EXCEL MANUAL

for Moore, McCabe, and Craig's

Introduction to the Practice of Statistics
Sixth Edition

Betsy Greenberg
University of Texas – Austin

W.H. Freeman and Company
New York

Contents

PREFACE

This Excel Manual is a supplement to Statistics textbooks by David S. Moore, et al. that are published by W.H. Freeman. This manual is intended to help the student perform the analysis described in those textbooks.

Excel is widely available as part of Microsoft Office. It contains some statistical functions in its basic installation. It also comes with statistical routines in the Data Analysis Toolpak, an add-in found separately on the Microsoft Office CD. Excel is a useful teaching and learning tool, however it is not meant to replace more sophisticated statistical tools such as SPSS or SAS. People often use Excel as their everyday statistics software because they have already purchased it. This manual helps students understand the capabilities of Excel for statistical analysis.

In addition to describing the standard features of Excel, this manual also illustrates the capabilities of the WHFStat Add-In module. The WHFStat Add-In module is available from W.H. Freeman. The module is programmed to include the following procedures and graphical analyses under the umbrella of a single menu.

- Descriptive statistics
- Probability calculations
- Discrete probability Distributions
- Estimating and Testing Means
- Proportion Testing
- Correlation and Regression
- Time Series Forecasting
- Two-way table and Chi-squared test
- Analysis of variance
- Graphs including normal quantile plots, boxplots, and control charts

Betsy Greenberg
University of Texas at Austin
August 2008

CHAPTER

0

Introduction

Microsoft Excel is a widely used spreadsheet application that millions of people use in their personal and professional lives to store, analyze, and present information. This manual describes how Microsoft Excel can be used effectively in your statistics course.

Using Excel

Microsoft Excel, commonly referred to as just Excel, is a spreadsheet program that organizes data in columns and rows, much like an accounting worksheet or table of data. Excel can also perform statistical analysis using built-in functions.

The WHFStat Add-In software works within Excel to group all of the statistics functions into one menu. This software is described in section 0.8, and is available on StatsPortal, your Online Study Center, or packaged with this manual.

Versions of Excel

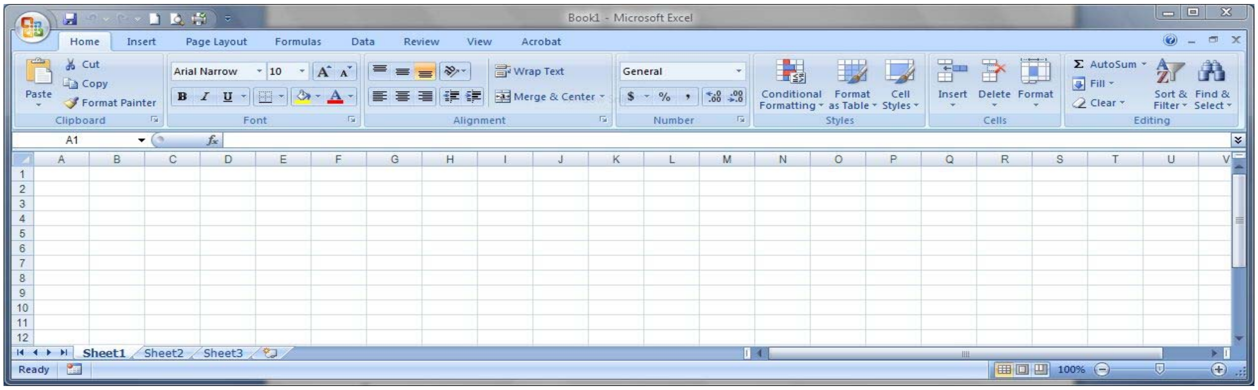
The examples in this book were written using Microsoft Excel 2007. The WHFStat Add-In module operates with Excel 2003 or Excel 2007 under either the Windows Vista or Windows XP operating systems. It is also compatible with Excel 2004 for Macintosh. Versions of Excel prior to version 2003 cannot be used with this software.

Prior Knowledge

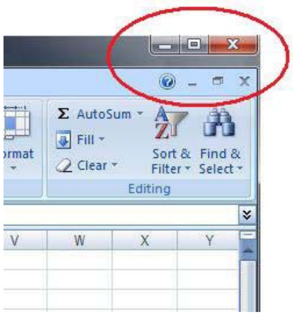
It is not necessary to have any prior knowledge of Excel to use this manual. However, it will be helpful to become familiar with Excel before using it for statistical analysis.

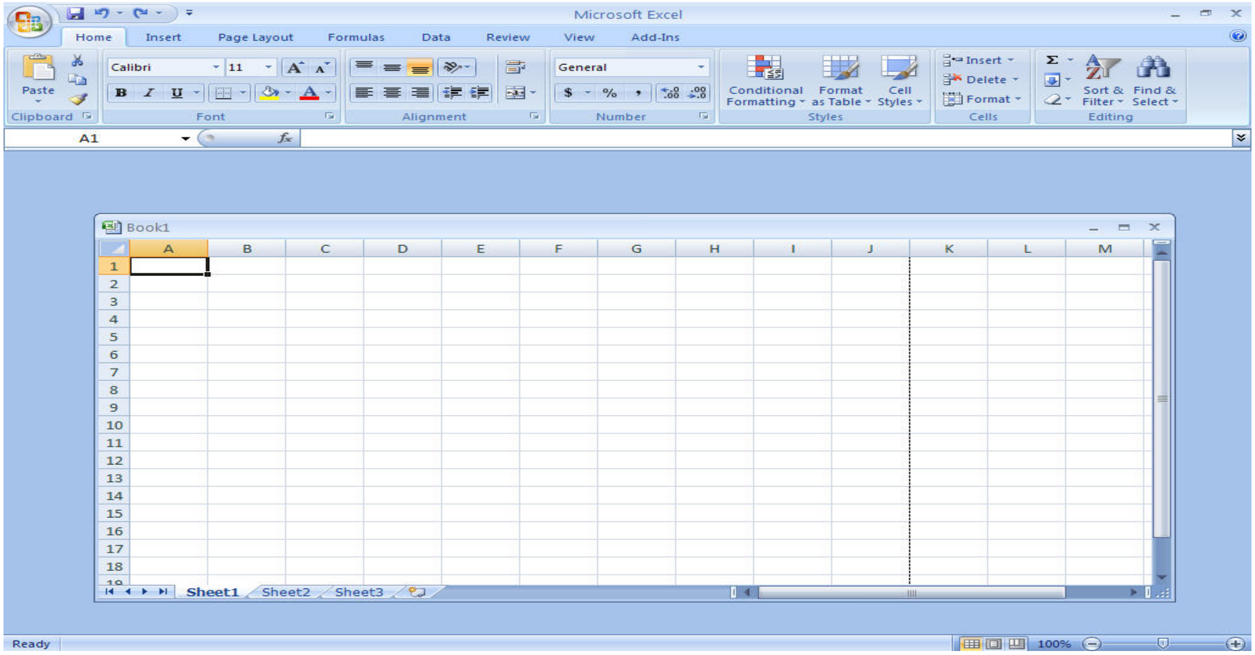
Worksheet Basics

When Excel is launched, a new file opens with a series of blank worksheets, also known as "sheets." The file itself is called a "workbook," which refers to the entire collection of spreadsheets, graphs, and user-developed programming code in the file. The figure below is a screenshot of a blank sheet in the Excel 2007 application.

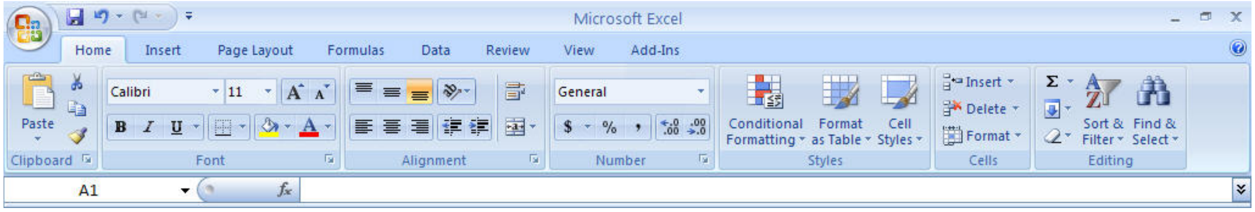


In the upper-right corner of the application window are three buttons that allow the user to minimize, maximize, or close the window. Notice that there are two sets of these buttons, one in the top grey portion of the window and one in the lighter blue area. This is because the Excel application is actually displaying one window within another. Clicking the middle of the three buttons (ignore the question mark button for now) in the light-blue area will make the two windows more prominent, as shown below.

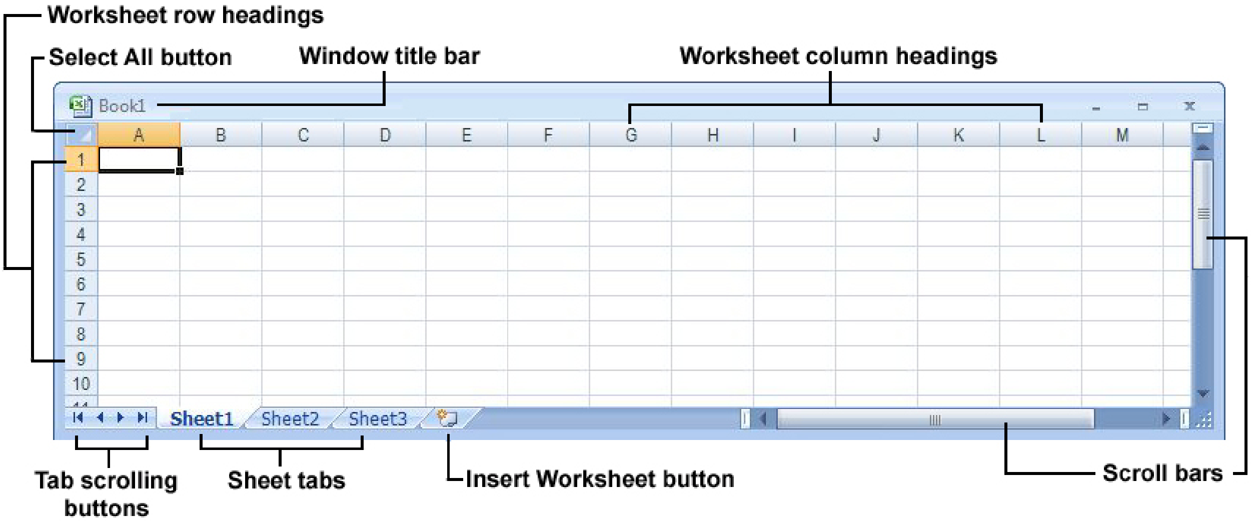




The outer window is the Excel application window, which contains all of the buttons and menus that control the functionality of the program. The inner window contains the workbook with all of its sheets.



Looking more closely at this inner window reveals a number of controls that allow the user to navigate around the active worksheet or to display other sheets in the workbook.

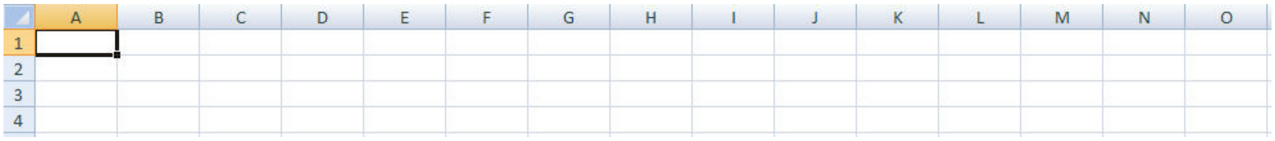


Sheet Tabs

Each worksheet is labeled with a tab at the bottom of the workbook, and individual sheets are activated by clicking these tabs. More than one sheet can be activated by selecting the first sheet, holding the Control Key down, and selecting additional sheets as required. If there are too many sheet tabs to be displayed all at once, the tab scrolling buttons can be used to bring a particular tab into view. Double clicking or right-clicking the name on a tab allows the sheet to be renamed. Sheets can be rearranged by dragging and dropping a given sheet tab to a new location within the group of tabs as a whole. Clicking the Insert Worksheet button adds a new blank sheet to the workbook. The scroll bars allow the portion of the spreadsheet currently displayed to be moved left or right, up or down.

Rows, Columns, and Cells

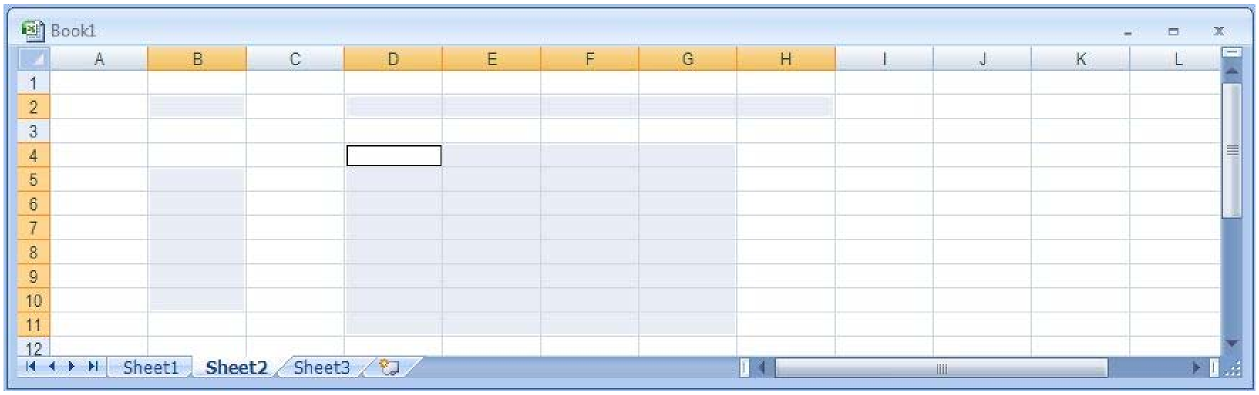
Notice that the worksheet is divided into a series of columns labeled with letters at the top, and a series of rows labeled with numbers on the far left. At the intersection of any column and row is a discrete portion of the sheet called a cell. All numeric and text data for a worksheet is housed within these cells. An individual cell is identified by the row and column in which it resides. For example, the cell located at the intersection of column A and row 1 is identified as A1, which is also known as the cell's address.



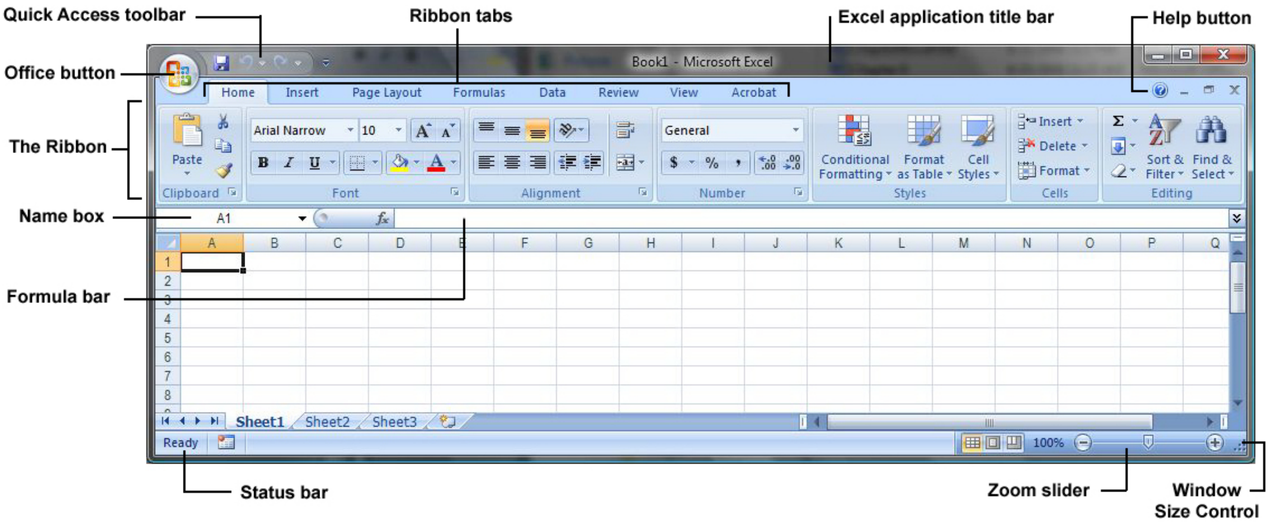
Selecting Cells and Ranges

In order to enter data in a cell, the target cell must first be selected. A cell is selected by using the mouse to click on a specific cell's location or by typing the arrow keys until the proper cell is reached. When a cell is selected, it is surrounded by a heavy black outline and the row and column headings corresponding to that cell are highlighted, as shown for cell A1 in the picture above. The highlighted cell is known as the active cell, and any numbers or text revisions to the spreadsheet will always be added to this active cell. When a new workbook is created, cell A1 in Sheet1 is automatically selected as the active cell.

Clicking and dragging across more than one cell selects all the cells across the entire region, known as a cell range. To select multiple ranges, select the first range, hold the Control key down, and select any other ranges of cells. Clicking the Select All button selects all cells in the current worksheet. The worksheet shown below illustrates four types of ranges: an individual cell, a partial row of cells, a partial column of cells, and a range crossing multiple columns and rows. A range must be a rectangular shape or a group of adjacent cells. The active cell is always the upper-left cell of the last selected range. Just like individual cells, ranges have addresses that describe the cells they contain. A range address is composed of the upper-left cell in the range, a colon, and the lower-right cell in the range. The pictured ranges have the following range addresses: B2, B5:B10, D2:H2, D4:G11. The current active cell as pictured is D4.



Using Excel's Functionality



The Ribbon

Excel 2007 introduced a new interface for accessing Excel's various controls and functionality called the Ribbon. The ribbon provides a series of context specific commands, grouped together so that similar commands display at the same time. The various groups of commands are accessed by selecting a ribbon tab near the top of the ribbon. Each tab displays a series of related commands.

Brief overviews of the commands available in the ribbon tabs are outlined in the table below:

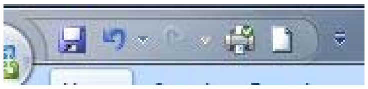
Home	The most commonly used functions — cut, copy, and paste; font formatting and alignment, number formats, cell background color and borders, inserting and deleting cells, sorting, filtering, and finding/replace functions
Insert	Functions to insert tables, charts, artwork, graphics or specialized text

Page Layout	Printing options, workbook themes and colors, margins, page breaks, and scaling
Formulas	Controls to assist the user in creating, editing, and auditing formulas and calculation options
Data	Sorting and filtering, data validation, outlining, connecting to data in external sources such as databases or the internet
Review	Spellcheck and proofing tools, protecting and sharing workbooks, adding comments, tracking changes
View	Display the workbook in various ways, hide/display gridlines and headings, arrange and size windows

The Office Button

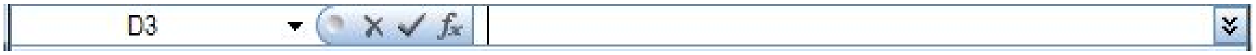
Commands to open a new or existing workbook, save changes, and print can be found by clicking the circular Office button in the upper-left corner of the application window. For users accustomed to Excel versions prior to 2007, these commands correspond to the old File menu, which is not part of the new Excel ribbon.

The Quick Access Toolbar

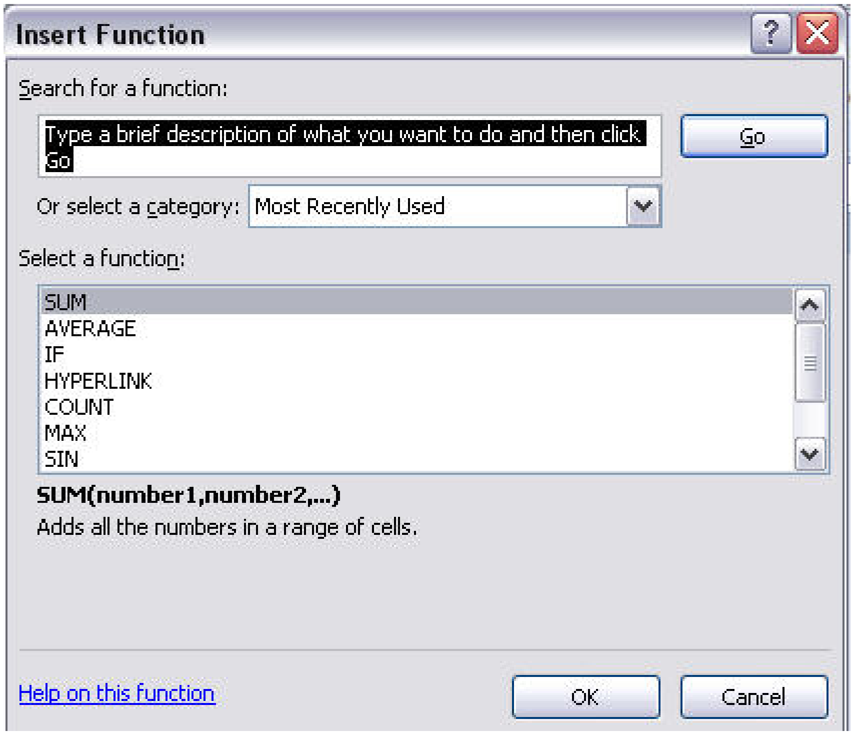


Excel 2007 also gives the user the opportunity to place some of the most commonly used commands in a special Quick Access toolbar that is always available, regardless of which ribbon tab is currently selected. It is located just right of the Office button in the upper-left corner of the application window. Saving a file, undoing or redoing a change, previewing or printing can easily be added to this menu by selecting options in the small drop-down arrow at the right end of the toolbar. Excel also provides the ability to add practically any built-in command to this toolbar, so if a particular command is frequently used, it can be added here, rather than constantly selecting it on the ribbon.

The Formula Bar and Name Box



Although cell contents can be edited directly in the active cell, it is generally easier to edit cell contents by using the formula bar, located below the ribbon and just above the worksheet window. This is particularly useful for long formulas or text. When cell contents are being edited, two buttons appear, which allow the user to cancel the changes being made (the Cancel button, marked with an ✕) or accept the changes as entered (the Enter button, marked with a check mark ✓). Always available is the Insert Function button fx, which easily allows the user to select one of many pre-existing Excel functions for use in the cell being edited (see **Entering Data: Formulas** below). Once a function has been selected using this tool, Excel displays a helpful interface to assist the user in building the formula correctly.



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Most Recently Used

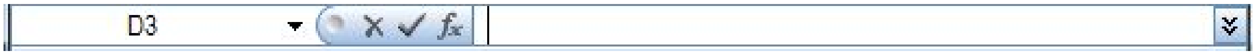
Select a function:

SUM
AVERAGE
IF
HYPERLINK
COUNT
MAX
SIN

SUM(number1,number2,...)

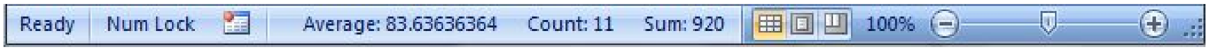
Adds all the numbers in a range of cells.

Sometimes the data displayed in the Formula bar is too long to be displayed in a single line. The height of the formula bar can be adjusted to display multiple lines by dragging the bottom portion of the formula bar downward. Once adjusted, one can toggle between the single line and multiple line displays by clicking the double downward arrow button at the end of the Formula bar.



To the left of the Formula bar is the Name box, which displays the address of the current active cell D3 in the picture above. This can also be used to quickly navigate to a specific cell by typing the cell address into the Name box and hitting Return. The requested cell is selected and the spreadsheet is scrolled to the appropriate location. When editing formulas, the Name box displays the most frequently used Excel functions, allowing the user to easily add them to the current formula by selecting from a drop-down menu.

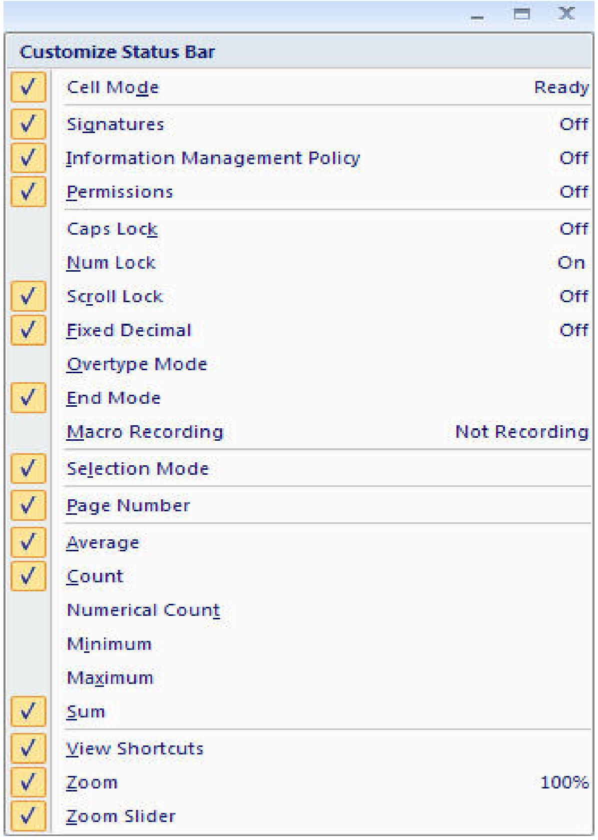
The Status Bar, Zoom Slider, and Window Size Control



At the bottom of the Excel application window are several more useful features. The status bar, located at the far left, displays messages about the current status of the Excel application. Right-clicking the status bar allows the user to select a number of options for what is displayed, such as whether or not Caps Lock is turned on or quick sums, counts, and averages of the currently selected cells.

Just to the right of the status bar are three buttons that allow the user to switch between Normal, Page Layout, and Page Break preview views.

Just to the right is a slider that controls the current Zoom setting for the worksheet. This can be adjusted to make a greater or lesser portion of the spreadsheet be displayed in the current window. In the bottom-right corner of the window is the Window Size Control, which can be used to adjust the size of the application window.

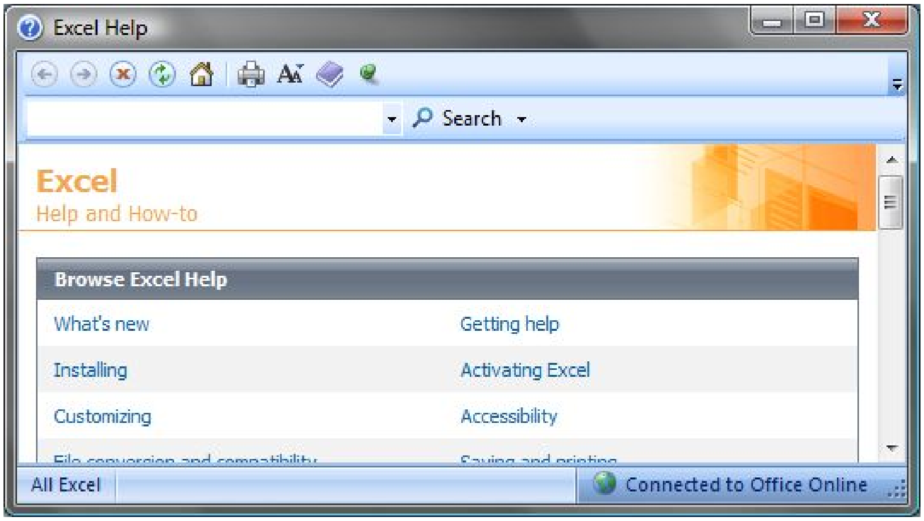


Customize Status Bar

✓	Cell Mode	Ready
✓	Signatures	Off
✓	Information Management Policy	Off
✓	Permissions	Off
	Caps Lock	Off
	Num Lock	On
✓	Scroll Lock	Off
✓	Fixed Decimal	Off
	Overtype Mode	
✓	End Mode	
	Macro Recording	Not Recording
✓	Selection Mode	
✓	Page Number	
✓	Average	
✓	Count	
	Numerical Count	
	Minimum	
	Maximum	
✓	Sum	
✓	View Shortcuts	
✓	Zoom	100%
✓	Zoom Slider	

The Help Button

Near the top-right corner of the application window is a blue circle containing a question mark. This Help button, activates Excel's Help system. An extensive amount of information comes pre-loaded with the Excel application. Excel also automatically searches for the most up-to-date information on Microsoft's Excel website. While this manual provides a quick summary of the most basic Excel functionality, the Help system will provide more detailed information on specific topics as you need them.



Excel
Help and How-to

Browse Excel Help

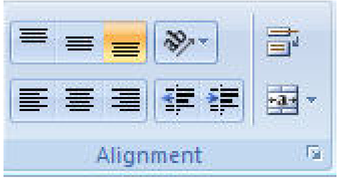
What's new	Getting help
Installing	Activating Excel
Customizing	Accessibility
File conversion and compatibility	Saving and printing

Entering Data

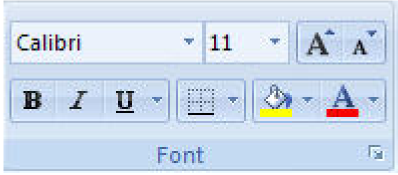
Three types of information can be entered into a cell: text, numeric values, and formulas.

Text

Text can be entered in any combination of letters, numbers, or special characters. By default, text is aligned to the left within the cell. This can be changed via the Alignment group of buttons on the Home tab of the Ribbon **(Ribbon ► Home ► Alignment).**

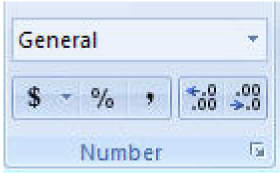


Although an individual cell can contain 32,767 characters, generally large strings of text are broken up into smaller pieces and spread across multiple cells. If the text entered into a cell is longer than the width of the cell allows to be displayed, the display is truncated. The completed text is still housed within the cell, and can be viewed in the formula bar. See the appropriate section above for instructions on how to adjust the amount of lines displayed in the Formula bar. The font type, size, color, and other font formatting features are adjusted using the controls in the **Ribbon ► Home ► Font group**.



Numeric Values

Excel is used primarily to perform calculations, so typically many of the cells in a spreadsheet contain numbers. Excel can be instructed to interpret the number in a specific cell as a date or time, a fraction, an amount of currency, a percentage, a phone number, or just a regular number. This is controlled with the buttons in the Number group on the Home tab of the Ribbon **(Ribbon ► Home ► Number).**



If you enter a number and it appears differently than expected, try changing the cell's number format settings. For example, when entering **1/4** into an unformatted cell, Excel

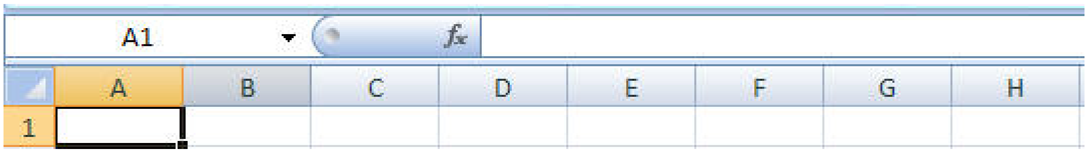
displays this as **4-Jan**. Excel has interpreted the entry as the short format for a date and displayed it in the default date format. Once the number format for a cell has been specified by the user, it retains that format until changed. By default, numbers are right aligned, but this can be changed with the Alignment controls as described above for Text entries.
A few rules to keep in mind when entering numeric values:

- No spaces allowed,
- The first character of a number must be 0 through 9, +, –, or $.
- The number can include commas, decimal points (using the period key) or forward slashes (such as with dates or fractions).
- Negative numbers are designated with a preceding negative sign (-) or by surrounding the number with parentheses.

