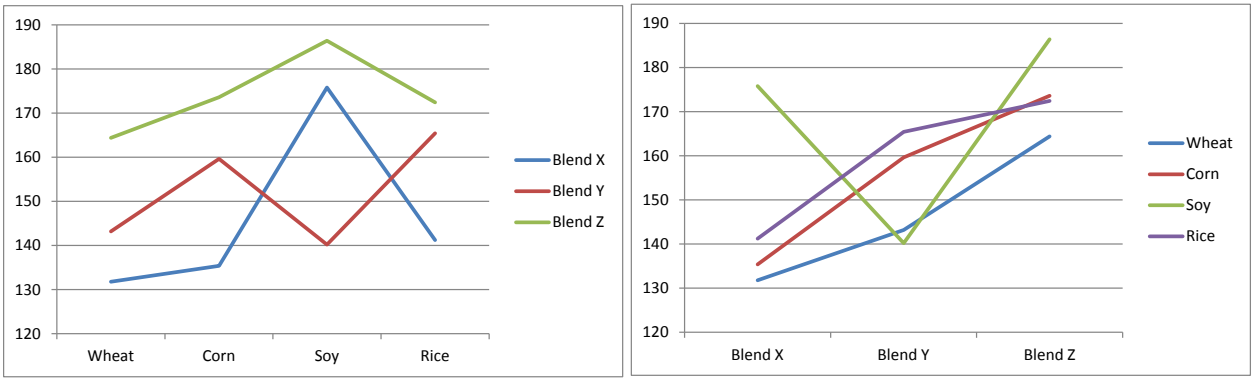


Figure 58: Interaction plots

Follow-up tests are carried out as for one-way ANOVA, although there are more possibilities.

ANOVA with Repeated Measures

Basic Concepts

ANOVA with repeated measures is an extension of the t test with paired samples test to more than two variables. As for the paired samples t test, the key difference between this type of ANOVA and the tests we have considered until now is that the variables under consideration are not independent.

Some examples of this type analysis are:

- A study is made of 30 subjects, each of whom is asked to take a test on driving a car, a boat, and an airplane
- A study is made of 30 rats, each of whom is given training once a day for 10 days with their scores recorded each day
- A study is made of 30 married couples and the husband's IQ is compared with his wife's

The important characteristic of each of these examples is that the treatments are not independent of each other. The most common of these analyses is to compare the results of some treatment given to the same participant over a period of time (as in the second example).

Example

A program has been developed to reduce the levels of anxiety for new mothers. In order to determine whether or not the program is successful, a sample of 15 women was selected and their level of anxiety was measured (low scores indicate higher levels of anxiety) before the program. Their level of anxiety was also measured one, two, and three weeks after the beginning of the program. Based on the data in Figure 59, determine whether the program is effective in reducing anxiety.

	A	B	C	D	E
3	Subject	Before	1 week	2 weeks	3 weeks
4	1	16	22	23	25
5	2	12	18	21	29
6	3	19	24	26	27
7	4	8	19	23	30
8	5	6	12	13	17
9	6	7	13	11	18
10	7	19	18	17	26
11	8	22	22	23	30
12	9	12	20	22	24
13	10	16	22	23	29
14	11	14	25	16	23
15	12	24	26	30	27
16	13	9	12	11	20
17	14	10	9	13	23
18	15	2	8	6	13

Figure 59: Sample data

Analysis

To perform ANOVA with repeated measures, we use Excel's **ANOVA: Two Factor without Replication** data analysis tool.

To access this tool, press **Data > Analysis|Data Analysis** and fill in the dialog box that appears as in Figure 60:

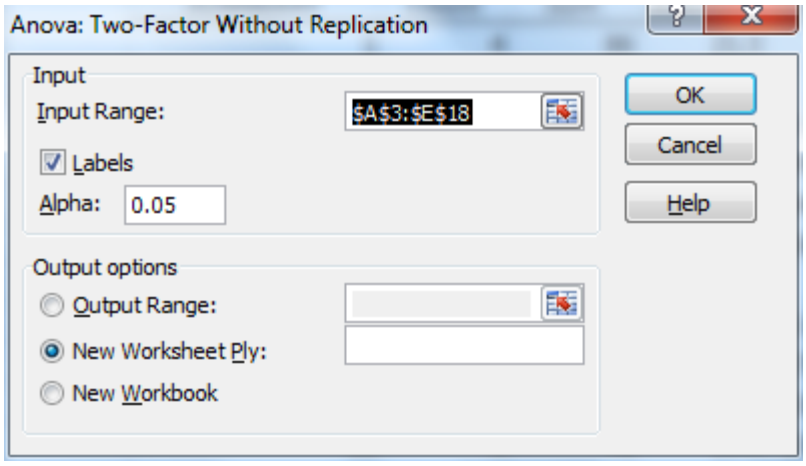


Figure 60: Dialog box for ANOVA: Two-Factor without replication

After pressing **OK**, the output shown in Figures 61 and 62 appears:

ANOVA: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1	4	86	21.5	15
2	4	80	20	50
3	4	96	24	12.66667
4	4	80	20	84.66667
5	4	48	12	20.66667
6	4	49	12.25	20.91667
7	4	80	20	16.66667
8	4	97	24.25	14.91667
9	4	78	19.5	27.66667
10	4	90	22.5	28.33333
11	4	78	19.5	28.33333
12	4	107	26.75	6.25
13	4	52	13	23.33333
14	4	55	13.75	40.91667
15	4	29	7.25	20.91667
Before	15	196	13.06667	39.35238
1 week	15	270	18	34.28571
2 weeks	15	278	18.53333	44.69524
3 weeks	15	361	24.06667	26.35238

Figure 61: ANOVA descriptive statistics

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1702.833	14	121.631	15.82722	1.95E-12	1.935009
Columns	910.9833	3	303.6611	39.51389	2.69E-12	2.827049
Error	322.7667	42	7.684921			
Total	2936.583	59				

Figure 62: Two-Factor ANOVA with repeated measures

For ANOVA with repeated measures, we aren't interested in the analysis of the rows, only with the columns, which correspond to variations by time. Since the test statistic F = 29.13 > 2.83 = F-crit (or p-value = 2.69E-12 < .05 = α) in Figure 62, we reject the null hypothesis. We can conclude that there are significant differences in mean anxiety levels by week.