Numeric values that do not follow the guidelines listed above, or that contain letters or other characters are interpreted as text. To force a number to be interpreted as text, precede the number with an apostrophe (single quote). If a cell is too narrow to display the entire number it contains, Excel instead displays a series of # signs. To display the number correctly, adjust the column width as described in the **Formatting a Worksheet Section** below.

Formulas

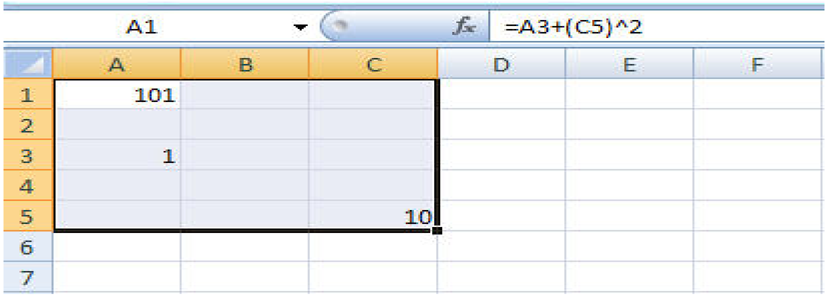
Formulas are mathematical expressions that can use values or formulas in other cells to calculate new values. Formulas can include numbers, cell addresses, multi-cell ranges, functions, and text. Upon entering a formula into a cell, the result of the formula is displayed in the cell itself and the equation is displayed in the Formula bar.



To create a formula, make sure the first character within the cell is an equal sign (=). This alerts Excel that the following data entered in the cell should be interpreted as a formula.
Excel uses the following symbols for these most common mathematical operations:

- the plus sign (+) for addition
- the minus sign (-) for subtraction
- the asterisk (*) for multiplication
- the forward slash (/) for division
- the caret symbol (^) for exponentiation
- the open and close parentheses ( ) for grouping parts of the formula

Example formula =A3+ (C5) ^2



A1 fx =A3+(C5)^2

	A	B	C	D	E	F
1	101					
2						
3	1					
4						
5			10			
6						
7						

If a formula refers to a cell address for a value, and the value in that cell is changed, the formula is automatically updated and the new value displayed. This allows the user to continually update values throughout the spreadsheet and immediately see the resulting changes in the formulas as those value changes are made.

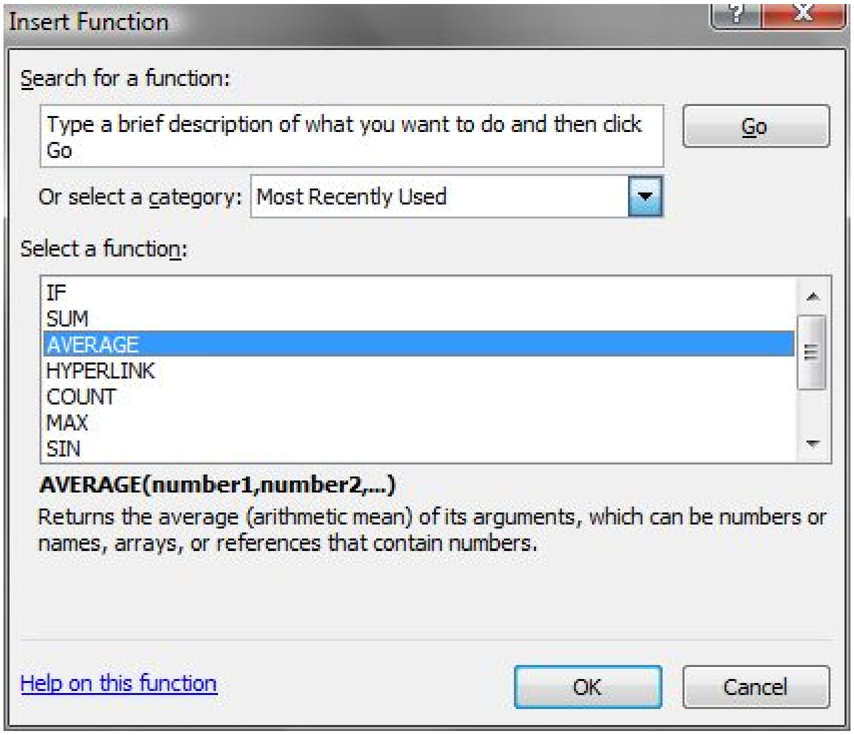
Functions

Functions can be used for arithmetic, statistical, scientific, logical or financial calculations, or even to manipulate text and find values within the spreadsheet.

The most common functions are SUM, COUNT, AVERAGE, MAX, and MIN, but there are hundreds of functions available for your use. The general format for a function is an equal sign (just as with any formula), the capitalized name of the function itself, an open parenthesis, one or more arguments, and a close parenthesis. Arguments are the specific pieces of data required by that function to do the calculation.

For example, a formula using the AVERAGE function would typically be of the form **=AVERAGE (A2:A25)**. We have supplied the cell range A2:A25 for the argument. The cell range can either be typed into the formula, or it can be entered by dragging the mouse across the appropriate cell range when that portion of the formula is reached. The function name itself is not case sensitive and will be capitalized automatically when entry of the formula has been completed.

While a function can be typed directly into a cell, it is much easier to use the built-in Insert function button, located in the Formula bar fx. Clicking this displays the following interface, which guides the user through searching for an appropriate function and entering the data for any required arguments.



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Most Recently Used

Select a function:

IF
SUM
AVERAGE
HYPERLINK
COUNT
MAX
SIN

AVERAGE(number1,number2,...)

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

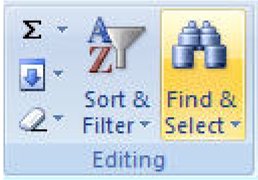
Modifying Data

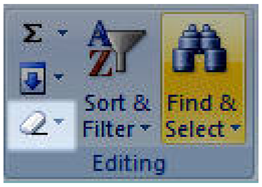
Editing

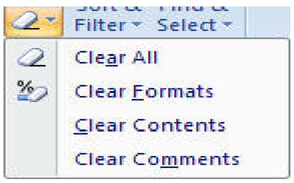
Once a cell's content has begun to be entered, the backspace or delete keys can be used to modify the contents. To discard the changes completely, type the ESC key or click on the Cancel button ✕ in the Formula Bar. If the cell's content has been entered previously, it can be revised by double clicking the cell and moving the cursor to the appropriate location within the contents of the cell. Cell contents can also be edited by clicking the cell and changing the contents displayed in the Formula bar.

Deleting or Clearing Data

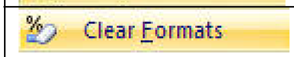
To delete cell contents, select the range of cells to be deleted and type the Delete key. This does not remove the actual cell from the spreadsheet, just its contents. Any cell formatting will remain. To have the option to remove cell contents, formatting, comments, or all three at once, select the range of cells to be cleared and click the Clear button within the Editing group on the Home tab of the Ribbon **(Ribbon ▶ Home ▶ Editing ▶ Clear).**





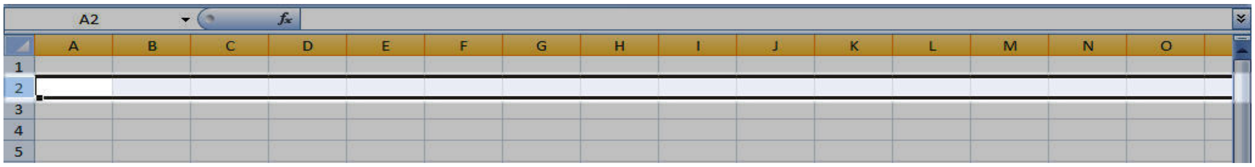


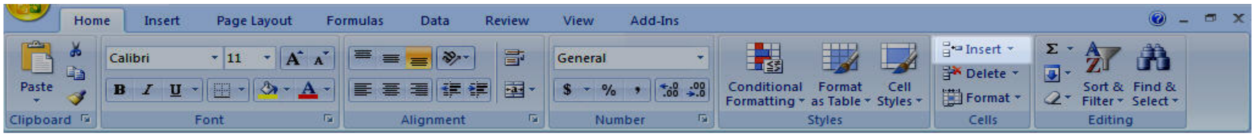
The following options will be displayed:

	Clears formats, contents, and comments as described below
	Clears any background or border coloring, specific font styles or number formats, conditional formatting, cell alignment, etc.
	Clears data entered in the cell, similar to typing the Delete key
	Removes any comments attached to the cell

Inserting and Deleting Rows and Columns

Sometimes after data has been entered into a series of rows, it becomes necessary to insert new data between two of the existing rows. To insert a row, click the numbered row heading of the row beneath where you want to add the row and click the Insert button within the Cells group on the Home tab **(Ribbon ▶ Home ▶ Cells ▶ Insert).**

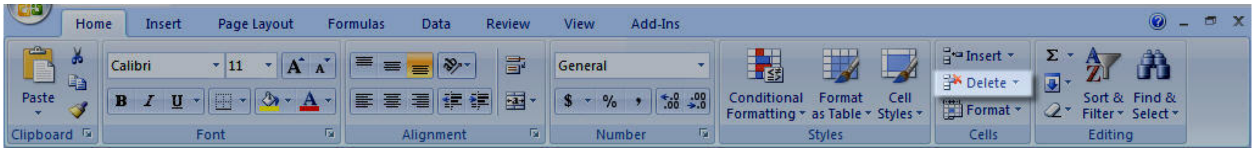




A row will be inserted and any data previously in the selected row or below is shifted down. To insert more than one row, click and drag on more than one row heading and click the Insert button. New rows are added and old rows are shifted as appropriate.

Inserting columns functions in much the same way, except one clicks on the desired number of column lettered headings immediately to the right of where the new columns should be inserted. Clicking the same Insert button executes the action.

To delete rows or columns, select the specific rows or columns to be deleted and click the Delete button **(Ribbon ▶ Home ▶ Cells ▶ Delete),** which is right next to the Insert button. As rows or columns are deleted, all rows beneath or all columns to the right of the deleted section are shifted to fill the gap.

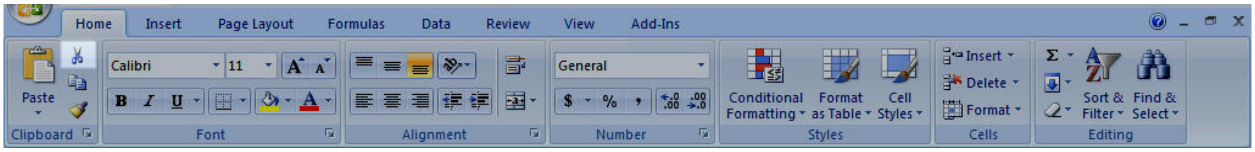


Moving, Copying, and Filling Information

Once cells contain content, that content can easily be moved or copied to another location within the same sheet, to another sheet, to a sheet in a different workbook, or even to another application.

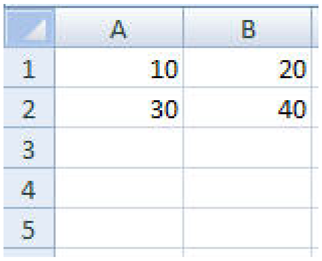
Cut and Paste Cell Content

You may be familiar with the practice of cutting and pasting data in other applications, and Excel provides this functionality as well. Select the range of cells that contain the information to be moved and click the Cut button **(Ribbon ▶ Home ▶ Clipboard ▶ Cut).**

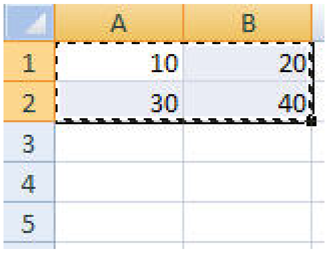


Alternately, after selecting the target range of cells, right-click and select Cut from the pop-up menu or type Ctrl+X. All three methods of "cutting" place the entire contents of the selected cell range in Excel's memory (referred to in all Microsoft Office products as the Clipboard). At this point, the data has not yet been moved from the cells, but the selected cut range is indicated with a flashing dotted line surrounding it.

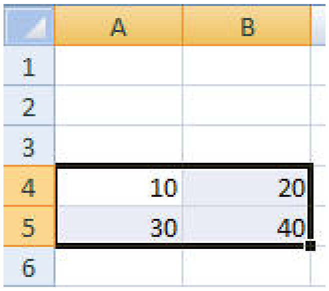
Next, select the upper-left cell of the new area where you want the data you have just cut to be "pasted." Click the Paste button **(Ribbon ▶ Home ▶ Clipboard ▶ Paste)** and the data from the old cells is placed within the new ones.



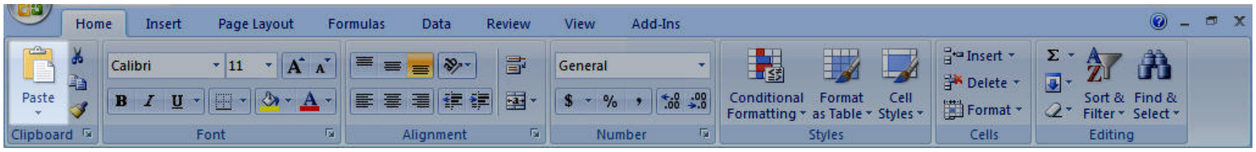
	A	B
1	10	20
2	30	40
3		
4		
5		



	A	B
1	10	20
2	30	40
3		
4		
5		

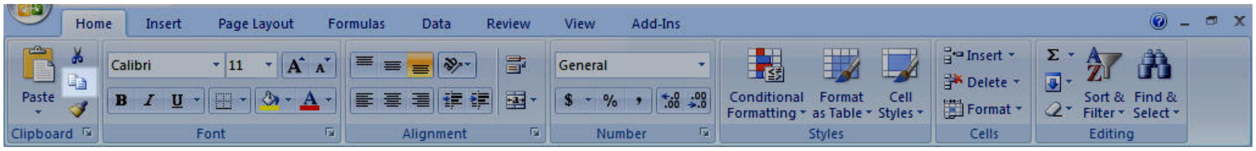


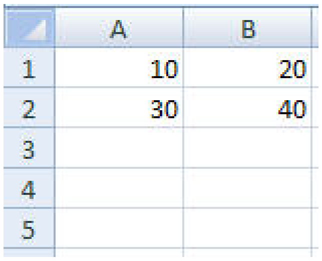
	A	B
1		
2		
3		
4	10	20
5	30	40
6		



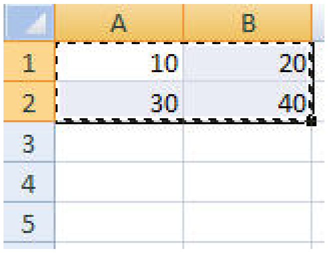
Copy and Paste Cell Content

To copy a target range to another location, with the old data remaining where it was, use the same basic method as described above, but select the Copy button or menu option instead of Cut.

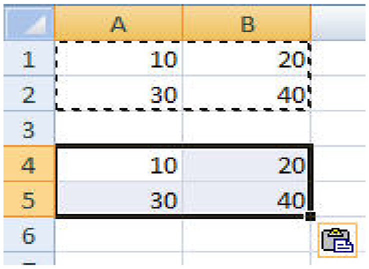




	A	B
1	10	20
2	30	40
3		
4		
5		



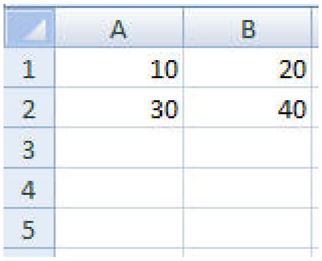
	A	B
1	10	20
2	30	40
3		
4		
5		



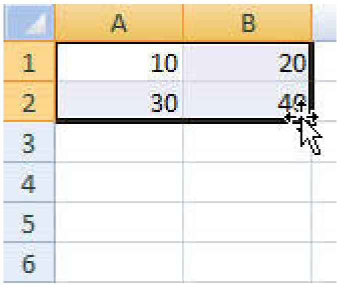
	A	B
1	10	20
2	30	40
3		
4	10	20
5	30	40
6		

Drag and Drop Cell Content

A target range of cells can also be dragged and dropped to another location on the same sheet. To do so, select the cell range to be moved. Notice that when the mouse pointer is placed directly over the heavy black line surrounding the selected range that the pointer changes to a small cross with four arrows.

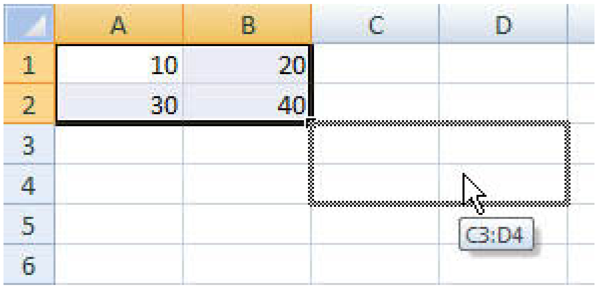


	A	B
1	10	20
2	30	40
3		
4		
5		

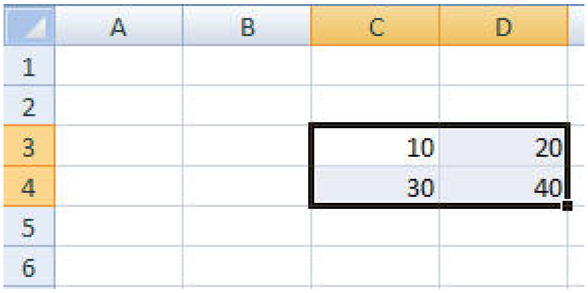


	A	B
1	10	20
2	30	40
3		
4		
5		
6		

When the four arrow cross is displayed, click and hold the mouse, dragging the mouse to another location on the spreadsheet. The entire selected range of cells moves along with it, including all content and formatting.

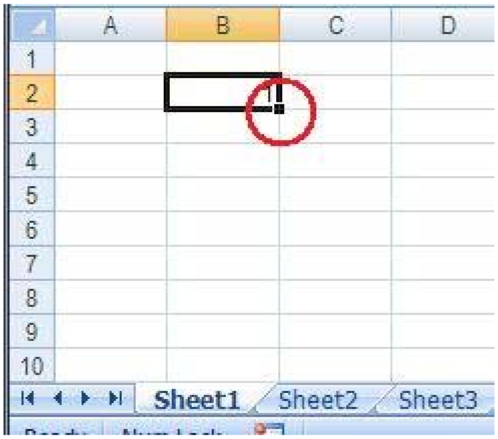


	A	B	C	D
1	10	20		
2	30	40		
3				
4				
5				
6				



	A	B	C	D
1				
2				
3			10	20
4			30	40
5				
6				

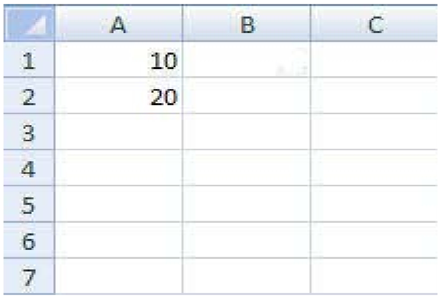
AutoFill Cell Content



Excel also provides a simple way to populate data or formulas across a range of cells, or to create an incremental data series. To simply copy data or a formula across a range, select the cell to be copied. Notice that there is a small square (called a "fill handle") in the bottom-right corner of the heavy line surrounding the selected cell (circled in the image to the left).

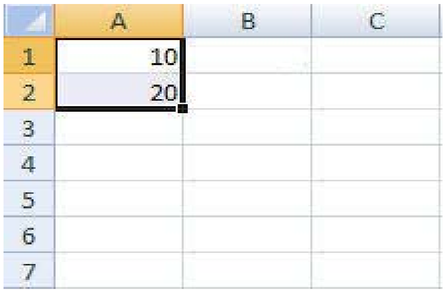
When the mouse pointer is placed over this handle, the arrow pointer becomes a crosshair. As that crosshair is displayed, click on the fill handle and drag the mouse down or to the right across the cells to be filled. Upon releasing the mouse, the data or formula in the original cell is copied across the range. Any formatting in the original cell is copied as well.

To create an incremental series, type the first two numbers in the series in adjacent cells.

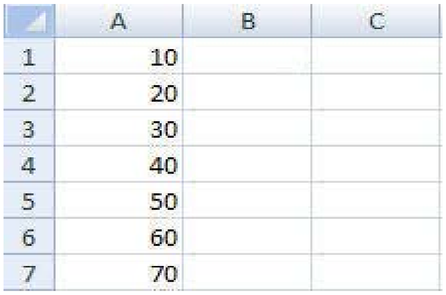


	A	B	C
1	10		
2	20		
3			
4			
5			
6			
7			

Following the same procedure as described for copying above, select the two cells containing the first two data points in the series and drag the fill handle across the appropriate number of cells for the whole series. Based upon the first two numbers entered, Excel AutoFills the remainder of the series.



	A	B	C
1	10		
2	20		
3			
4			
5			
6			
7			

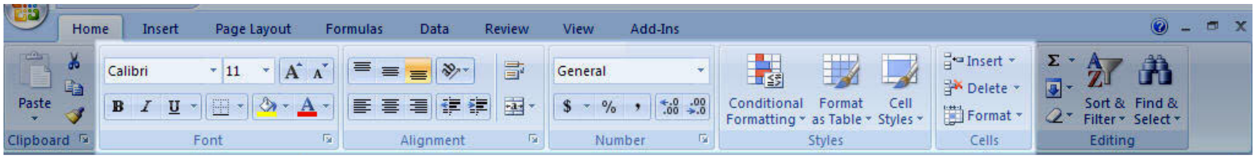


	A	B	C
1	10		
2	20		
3	30		
4	40		
5	50		
6	60		
7	70		

Excel can also AutoFill the names of months. Simply enter January or Jan, and using the AutoFill method described above, Excel fills in the remainder of the months in the format entered. After December, Excel continues on with January again, filling in each successive month over the entire dragged range.

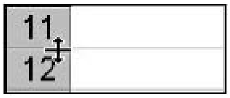
Formatting a Worksheet

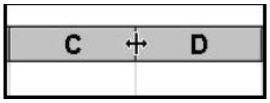
Excel provides a wealth of tools to customize the look and feel of spreadsheets. First, select a cell or a range of cells to be formatted. Using the buttons on the Home tab in the Font, Alignment, Number, Styles, and Cells groups, the background color, border colors, row height, column width, fonts styles, and size and number formats can all be changed. Individual cells can be merged using the Merge and Center options in the Alignment group. Preprogrammed formats can be applied using the Cell Styles options in the Styles group.



Adjusting Column Width and Row Height

The height and width of an individual row or column can be changed by clicking and dragging the line between the rows or column in their respective headers. When pointing the mouse directly at the line between row headings or column headings, the pointer arrow changes to a line with arrows pointing in two directions (see images below).

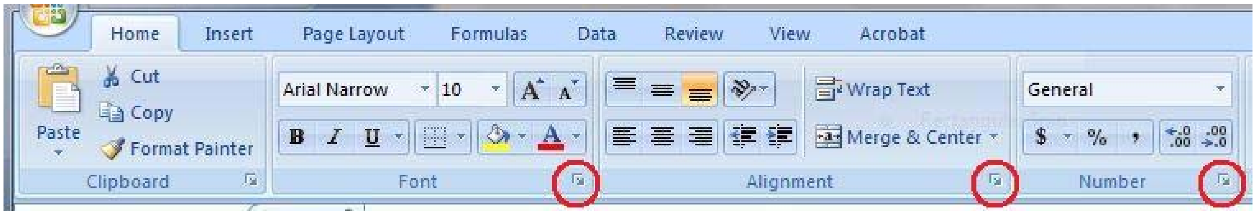




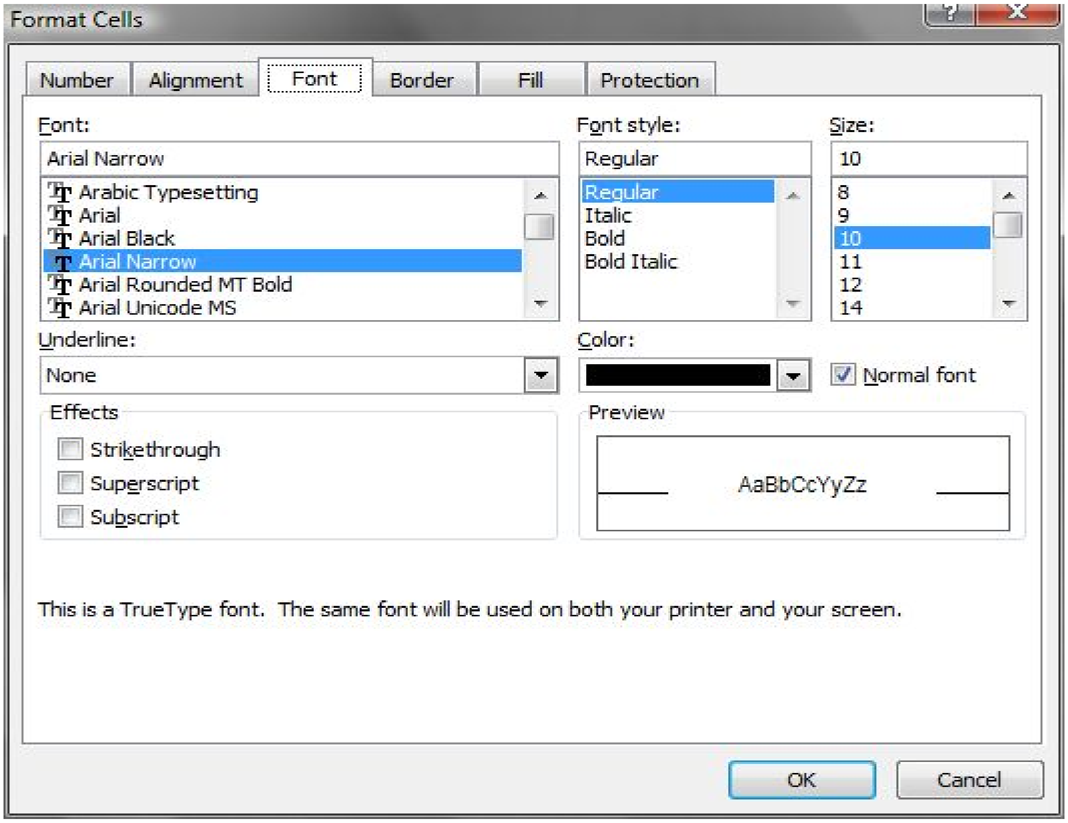
With this double-arrowed pointer displaying, click and hold the mouse. Light gray lines show the current boundaries of the row or column being adjusted. Dragging the mouse widens or narrows these boundaries to display the proposed width or height. If multiple rows or columns are selected at one time, the height or width is adjusted for the entire selection. Alternately, you can double click on the line between row or column headings for a “best fit” option for the selection. Row or column headings can also be right-clicked to display a menu that includes a Row Height or Column Width option.

Formatting Cells

Many cell, font, and number formatting options are available directly from the Home tab of the Ribbon, the Format Cells feature provides additional formatting options. It can be accessed by clicking the arrow-within-square button located in the bottom right of the Font, Alignment, and Number groups on the Home tab (see the circles in the image below).



Adjusting the settings in the format cells will change the format in the selected cells.



Font: Arial Narrow

Font style: Regular

Size: 10

Underline: None

Color:

☑ Normal font

□ Strikethrough

□ Superscript

□ Subscript

This is a TrueType font. The same font will be used on both your printer and your screen.

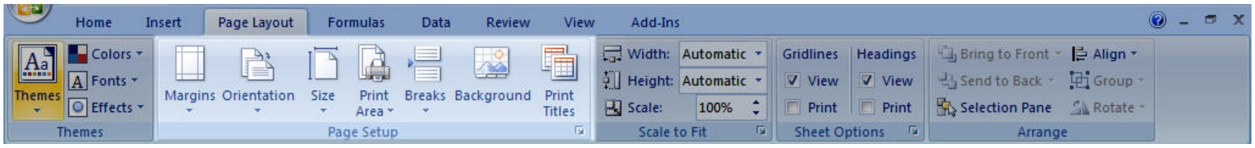
Tabs at the top provide the following formatting controls:

Number	Select number format styles of general, currency, date, time, percentage, etc. Also controls the number of decimal places displayed, whether or not commas or currency symbols are displayed, and how negative values are differentiated.
Alignment	Horizontal and vertical text alignment, direction of text, such as at a 45^{o} angle, whether to wrap text, indent, etc.
Font	Select font family and size, options for bold, italic, underline, strikethrough, superscript, subscript, and font color.
Border	Turn on or off borders at each of the four sides of a cell or the two diagonals. Weight, style, and color of each border segment can be adjusted independently.
Fill	Control the color and pattern of cell interior backgrounds.
Protection	Control whether users can edit the contents of cells or view cells' formulas in the Formula bar. As with all formatting options, this can be controlled on a cell-by-cell basis.

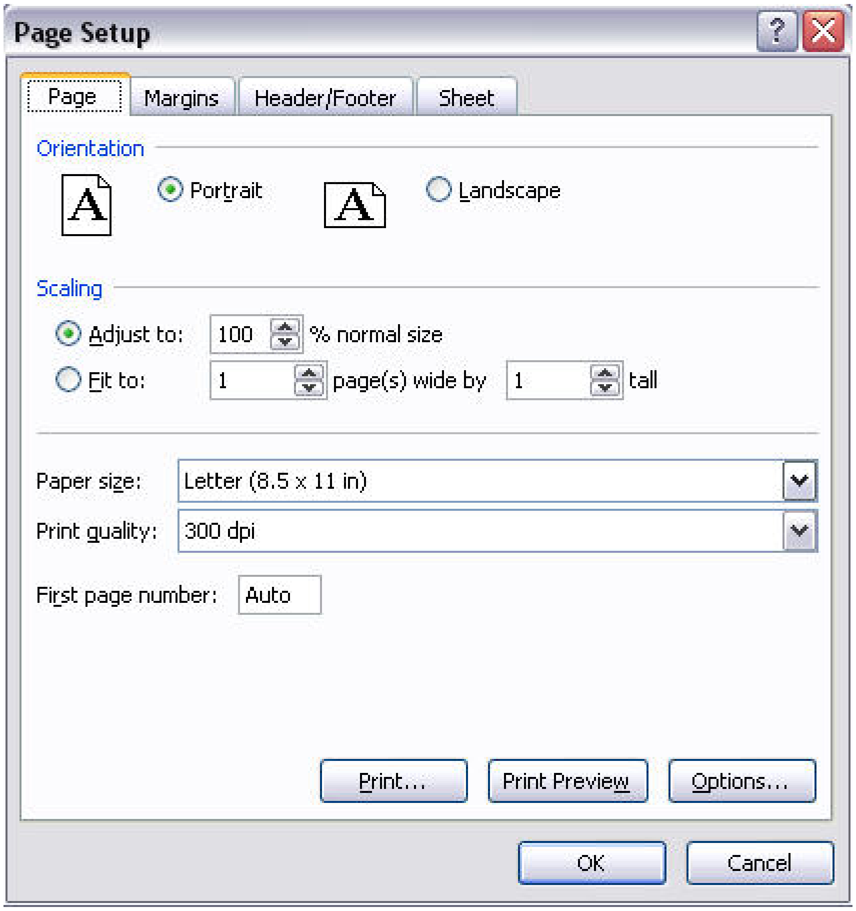
Printing

Page Setup Options

Excel provides a number of tools to configure how the spreadsheet will look when it is printed. The primary printing options are located on the Page Layout tab of the Ribbon in the Page Setup group **(Ribbon ▶ Page Layout ▶ Page Setup).** Options include controls to adjust the page margins, page orientation, the expected size of the paper being used for printing, background images, where page breaks occur, which cells in the spreadsheet will be printed, and whether or not to repeat certain rows at the top or certain columns at the left of each page.



Clicking the arrow-within-square button in the bottom right of the Page Setup group opens up a more detailed Page Setup interface with a greater level of control for these options as well as the ability to specify page headers and footers.



Orientation: ● Portrait ○ Landscape

● Adjust to: 100 % normal size

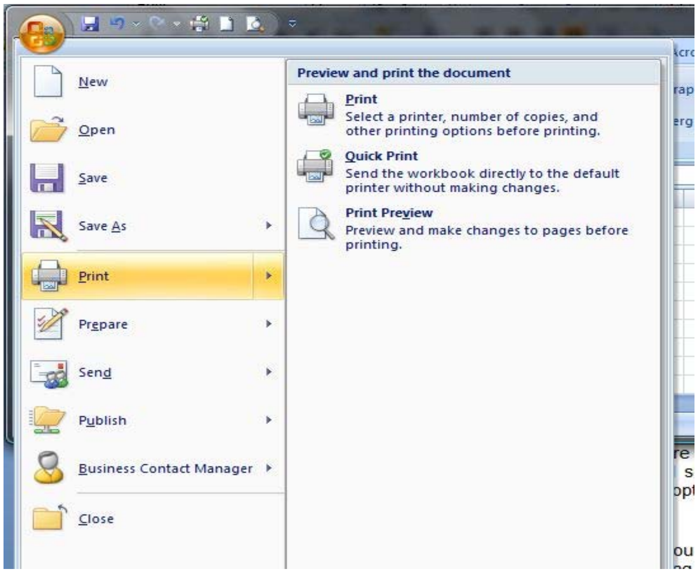
○ Fit to: 1 page(s) wide by 1 tall

Paper size: Letter (8.5 x 11 in)

Print quality: 300 dpi

First page number: Auto

Print, Quick Print, and Print Preview



Preview and print the document

Print
Select a printer, number of copies, and other printing options before printing.

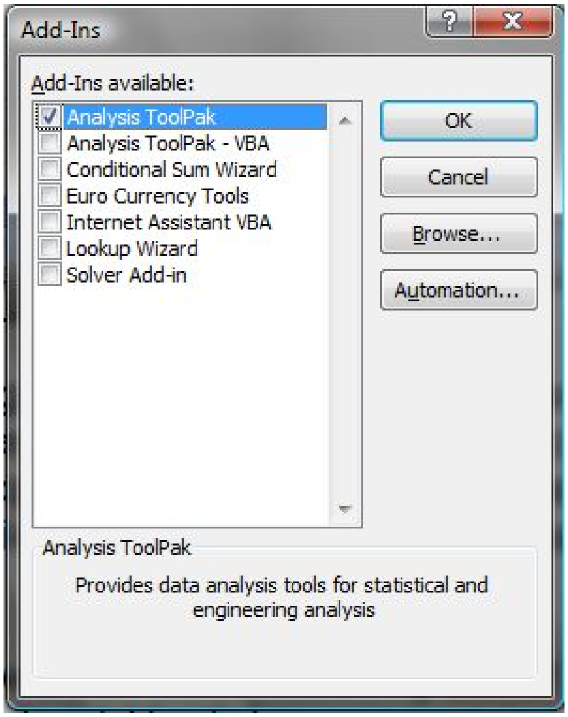
Quick Print
Send the workbook directly to the default printer without making changes.

Print Preview
Preview and make changes to pages before printing.

To access printing options, click the circular Office button in the top left of the Excel application window. From the Office menu, select the Print sub-menu, as displayed on the left.

Selecting the Print command will display an interface that allows the user to select a printer and printing options. The Quick Print command will print the current spreadsheet using the default printer and default print options. The Print Preview command allows the user to see how the spreadsheet will appear before printing it.

Using Excel's Statistical Tools



Add-Ins available:

☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Conditional Sum Wizard
□ Euro Currency Tools
□ Internet Assistant VBA
□ Lookup Wizard
□ Solver Add-in

Analysis ToolPak

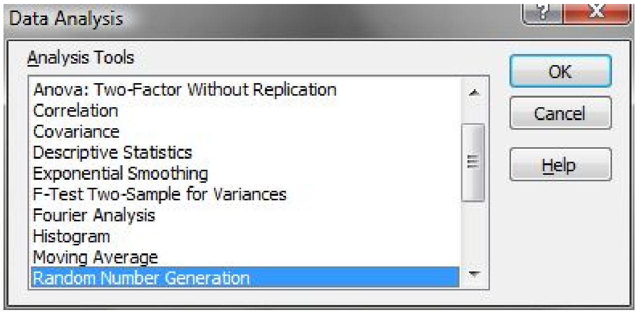
Provides data analysis tools for statistical and engineering analysis

Excel contains a set of pre-built statistical analysis tools as part of the Analysis Toolpak add-in included with the Excel software. For some Excel installations, it will need to be "turned on." Doing so requires the following steps:

1. Click the circular Office button in the upper-left corner of the Excel application window.
2. In the light-blue border at the bottom of the Office menu, click the Excel Options button.
3. In the pop-up interface that displays, select Add-ins from the navigation bar on the left.
4. A list of available add-ins will be displayed. It may take a few moments for Excel to collect this information. At the bottom of the list, there should be a drop-down menu with add-ins selected. Click the Go button next to this.
5. Make sure that Analysis ToolPak is checked.

Note: If the Analysis ToolPak is not listed, it will need to be added from the Excel installation software.

Once the Analysis ToolPak add-in is installed, there should be a new analysis group available on the Data tab of the Ribbon **(Ribbon ► Data ► Analysis).** Click the Data Analysis button, and the following list of available analysis tools will be displayed.



Many of the included Excel statistical analysis tools are detailed where appropriate in the exercises in the following chapters.

Where appropriate, exercises taken from the textbook are solved using both the Excel analysis tools and the WHFStat add-in module packaged with this manual.

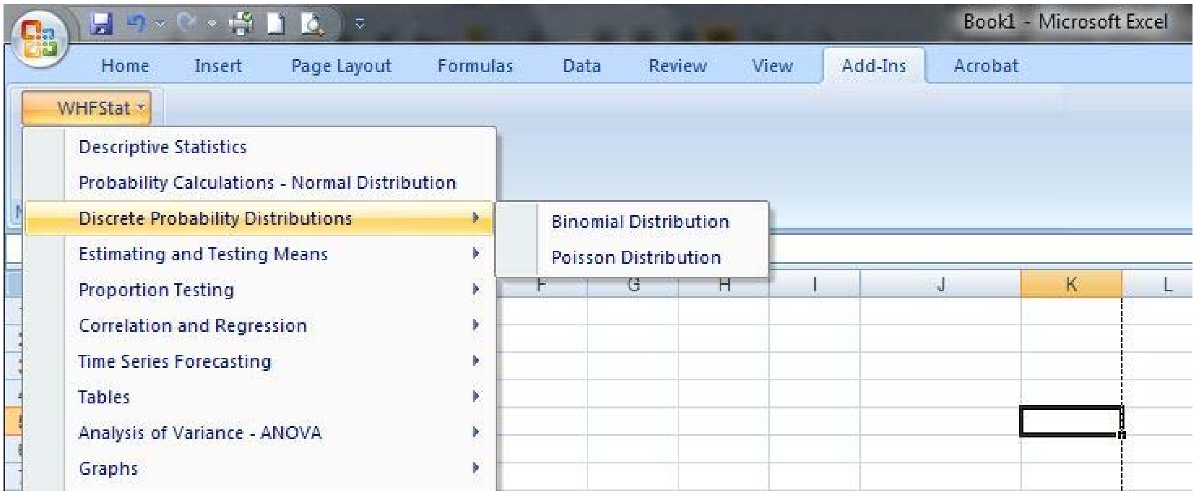
The Excel solutions are identified by this icon

The WHFStat solutions are identified by this icon

Using the WHFStat Add-In Module

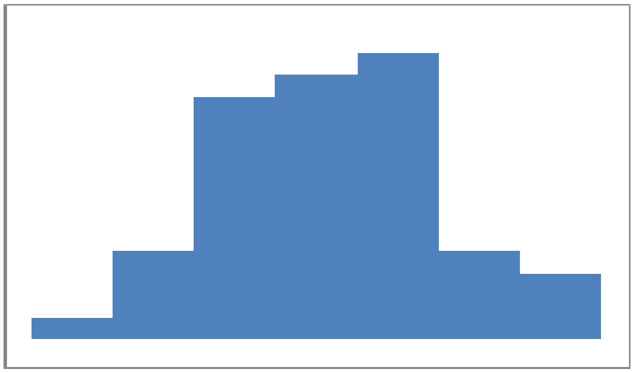
WHFStat is an Excel Add-in, software that makes it easier to use Excel to do most statistical operations. The software is available on StatsPortal, your Online Study Center or packaged with this manual.

Once installed, the WHFStat Add-In module will be integrated into your Excel application and will automatically load every time you open Excel. You will notice a new Add-Ins tab on the Ribbon, upon which the Menu Commands group will have a button labeled WHFStat. Clicking this will display the various menu options available for the add-in.



CHAPTER

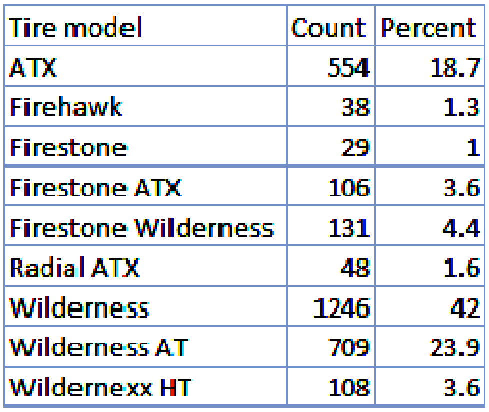
1



Looking at Data: Exploring Distributions

Bar Charts

Excel allows us to examine the distribution of variables with graphs. Bar charts are useful for categorical data. The following data provides the tire model reported for 2969 accidents that involved Firestone tires.

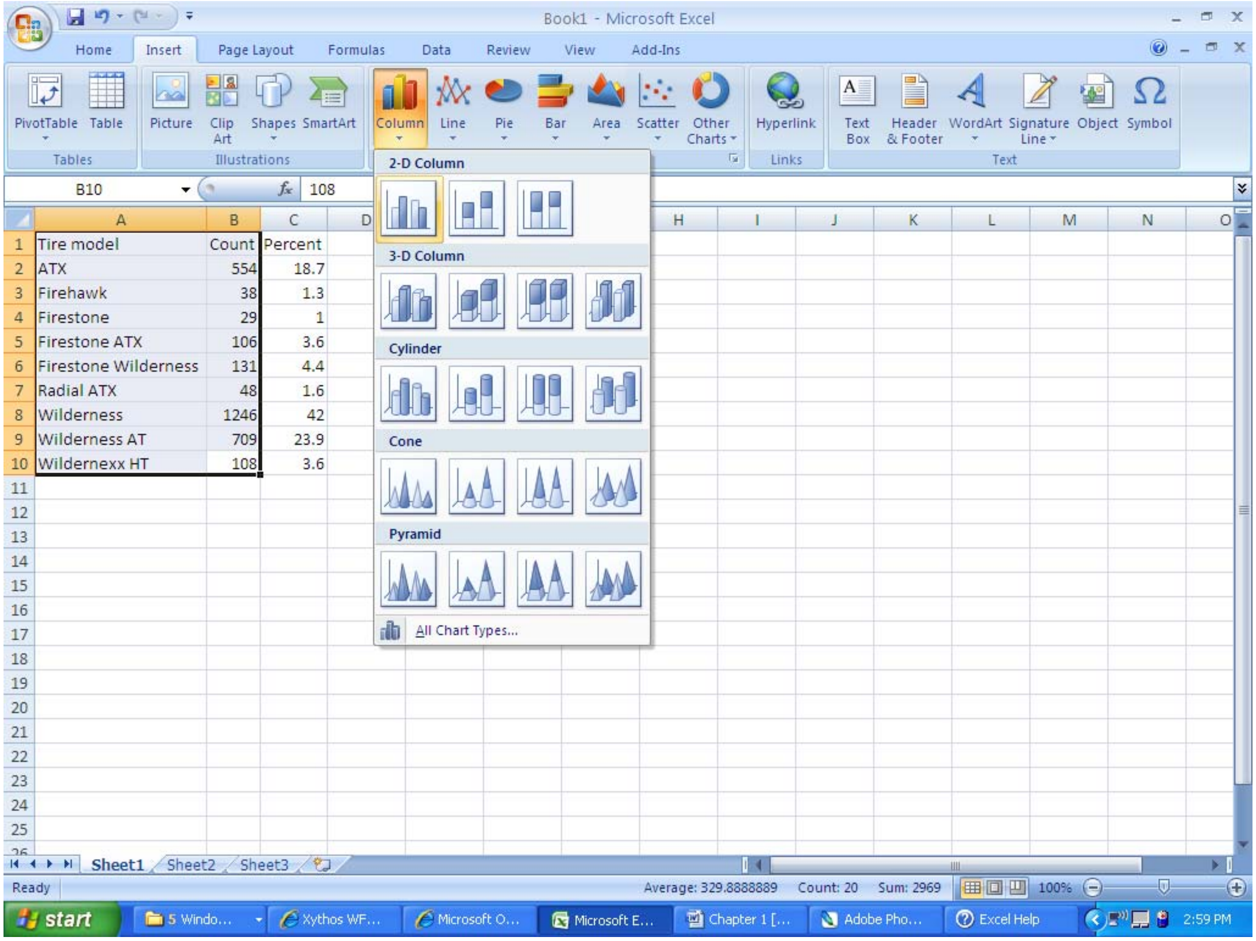


Tire model	Count	Percent
ATX	554	18.7
Firehawk	38	1.3
Firestone	29	1
Firestone ATX	106	3.6
Firestone Wilderness	131	4.4
Radial ATX	48	1.6
Wilderness	1246	42
Wilderness AT	709	23.9
Wildernexx HT	108	3.6

We will use this data to make a Bar Chart with Excel. Highlight the data and select

Insert ➤ Column ➤ 2-D Column

as shown below.

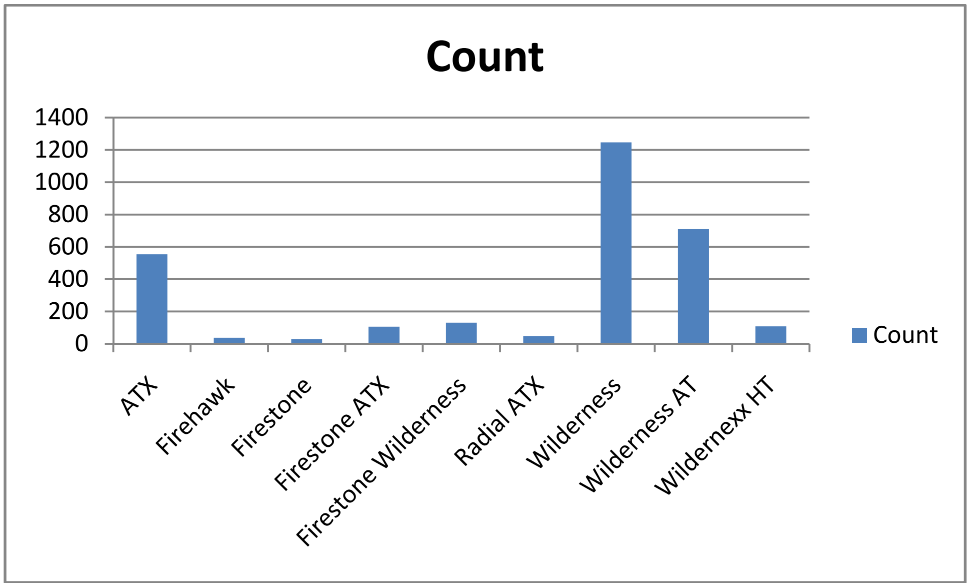


B10 | 108

	A	B	C
1	Tire model	Count	Percent
2	ATX	554	18.7
3	Firehawk	38	1.3
4	Firestone	29	1
5	Firestone ATX	106	3.6
6	Firestone Wilderness	131	4.4
7	Radial ATX	48	1.6
8	Wilderness	1246	42
9	Wilderness AT	709	23.9
10	Wildernexx HT	108	3.6

Average: 329.8888889 Count: 20 Sum: 2969

Clicking on the 2-D Column will produce the following bar chart.

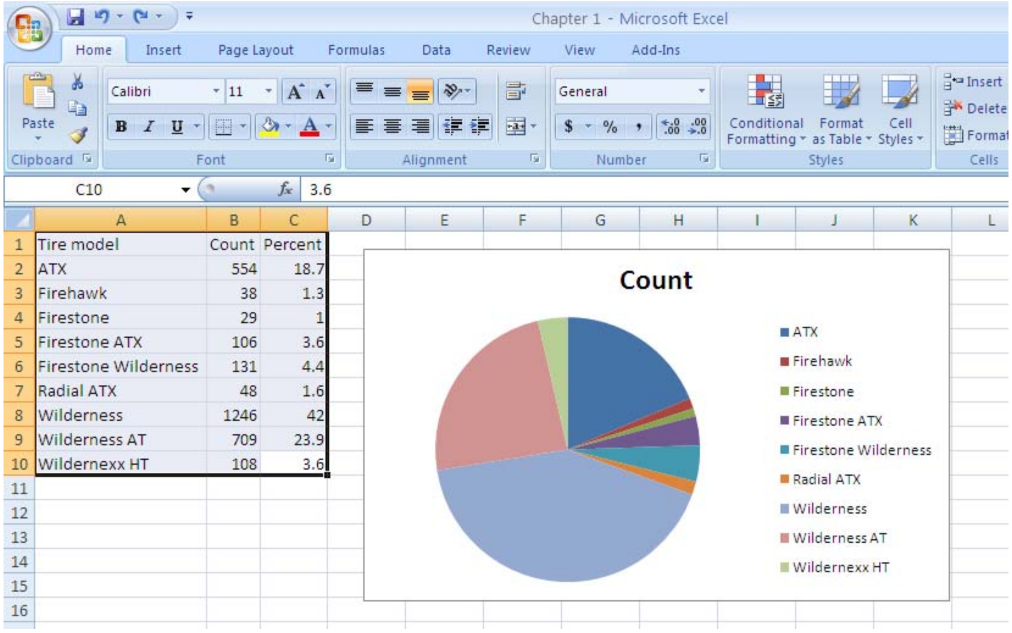


Pie Charts

Another way to examine distributions of categorical variables is with a pie chart. We will continue to use the data from the previous example to show how to make a pie chart with Excel. To make a pie chart of the waste data, select

Insert ➤ Pie ➤ 2-D Pie

from the menu.



C10 *fx* 3.6

	A	B	C
1	Tire model	Count	Percent
2	ATX	554	18.7
3	Firehawk	38	1.3
4	Firestone	29	1
5	Firestone ATX	106	3.6
6	Firestone Wilderness	131	4.4
7	Radial ATX	48	1.6
8	Wilderness	1246	42
9	Wilderness AT	709	23.9
10	Wildernexx HT	108	3.6

Although we highlighted three columns, Excel used only the first two when constructing the pie chart. The values in the column labeled percent can be used by changing the data. If you right click on the pie chart and choose Select data, the Select Data Source dialog box pops up. In that dialog box, you can delete the Count series so that the Percent series will be used instead.

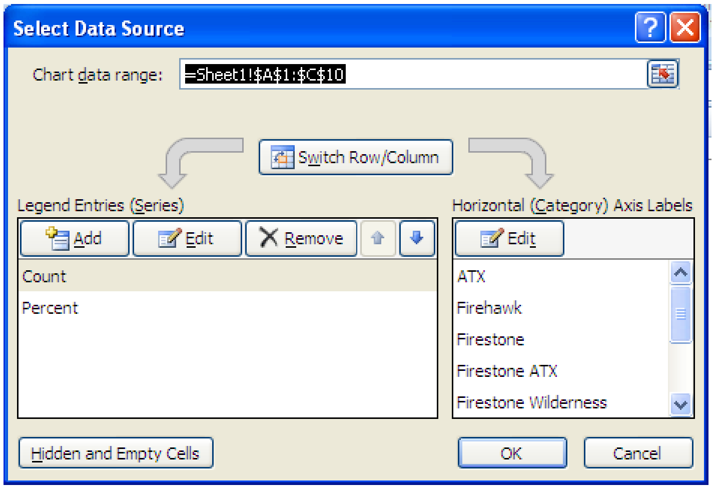


Chart data range: =Sheet1!$A$1:$C$10

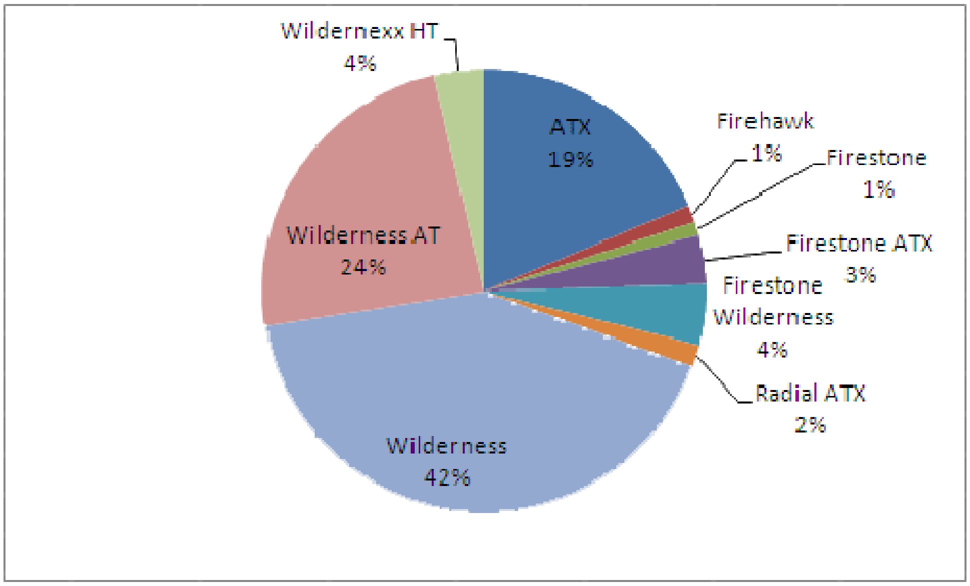
Legend Entries (Series)

Count
Percent

Horizontal (Category) Axis Labels

ATX
Firehawk
Firestone
Firestone ATX
Firestone Wilderness

Alternatively, you can click on the graph and select the Design tab from the menu to select an alternative presentation such as the one shown below.

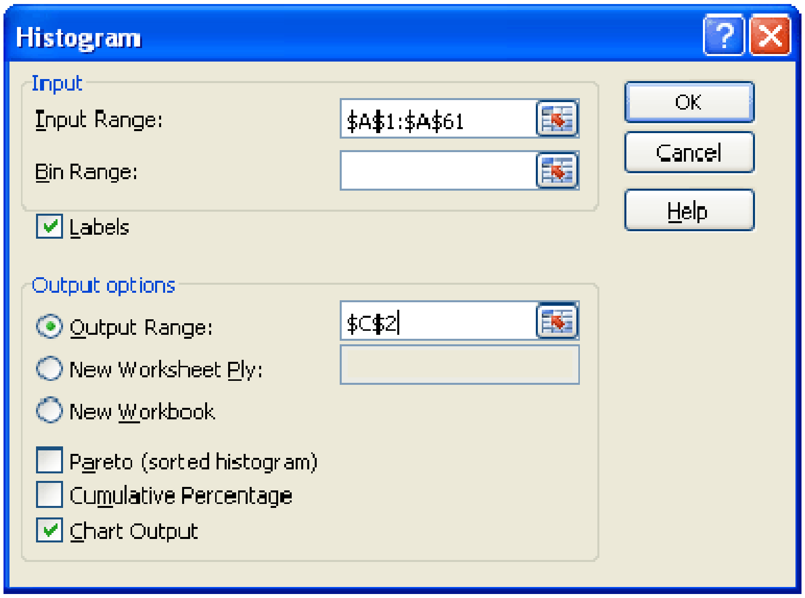


Histograms

The most common graph for the distribution of a quantitative variable is a histogram. We will illustrate this with IQ test scores for 60 fifth-grade students. To create a histogram, select

Data ➤ Data Analysis ➤ Histogram

from the menu. In the dialog box, specify the input range to be where the data is located as shown below. If the first cell is a label, check the Labels box. The Output Range specifies where the output will be placed. Finally, check the Chart Output box and then click OK.



Histogram

Input

Input Range: $A$1:$A$61

Bin Range:

☑ Labels

Output options

● Output Range: $C$2

○ New Worksheet Ply:

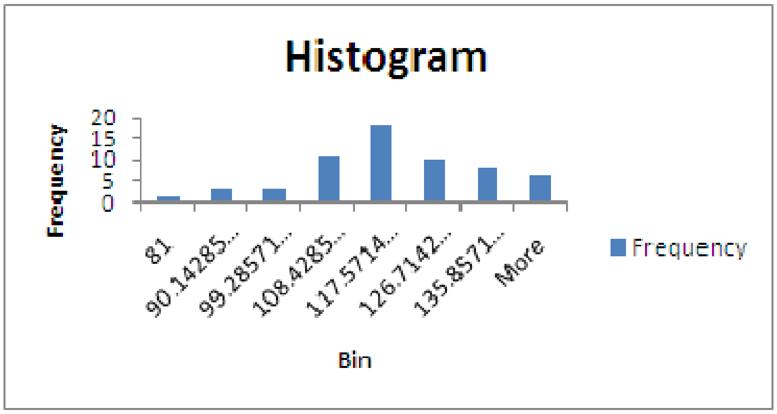
○ New Workbook

☐ Pareto (sorted histogram)

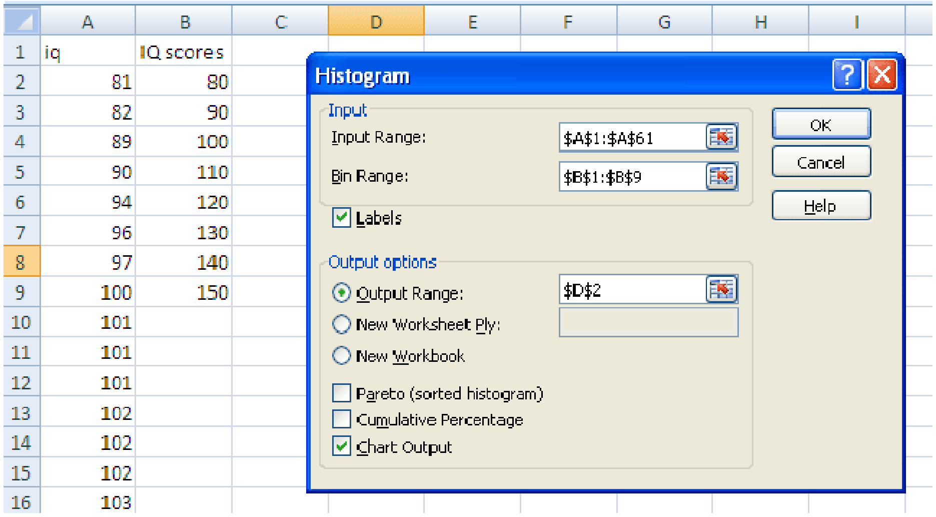
☐ Cumulative Percentage

☑ Chart Output

The default histogram appears as follows.



You can also specify alternative bin ranges to avoid the default values selected by Excel. It is helpful to first select **Data ➤ Sort** from the menu to sort the data before deciding on the Bin Ranges. The new values are then typed into a column on the Excel worksheet. If the data has a label and the Label box will be checked, then this new column should also have a label. The new column is then entered into the Histogram dialog box next to Bin Range.

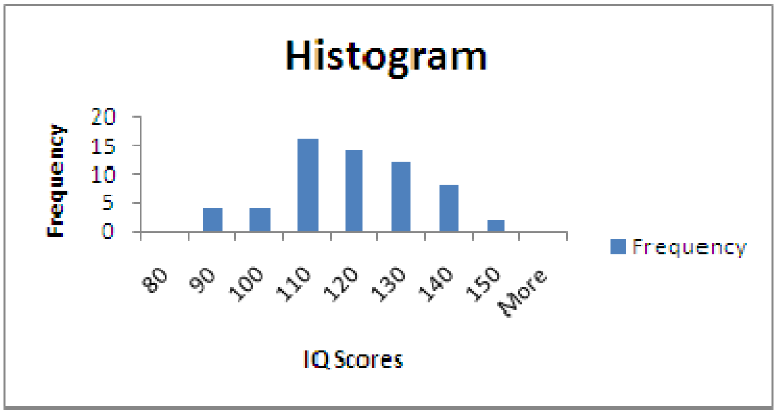


	A	B	C	D	E	F	G	H	I
1	iq	IQ scores							
2	81	80							
3	82	90							
4	89	100							
5	90	110							
6	94	120							
7	96	130							
8	97	140							
9	100	150							
10	101								
11	101								
12	101								
13	102								
14	102								
15	102								
16	103								

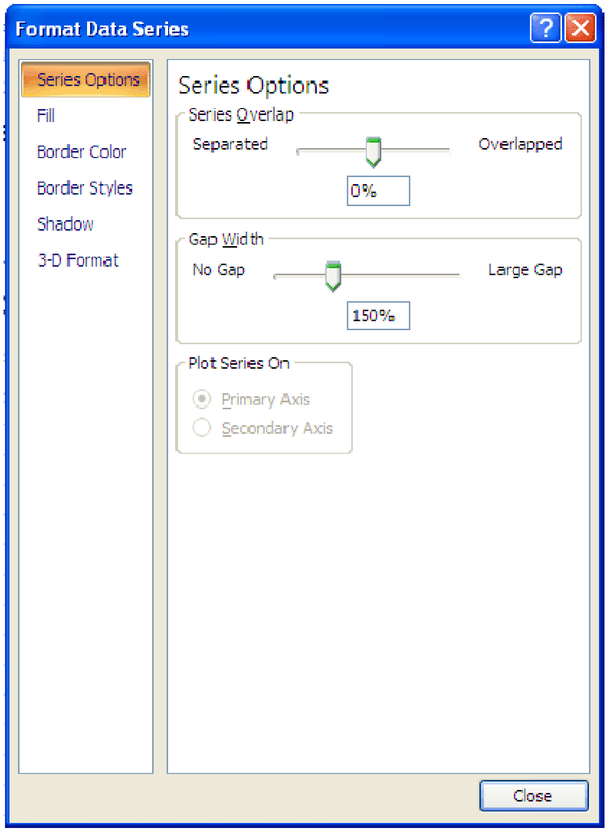
Input
Input Range: $A$1:$A$61
Bin Range: $B$1:$B$9
☑ Labels

Output options
● Output Range: $D$2
○ New Worksheet Ply:
○ New Workbook
□ Pareto (sorted histogram)
□ Cumulative Percentage
☑ Chart Output

The new histogram will use the selected bin ranges, but not be entirely satisfactory. For example, the bin ranges in the histogram below appear to be interval midpoints instead of cutpoints. It appears from this histogram as though there are no observations below 85, when in fact there are.