Assumptions

The requirements for ANOVA with matched samples are as follows:

- Subjects are independent and randomly selected from the population
- Normality (or at least symmetry)–although overly restrictive, it is sufficient for this condition to be met for each treatment (that is, for each column in the data range)
- Sphericity

Data satisfies the **sphericity** property when the pairwise differences in variance between the samples are all equal. This assumption is stronger than the homogeneity assumption for the other versions of ANOVA previously described.

We perform the test for sphericity in Figure 63:

	T	U	V	W
3	Sphericity			
4				
5	T0-T1	14.06667		=VAR.S(B4:B18-C4:C18)
6	T0-T2	17.40952		=VAR.S(B4:B18-D4:D18)
7	T0-T3	19.42857		=VAR.S(B4:B18-E4:E18)
8	T1-T2	11.12381		=VAR.S(C4:C18-D4:D18)
9	T1-T3	17.20952		=VAR.S(C4:C18-E4:E18)
10	T2-T3	12.98095		=VAR.S(D4:D18-E4:E18)

Figure 63: Sphericity

Although the values in column U of Figure 63 are not equal, they aren't too far apart. Also, especially since the p-value obtained in Figure 62 is so low, we don't expect sphericity to be much of a problem for this example.

When sphericity is a problem, we can use a variety of correction factors to improve the ANOVA results. Unfortunately, explaining this will take us too far afield for our purposes here. If you would like more information, click here.

Follow-up Analysis

Since the mean score of 24.07 in Week 3 is much higher than the mean Before score of 13.07 (see Figure 61), it appears that there is a positive effect to be had from taking the program. We can confirm this by using a paired sample t test comparing Before with three weeks after, as shown in Figure 64:

t-Test: Paired Two Sample for Means

	Before	*3 weeks*
Mean	13.06667	24.06667
Variance	39.35238	26.35238
Observations	15	15
Pearson Correlation	0.718509	
Hypothesized Mean Difference	0	
Df	14	
t Stat	-9.66536	
P(T<=t) one-tail	7.11E-08	
t Critical one-tail	1.76131	
P(T<=t) two-tail	1.42E-07	
t Critical two-tail	2.144787	

Figure 64: Paired t test: Before vs. three weeks after

We see that p-value = 1.42E-07, which is considerably less than .05, thus confirming that there is a significant difference between Before and three weeks after.

In fact, we can see from Figure 65 that there is a significant improvement in the scores even after one week:

t-Test: Paired Two Sample for Means

	Before	*1 week*
Mean	13.06667	18
Variance	39.35238	34.28571
Observations	15	15
Pearson Correlation	0.810897	
Hypothesized Mean Difference	0	
Df	14	
t Stat	-5.09437	
P(T<=t) one-tail	8.17E-05	
t Critical one-tail	1.76131	
P(T<=t) two-tail	0.000163	
t Critical two-tail	2.144787	

Figure 65: Paired t test: Before vs. one week

In fact, even if we apply Bonferroni's correction factor using .05/2 = .025 (because we have now performed two t tests), we can see that p-value = .000163 < .025) which confirms that there is a significant improvement after only one week.

Chapter 10 Correlation and Covariance

Basic Definitions

The **sample covariance** between two sample random variables x and y is a measure of the linear association between the two variables and is defined by the formula:

$$cov(x,y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Big/ (n-1)$$

The covariance is similar to the variance, except that the covariance is defined for two variables (x and y) whereas the variance is defined for only one variable. In fact, $cov(x,x) = var(x)$.

The **correlation coefficient** between two sample variables x and y is a scale-free measure of linear association between the two variables and is defined by the formula:

$$r = cov(x,y)/s_x s_y$$

If necessary, we can write r as r_{xy} to explicitly show the two variables. We also use the term **coefficient of determination** for r^2.

If r is close to 1, then x and y are positively correlated. A **positive linear correlation** means that high values of y are associated with high values of x, and low values of y are associated with low values of x.

If r is close to -1, then x and y are negatively correlated. A **negative linear correlation** means that high values of y are associated with low values of x, and low values of y are associated with high values of x.

When r is close to 0, there is little linear relationship between x and y, that is, x and y are **independent**.

Scatter Diagram

To better visualize the association between two data sets $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, we can employ a chart called a **scatter diagram** (also called a **scatter plot**).

Example: A study is designed to check the relationship between smoking and longevity. A sample of 15 men 50 years or older was taken. The average number of cigarettes smoked per day and the age at death were recorded, as summarized on the left side of Figure 66. Create a scatter diagram to show the level of association between the number of cigarettes smoked and a person's longevity.