The gap width between the bars can be changed or eliminated by right clicking on a bar and selecting Format Data Series and changing the option in the following dialog box.



Series Overlap: 0%

Gap Width: 150%

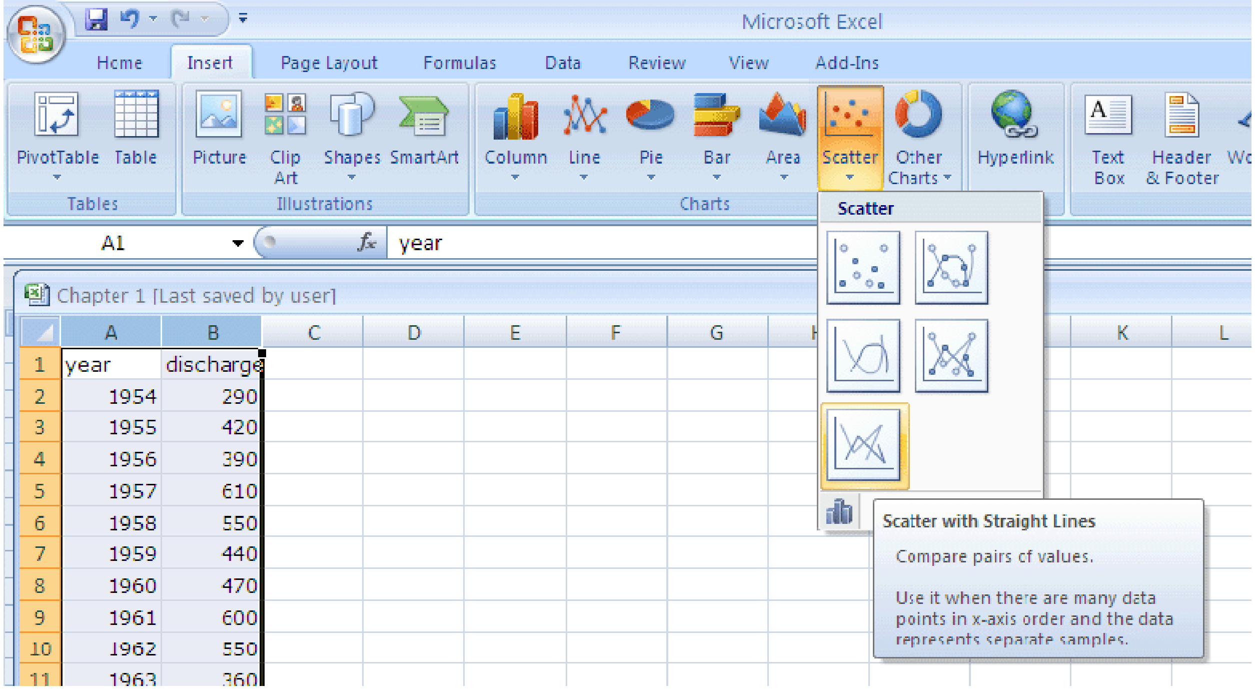
Alternatively, you can construct histograms by selecting **Add-Ins ➤ WHFStat ➤ Graphs ➤ Histogram** and filling out the dialog box.

Time Series Plots

When quantitative data are collected over time, it is a good idea to plot the observations in the order they were collected. For example, the following data lists the volume of water discharged by the Mississippi River in the Gulf of Mexico for each year from 1954 to 2001.

Year	Discharger	Year	Discharge	Year	Discharger	Year	Discharge
1954	290	1966	410	1978	560	1990	680
1955	420	1967	460	1979	800	1991	700
1956	390	1968	510	1980	500	1992	510
1957	610	1969	560	1981	420	1993	900
1958	550	1970	540	1982	640	1994	640
1959	440	1971	480	1983	770	1995	590
1960	470	1972	600	1984	710	1996	670
1961	600	1973	880	1985	680	1997	680
1962	550	1974	710	1986	600	1998	690
1963	360	1975	670	1987	450	1999	580
1964	390	1976	420	1988	420	2000	390
1965	500	1977	430	1989	630	2001	580

To make a time series plot of this data, highlight the data and select **Insert ➤ Scatter and select a graph design** from the menu.



A1: year

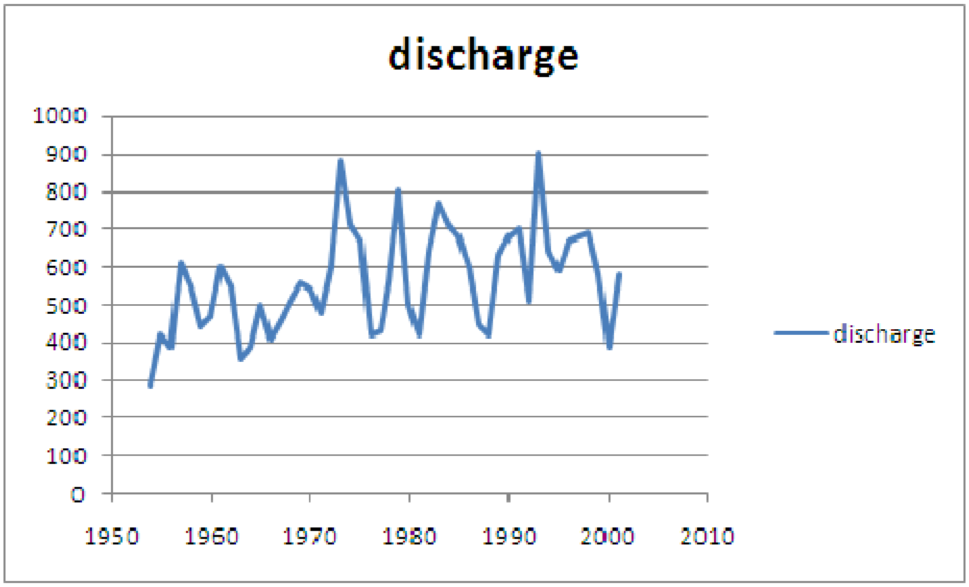
	A	B
1	year	discharge
2	1954	290
3	1955	420
4	1956	390
5	1957	610
6	1958	550
7	1959	440
8	1960	470
9	1961	600
10	1962	550
11	1963	360

Scatter with Straight Lines

Compare pairs of values.

Use it when there are many data points in x-axis order and the data represents separate samples.

The time series plot will appear as soon as you click on the plot design of your choice. If Scatter with Straight Lines is selected, the plot will appear as follows. As usual, the plot can be altered by clicking on the graph and selecting the Design tab.



Since the dates appeared in one column with the data in the next column to the right, the time plot has the dates on the *x*-axis and the data on the *y*-axis. If the data appears in a different order, you can right click on a data point and choose Select Data. The dialog box that is shown below allows you to switch columns, and add or delete a series so that you can use the appropriate data.

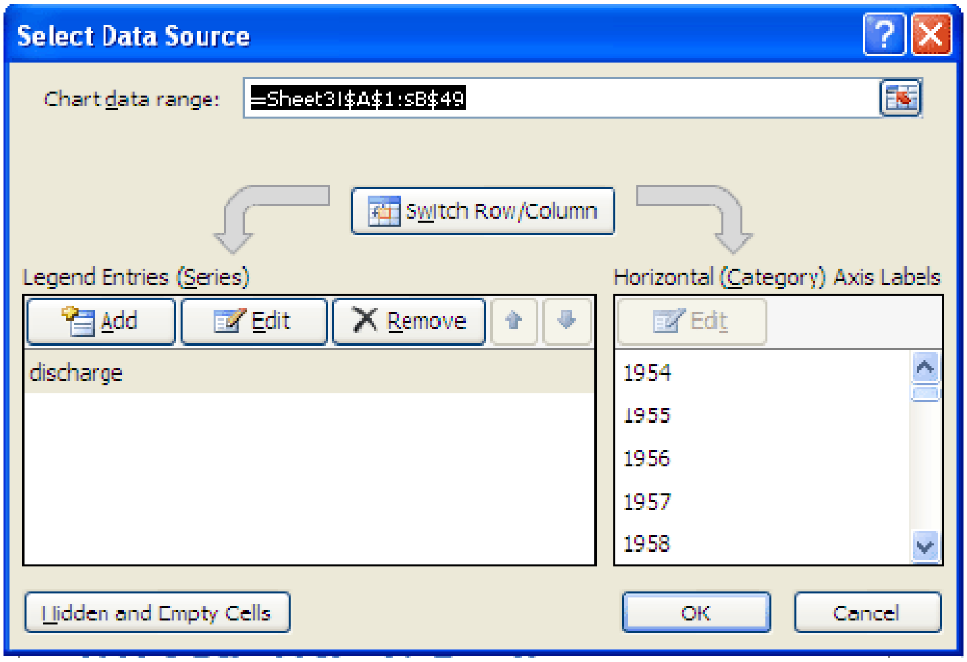


Chart data range: =Sheet3!$A$1:$B$49

Legend Entries (Series): discharge

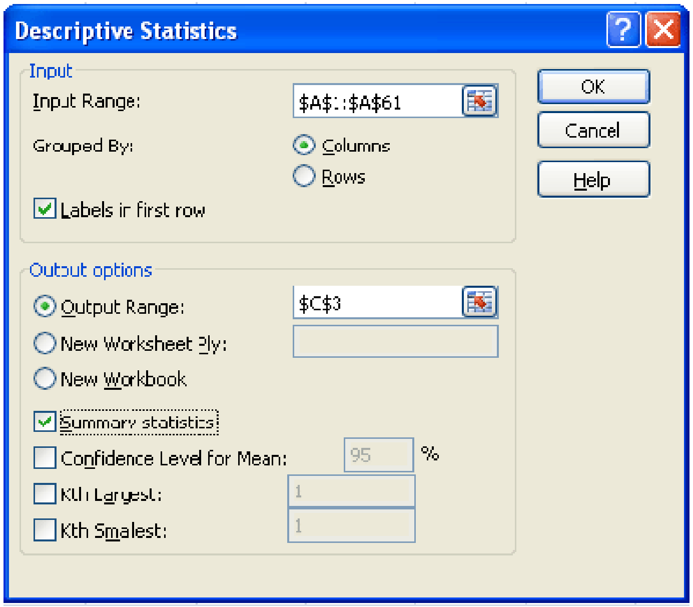
Horizontal (Category) Axis Labels: 1954, 1955, 1956, 1957, 1958

If your data is not accompanied by a column of dates, highlight only that data and select **Insert ➤ Line** from the menu. Select the design that you prefer to obtain a time plot. In this case, the *x*-axis will be labeled with consecutive numbers instead of dates.

Numerical measures are often used to describe distributions. Select

Data ➤ Data Analysis ➤ Descriptive Statistics

from the menu to obtain descriptive statistics. Enter the input range for the data. If the data includes a label in the first row, check the appropriate box. Specify where the output will appear, check the box next to Summary Statistics, and click OK in the following dialog box.


Descriptive Statistics

Input

Input Range: A::$A$61

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $C$3

○ New Worksheet Ply:

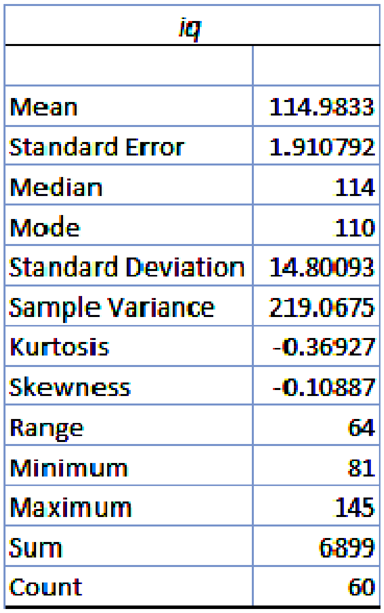
○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1


iq	
Mean	114.9833
Standard Error	1.910792
Median	114
Mode	110
Standard Deviation	14.80093
Sample Variance	219.0675
Kurtosis	-0.36927
Skewness	-0.10887
Range	64
Minimum	81
Maximum	145
Sum	6899
Count	60

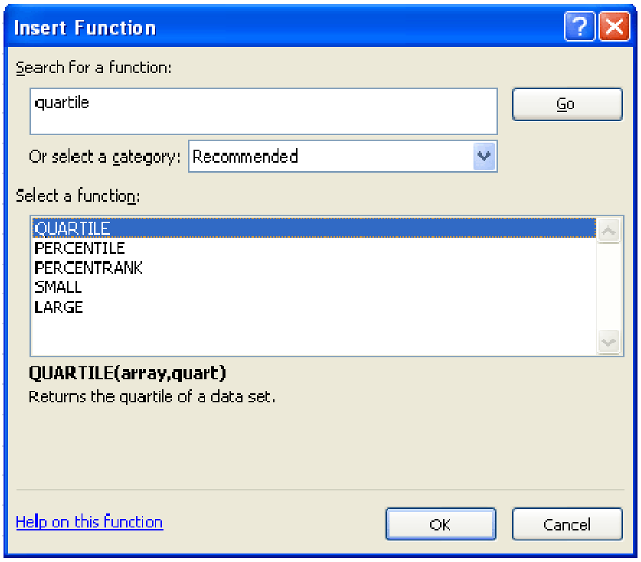
The command summarizes several different measures of both the center and spread of a distribution. The command prints the statistics Mean, Standard Error, Median, Mode, Standard Deviation, Sample Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count, for each column specified.

Count is the number of actual values in the column (missing values are not counted). Mean is the average of the values. To find the median, the data first must be ordered. If N is odd, the median is the value in the middle. If N is even, the median is the average of the two middle values. StDev is the standard deviation computed as

$$\text{StDev} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{\text{N} - 1}}$$

Standard Error is the standard error of the mean. It is calculated as $\text{StDev}/\sqrt{\text{N}}$.

The same results can be obtained using functions in Excel. For example, typing =AVERAGE(A2:A61) into a cell gives the mean, =STDEV(A2:A61) gives the standard deviation, and =COUNT(A2:A61) gives the sample size. In addition, we can obtain the quartiles needed for the five-number summary using the QUARTILE function. If you click on the Insert Function, *fx*, you can search for the appropriate function as shown below.



Search for a function: quartile

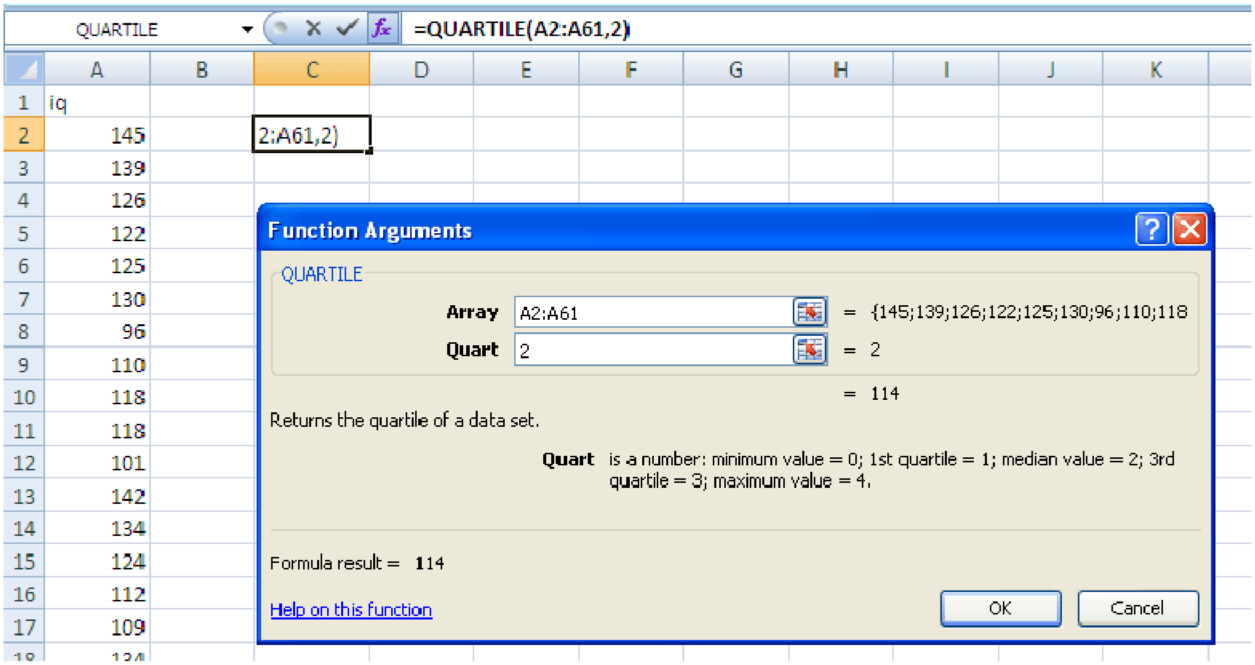
Or select a category: Recommended

Select a function:

- QUARTILE
- PERCENTILE
- PERCENTRANK
- SMALL
- LARGE

QUARTILE(array,quart)
Returns the quartile of a data set.

As shown below, you can obtain the function either by typing the formula int0 an empty cell, clicking on an empty cell and then typing the formula into the formula bar to the right of the Insert Function button, or by filling in a dialog box. Either way, you must specify the Array that holds the data and then the Quart, where Quart = 0 is the minimum value, 1 is the first quartile, 2 is the median, 3 is the third quartile, and 4 is the maximum. Excel doesn't use exactly the same algorithm to calculate quartiles as your textbook, so minor differences in results will sometimes occur.



=QUARTILE(A2:A61,2)

	A	B	C
1	iq		
2	145		2:A61,2)
3	139		
4	126		
5	122		
6	125		
7	130		
8	96		
9	110		
10	118		
11	118		
12	101		
13	142		
14	134		
15	124		
16	112		
17	109		

QUARTILE

Array A2:A61 = {145;139;126;122;125;130;96;110;118

Quart 2 = 2

= 114

Returns the quartile of a data set.

Quart is a number: minimum value = 0; 1st quartile = 1; median value = 2; 3rd quartile = 3; maximum value = 4.

Formula result = 114

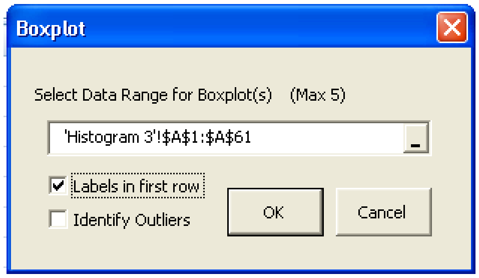
The five-number summary consisting of the median, quartiles, and minimum and maximum values provides a quick overall description of a distribution. If you select **Add-Ins ➤**

WHFStat ➤ Descriptive Statistics from the Excel menu, the descriptive statistics includes all of the values needed for the five-number summary.

Boxplots based on the five-number summary display the main features of a column of data. Boxplots can be obtained by selecting

Add-Ins ➤ WHFStat ➤ Graph ➤ Boxplot

from the menu and then filling in the dialog box as shown below.



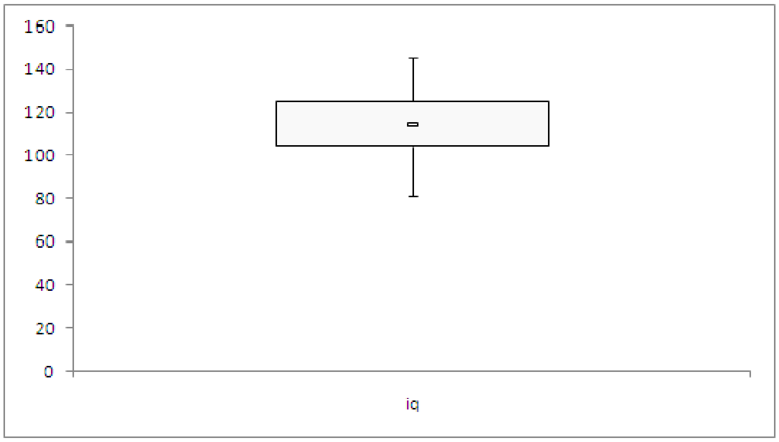
Select Data Range for Boxplot(s) (Max 5)

'Histogram 3'!$A$1:$A$61

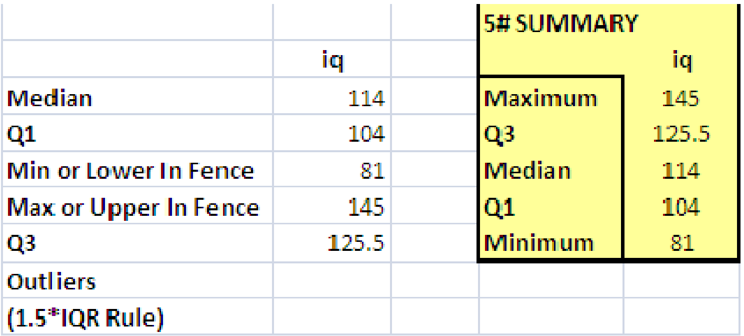
☑ Labels in first row

☐ Identify Outliers

A boxplot graphically displays the main features of data from a single variable. A box-plot illustrated for the IQ data.



The boxplot consists of a box, whiskers, and outliers. Excel draws a line across the box at the median. The bottom of the box is at the first quartile (Q1) and the top is at the third quartile (Q3). The whiskers are the lines that extend from the top and bottom of the box to the adjacent values. If the Identify Outliers on Graph box is not checked, the whiskers extend to the lowest and highest observations. If the box is checked, the whiskers extend only to the lowest and highest observations inside the region defined by the lower limit $Q1-1.5(Q3-Q1)$ and the upper limit $Q3+1.5(Q3-Q1)$. Points outside the lower and upper limits are identified as outliers and listed on the worksheet. As shown below, the IQ data did not have outliers identified by this criteria.



			5# SUMMARY	
	iq			iq
Median	114		Maximum	145
Q1	104		Q3	125.5
Min or Lower In Fence	81		Median	114
Max or Upper In Fence	145		Q1	104
Q3	125.5		Minimum	81
Outliers				
(1.5*IQR Rule)				

To construct side-by-side boxplots comparing different distributions, simply enter up to five adjacent columns of data into the Input Range.

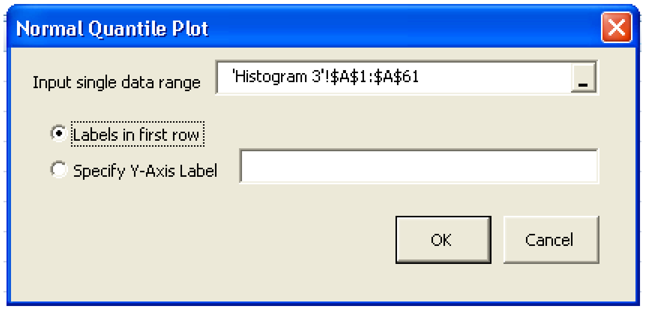
Normal Calculations

Sometimes the Normal density can describe the overall pattern of a distribution. A histogram may be helpful in deciding when this is appropriate. Normal quantile plots are also useful in determining whether a distribution is approximately Normal. If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data are Normal.

Boxplots can be obtained by selecting

Add-Ins ➤ WHFStat ➤ Graphs ➤ Normal Quantile Plot

from the menu and then filling in the dialog box as shown below.



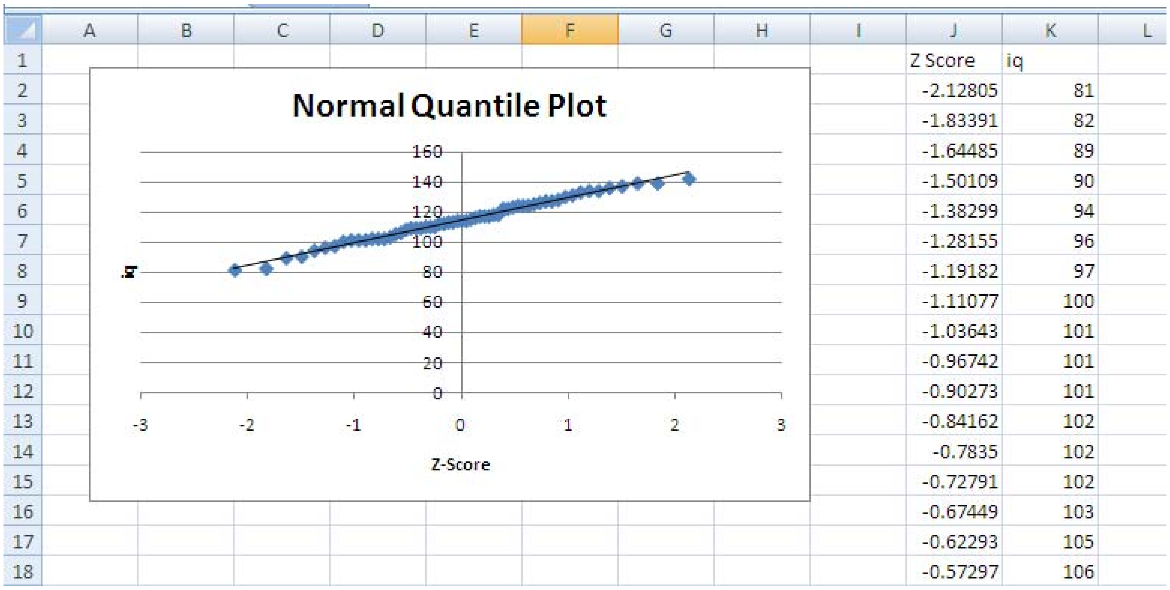
Normal Quantile Plot

Input single data range: 'Histogram 3'!$A$1:$A$61

● Labels in first row

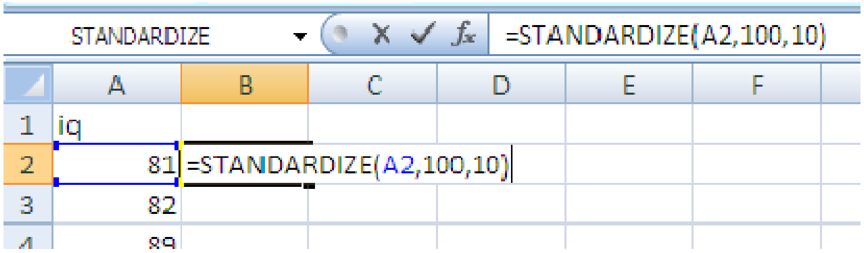
○ Specify Y-Axis Label

The normal quantile plot below is very close to a straight line, indicating that the IQ scores are normally distributed.



	J	K
1	Z Score	iq
2	-2.12805	81
3	-1.83391	82
4	-1.64485	89
5	-1.50109	90
6	-1.38299	94
7	-1.28155	96
8	-1.19182	97
9	-1.11077	100
10	-1.03643	101
11	-0.96742	101
12	-0.90273	101
13	-0.84162	102
14	-0.7835	102
15	-0.72791	102
16	-0.67449	103
17	-0.62293	105
18	-0.57297	106

The Normal distribution is a good description of the overall pattern of the data. Excel can be used to perform Normal distribution calculations. If data in a column are Normally distributed, then the data can be standardized to obtain data with a standard Normal distribution, that is, those with mean equal to zero and standard deviation equal to one. The STANDARDIZE(x,mean,standard_dev) function can be used to do this. The function requires that you specify the *x* value that you want to standardize, the mean of the distribution and the standard deviation of the distribution. The function returns the standardized value, $z = (x - \overline{x})/s$. For example, if the IQ scores are normally distributed with a mean equal to 100 and a standard deviation equal to 10, then we can standardize as shown below.



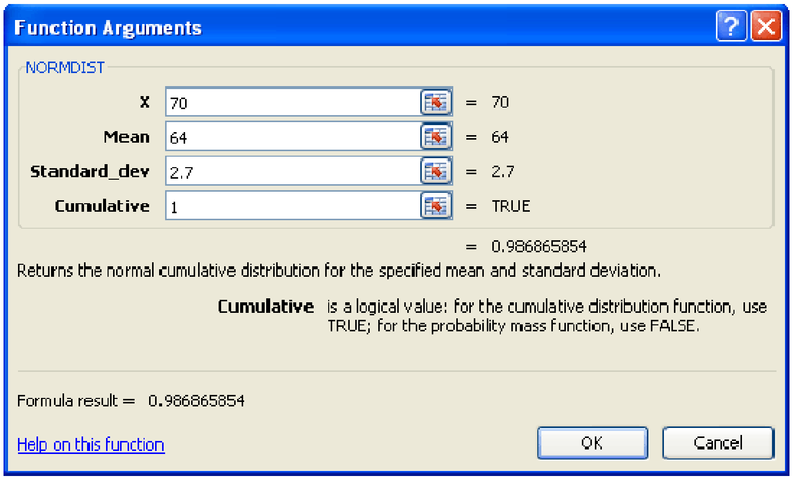
=STANDARDIZE(A2,100,10)

	A	B	C	D	E	F
1	iq					
2	81	=STANDARDIZE(A2,100,10)				
3	82					
4	89					

If we copy the formula down the column, we can obtain the standardized values for all of the vocabulary scores. The standardized IQ scores (or *z*-score) will tell how far above or below the mean a particular score falls. The measure is in units of standard deviations. The first student has a score of 81, a value that is below the mean (z = –1.9). Another score, 117 is above the mean (z = 1.7).

We could examine the standardized values to see how well they obey the 68-95-99.7 rule. Approximately 68% of the standardized values should have values between –1 and +1, 95% should have values between –2 and +2, and 99.7% should have values between –3 and +3.

You can use Excel to do probability calculations for the Normal distribution using the **NORMDIST**(**x**,**mean**,**standard_dev**,**cumulative**) function. When cumulative=1, this function returns the normal distribution for the specified mean and standard deviation. In addition to the 1 for cumulative, you must specify the *x* value for the distribution along with the mean and standard deviation. For example, the heights are approximately Normal with a mean of about 64 inches and a standard deviation of 2.7 inches. To find the proportion of women who are less than 70 inches tall, we select type =NORMDIST(70,64,2.7,1) into a cell or the formula bar *fx*. Alternatively, click **Function Wizard** *fx* on the formula bar to select the NORMDIST function and fill in the dialog box.



Function Arguments

NORMDIST

X 70 = 70

Mean 64 = 64

Standard_dev 2.7 = 2.7

Cumulative 1 = TRUE

= 0.986865854

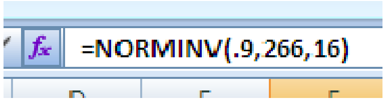
Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.986865854

The result says that the proportion of women who are less than 70 inches tall is .986866, or nearly 99%. This is slightly different from the result that would be obtained using Table A since it is not required to round the standardized value.

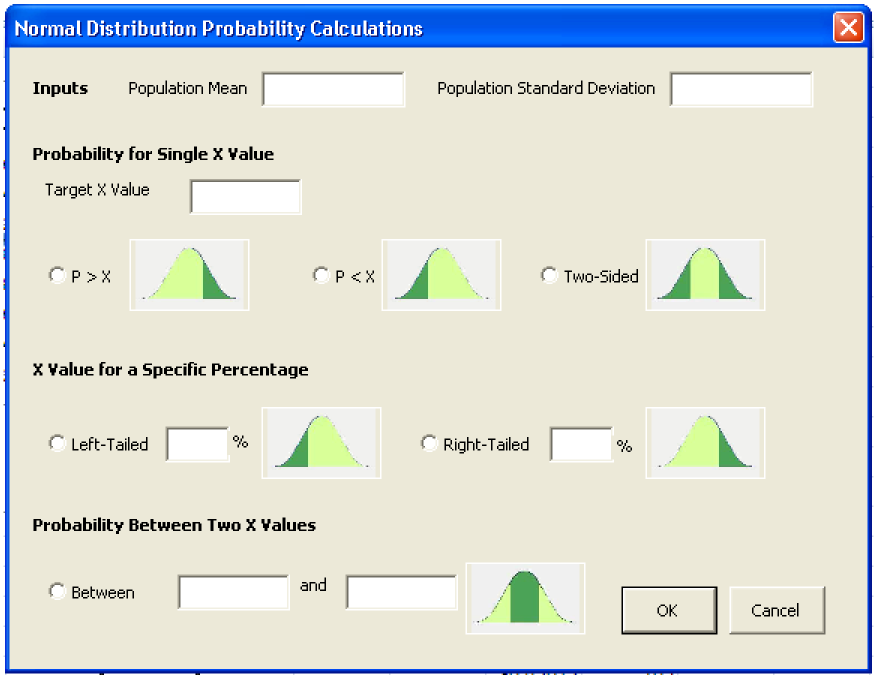
We can also use Excel to do backward calculations. The length of human pregnancies in days from conception to birth follow approximately the *N*(266,16) distribution. To find the length of the longest 10% of pregnancies, we can use the NORMINV function. The function requires that we specify the appropriate probability along with the mean and standard deviation of the distribution. Since we want the value for the top 10%, the input constant is 0.9 corresponding to 90% below the calculated value. As for all functions, we click on the cell where you want the results and then type in that cell or on the formula bar as shown below, or click on the function wizard.



=NORMINV(.9,266,16)

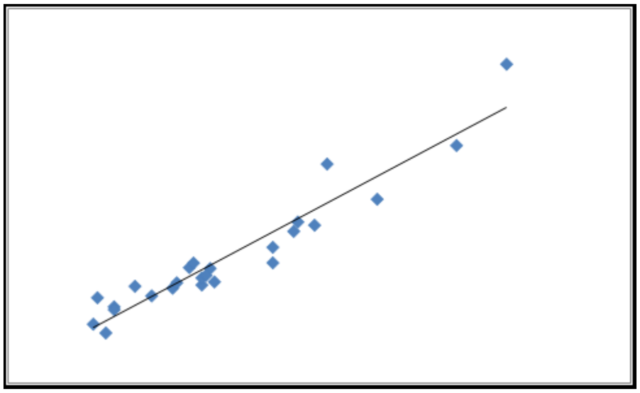
The function returns the inverse of the normal cumulative distribution for the specified mean and standard deviation. In this example the value returned is 286.5, indicating the the longest 10% of human pregnancies last at least 286.5 days.

Alternatively, we can select **Add-Ins ➤ WHFStat ➤ Graphs ➤ Normal Quantile Plot** to do either forward or backward calculations. In the Inputs section, specify the Population Mean and Population Standard deviation. To Calculate an upper or lower tail probability, or even the probability of both tails, fill in the first section with a Target X Value and click the appropriate radio button. To do a backward calculation fill out the second section with the Left or Right-Tailed probability. The third section of the dialog box can be used if you wish to calculate the probability between two values.


Normal Distribution Probability Calculations
Inputs
Population Mean
Population Standard Deviation
Probability for Single X Value
Target X Value
○ P > X
○ P < X
○ Two-Sided
X Value for a Specific Percentage
○ Left-Tailed
%
○ Right-Tailed
%
Probability Between Two X Values
○ Between
and
OK
Cancel

CHAPTER

2



Looking at Data: Exploring Relationships

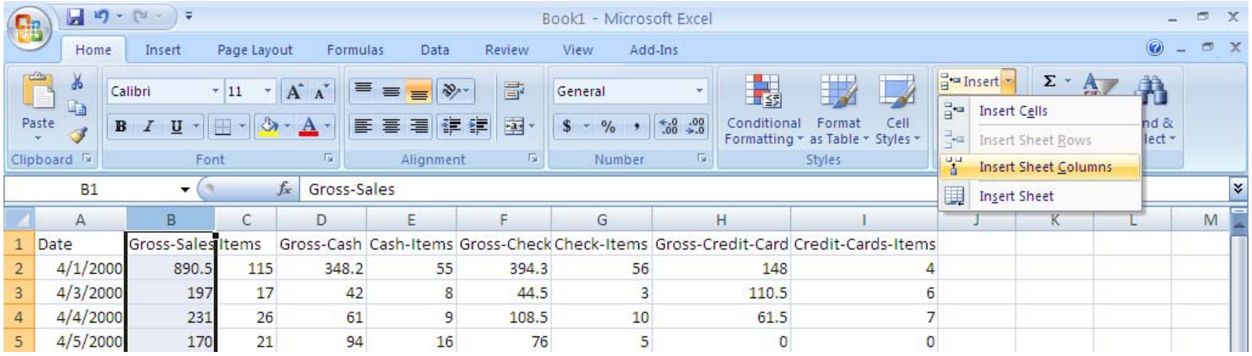
Scatterplots

Often we are interested in illustrating the relationships between two variables, such as the relationship between height and weight, between smoking and lung cancer, or between advertising expenditures and sales. For illustration, we will consider the relationship between the number of items sold and gross sales at Duck Worth Wearing, a shop selling high-quality, second-hand children's clothing, toys, and furniture.

If both variables are quantitative, the most useful display of their relationship is the scatterplot. Scatterplots can be produced by highlighting the variables in the scatterplot and selecting

Insert ➤ Scatter ➤ Scatter with only Markers

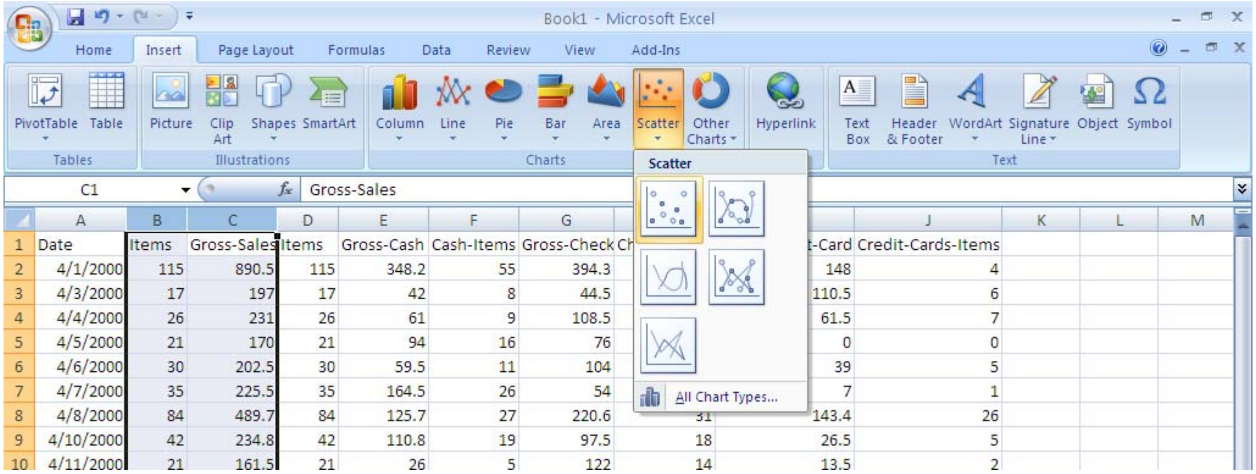
from the menu. The explanatory variable should be plotted on the x-axis and the response variable should be plotted on the y-axis. The highlighted columns should have the x variable on the left and the y variable on the right, so you may need to rearrange your data. If so, you can highlight the column with the y variable and select **Home ➤ Insert ➤ Insert Sheet Columns** to add space for a column to the left of the y variable. Copy the x variable into the empty space.



B1 *fx* Gross-Sales

	A	B	C	D	E	F	G	H	I
1	Date	Gross-Sales	Items	Gross-Cash	Cash-Items	Gross-Check	Check-Items	Gross-Credit-Card	Credit-Cards-Items
2	4/1/2000	890.5	115	348.2	55	394.3	56	148	4
3	4/3/2000	197	17	42	8	44.5	3	110.5	6
4	4/4/2000	231	26	61	9	108.5	10	61.5	7
5	4/5/2000	170	21	94	16	76	5	0	0

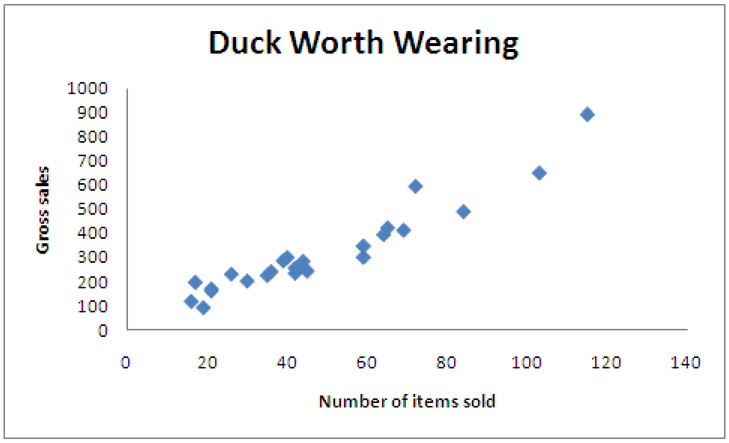
Once the data are correctly arranged and you click on the Scatter with only Markers button, your scatterplot will appear.



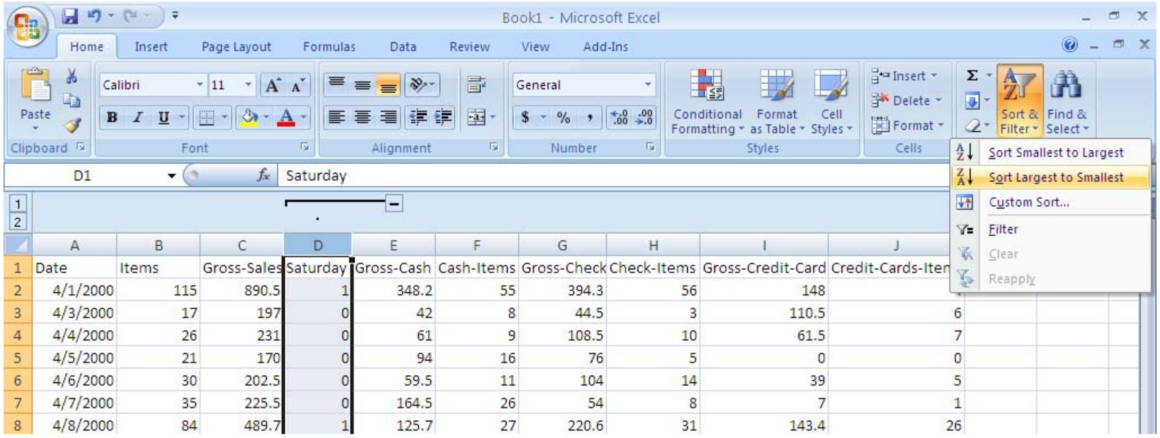
C1 *fx* Gross-Sales

	A	B	C	D	E	F	G	H	I	J
1	Date	Items	Gross-Sales	Items	Gross-Cash	Cash-Items	Gross-Check	Ch	-Card	Credit-Cards-Items
2	4/1/2000	115	890.5	115	348.2	55	394.3		148	4
3	4/3/2000	17	197	17	42	8	44.5		110.5	6
4	4/4/2000	26	231	26	61	9	108.5		61.5	7
5	4/5/2000	21	170	21	94	16	76		0	0
6	4/6/2000	30	202.5	30	59.5	11	104		39	5
7	4/7/2000	35	225.5	35	164.5	26	54		7	1
8	4/8/2000	84	489.7	84	125.7	27	220.6	31	143.4	26
9	4/10/2000	42	234.8	42	110.8	19	97.5	18	26.5	5
10	4/11/2000	21	161.5	21	26	5	122	14	13.5	2

Initially, your scatterplot may not look the way you want it to. If you are clicked on the chart, Chart Tools will also appear at the top of the Excel menu. These tools allow you to modify the chart. Choose layout and modify as desired. For example, select **Layout ➤ Axis Titles ➤ Primary Horizontal Axis Title ➤ Title Below Axis** to add an x-axis title. Click the axis title and type the text that you want. The data in the scatterplot are positively associated, in a roughly linear pattern with no clear outliers.



We can add information about a third categorical variable to a scatterplot by using different symbols for different points. The Duck Worth Wearing store is open Monday through Saturday. The five Saturdays in April 2000 (04/01, 04/08, 04/15, 04/22 and 04/29) are the days with the highest numbers of items sold. We can improve the scatterplot by plotting the Saturdays with a different plot symbol. First, we add a categorical variable Saturday to the Excel spreadsheet. This variable has only two values: “1” for the Saturdays and “0” for the weekdays. The Saturday data is easily separated from the weekdays by sorting as shown.



D1 Saturday

	A	B	C	D	E	F	G	H	I	J
1	Date	Items	Gross-Sales	Saturday	Gross-Cash	Cash-Items	Gross-Check	Check-Items	Gross-Credit-Card	Credit-Cards-Iten
2	4/1/2000	115	890.5	1	348.2	55	394.3	56	148	
3	4/3/2000	17	197	0	42	8	44.5	3	110.5	6
4	4/4/2000	26	231	0	61	9	108.5	10	61.5	7
5	4/5/2000	21	170	0	94	16	76	5	0	0
6	4/6/2000	30	202.5	0	59.5	11	104	14	39	5
7	4/7/2000	35	225.5	0	164.5	26	54	8	7	1
8	4/8/2000	84	489.7	1	125.7	27	220.6	31	143.4	26

A labeled scatterplot can then be obtained by selecting **Design ➤ Select Data ➤ Add** to add a series with only the Saturday data.

.

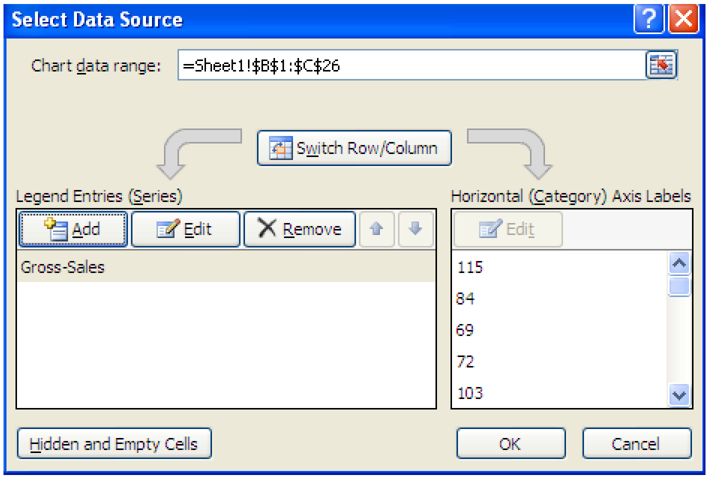


Chart data range: =Sheet1!$B$1:$C$26

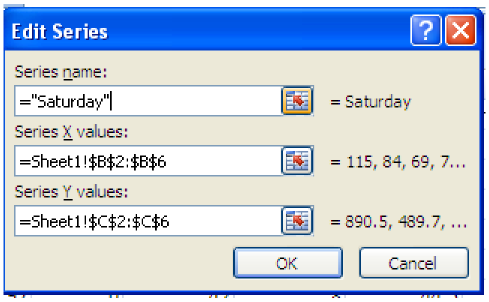
Legend Entries (Series)

Gross-Sales

Horizontal (Category) Axis Labels

115
84
69
72
103

Specify the x and y values and a series name by clicking on the small spreadsheet icons.

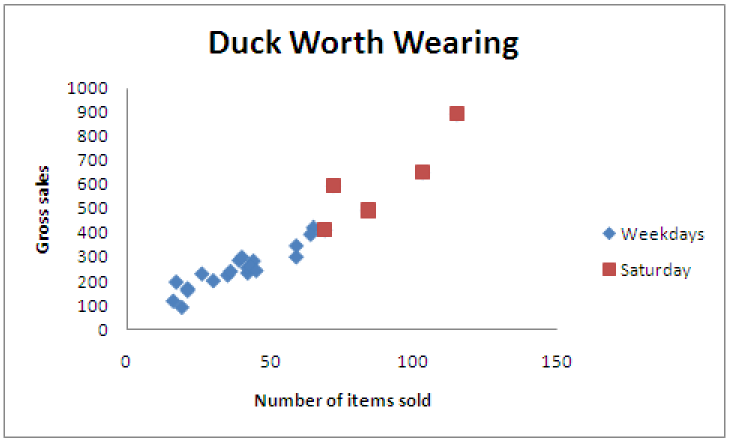


Series name: ="Saturday" = Saturday

Series X values: =Sheet1!$B$2:$B$6 = 115, 84, 69, 7…

Series Y values: =Sheet1!$C$2:$C$6 = 890.5, 489.7, …

The additional series will appear in the graph. You will probably want to add a legend to the scatterplot by clicking on the chart and selecting **Layout ➤ Legend**.



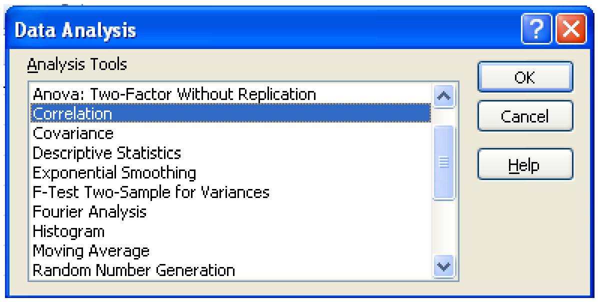
The scatterplot with Saturdays identified shows that the company is busier on Saturdays.

Correlation

We can compute the correlation coefficient between two quantitative variables using Excel. The correlation coefficient can be calculated by selecting

Data ➤ Data Analysis ➤ Basic Statistics ➤ Correlation

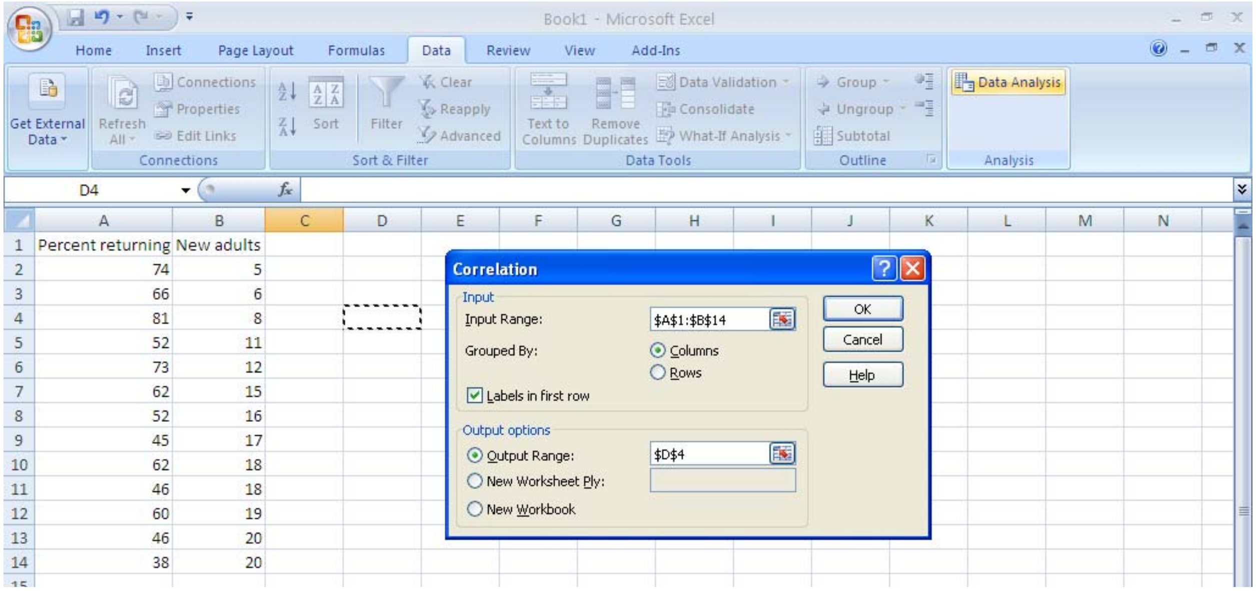
from the menu.



Analysis Tools
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation

Below we illustrate a correlation calculation with bird colony data. The data gives, for 13 colonies of sparrow haws, the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony.

To calculate the correlation, the input range should be the variable for which you wish you are needing the correlation. If there are labels in the first row, check the appropriate box. Select a location for the output and click on the OK button.

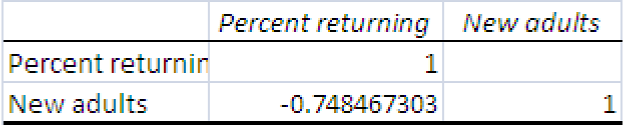


	A	B
1	Percent returning	New adults
2	74	5
3	66	6
4	81	8
5	52	11
6	73	12
7	62	15
8	52	16
9	45	17
10	62	18
11	46	18
12	60	19
13	46	20
14	38	20

Input
Input Range: $A$1:$B$14
Grouped By: ● Columns ○ Rows
☑ Labels in first row

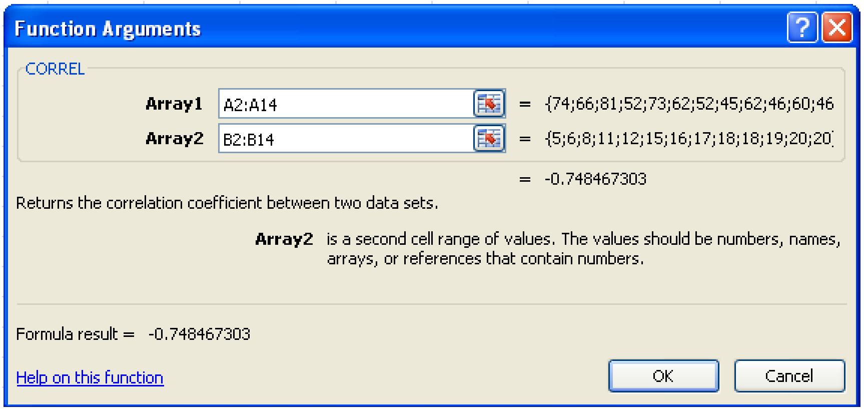
Output options
● Output Range: $D$4
○ New Worksheet Ply:
○ New Workbook

The correlation of the two variables is shown in the table below to be -0.748. If more than two variables are selected in the Input Range, Excel will include the correlation coefficients between all pairs of variables.



	Percent returning	*New adults*
Percent returnir	1	
New adults	-0.748467303	1

Alternatively, the correlation of two variables can be calculated using the **CORREL** function. Type =CORREL(data range) into any cell or click on the Insert Function button and type CORREL to obtain the dialog box shown below.



Array1 A2:A14 = {74;66;81;52;73;62;52;45;62;46;60;46

Array2 B2:B14 = {5;6;8;11;12;15;16;17;18;18;19;20;20

= -0.748467303

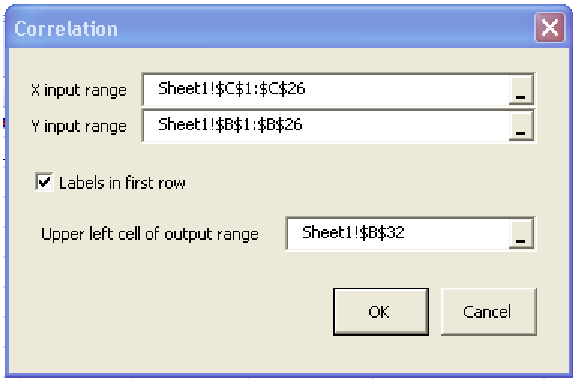
Returns the correlation coefficient between two data sets.

Array2 is a second cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result = -0.748467303

Help on this function

Correlation can also be calculated by selecting **Add-Ins ➤ WHFStat ➤ Correlation and Regression ➤ Correlation** from the Excel menu. The dialog box is filled out as shown to obtain the correlation coefficient between two variables.



X input range: Sheet1!$C$1:$C$26

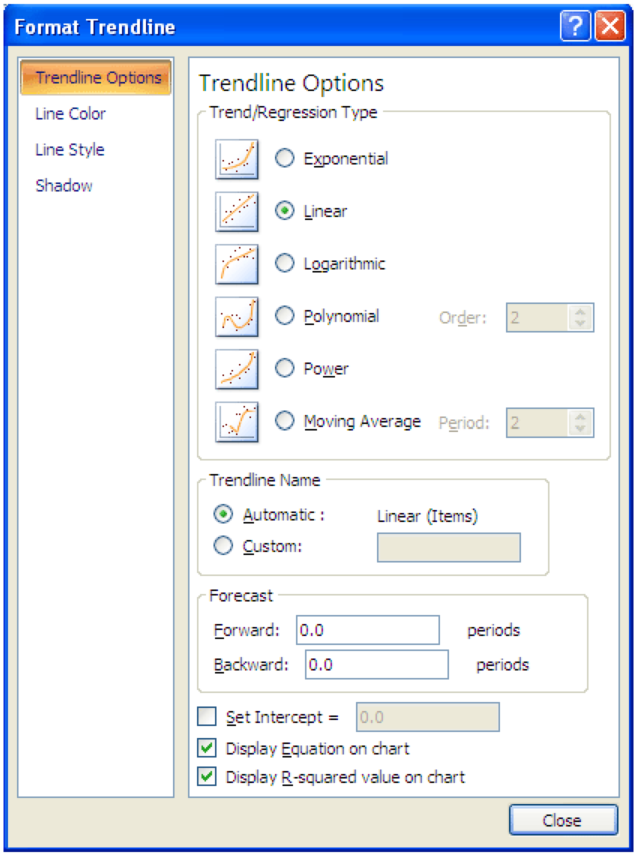
Y input range: Sheet1!$B$1:$B$26

☑ Labels in first row

Upper left cell of output range: Sheet1!$B$32

Least-Squares Regression

The scatterplot for Duck Worth Wearing shows that there is a strong linear relationship between the number of items sold and the gross sales. To calculate the least-squares line of the form $y = a + bx$ from data, right click on a point on the scatterplot, select Add Trendline from the list, select Linear on the Trendline Options, check the box next to Display Equation on the chart, and Display R-squared value on the chart if desired.



Format Trendline

Trendline Options
Line Color
Line Style
Shadow

Trendline Options

Trend/Regression Type

○ Exponential
● Linear
○ Logarithmic
○ Polynomial Order: 2
○ Power
○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Items)
○ Custom:

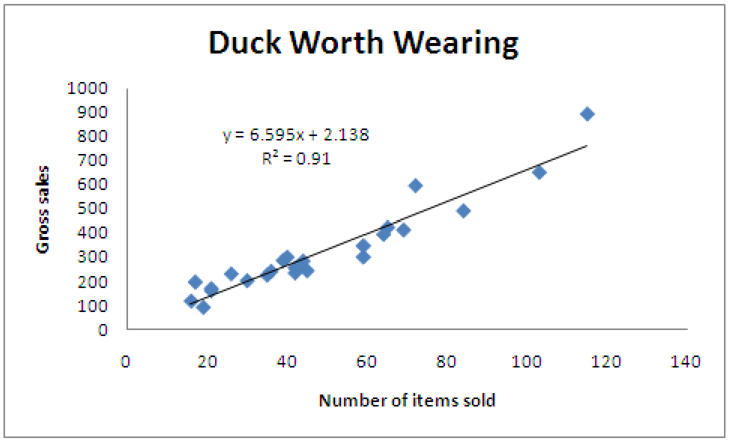
Forecast

Forward: 0.0 periods
Backward: 0.0 periods

□ Set Intercept = 0.0
☑ Display Equation on chart
☑ Display R-squared value on chart

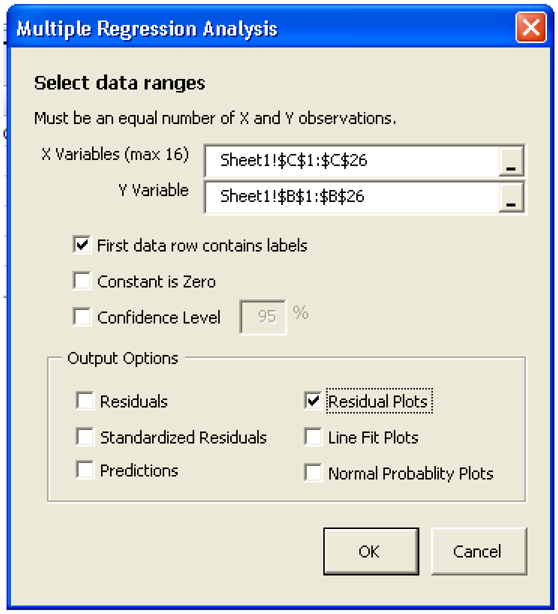
Close

The scatterplot now shows the least-sqaures line, the equation for the line (y = 6.595x + 2.138) and that r^2 = 0.91. The slope and intercept can also be found using Excel's **SLOPE** and **INTERCEPT** functions. These functions can be typed into a cell or you can click on the Insert Function button. Both **SLOPE** and **INTERCEPT** require that you specify the known values of y and x.