	A	B
3	Cigarettes	Longevity
4	5	80
5	23	78
6	25	60
7	48	53
8	17	85
9	8	84
10	4	73
11	26	79
12	11	81
13	19	75
14	14	68
15	35	72
16	29	58
17	4	92
18	23	65

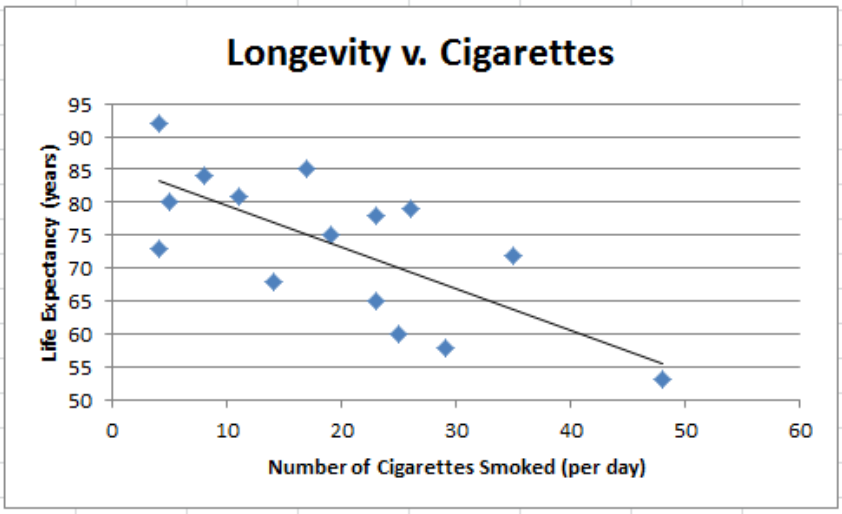


Figure 66: Scatter diagram

This is done as follows:

1. Highlight the range A4:B18.and select **Insert > Charts|Scatter**
2. The **Design**, **Layout**, and **Format** ribbons now appear to help you refine the chart. At any time, you can click the chart to get access to these ribbons.

The resulting chart is shown on the right side of Figure 66 although, initially, the chart does not contain a chart title or axes titles. We add these as follows:

3. To add a chart title, click the chart, select **Layout > Labels|Chart Title**, then choose **Above Chart**, and enter the title Longevity v. Cigarettes.
4. The title of the horizontal axis can be added in a similar manner by selecting **Layout > Labels|Axis Titles > Primary Horizontal Axis Title > Title Below Axis** and entering Number of Cigarettes Smoked (per day).
5. The title of the vertical axis is added by selecting **Layout > Labels|Axis Titles > Primary Vertical Axis Title > Rotated Title** and entering the Life Expectancy (years)
6. The default legend is not particularly useful and so you can delete it by selecting **Layout > Labels|Legend > None**

Since the smallest value of Longevity is 53, the chart will display better if we raise the horizontal axis. This is done as follows:

7. Double-click the vertical axis (0 to 100). **Format Axis** will appear as shown in Figure 67. Make sure the **Axis Options** tab is selected and change the **Minimum** field from **Auto** to **Manual** and then enter the value 50. Click **Close**:

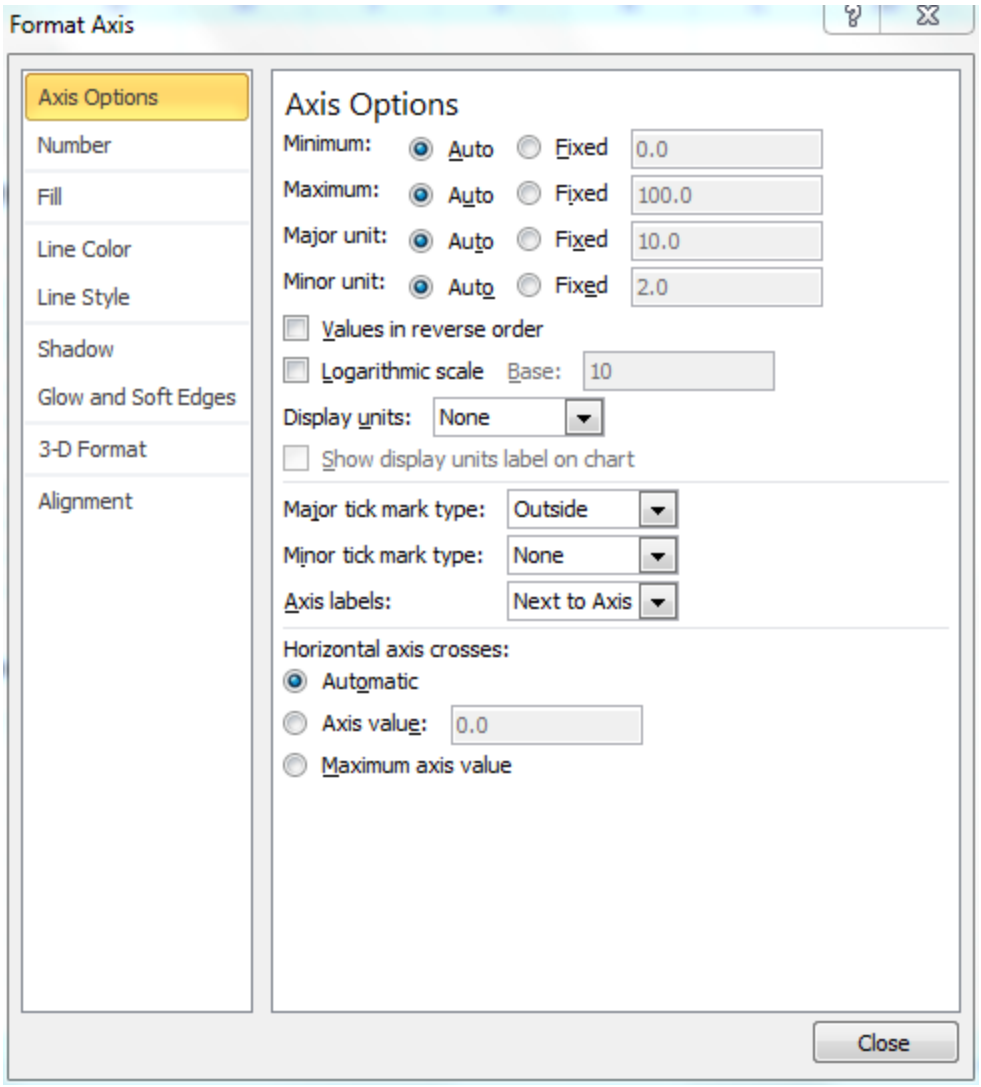


Figure 67: Format axis dialog box

Finally, we add the straight line that best fits the data points, as follows:

8. Select **Layout > Analysis|Trendline > Linear Trendline**

If desired, you can move the chart within the worksheet by left-clicking the chart and dragging it to the desired location. You can also resize the chart, making it a little smaller or bigger, by clicking one of the corners of the chart and dragging the corner to change the dimensions. To ensure that the aspect ratio (that is, the ratio of the length to the width) doesn't change, it is important to hold **Shift** down while dragging the corner.

It appears from the scatter diagram in Figure 66 that there is a negative linear correlation between longevity and cigarette smoking.

Excel Functions

Excel supplies the following worksheet functions for ranges R1 and R2 with the same number of elements:

COVARIANCE.P(R1, R2) = the population covariance between the data in range R1 and the data in range R2

COVARIANCE.S(R1, R2) = the sample covariance between the data in range R1 and the data in range R2

CORREL(R1, R2) = the correlation coefficient between the data in range R1 and the data in range R2

RSQ(R1, R2) = the coefficient of determination between the data in ranges R1 and R2; this is equivalent to the formula =CORREL(R1, R2) ^ 2

Note that no distinction needs to be made between the population and sample correlation coefficients since they have the same values.

COVARIANCE.P and COVARIANCE.S are only valid in versions of Excel starting with Excel 2010. For previous releases, Excel only supplies a function for the population covariance, namely, COVAR(R1, R2) which is equivalent to COVARIANCE.P(R1, R2). The sample covariance can be calculated using the formula:

=n * COVAR(R1, R2) / (n – 1)

where n = the number of elements in R1 (or R2)

For the data shown in Figure 66, the covariance is -97.44 and the correlation coefficient is -0.71343. These values are calculated by the formulas:

=COVARIANCE.S(A4:A18,B4:B18)

=CORREL(A4:A18,B4:B18)

Excel Functions with Missing Data

Note that in computing CORREL(R1, R2) or any of the covariance functions described in this chapter, any pairs for which one or both of the elements in the pair are non-numeric are simply ignored.

Example: Calculate the correlation between smoking and longevity based on the data in Figure 68:

	A	B
3	Cigarettes	Longevity
4	5	80
5	23	78
6	25	60
7	48	53
8	17	85
9	8	84
10	4	73
11	9	
12	26	79
13	11	81
14	19	75
15		80
16	14	68
17	35	72
18	29	58
19	4	92
20	23	65
21		
22	CORREL	-0.71343

Figure 68: Correlation coefficient with missing data

The data in Figure 68 is identical to that in Figure 66 except that two sample elements have been added (in rows 11 and 15) where one of the elements in the pair is blank. Since these elements are ignored, =CORREL(A4:B20) yields the value -.71343, the same as we calculated from the data in Figure 66.

Hypothesis Testing

It would be useful to know whether or not there is a significant correlation between smoking and longevity. A correlation of 0 would indicate that smoking and longevity are independent (that is, no association exists). Since the observed correlation coefficient of -.71343 is pretty far from 0, it appears from the data that there is a negative correlation. But we would like to know whether or not this is significant or due to random effects.

It turns out that, when the population correlation is 0, written $\rho = 0$ (that is, x and y are independent), then for sufficiently large values of the sample size n, the sample correlation r will be approximately normally distributed. So we can use a t test with n – 2 degrees of freedom where:

$$t = \frac{r}{s_r} \text{ and } s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Here s_r is the standard error of r.

Example: Determine whether or not there is a significant correlation between smoking and longevity.

The analysis is shown in Figure 69:

	A	B	C	D	E	F	G
2							
3	# Cig	Life Exp		t test (two-tail)			
4	5	80					
5	23	78		r	-0.71343		=CORREL(A4:A18,B4:B18)
6	25	60		ρ	0		
7	48	53		n	15		=COUNT(A4:A18)
8	17	85		df	13		=E7-2
9	8	84		s_r	0.194347		=SQRT((1-E5^2)/E8)
10	4	73		t	-3.67092		=E5/E9
11	26	79		α	0.05		
12	11	81		t-crit	2.160369		=T.INV.2T(E11,E8)
13	19	75		p-value	0.002822		=T.DIST.2T(ABS(E10),E8)
14	14	68		sig	yes		=IF(E13<E11,"yes","no")
15	35	72					
16	29	58		lower	-1.13329		=E5-E12*E9
17	4	92		upper	-0.29357		=E5+E12*E9
18	23	65					

Figure 69: Hypothesis testing of the correlation coefficient

Based on the results in Figure 69 (p-value = 0.002822 < .05 = α), we reject the null hypothesis that $\rho = 0$ and conclude that there is a significant correlation between smoking and longevity. It is important to note that such a correlation doesn't imply that we can claim that smoking causes lower longevity. For example, perhaps there is some other factor that causes a person to smoke and to have a shorter life expectancy.

If $\rho \neq 0$, then the approach previously described won't work since, even for large values of n, the sample correlation r is not approximately normally distributed. Fortunately, there is a transformation of the correlation coefficient which will address this defect.