To find the residual for each point, first calculate the fitted value for each point, then calculate the value of the residual. For each point, the fitted value, $\hat{y} = a + bx$ and the residual is $y - \hat{y}$. Once the residuals have been calculated, the residual plot is just a scatterplot of the residual versus the x variable.

Alternatively, residual plots can be obtained by selecting **Add-Ins ➤ WHFStat ➤ Correlation and Regression ➤ Correlation** from the Excel menu. The dialog box is filled out as shown.



Multiple Regression Analysis

Select data ranges

Must be an equal number of X and Y observations.

X Variables (max 16) Sheet1!$C$1:$C$26

Y Variable Sheet1!$B$1:$B$26

☑ First data row contains labels

☐ Constant is Zero

☐ Confidence Level 95 %

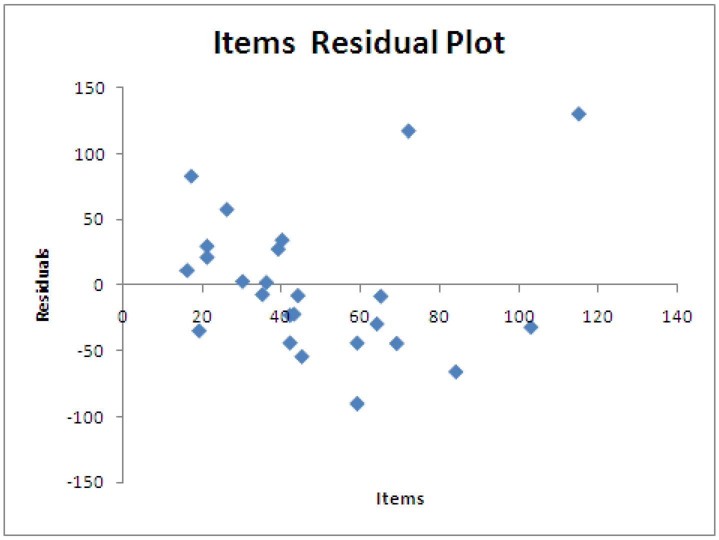
Output Options

☐ Residuals ☑ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

☐ Predictions ☐ Normal Probablity Plots

The following residual plot was produced by the Add In and shows some tilt due to the two large residuals that are somewhat influential.



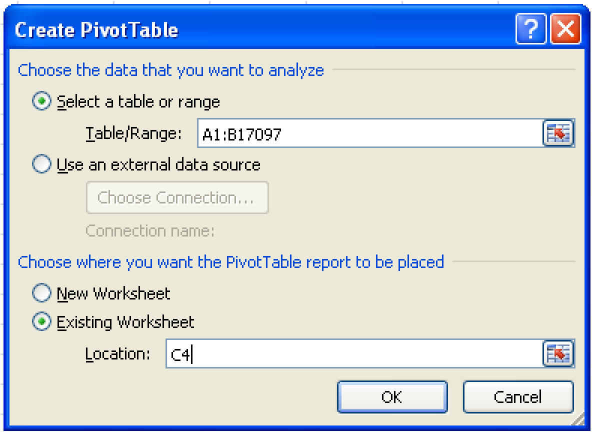
Tables for Categorical Variables

We can describe relationships between two or more categorical variables using two- or three-way tables in Excel. We will use the data on binge drinking by college students. In this data set, we have stored information on 17,096 students classified by gender and whether or not they are frequent binge drinkers.

To make a two-way table in Excel select

Insert ➤ Pivot Table

from the menu. In the dialog box, select the range of input data and the location where you want the Pivot Table report to be placed as shown below.



Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: A1:B17097

○ Use an external data source

Connection name:

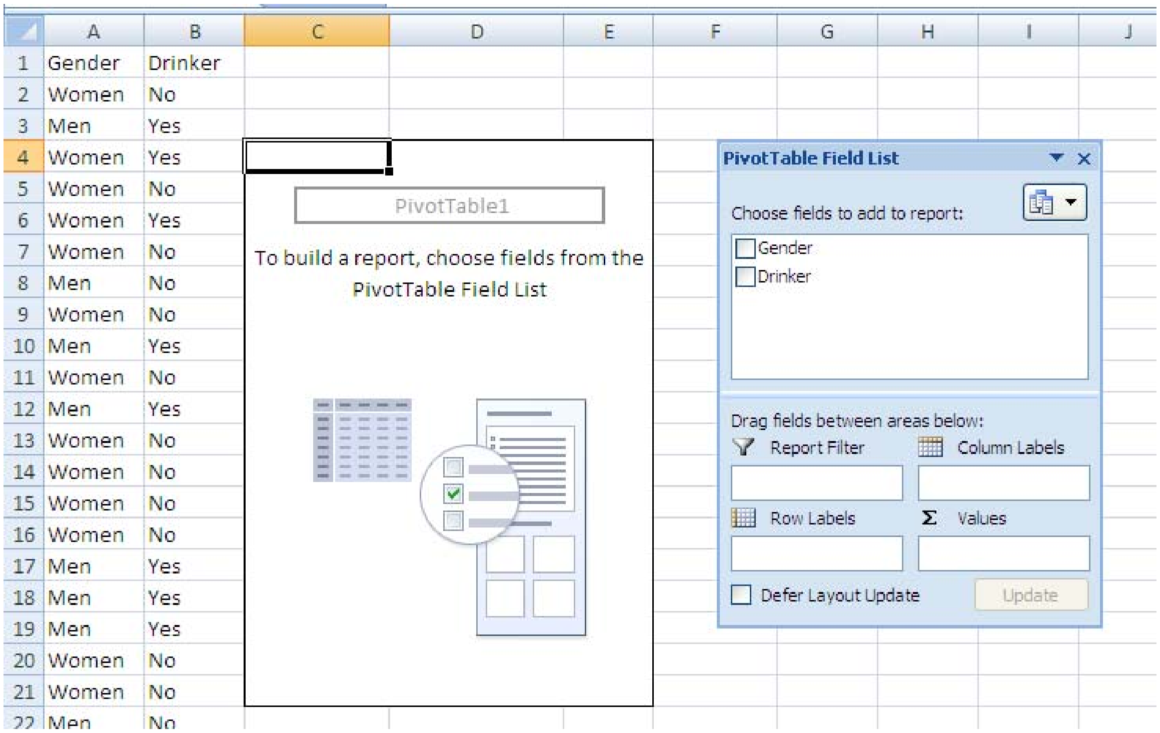
Choose where you want the PivotTable report to be placed

○ New Worksheet

● Existing Worksheet

Location: C4

Click OK and the blank Pivot Table will appear. The Pivot Table Field List will also appear as long as a cell within the Pivot Table is selected.



	A	B
1	Gender	Drinker
2	Women	No
3	Men	Yes
4	Women	Yes
5	Women	No
6	Women	Yes
7	Women	No
8	Men	No
9	Women	No
10	Men	Yes
11	Women	No
12	Men	Yes
13	Women	No
14	Women	No
15	Women	No
16	Women	No
17	Men	Yes
18	Men	Yes
19	Men	Yes
20	Women	No
21	Women	No
22	Men	No

PivotTable1

To build a report, choose fields from the PivotTable Field List

PivotTable Field List

Choose fields to add to report:

□ Gender
□ Drinker

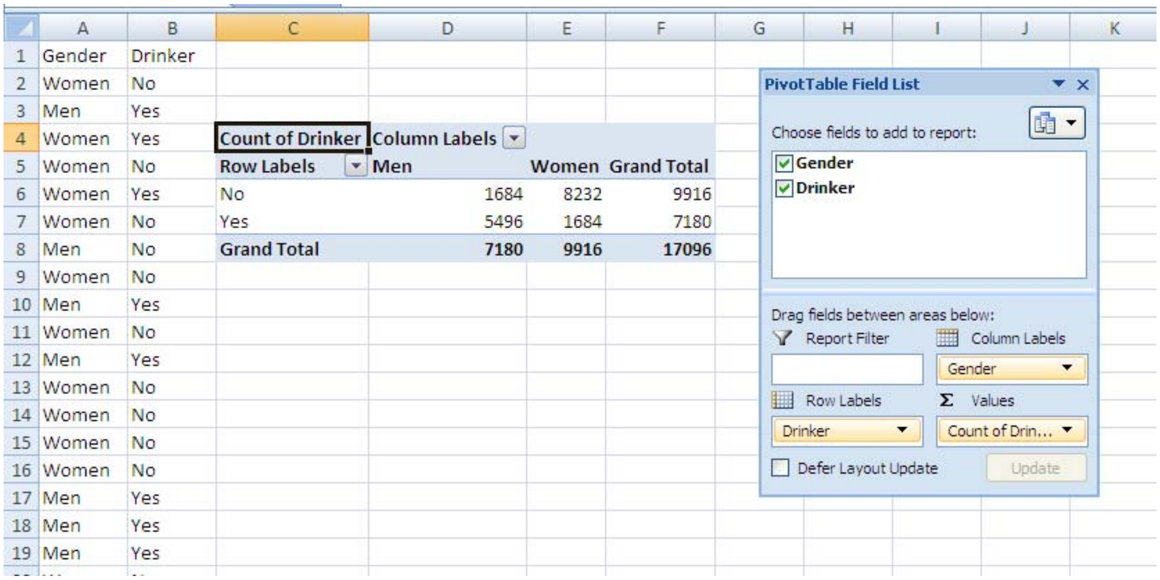
Drag fields between areas below:

Report Filter | Column Labels

Row Labels | Values

□ Defer Layout Update

If we view gender as the explanatory variable and frequent binge drinking as the response variable, then we put gender in the columns and frequent binge drinking in the rows. This is easily done by dragging the word Gender into the Column Labels field and the word Drinker into the Row Labels field. Once either Gender or Drinker is dragged into the Values field, the data will appear in the Pivot Table as shown below.



	A	B
1	Gender	Drinker
2	Women	No
3	Men	Yes
4	Women	Yes
5	Women	No
6	Women	Yes
7	Women	No
8	Men	No
9	Women	No
10	Men	Yes
11	Women	No
12	Men	Yes
13	Women	No
14	Women	No
15	Women	No
16	Women	No
17	Men	Yes
18	Men	Yes
19	Men	Yes

Count of Drinker	Column Labels		
Row Labels	Men	Women	Grand Total
No	1684	8232	9916
Yes	5496	1684	7180
Grand Total	7180	9916	17096

PivotTable Field List

Choose fields to add to report:

☑ Gender
☑ Drinker

Drag fields between areas below:

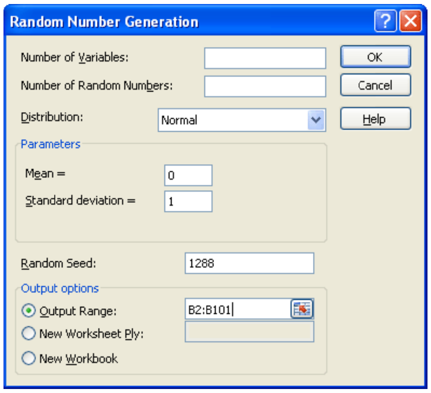
Report Filter:
Column Labels: Gender
Row Labels: Drinker
Values: Count of Drin...

□ Defer Layout Update

For three-way tables, an additional variable would be included and dragged into the Report Filter field.

Once the two-way table has been constructed, marginal and conditional probabilities can be constructed by typing the appropriate formulas into cells. For example, to calculate the proportion of men that are frequent binge drinkers, the formula would be =D7/D8.

CHAPTER 3



Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Normal

Parameters

Mean = 0

Standard deviation = 1

Random Seed: 1288

Output options

● Output Range: B2:B101

○ New Worksheet Ply:

○ New Workbook

Producing Data

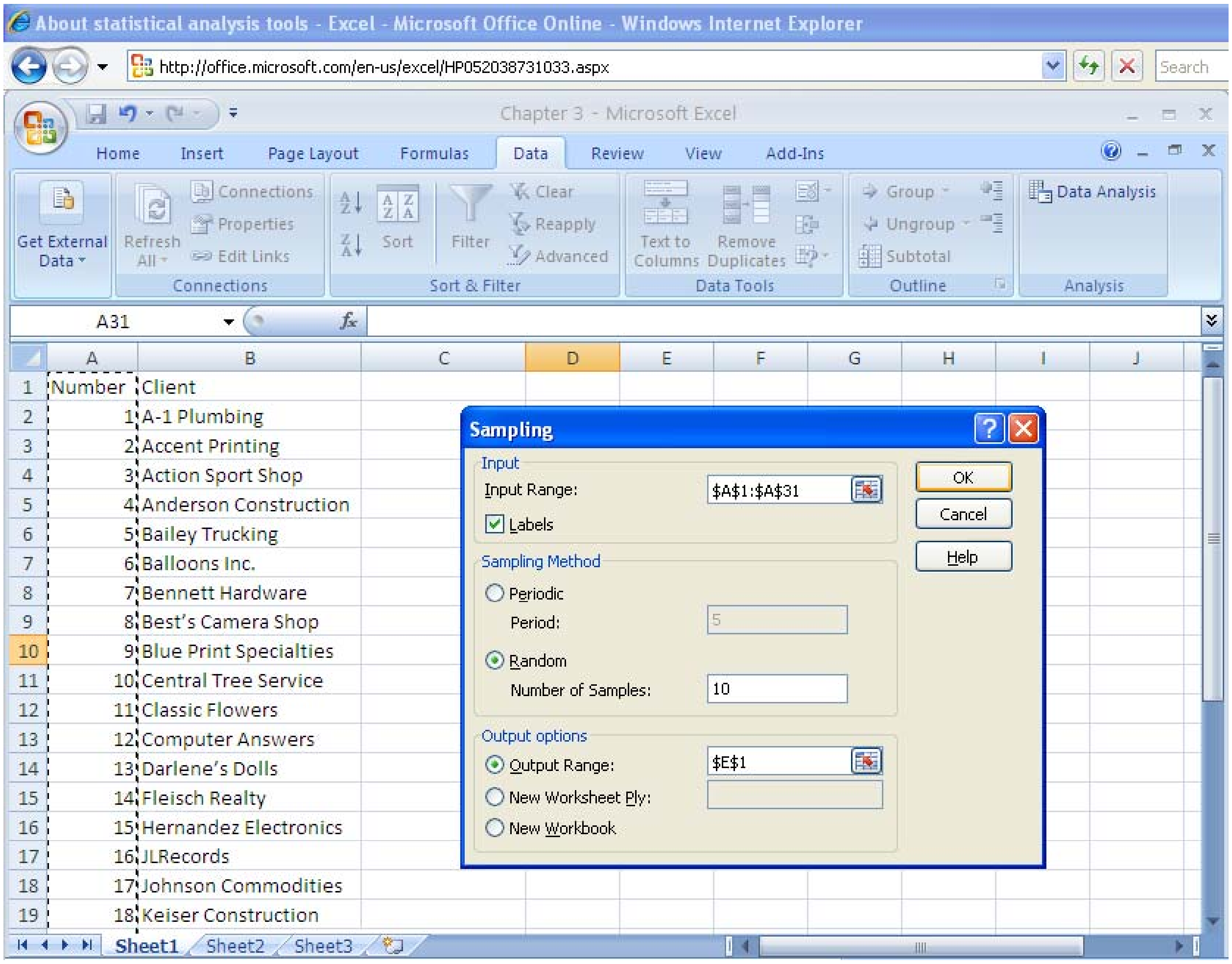
Random Samples

Excel allows us to select a simple random sample from a population. To choose a random sample, select

Data ➤ Data Analysis ➤ Sampling

from the menu. Specify the input range from which you are sampling, click on the radio button for Random, specify the number of samples, and the output range. In the example below, we wish to select a sample of five randomly selected small business clients for a customer satisfaction survey. The input range for Sampling must be numeric. If you have a list of names instead of numbers, you must create a corresponding list of numbers. To enter a list of numbers into Excel, enter the first few numbers to establish the pattern. Highlight these numbers and then use the fill handle to automatically fill data in worksheet cells.

The sample you select may have repeated numbers. You can select a new sample if this is not what you want or you can select a sample larger than needed so that you can skip any repeats.



	A	B
1	Number	Client
2	1	A-1 Plumbing
3	2	Accent Printing
4	3	Action Sport Shop
5	4	Anderson Construction
6	5	Bailey Trucking
7	6	Balloons Inc.
8	7	Bennett Hardware
9	8	Best's Camera Shop
10	9	Blue Print Specialties
11	10	Central Tree Service
12	11	Classic Flowers
13	12	Computer Answers
14	13	Darlene's Dolls
15	14	Fleisch Realty
16	15	Hernandez Electronics
17	16	JLRecords
18	17	Johnson Commodities
19	18	Keiser Construction

Sampling

Input
Input Range: $A$1:$A$31
☑ Labels

Sampling Method
○ Periodic
Period: 5
● Random
Number of Samples: 10

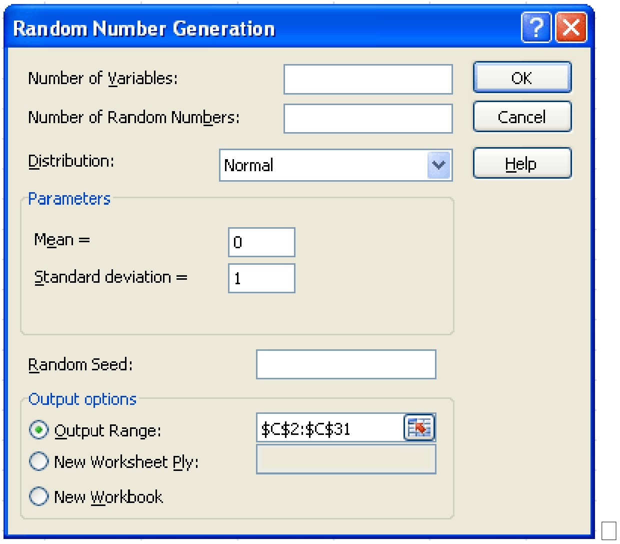
Output options
● Output Range: $E$1
○ New Worksheet Ply:
○ New Workbook

Alternatively, a sample can be chosen by assigning random numbers to each item or person in the population and then sorting the population to select the items with either the smallest or largest random numbers. To assign random numbers, select

Data ➤ Data Analysis ➤ Random Number Generation

from the menu.

To assign a random number to each client on the list, choose **Data ➤ Data Analysis ➤ Random Number Generation** from the menu. Specify that you wish to select the random numbers from a the normal distribution (or uniform distribution) and specify the cells where the results will be stored as shown below.



Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Normal

Parameters

Mean = 0

Standard deviation = 1

Random Seed:

Output options

● Output Range: $C$2:$C$31

○ New Worksheet Ply:

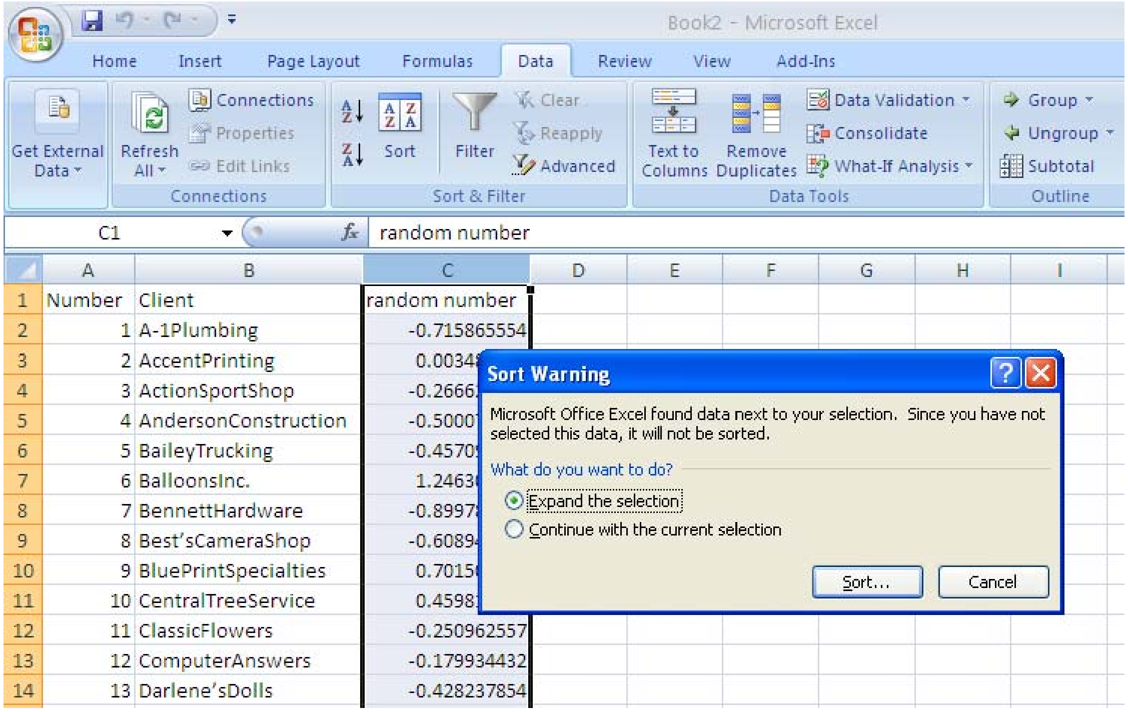
○ New Workbook

Sorting Data

Rather than searching through the list of random numbers to select the clients with the smallest (or largest) random numbers, it is convenient to sort the clients. Highlight the column with the random numbers and select

Data ➤ Sort

from the menu. This command orders the data in a column in numerical sequence. If Excel finds data next to your selected column, you will get the Sort Warning shown below and have a choice to "Expand the selection" or "Continue with the current selection." Since you want to keep client name and number associated with the random numbers, you should select "Expand the selection." If you select "Continue with the current selection," only the highlighted column with be sorted and the adjacent columns will remain as they are.



C1 *fx* random number

	A	B	C
1	Number	Client	random number
2	1	A-1Plumbing	-0.715865554
3	2	AccentPrinting	0.0034
4	3	ActionSportShop	-0.2666
5	4	AndersonConstruction	-0.5000
6	5	BaileyTrucking	-0.4570
7	6	BalloonsInc.	1.2463
8	7	BennettHardware	-0.8997
9	8	Best'sCameraShop	-0.6089
10	9	BluePrintSpecialties	0.7015
11	10	CentralTreeService	0.4598
12	11	ClassicFlowers	-0.250962557
13	12	ComputerAnswers	-0.179934432
14	13	Darlene'sDolls	-0.428237854

Sort Warning

Microsoft Office Excel found data next to your selection. Since you have not selected this data, it will not be sorted.

What do you want to do?

● Expand the selection

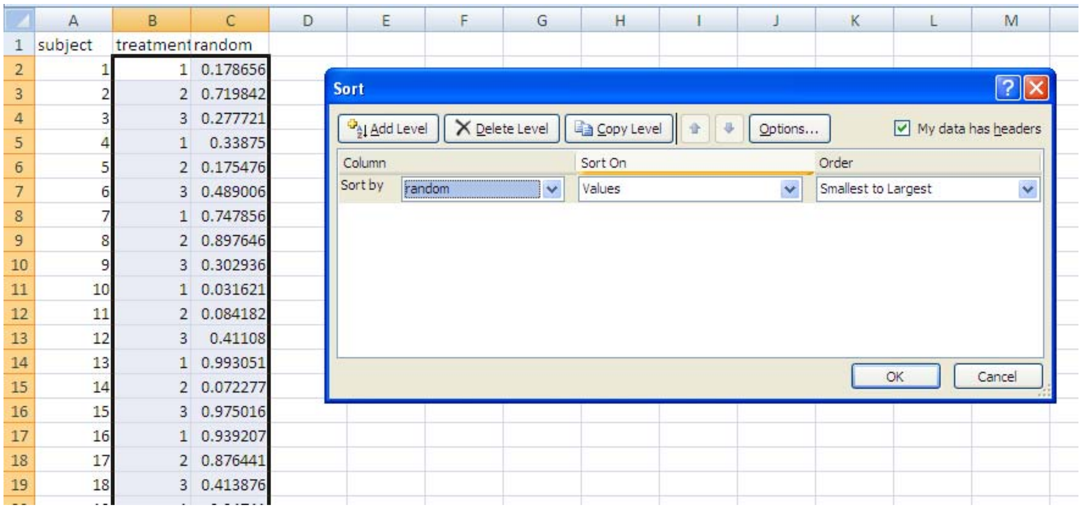
○ Continue with the current selection

Randomization in Experiments

Sampling can be used to randomly select treatment groups in an experiment. If you have a list of subjects and a list of treatments, random numbers can be used to reorder one of these lists to make the random assignments.

For example, suppose that 60 subjects are to be assigned to 3 treatments. You could have the numbers 1 through 60 in one column and the numbers 1, 2, and 3, each repeated 20 times, listed in another column. A third column of random numbers should also be added. The random numbers can be selected by choosing **Data ➤ Data Analysis ➤ Random Number Generation** from the menu as explained earlier, or by using either the **RAND()** or **RANDBETWEEN**(bottom,top). **RAND**() returns a random number between 0 and 1, evenly distributed. Although the parenthesis are required, this function has no arguments. **RANDBETWEEN**(bottom,top) returns a random number between the bottom and top numbers that you specify. If you use these functions, you must copy the numbers and then do a special paste of only the values. To copy the random numbers, select the Home tab, in the **Clipboard** group, click **Copy** . Alternatively, you could select Control C or right click on the data and select Copy. To paste only the values without the formula, on the **Home** tab, in the **Clipboard** group, click **Paste** , and then click **Paste Values**.

Now highlight the column with the treatments and the random numbers. Select **Data ➤ Sort** from the menu. In the dialog box, you will specify the column that you wish to sort. The other column that you highlighted will be carried along.



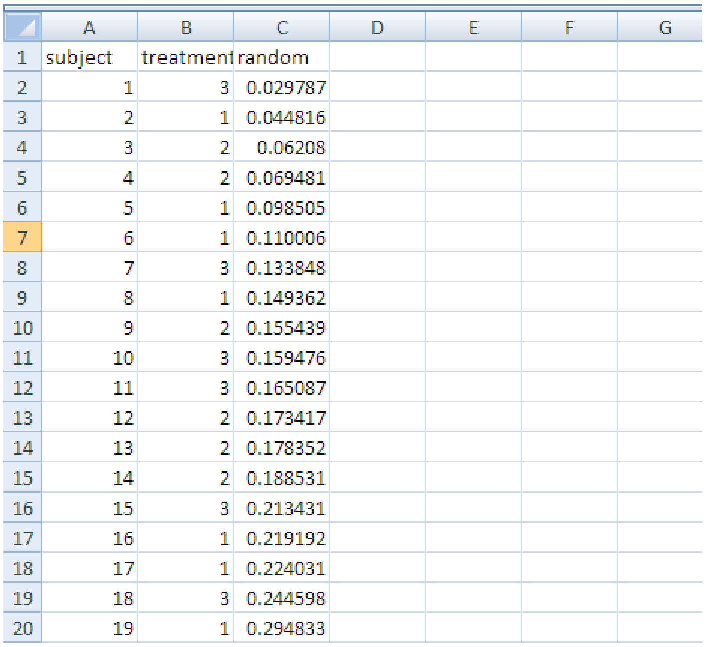
	A	B	C
1	subject	treatment	random
2	1	1	0.178656
3	2	2	0.719842
4	3	3	0.277721
5	4	1	0.33875
6	5	2	0.175476
7	6	3	0.489006
8	7	1	0.747856
9	8	2	0.897646
10	9	3	0.302936
11	10	1	0.031621
12	11	2	0.084182
13	12	3	0.41108
14	13	1	0.993051
15	14	2	0.072277
16	15	3	0.975016
17	16	1	0.939207
18	17	2	0.876441
19	18	3	0.413876

Sort

☑ My data has headers

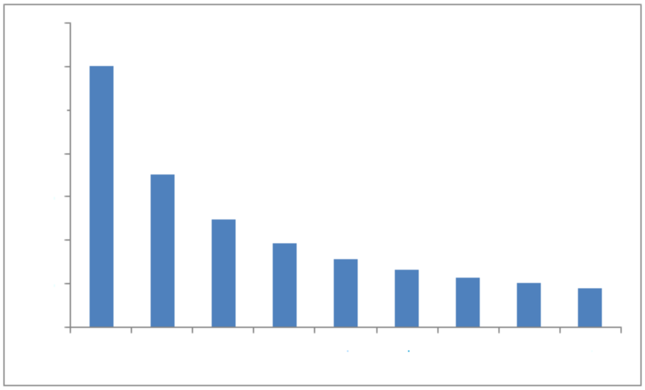
Column	Sort On	Order
Sort by random	Values	Smallest to Largest

After sorting, the treatments will appear in a random order and the worksheet will show which subjects get each treatment.



	A	B	C
1	subject	treatment	random
2	1	3	0.029787
3	2	1	0.044816
4	3	2	0.06208
5	4	2	0.069481
6	5	1	0.098505
7	6	1	0.110006
8	7	3	0.133848
9	8	1	0.149362
10	9	2	0.155439
11	10	3	0.159476
12	11	3	0.165087
13	12	2	0.173417
14	13	2	0.178352
15	14	2	0.188531
16	15	3	0.213431
17	16	1	0.219192
18	17	1	0.224031
19	18	3	0.244598
20	19	1	0.294833

CHAPTER 4



Probability

Simulating Random Data

Excel can be used to simulate random data. To simulate random data, select

Data ➤ Data Analysis ➤ Random Number Generation

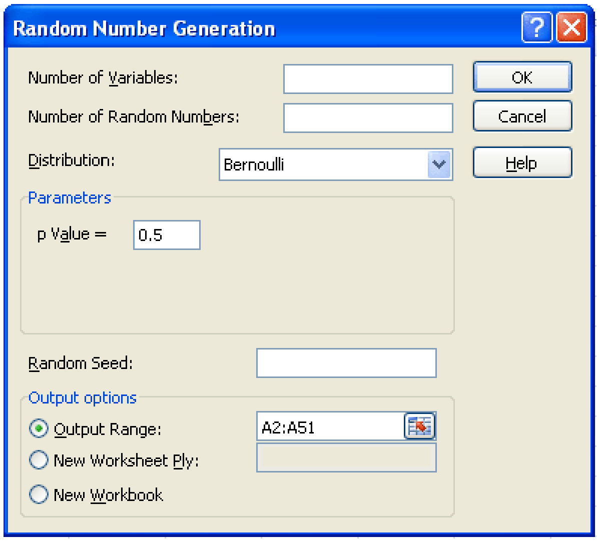
from the menu. Then select a distribution from the dialog box shown on the next page. For example, if you select **Bernoulli** from the menu, you can generate a random sequence of 0s and 1s. In the dialog box, select the distribution. If the distribution is Bernoulli, specify also a *P*-value (probability of success) and where the results are to be stored. For example, to simulate a sequence of 50 coin tosses (with equal probability of heads and tails), the dialog box must be filled in as shown on the following page.

For example, the sequence of 0s and 1s may look like:

1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0

The 1s correspond to the variable for which you input the probability of success. In this case, we may consider the 1s to be "heads" and the 0s to be "tails." Thus, our sequence of coin flips would be:

H, T, H, T, T, T, H, T, H, H, H, T, T, H, H, T, T, T, T, H, H, H, T, T, H, T, H, H, H, H, T, T, H, T, H, T, H, T, H, T, T, T, H, H, H, T, H, H, H, T.



Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Bernoulli

Parameters

p Value = 0.5

Random Seed:

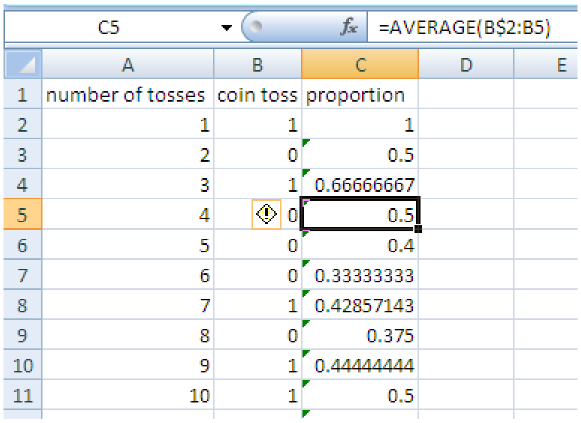
Output options

● Output Range: A2:A51

○ New Worksheet Ply:

○ New Workbook

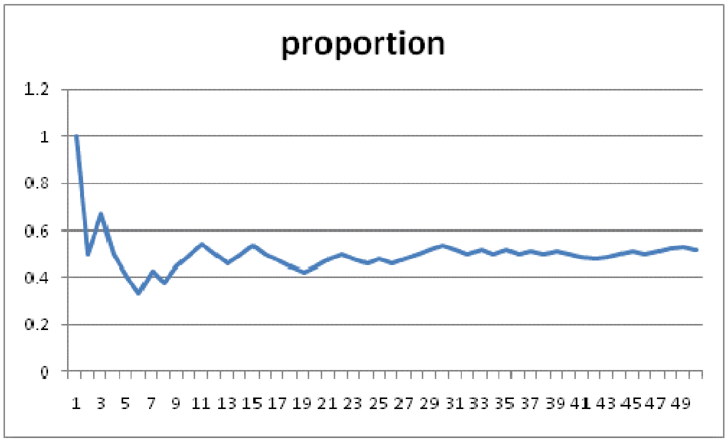
To graph the proportion of heads versus the number of tosses, we need to calculate the proportion of heads as shown below. Notice that the formula includes a $ sign to indicate that the first number in the average is always in row 2, even when the formula is copied.



C5 =AVERAGE(B$2:B5)

	A	B	C	D	E
1	number of tosses	coin toss	proportion		
2	1	1	1		
3	2	0	0.5		
4	3	1	0.66666667		
5	4	0	0.5		
6	5	0	0.4		
7	6	0	0.33333333		
8	7	1	0.42857143		
9	8	0	0.375		
10	9	1	0.44444444		
11	10	1	0.5		

To construct a graph of the data, highlight the entire column of proportions and select **Insert ➤ Line** from the menu. The proportion of tosses that give a head changes as we make tosses. Eventually, the proportion approaches 0.5.

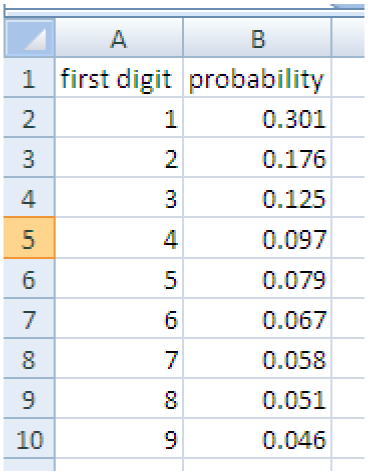


Simulating from Other Distributions

In addition to Bernoulli, Excel can be used to simulate data from many other distributions. These distributions are Uniform, Normal, Bernoulli, Binomial, Poisson, Patterned, and Discrete. For example, discrete distributions can be simulated with Excel. Benford's law describes a distribution that is often observed in the first digit of numerical records. Here is the distribution for Benford's law.

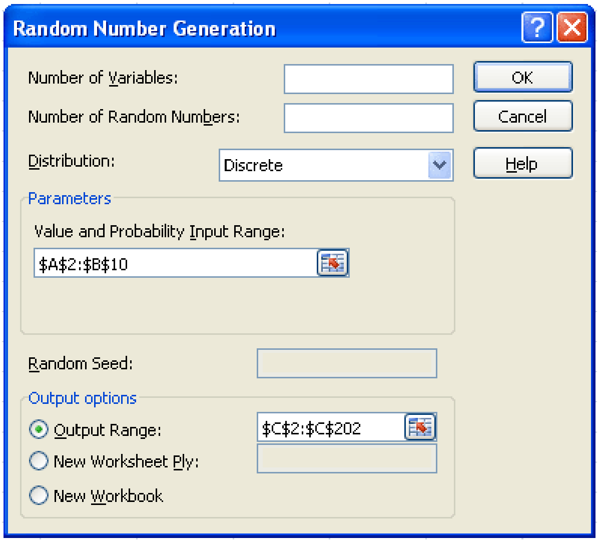
First digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Any discrete distribution can be specified by putting the values and corresponding probabilities into two columns. We will simulate observations from the distribution following Benford's law. First we enter the sizes and probabilities into an Excel spreadsheet as shown below.



	A	B
1	first digit	probability
2	1	0.301
3	2	0.176
4	3	0.125
5	4	0.097
6	5	0.079
7	6	0.067
8	7	0.058
9	8	0.051
10	9	0.046

To simulate data from the specified discrete distribution, select **Data ➤ Data Analysis ➤ Random Number Generation** from the menu and select **Discrete** in the dialog box. As shown, you must specify where the data are to be stored, the column specifying the values, and the column specifying the probabilities.



Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Discrete

Parameters

Value and Probability Input Range: $A$2:$B$10

Random Seed:

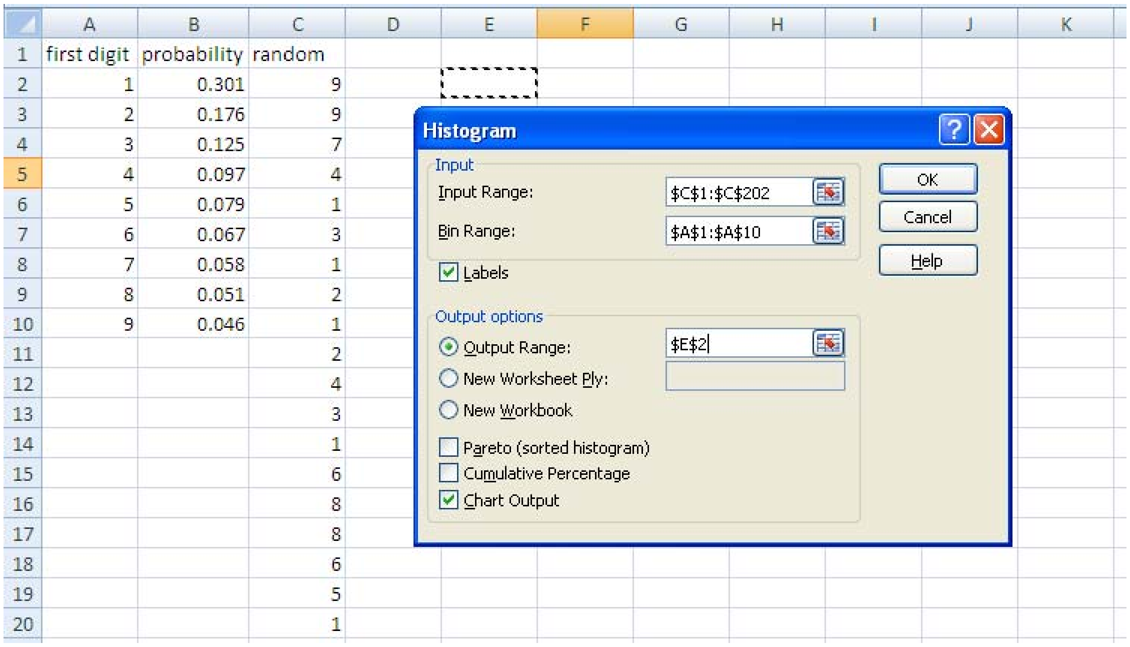
Output options

● Output Range: $C$2:$C$202

○ New Worksheet Ply:

○ New Workbook

Simulated data will not look exactly like the distribution from which they are selected. To see the difference between the exact distribution and the simulated data, we can compare graphs. To graph the simulated data, select **Data ➤Data Analysis ➤ Histogram** from the menu. In the dialog box, select the random data for the Input Range and the column listing the first digits for the Bin Range as shown below.



	A	B	C
1	first digit	probability	random
2	1	0.301	9
3	2	0.176	9
4	3	0.125	7
5	4	0.097	4
6	5	0.079	1
7	6	0.067	3
8	7	0.058	1
9	8	0.051	2
10	9	0.046	1
11			2
12			4
13			3
14			1
15			6
16			8
17			8
18			6
19			5
20			1

Histogram

Input

Input Range: $C$1:$C$202

Bin Range: $A$1:$A$10

☑ Labels

Output options

● Output Range: $E$2

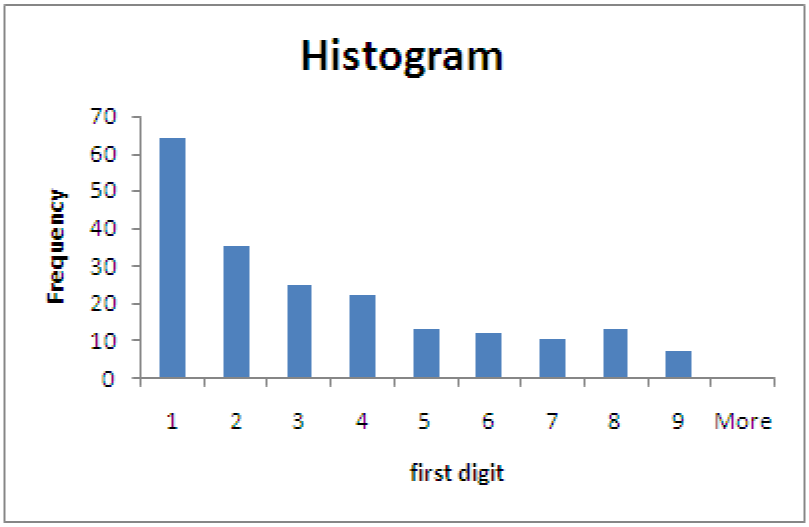
○ New Worksheet Ply:

○ New Workbook

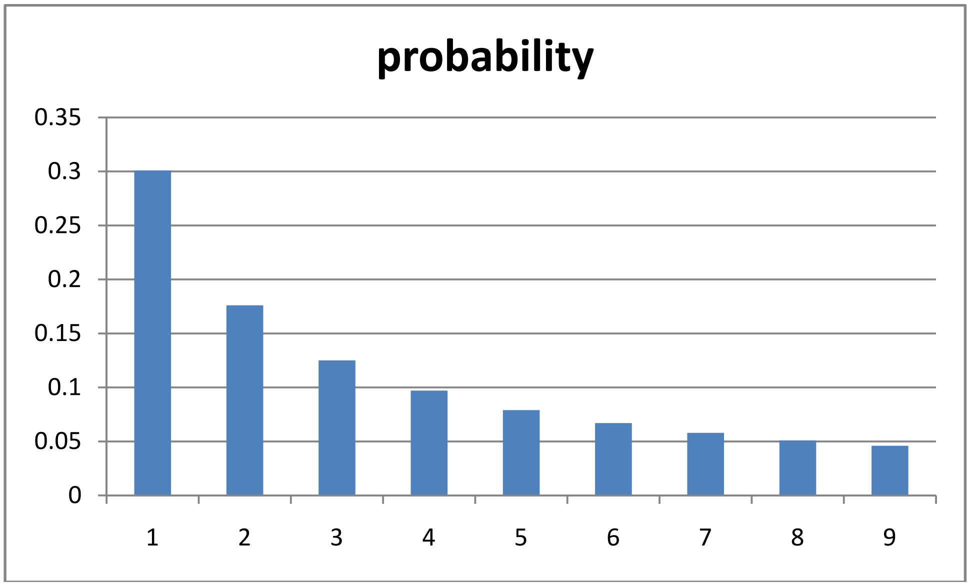
☐ Pareto (sorted histogram)

☐ Cumulative Percentage

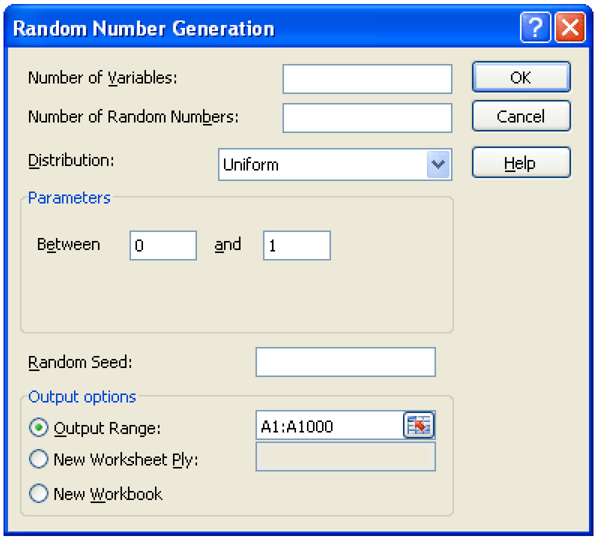
☑ Chart Output



To compare the simulated data with the specified distribution, highlight the column of probabilities and select **Insert ➤ Column** from the menu. Both graphs are skewed to the right, but the simulated data in the histogram are not as smoothly distributed as the probability distribution illustrated in the following bar chart. The randomness illustrated in the histogram is typical of simulated data.



To generate random numbers that are spread out uniformly between two numbers, select **Data ➤ Data Analysis ➤ Random Number Generation** from the menu and select **Uniform** in the dialog box. For example to generate 1000 random numbers uniformly across the interval from 0 to 1, the dialog box would be filled out as follows.



Distribution: Uniform

Parameters

Between 0 and 1

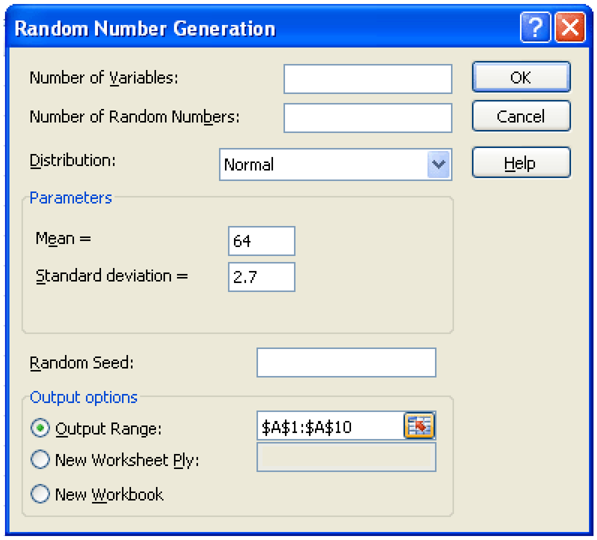
Output options

● Output Range: A1:A1000

○ New Worksheet Ply:

○ New Workbook

Select **Data ➤ Data Analysis ➤ Random Number Generation** from the menu and select **Normal** in the dialog box to simulate observations from a Normal distribution. To simulate the heights of ten young women with the *N*(64, 2.7) distributions, the dialog box would be filled out as follows.



Distribution: Normal

Parameters

Mean = 64

Standard deviation = 2.7

Output options

● Output Range: $A$1:$A$10

○ New Worksheet Ply:

○ New Workbook

Probability Calculations

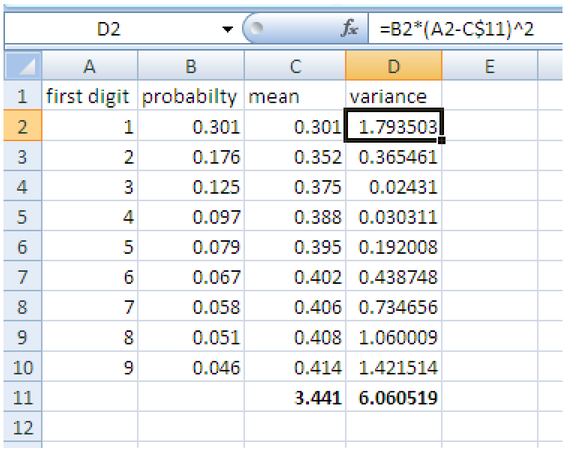
Excel lets you perform mathematical operations and functions. The results of a calculation can be stored in a cell. For example, if we wish to calculate the probability that a first digit is equal to or greater than 6, we can use our Benford's Law data and use Excel's SUM function to add the appropriate probabilities.

We can also use Excel to calculate the mean and variance of a discrete random variable using the following equations.

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_k - \mu_X)^2 p_k$$

To calculate the mean and variable of *X* we arrange the calculation in the form of a table as shown below. The third column gives the value of $x_1 p_1$, i.e., the product of the first two columns. The values in this column are added up using the SUM function to find the mean. The next column gives $(x_i + \mu_X)^2$. If this formula is to be copied, the value for μ_X should have $ in front of the row number so that the value doesn't change. The values in the fourth column are added up to obtain the variance. The standard deviation of *X* can be found using the SQRT function.



D2 =B2*(A2-C$11)^2

	A	B	C	D	E
1	first digit	probabilty	mean	variance	
2	1	0.301	0.301	1.793503	
3	2	0.176	0.352	0.365461	
4	3	0.125	0.375	0.02431	
5	4	0.097	0.388	0.030311	
6	5	0.079	0.395	0.192008	
7	6	0.067	0.402	0.438748	
8	7	0.058	0.406	0.734656	
9	8	0.051	0.408	1.060009	
10	9	0.046	0.414	1.421514	
11			**3.441**	**6.060519**	
12					

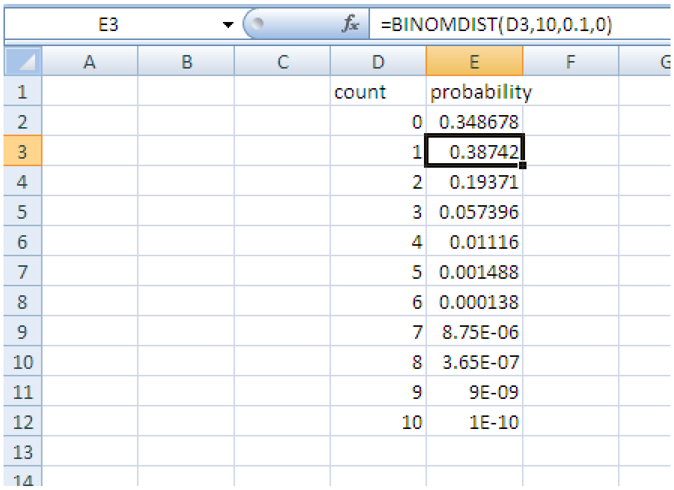
Binomial Probabilities

Suppose a music inspector inspects a sample of ten CDs from a shipment of 10,000 music CDs. Suppose that 10% of the CDs in the shipment are bad. The inspector will count the number *X* of bad CDs. Earlier in this chapter, we learned to generate random numbers for this situation by selecting **Data ➤ Data Analysis ➤ Random Number Generation** from the menu and selecting the Bernoulli distribution to generate a sequence of ten 1's and 0's to represent the bad and good CDs.

If we are interested only in the number of bad CDs, we can generate the number *X* by selecting **Data ➤ Data Analysis ➤ Random Number Generation** from the menu and selecting the Binomial distribution. Instead of generating only one value for *X*, we can simulate a large number of repetitions of the sample.

In addition to simulating binomial data, we can use Excel to calculate exact binomial probabilities. The **BINOMDIST(number_s,trials,probability_s,cumulative)** function calculated the probability of *h*. Specify the number of successes, number of trials, probability of success, and 0 for the probability. For example, BINOMDIST(1,10,0.1,0) gives the value 0.38742. This is the probability that exactly one CD out of 10 is bad.

If you want the entire probability distribution, enter the numbers 0 through 10 in a column on the Excel worksheet and use the same formula to obtain the probability of each possible outcome for a binomial distribution with $n = 10$ and $p = 0.1$ as shown below.

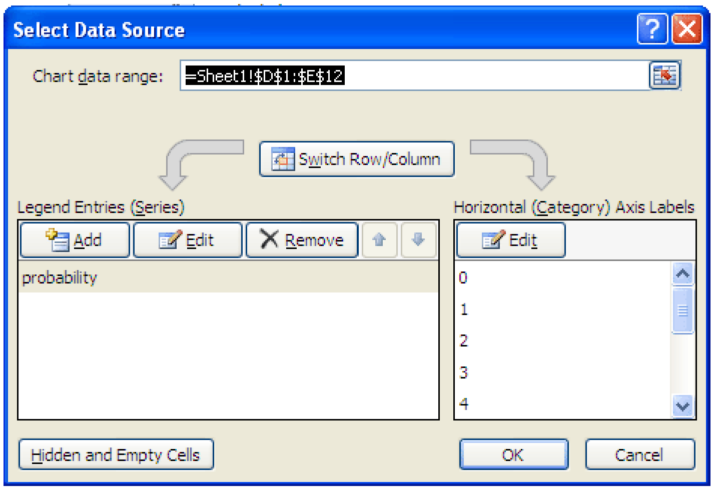


E3 =BINOMDIST(D3,10,0.1,0)

	A	B	C	D	E	F
1				count	probability	
2				0	0.348678	
3				1	0.38742	
4				2	0.19371	
5				3	0.057396	
6				4	0.01116	
7				5	0.001488	
8				6	0.000138	
9				7	8.75E-06	
10				8	3.65E-07	
11				9	9E-09	
12				10	1E-10	
13						
14						

If you enter a 1 instead of a 0 in the last position of the BINOMDIST function, Excel calculates $P(X \leq x)$ instead of $P(X = x)$. If you wish to calculate $P(X \geq x)$, then it is necessary to realize that $P(X \geq x) = 1 - P(X \leq x{-}1)$.

To graph the distribution, highlight the column of probabilities and select **Insert ➤ Column ➤ 2-D Column** from the menu. The default axis labels will be incorrect, so you need to click on the graph and then select Design ➤ **Select Data** from the menu and then edit the horizontal (category) axis labels on the right side of the dialog box.



Select Data Source

Chart data range: =Sheet1!$D$1:$E$12

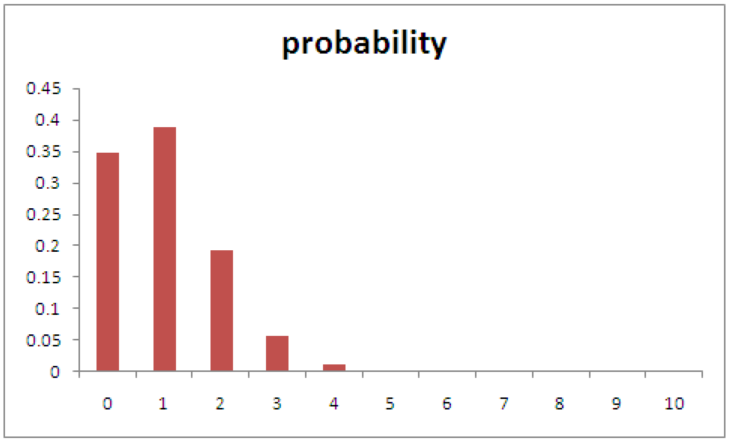
Legend Entries (Series)

probability

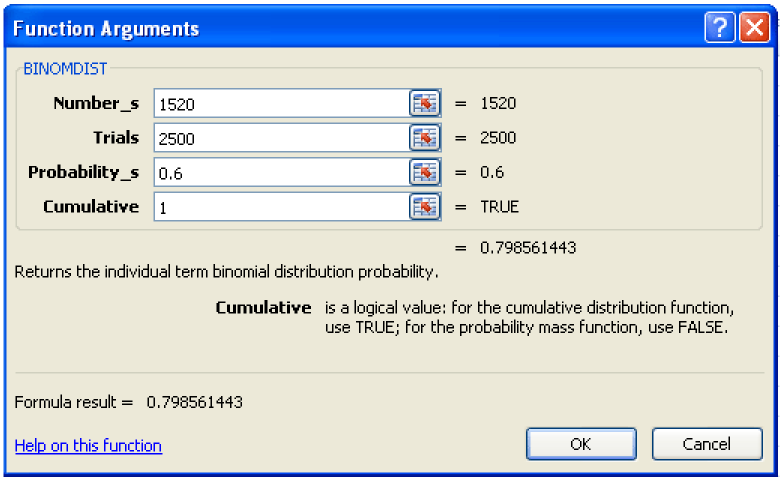
Horizontal (Category) Axis Labels

0
1
2
3
4

The resulting graph would look like the one below.



Suppose that an opinion poll asks 2500 adults whether they agree or disagree that "I like buying new clothes, but shopping is often frustrating and time-consuming." Suppose also that 60% of all adult U.S. residents would say "Agree." To find $P(X \leq 1520)$, the probability that at least 1520 adults agree, use the BINOMDIST function with Number_s set to 1520, 2500 trials, 0.6 probability of success, and Cumulative set to 1.



BINOMDIST

Number_s	1520	= 1520
Trials	2500	= 2500
Probability_s	0.6	= 0.6
Cumulative	1	= TRUE

= 0.798561443

Returns the individual term binomial distribution probability.

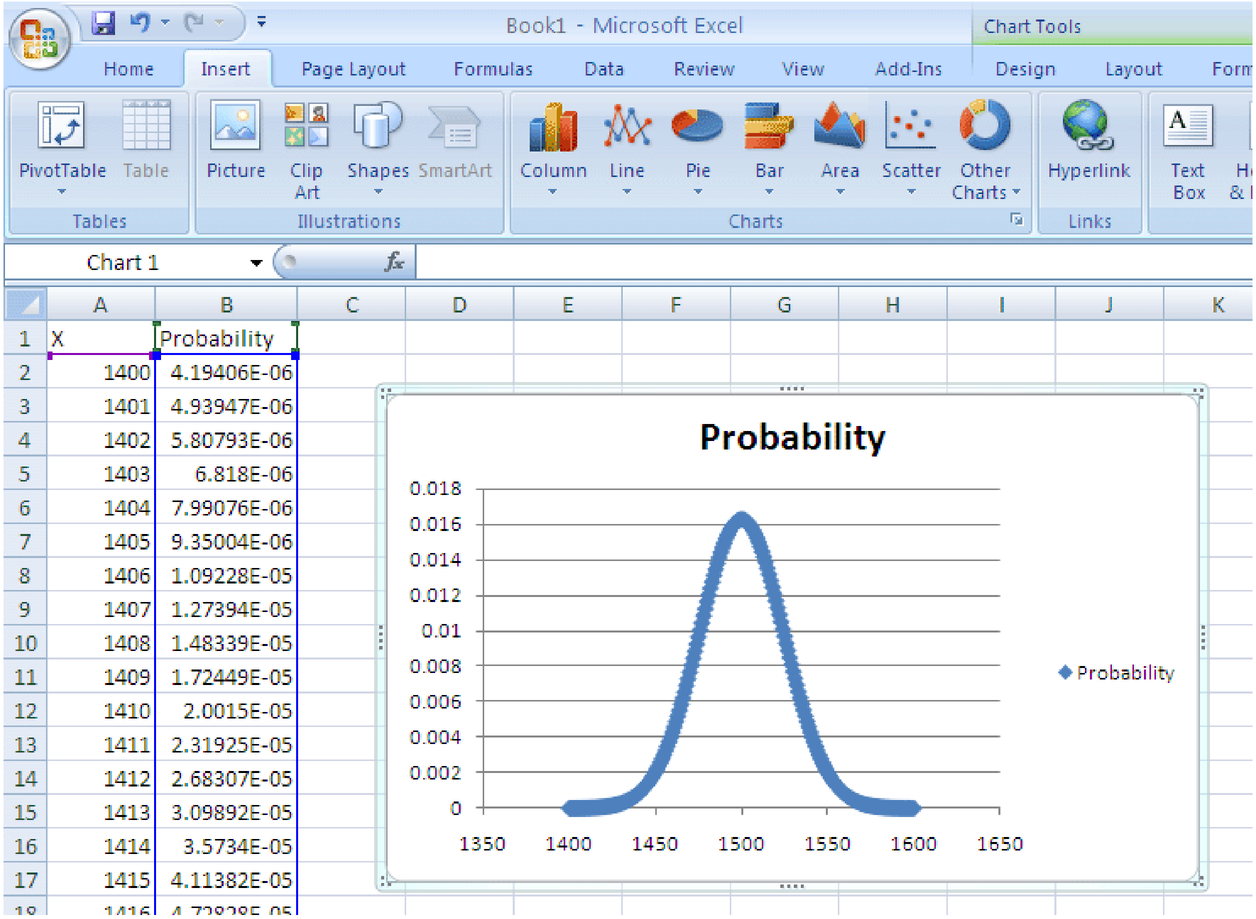
Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.798561443

As shown, the result is 0.7986

Normal Approximation to the Binomial

To illustrate the shape of the distribution on the number of adults that would say "Agree" out of the 2500 adults polled, enter the numbers from 1400 to 1600 in steps of 1 to be stored in a column of your choice. Remember that you can enter the first few numbers and then select these cells and drag the fill handle . down the cells that you want to fill. Next use the BINOMDIST function to compute the probability for 2500 trials with 0.6 probability of success for each value. Finally, highlight the two columns and select **Insert ➤ Scatter** from the menu to illustrate the shape. The numbers 1400 to 160 are the X values and the probabilities are the Y values as shown below.



	A	B
1	X	Probability
2	1400	4.19406E-06
3	1401	4.93947E-06
4	1402	5.80793E-06
5	1403	6.818E-06
6	1404	7.99076E-06
7	1405	9.35004E-06
8	1406	1.09228E-05
9	1407	1.27394E-05
10	1408	1.48339E-05
11	1409	1.72449E-05
12	1410	2.0015E-05
13	1411	2.31925E-05
14	1412	2.68307E-05
15	1413	3.09892E-05
16	1414	3.5734E-05
17	1415	4.11382E-05
18	1416	4.72828E-05

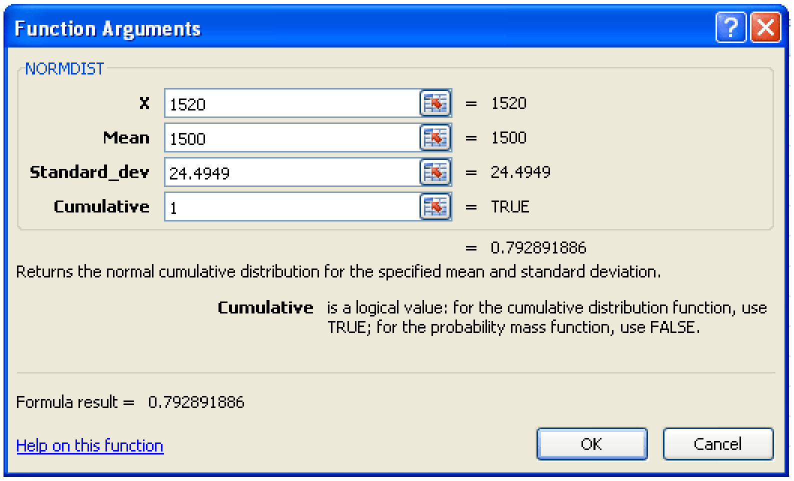
As the figure shows, the binomial probabilities will be approximated well by a normal distribution. The values for the mean and standard deviation are equal to

$$\mu = np = 2500 \times 0.6 = 1500$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{2500 \times 0.6 \times 0.4} = 24.4949.$$

The values are easily calculated using Excel. Enter "=2500*.6" into a cell for the mean or "=sqrt(250*.6*.4)" for the standard deviation.