For any r, define the **Fisher transformation** of r as follows:

$$r' = \frac{1}{2}\ln\frac{1+r}{1-r} = (\ln(1+r) - \ln(1-r))/2$$

For n is sufficiently large, the Fisher transformation r' of the correlation coefficient r for samples of size n has a normal distribution with standard deviation:

$$s_{r'} = \frac{1}{\sqrt{n-3}}$$

Excel provides functions that calculate the Fisher transformation and its inverse.

FISHER (r) = .5 * LN((1 + r) / (1 – r))

FISHERINV(z) = (EXP(2 * z) – 1) / (EXP(2 * z) + 1)

Example: Test whether or not the correlation coefficient for the data in Figure 66 is significantly different from -0.5:

	A	B	C	D	E	F	G
3	Cigarettes	Longevity					
4	5	80		n	15		=COUNT(A4:A18)
5	23	78		r	-0.71343		=CORREL(A4:A18,B4:B18)
6	25	60		r'	-0.89414		=FISHER(E5)
7	48	53		ρ	-0.5		
8	17	85		ρ'	-0.54931		=FISHER(E7)
9	8	84		s.e.	0.267261		=1/SQRT(E4-1)
10	4	73		z	1.290232		=ABS(E6-E8)/E9
11	26	79		p-value	0.098485		
12	11	81		alpha	0.05		
13	19	75		sig	no		=IF(E11<E12/2,"yes","no")
14	14	68		z-crit	1.959964		=NORM.S.INV(1-E12/2)
15	35	72		lower'	-1.41796		=E6-E14*E9
16	29	58		upper'	-0.37031		=E6+E14*E9
17	4	92		lower'	-0.88917		=FISHERINV(E15)
18	23	65		upper'	-0.35427		=FISHERINV(E16)

Figure 70: Hypothesis testing using the Fisher transformation

From Figure 70, we see that there is no significant difference between the population correlation coefficient and -.5 (p-value = 0.098485 > .05 = α). While this is not particularly useful information, the calculation of the 95 percent confidence interval for the population correlation coefficient is. This confidence interval is calculated as follows:

$$\rho' = r' \pm z_{crit} \cdot s.e.$$

$$\rho_{lower} = \text{FISHERINV}(\rho'_{lower}) \qquad \rho_{upper} = \text{FISHERINV}(\rho'_{upper})$$

We see that the 95 percent confidence interval is (-.889,-.354). Note that 0 doesn't lay in this interval, once again demonstrating that the correlation coefficient is significantly different from 0 (as previously shown). Since -.5 is in this interval, we can't rule out that the actual population correlation coefficient is -.5 (with 95 percent confidence).

Data Analysis Tools

Excel provides the **Correlation** data analysis tool to calculate the pairwise correlation coefficients for a number of variables.

Example: Calculate the pairwise correlation coefficients and covariance for the data in Figure 71:

	A	B	C	D	E	F	G	H	I	J
3		Poverty	Infant Mort	White	Crime	Doctors	Traf Deaths	University	Unemployed	Income
4	Alabama	15.7	9.0	71.0	448	218.2	1.81	22.0	5.0	42,666
5	Alaska	8.4	6.9	70.6	661	228.5	1.63	27.3	6.7	68,460
6	Arizona	14.7	6.4	86.5	483	209.7	1.69	25.1	5.5	50,958
7	Arkansas	17.3	8.5	80.8	529	203.4	1.96	18.8	5.1	38,815
8	California	13.3	5.0	76.6	523	268.7	1.21	29.6	7.2	61,021
9	Colorado	11.4	5.7	89.7	348	259.7	1.14	35.6	4.9	56,993
10	Connecticut	9.3	6.2	84.3	256	376.4	0.86	35.6	5.7	68,595
11	Delaware	10.0	8.3	74.3	689	250.9	1.23	27.5	4.8	57,989
12	Florida	13.2	7.3	79.8	723	247.9	1.56	25.8	6.2	47,778
13	Georgia	14.7	8.1	65.4	493	217.4	1.46	27.5	6.2	50,861
14	Hawaii	9.1	5.6	29.7	273	317.0	1.33	29.1	3.9	67,214
15	Idaho	12.6	6.8	94.6	239	168.8	1.60	24.0	4.9	47,576

Figure 71: Sample data

We use Excel's **Correlation** data analysis tool by selecting **Data > Analysis | Data Analysis** and selecting **Correlation** from the menu that is displayed. The dialog box shown in Figure 72 appears:

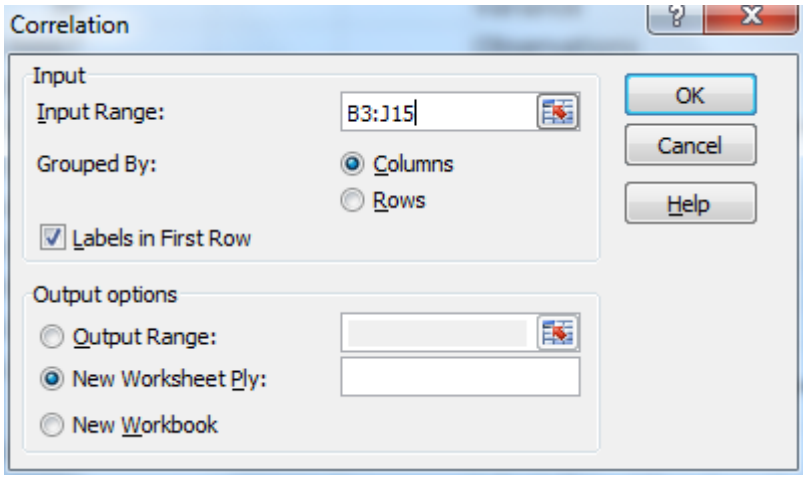


Figure 72: Dialog box for the correlation data analysis tool

Using Excel's **Correlation** data analysis tool, we can compute the pairwise correlation coefficients for the various variables shown in the table in Figure 71. The results are shown in Figure 73:

	L	M	N	O	P	Q	R	S	T	U
3		*Poverty*	*Infant Mort*	*White*	*Crime*	*Doctors*	*Traf Deaths*	*University*	*Unemployed*	*Income*
4	Poverty	1								
5	Infant Mort	0.47731411	1							
6	White	0.29304347	0.05942314	1						
7	Crime	0.12282942	0.40465819	0.04179228	1					
8	Doctors	-0.5952193	-0.4776715	-0.3427498	-0.2886721	1				
9	Traf Deaths	0.65927078	0.58554357	0.03498247	0.31209584	-0.7946938	1			
10	University	-0.6732119	-0.6659181	-0.006657	-0.3330991	0.74015762	-0.9299867	1		
11	Unemployed	0.07172913	-0.1582057	0.28175584	0.48393135	-0.0409456	-0.0526137	0.09333291	1	
12	Income	-0.900522	-0.6482486	-0.367791	-0.1594619	0.72263637	-0.7308936	0.75731823	0.14939997	1

Figure 73: Pairwise correlation coefficients

Similarly, we use Excel's **Covariance** data analysis tool to obtain the covariances as shown in Figure 74:

	L	M	N	O	P	Q	R	S	T	U
15		*Poverty*	*Infant Mort*	*White*	*Crime*	*Doctors*	*Traf Deaths*	*University*	*Unemployed*	*Income*
16	Poverty	7.51354167								
17	Infant Mort	1.59958333	1.49472222							
18	White	12.8291774	1.16032659	255.087292						
19	Crime	53.9952083	79.34125	107.046128	25719.4052					
20	Doctors	-86.505218	-30.963675	-290.24485	-2454.5853	2811.16188				
21	Traf Deaths	0.54175	0.21461111	0.16749725	15.0048333	-12.631516	0.08987222			
22	University	-8.666875	-3.82375	-0.4993558	-250.89521	184.312817	-1.3094167	22.0585417		
23	Unemployed	0.17520833	-0.1723611	4.0100905	69.159375	-1.9345795	-0.0140556	0.390625	0.79409722	
24	Income	-23837.104	-7653.475	-56726.022	-246958.62	369998.15	-2115.9417	34348.1708	1285.65417	93255319.6

Figure 74: Pairwise covariances

Chapter 11 Linear Regression

Linear Regression Analysis

The goal of linear regression is to create a model from observed data which captures the relationship between an independent variable x and a dependent y, and to use this model to predict the values of the dependent variable based on values of the independent variable (especially for values of the independent variable that were not originally observed). Even when we can make such predictions, this doesn't imply that we can claim any causal relationship between the independent and dependent variables.

Regression Line

Essentially, we are looking for a straight line that best fits the observed data $(x_1, y_1), \ldots, (x_n, y_n)$.

Recall that the equation for a straight line is $y = bx + a$, where:

b = the slope of the line
a = y-intercept, that is, the value of x where the line intersects with the y-axis

Using a technique called **ordinary least squares**, it turns out that:

$$b = cov(x, y) \,/\, var(x)$$

$$a = \bar{y} - b\bar{x}$$

Excel provides the following functions where R1 is a range containing y data values and R2 is a range with x data values:

SLOPE(R1, R2) = slope of the regression line as previously described

INTERCEPT(R1, R2) = y-intercept of the regression line as previously described

From our previous observation:

b = SLOPE(R1, R2) = COVARIANCE.S(R1, R2) / VAR.S(R2)

a = INTERCEPT(R1, R2) = AVERAGE(R1) – b * AVERAGE(R2)

Example: Find the regression line for the data in Figure 66 (repeated on the left side of Figure 75). Based on this model, what is the life expectancy of someone who smokes 10, 20, or 30 cigarettes a day?

The results are shown on the right side of Figure 75:

	A	B	C	D	E	F	G
3	Cigarettes	Longevity					
4	5	80		slope	-0.6282		=SLOPE(B4:B18,A4:A18)
5	23	78		intercept	85.720421		=INTERCEPT(B4:B18,A4:A18)
6	25	60					
7	48	53		cigarettes	longevity		
8	17	85		10	79.438417		=TREND(B4:B18,A4:A18,D8)
9	8	84		20	73.156413		=FORECAST(D9,B4:B18,A4:A18)
10	4	73		30	66.874409		=D10*E4+E5
11	26	79					
12	11	81					
13	19	75					
14	14	68					
15	35	72					
16	29	58					
17	4	92					
18	23	65					

Figure 75: Regression line and predictions

Therefore, the regression line is:

$$y = -0.6282x + 85.7204$$

For any value of x (= # of cigarettes smoked), the regression equation will generate a value for y which is a prediction of the life expectancy of a person who smokes that number of cigarettes.

For example, if a person smokes 30 cigarettes a day, the model forecasts that the person will live 66.87 years (cell E10 in Figure 75).

Excel provides the following functions to help carry out these forecasts, where R1 is a range containing y data values and R2 is a range with x data values:

FORECAST(x, R1, R2) = calculates the predicted value y for the given value x of x

TREND(R1, R2, R3) = array function which predicts the y values corresponding to the x values in R3

We show how to use FORECAST in cell E9 of Figure 75 and TREND in cell E8 of Figure 75.

Actually, TREND is an array function and so it can be used to carry out multiple predictions. In fact, if we highlight the range E8:E10, enter the formula =TREND(B4:B18,A4:A18,D8;D10), and then press **Ctrl-Shift-Enter**, we will get the same results as shown in Figure 75.

Residuals

If range R3 is omitted, then TREND will generate the forecasted values of y for the various x values in R1. We can use this function to see how well the regression model predicts the y values which we have observed. This will help us determine how good the regression model really is. The result is shown in Figure 76:

	A	B	C	D	E
3	Cigarettes	Longevity		Forecast	Residual
4	5	80		82.5794192	-2.5794192
5	23	78		71.2718119	6.7281881
6	25	60		70.0154111	-10.015411
7	48	53		55.5668017	-2.5668017
8	17	85		75.0410143	9.9589857
9	8	84		80.694818	3.305182
10	4	73		83.2076196	-10.20762
11	26	79		69.3872107	9.6127893
12	11	81		78.8102167	2.1897833
13	19	75		73.7846135	1.2153865
14	14	68		76.9256155	-8.9256155
15	35	72		63.733407	8.266593
16	29	58		67.5026094	-9.5026094
17	4	92		83.2076196	8.7923804
18	23	65		71.2718119	-6.2718119
19	mean	73.5333333		73.5333333	8.527E-15
20	var	120.266667			

Figure 76: Residuals

Here we calculate the forecasted values in the range D4:D18 using the array formula:

=TREND(B4:B18,A4:A18)

The **residuals**, shown in the range E4:E18, are the differences between the actual y values and the forecasted y values. These are the error terms. They can be calculated by inserting the formula =B4-D4 in cell E4, highlighting the range E4:E18 and pressing **Ctrl-D**.

Although the forecasts are not 100 percent accurate, the mean of the forecasted values (cell D19) is the same as the mean of the observed values (cell B19), and the mean of the error terms (cell E19) is zero.

Model Fit

The sample variance of the observed y values (Longevity) can be calculated by the formula =VAR.S(B4:B18) and has a value 120.27. For our purposes, we will call this variance the **total mean square** (abbreviated MS_T). As we have seen in Chapter 3, the sample variance can be calculated by the formula:

$$MS_T = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\text{DEVSQ(B4: B18)}}{\text{COUNT(B4: B18)} - 1} = \frac{1683.73}{14} = 120.27$$

The numerator is called the **total sum of squares** (abbreviated SS_T) and the denominator is the **total degrees of freedom** (abbreviated df_T). The only thing new is the terminology.

It turns out that we can segregate the value of SS_T (which represents the total variability of the observed y values) into the portion of variability explained by the model (the **regression sum of squares** SS_{Reg}) and the portion not explained by the model (the **residual sum of squares** SS_{Res}) and, similarly, for the degrees of freedom. Thus, we have:

$$SS_T = SS_{Reg} + SS_{Res}$$

$$df_T = df_{Reg} + df_{Res}$$

Now the values of each of these terms are:

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

$$df_T = n - 1 \qquad df_{Reg} = k \qquad df_{Res} = n - k - 1$$

where the y_i are the observed values for y, the $\hat{y}_i$ are values of y predicted by the model and k = the number of independent variables in the model (k = 1 for now). We can also introduce mean squares values:

$$MS_T = \frac{SS_T}{df_T} \qquad MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} \qquad MS_{Res} = \frac{SS_{Res}}{df_{Res}}$$

Clearly the model fits the observed data well if SS_{Res} is small in comparison to SS_{Reg}. Since $SS_T = SS_{Reg} + SS_{Res}$, this is equivalent to saying that $\frac{SS_{Reg}}{SS_T}$ is relatively close to 1. It turns out that this fraction is equal to the coefficient of determination, that is, r^2. It is common to use a capital R instead of a small r and so we have:

$$R^2 = \frac{SS_{Reg}}{SS_T}$$

All these statistics can be calculated in Excel as shown in Figure 77:

	G	H	I	J	K	L	M
3		SS	df	MS			
4	T	1683.733	14	120.2667		R^2	0.508983
5	Reg	856.991	1	856.991		R	0.71343
6	Res	826.7423	13	63.59556			

Figure 77: Calculation of regression parameters in Excel

The formulas used are shown in Table 13 (based on the data Figure 76):

Cell	Entity	Formula
H4	SS_T	=DEVSQ(B4:B18)
H5	SS_{Reg}	=DEVSQ(D4:D18)
H6	SS_{Res}	=DEVSQ(E4:E18)
I4	df_T	=COUNT(B4:B18)-1
I5	df_{Reg}	=I5
I6	df_{Res}	=I4-I5

Cell	Entity	Formula
J4	MS_T	=H4/I4
J5	MS_{Reg}	=H5/I5
J6	MS_{Res}	=H6/I6
M4	R^2	=H5/H4
M5	R	=SQRT(M4)

Table 13: Regression formulas

The residual mean squares term MS_{Res} is a sort of variance for the error not captured by the model. The **standard error of the estimate** is simply the square root of MS_{Res}. For the current example, this is the square root of 63.596 (cell J6 of Figure 77), that is, 7.975. Excel provides the following function to calculate the standard error of the estimate:

> **STEYX**(R1, R2) = standard error of the estimate for the regression model where R1 is a range containing the observed y data values and R2 is a range containing the observed x data values

Multiple Regression Analysis

We now extend the linear regression model previously introduced with one independent variable to the case where there are one or more independent variables:

If y is a dependent variable and $x_1, \ldots, x_k$ are independent variables, then the **general multiple regression model** provides a prediction of y from the x_i of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients of the model and ε is the random error.

We further assume that for any given values of the x_i the random error ε is normally and independently distributed with mean zero. Essentially, the last of these conditions means that the model is a good fit for the data even if we eliminate the random error term.

Example: A jeweler prices diamonds on the basis of quality (with values from 0 to 8, with 8 being flawless and 0 containing numerous imperfections) and color (with values from 1 to 10, with 10 being pure white and 1 being yellow). Based on the price per carat of the 11 diamonds, each weighing between 1.0 and 1.5 carats as shown in Figure 78, build a regression model which captures the relationship between quality, color, and price.

Based on this model, estimate the price for a diamond whose color is 4 and whose quality is 6:

	A	B	C	D
3	Id	Color	Quality	Price
4	1	7	5	65
5	2	3	7	38
6	3	5	8	51
7	4	8	1	38
8	5	9	3	55
9	6	5	4	43
10	7	4	0	25
11	8	2	6	33
12	9	8	7	71
13	10	6	4	51
14	11	9	2	49

Figure 78: Sample data

We use Excel's Regression analysis tool by pressing **Data > Analysis | Data Analysis** and selecting **Regression** from the menu that is displayed. The dialog box shown in Figure 79 appears:

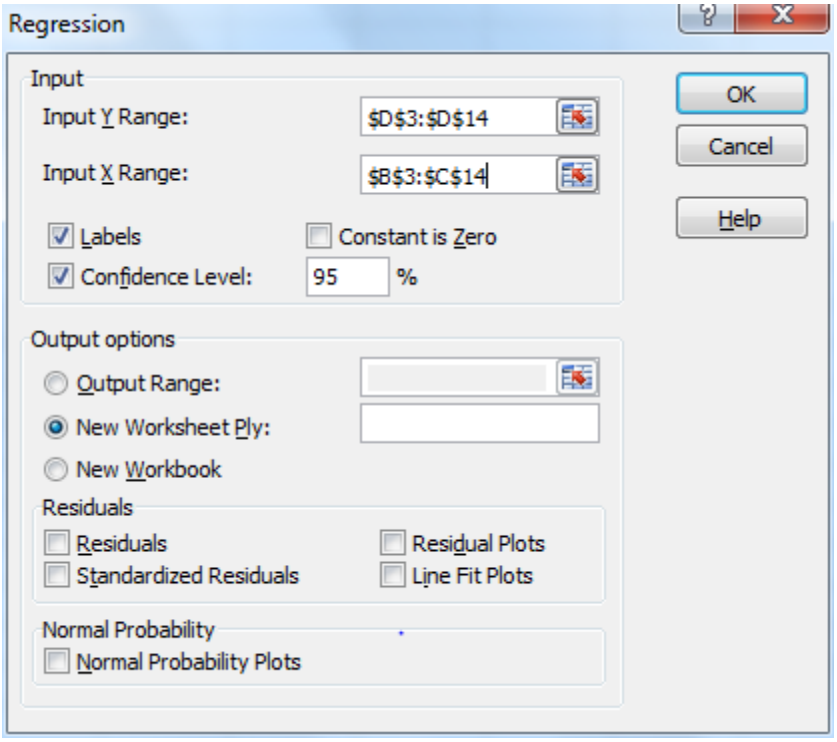


Figure 79: Dialog box for regression data analysis

After filling in the fields as shown in the figure and selecting **OK**, the output shown in Figure 80 will be displayed.

The calculations of the values in Figure 80 are extensions of the approaches described earlier in this chapter for the case where there is only one independent variable (see especially Table 13). **Adjusted R Square** (cell B6) is an attempt to create a better estimate of the true value of the population coefficient of determination, and is calculated by the formula:

=1-(1-B5)*(B8-1)/(B8-B12-1)

The calculations of the regression coefficients (B17:B19) and the corresponding standard errors (C17:C19) use the method of least squares but is beyond the scope of this book. The calculation of the p-values and confidence intervals for these coefficients is as described in Chapter 7.

One of the key results of the analysis in Figure 80 is that the regression model is given by the formula:

$$\text{Price} = 1.7514 + 4.8953 \cdot \text{Color} + 3.7584 \cdot \text{Quality}$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.922330727					
5	R Square	0.850693971					
6	Adjusted R Square	0.813367463					
7	Standard Error	5.888084465					
8	Observations	11					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	1580.280054	790.1400271	22.79061267	0.000496946	
13	Residual	8	277.3563093	34.66953867			
14	Total	10	1857.636364				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1.751403659	6.960202671	0.251631129	0.807669624	-14.29885248	17.8016598
18	Color	4.895288365	0.820229778	5.968191467	0.000335084	3.003835104	6.786741625
19	Quality	3.758415483	0.756510987	4.968091073	0.00109572	2.013898018	5.502932948

Figure 80: Regression data analysis

Since the p-value (cell F12) = 0.00497 < .05, we reject the null hypothesis. We conclude that the regression model is a good fit for the data.

Note that the Color and Quality coefficients are significant (p-value < .05), that is, they are significantly different from zero. Meanwhile, the intercept coefficient is not significantly different from zero (p-value = .8077 > .05).

That R square = .8507 indicates that a good deal of the variability of price is captured by the model, which supports the case that the regression model is a good fit for the data.

If a diamond has color 4 and quality 6, then we can use the model to estimate the price as follows:

$$\text{Price} = 1.75 + 4.90 \cdot \text{Color} + 3.76 \cdot \text{Quality} = 1.75 + 4.90 \cdot 4 + 3.76 \cdot 6 = 43.88305$$

We get the same result using the TREND function as shown in Figure 81:

	M	N	O	P	Q
2	color	quality	price		
3	4	6	43.88305		=TREND(C4:C14,A4:B14,M3:N3)

Figure 81: Prediction using regression model