When $np \geq 10$ and $n(1-p) \geq 10$, we can use the normal approximation to approximate binomial probabilities. Here, we approximate the probability that at least 1520 of the people in the sample find shopping frustrating when n = 2500 and p = 0.6. We act as though the count X has the N(1500, 24.4949) distribution. To obtain the normal approximation for this example, use NORMDIST function in Excel. We let X be 1520, specify a mean equal to 1500, a standard deviation equal to 24.4949, and Cumulative equal to 1. As with the binomial distribution, the cumulative probability is $P(X \leq x)$. To calculate $P(X > x)$, we must subtract the result from 1. As we see from the following, the normal approximation gives $P(X \leq 1520) = 0.7929$, approximately the same as the exact results we obtained previously.



NORMDIST

X 1520 = 1520

Mean 1500 = 1500

Standard_dev 24.4949 = 24.4949

Cumulative 1 = TRUE

= 0.792891886

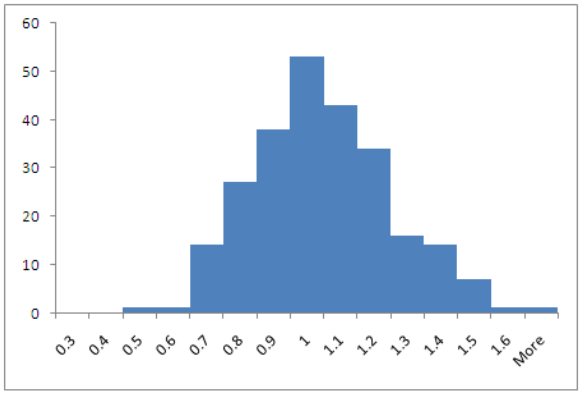
Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.792891886

CHAPTER

5

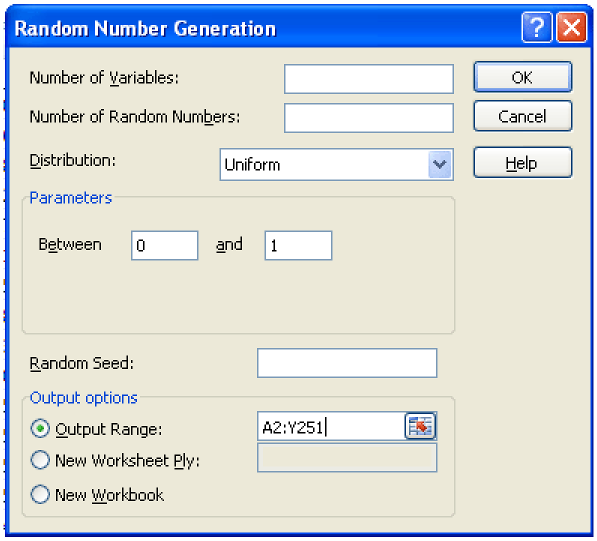


Sampling Distributions

The Central Limit Theorem

We can use Excel to illustrate the central limit theorem. The time a technician takes to service an air conditioning unit is exponentially distributed with mean μ = 1 hour and standard deviation σ = 1 hour. This distribution is strongly right skewed.

To generate 250 rows in 25 columns, select **Data ➤ Data Analysis ➤ Random Number Generation** from the menu. Specify an output location with 250 rows and 25 columns. Since the exponential distribution is not available in Excel, we will generate data from a Uniform distribution and then transform it into data that is exponentially distributed.


Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Uniform

Parameters

Between 0 and 1

Random Seed:

Output options

● Output Range: A2:Y251

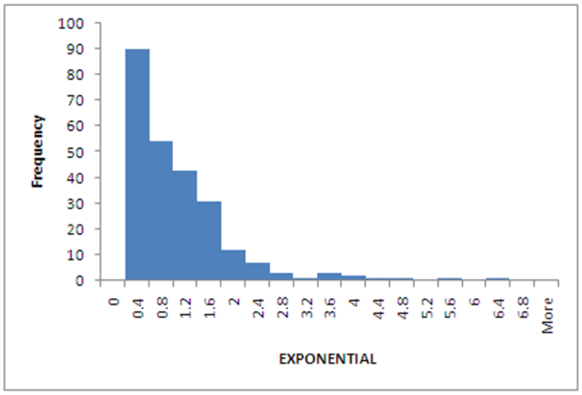
○ New Worksheet Ply:

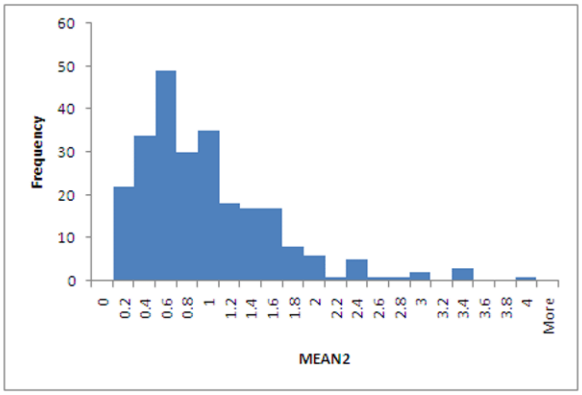
○ New Workbook

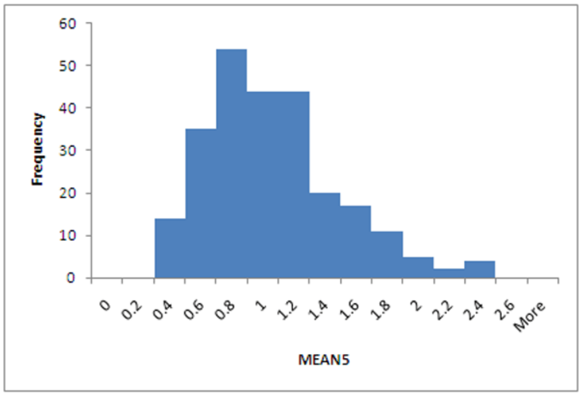
To transform the random numbers to values from an exponential distributions, we use the simple formula Y = -ln(U) where U is a uniformly distributed random variable on (0,1) and Y is an exponential random variable with mean equal to 1. For example if you type "=-ln(A2)" into cell Z2:AX251, you will create an observation that is exponentially distributed with with mean $\mu = 1$ hour and standard deviation $\sigma = 1$ hour. The formula can easily be copied to create 250 rows and 25 columns from this distribution.

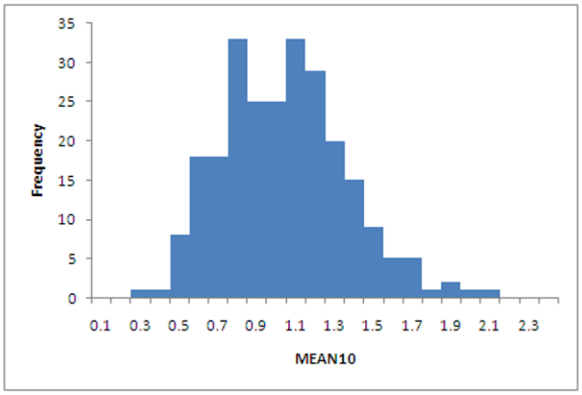
To illustrate the Central Limit Theorem, we create a column with the mean of 2 observations, the mean of 5 observations, the mean of 10 observations, and the mean of 25 observations. These are easily created using the **AVERAGE** function.

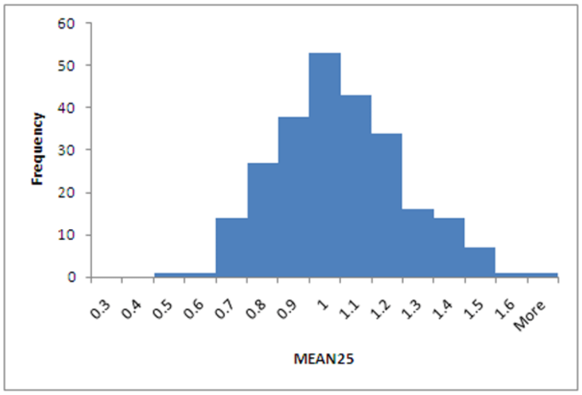
Select **Data ➤ Data Analysis ➤ Histogram** from the menu to produce a histogram of the original data or the mean of 2, 5, 10, or 25 observations. The histograms illustrate the right skewness of the original data and then sample means from 2, 5, 10, and 25 observations. As *n* increases, the shape of the distribution becomes more Normal. The mean stays at $\mu = 1$ and the standard deviation decreases.



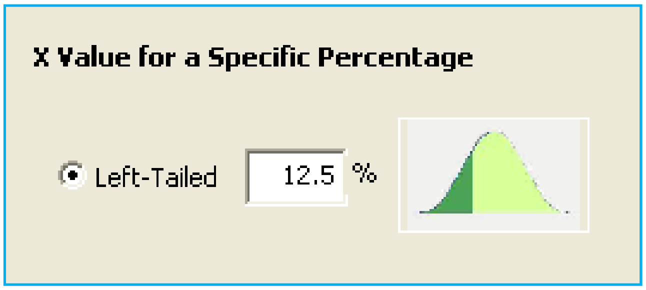



Frequency
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
More
MEAN5


Frequency
35
30
25
20
15
10
5
0
0.1
0.3
0.5
0.7
0.9
1.1
1.3
1.5
1.7
1.9
2.1
2.3
MEAN10


Frequency
60
50
40
30
20
10
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
More
MEAN25

CHAPTER 6



X Value for a Specific Percentage

● Left-Tailed 12.5 %

Introduction to Inference

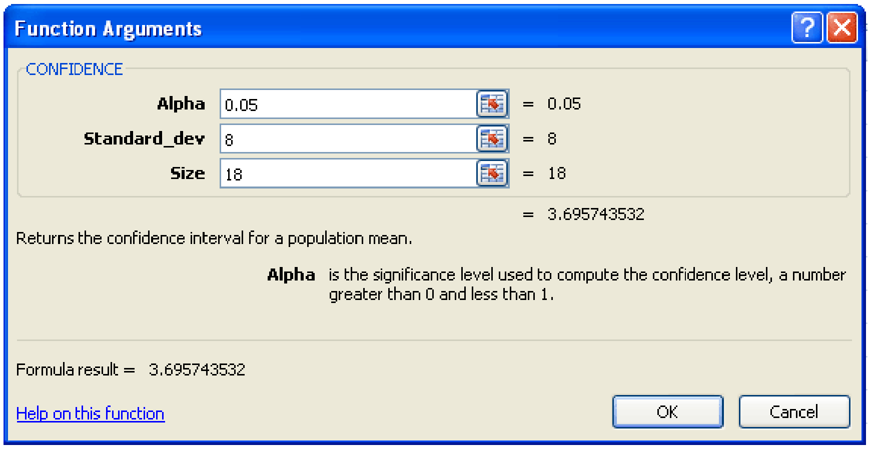
One-Sample *Z* Confidence Interval

Confidence intervals for the population mean μ, with σ known, can be calculated in Excel using the **AVERAGE** and **CONFIDENCE** functions. This interval goes from $\bar{x} - z^*(\sigma/\sqrt{n})$ to $\bar{x} + z^*(\sigma/\sqrt{n})$ where $\bar{x}$ is the mean of the data, n is the sample size, and z^* is the critical value from the normal table corresponding to the confidence level. $\bar{x}$ is calculated using the **AVERAGE** function and the margin of error, $z^*(\sigma/\sqrt{n})$, is calculated using the **CONFIDENCE** function.

To illustrate the confidence interval we consider biologists studying the healing of skin wounds. They measured the rate at which new cells closed a razor cut made in the skin of an anesthetized newt. Here are data from 18 newts, measured in micrometers (millionths of a meter) per hour:

29	27	34	40	22	28	14	35	26
35	12	30	23	18	11	22	23	33

We want a 95% confidence interval for the mean rate μ for all newts of this species. We enter the data into column A of an Excel spreadsheet and then calculate $\bar{x}$, the mean of these values using the **AVERAGE** function. The function arguments in the **CONFIDENCE** function are Alpha, Standard_dev, and Size. Alpha is 1-*C*, where *C* is the confidence level, Standard_dev is the known population standard deviation, and Size refers to the size of the sample as shown below.



CONFIDENCE

Alpha 0.05 = 0.05

Standard_dev 8 = 8

Size 18 = 18

= 3.695743532

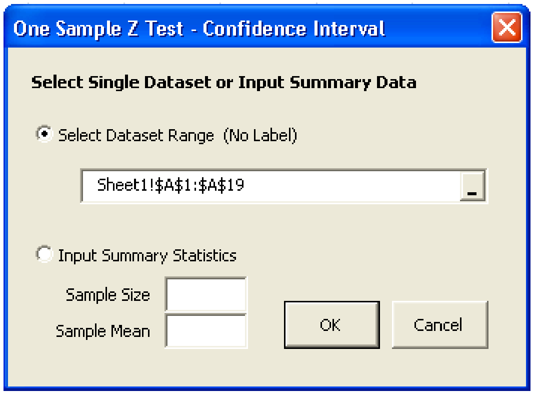
Returns the confidence interval for a population mean.

Alpha is the significance level used to compute the confidence level, a number greater than 0 and less than 1.

Formula result = 3.695743532

Excel calculates the mean to be 25.677 and the margin of error to be 3.696. Therefore the 95% confidence interval is (21.97, 29.36).

Alternatively, you can select **Add-Ins ➤ WHFStat ➤ Estimating and Testing Means ➤ 1 Sample z-Test** from the Excel menu. In the dialog box, specify the location of the data or the summary statistics. Click OK and another dialog box will appear.



Select Single Dataset or Input Summary Data

● Select Dataset Range (No Label)

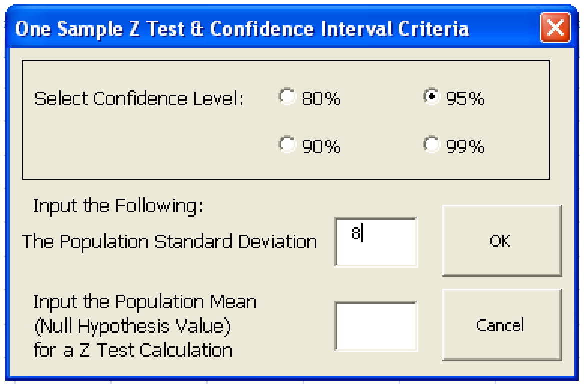
Sheet1!$A$1:$A$19

○ Input Summary Statistics

Sample Size

Sample Mean

In the following dialog box, specify the confidence level and population standard deviation and click OK.



One Sample Z Test & Confidence Interval Criteria

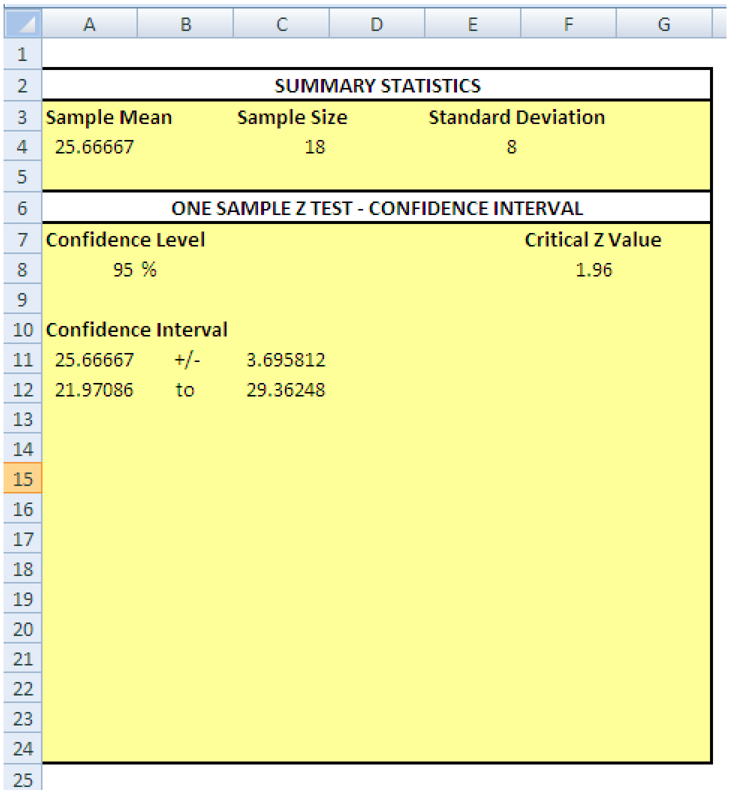
Select Confidence Level: ○ 80% ● 95% ○ 90% ○ 99%

Input the Following:

The Population Standard Deviation: 8

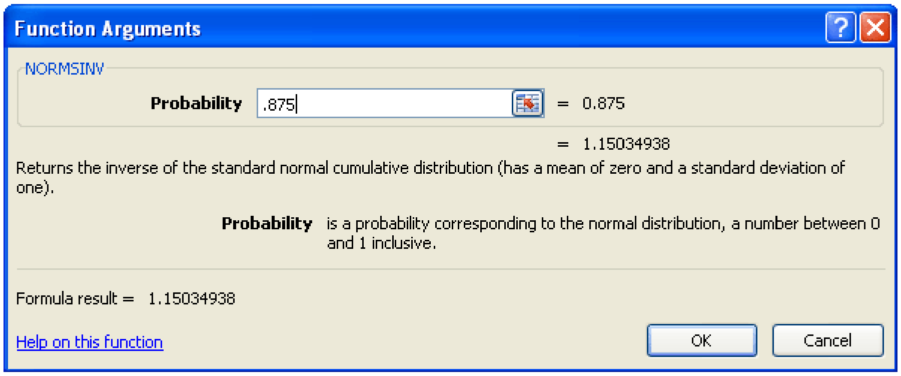
Input the Population Mean (Null Hypothesis Value) for a Z Test Calculation:

The results provide the sample mean, sample size, population standard deviation, and the margin of error along with the upper and lower confidence limits.



	A	B	C	D	E	F	G
1							
2	SUMMARY STATISTICS						
3	Sample Mean		Sample Size		Standard Deviation		
4	25.66667		18		8		
5							
6	ONE SAMPLE Z TEST - CONFIDENCE INTERVAL						
7	Confidence Level					Critical Z Value	
8	95	%				1.96	
9							
10	Confidence Interval						
11	25.66667	+/-	3.695812				
12	21.97086	to	29.36248				
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							

Excel can be used to find the critical value that is used for a specific level of confidence using the **NORSMINV** function. The **NORMSINV** function finds the value of *z* that has a specific area below it. For a level *C* confidence interval, we want to have an area of $(1-C)/2$ above and below the critical value. If we want the critical value for a 75% level of confidence, a value not included in Table D, we let $C = 0.75$. Therefore, $1-\{(1-C)/2\} = 0.875$ and the critical value z^* is 1.150.



Function Arguments

NORMSINV

Probability .875 = 0.875

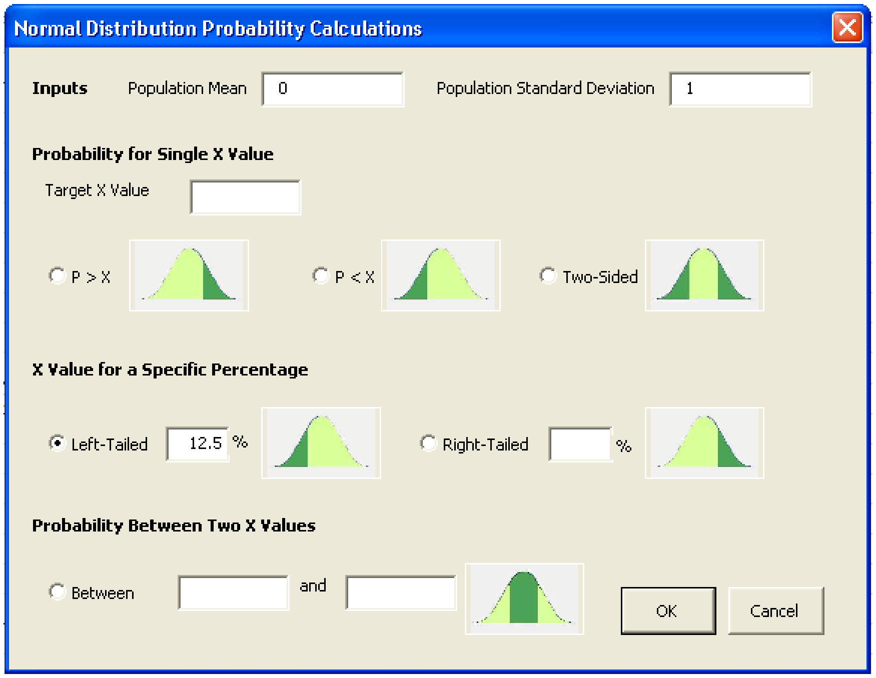
= 1.15034938

Returns the inverse of the standard normal cumulative distribution (has a mean of zero and a standard deviation of one).

Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.

Formula result = 1.15034938

Alternatively, you can select **Add-Ins ➤ WHFStat ➤ Probability Calculations-Normal Distribution** from the Excel menu. In the dialog box, specify a standard normal, i.e. the population mean is 0 and the population standard deviation is 1, and specify the area for the right or left tail. For a 75% confidence interval, each tail probability is 12.5%. Either can be specified, although if you specify the left tail, the critical value will be negative as shown on the following page.



Normal Distribution Probability Calculations

Inputs Population Mean 0 Population Standard Deviation 1

Probability for Single X Value

Target X Value

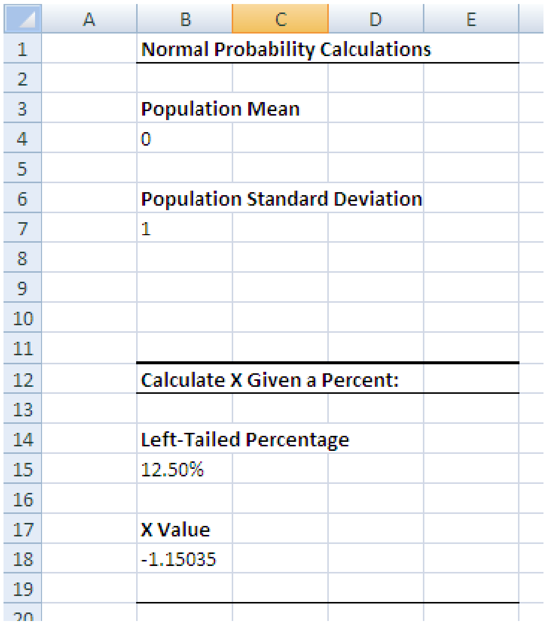
○ P > X ○ P < X ○ Two-Sided

X Value for a Specific Percentage

● Left-Tailed 12.5 % ○ Right-Tailed %

Probability Between Two X Values

○ Between and



	A	B	C	D	E
1		Normal Probability Calculations			
2					
3		Population Mean			
4		0			
5					
6		Population Standard Deviation			
7		1			
8					
9					
10					
11					
12		Calculate X Given a Percent:			
13					
14		Left-Tailed Percentage			
15		12.50%			
16					
17		X Value			
18		-1.15035			
19					

One-Sample *Z* Test

As with confidence intervals, we can use Excel to do a hypothesis test for a population mean μ, with σ known.

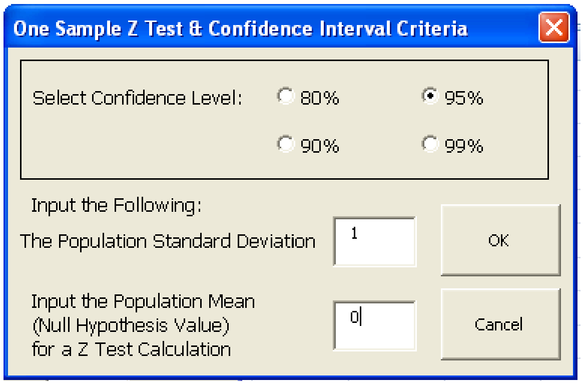
In this example, we consider sweetness loss scores for cola. Suppose we know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation σ = 1. The mean μ for all tasters measures loss of sweetness, and is different for different colas. Here are the sweetness losses for a new cola, as measured by 10 trained tasters:

$$2.0 \quad 0.4 \quad 0.7 \quad 2.0 \quad -0.4 \quad 2.2 \quad -1.3 \quad 1.2 \quad 1.1 \quad 2.3$$

We want to determine if there is significant evidence of sweetness loss. This calls for a test of the hypothesis that $\mu = 0$ against the alternative $\mu > 0$. We know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation σ = 1, so we enter that value into the dialog box. To test

$$H_0\colon \mu = 0$$
$$H_a\colon \mu > 0,$$

we select **Add-Ins ➤ WHFStat ➤ Estimating and Testing Means ➤ 1 Sample z-Test** from the Excel menu. In the dialog box, specify the location of the data or the summary statistics. Click OK and another dialog box will appear. In this dialog box, you must select a confidence level, specify σ, the population standard deviation, and the null hypothesis test value and click on OK.



One Sample Z Test & Confidence Interval Criteria

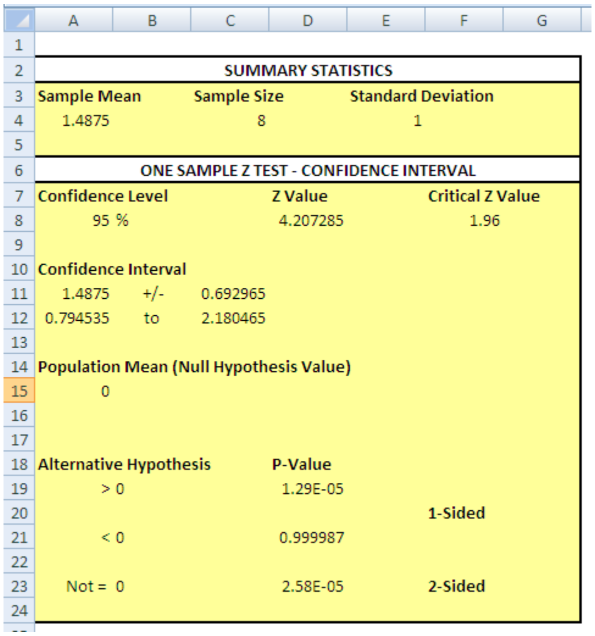
Select Confidence Level: ○ 80% ● 95% ○ 90% ○ 99%

Input the Following:

The Population Standard Deviation: 1

Input the Population Mean (Null Hypothesis Value) for a Z Test Calculation: 0

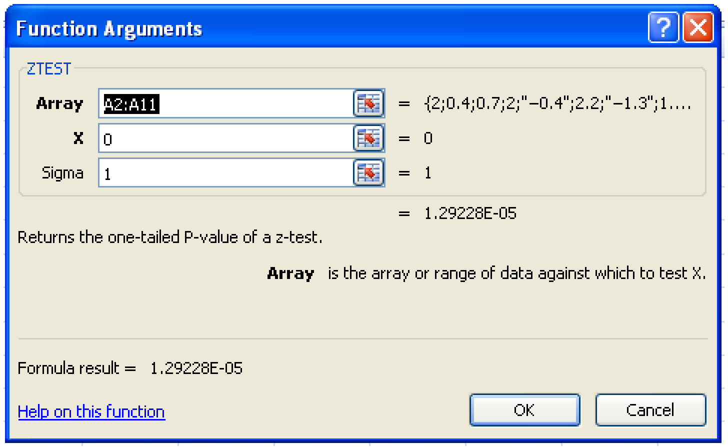
Excel provides the *P*-value for the ″less than″ (lower-tailed), “not equal” (two-tailed), and “greater than” (upper-tailed) hypotheses, in addition to the confidence interval.



	A	B	C	D	E	F	G
1							
2	SUMMARY STATISTICS						
3	Sample Mean		Sample Size		Standard Deviation		
4	1.4875		8		1		
5							
6	ONE SAMPLE Z TEST - CONFIDENCE INTERVAL						
7	Confidence Level			Z Value		Critical Z Value	
8	95 %			4.207285		1.96	
9							
10	Confidence Interval						
11	1.4875	+/-	0.692965				
12	0.794535	to	2.180465				
13							
14	Population Mean (Null Hypothesis Value)						
15	0						
16							
17							
18	Alternative Hypothesis			P-Value			
19	> 0			1.29E-05			
20						1-Sided	
21	< 0			0.999987			
22							
23	Not = 0			2.58E-05		2-Sided	
24							

The *P*-value for the “greater than” hypothesis is 1.29×10^{-5}. This is a small value, so the null hypothesis should be rejected.

The same upper-tail probability can be obtained for the alternative hypothesis, H_a: $\mu > 0$, using Excel’s **ZTEST** function. The function arguments are the Array or data, the *X* value, which is the value from the null hypothesis, and Sigma, the population standard deviation. If the lower-tail or two-tail *P*-value is needed, you must subtract the calculated value and/or double the value.



Function Arguments

ZTEST

Array A2:A11 = {2;0.4;0.7;2;"−0.4";2.2;"−1.3";1....

X 0 = 0

Sigma 1 = 1

= 1.29228E-05

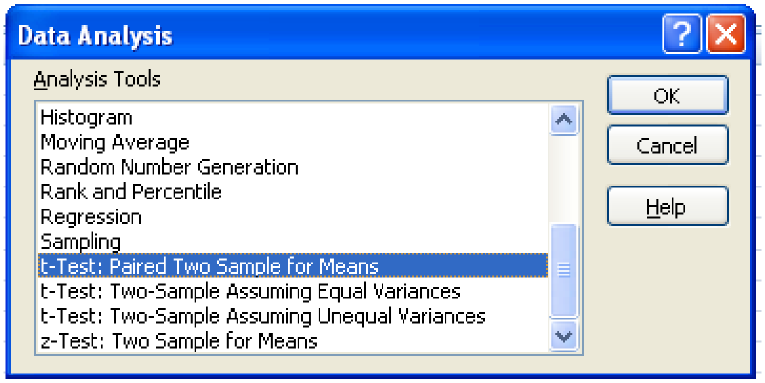
Returns the one-tailed P-value of a z-test.

Array is the array or range of data against which to test X.

Formula result = 1.29228E-05

CHAPTER

7


Data Analysis
?
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

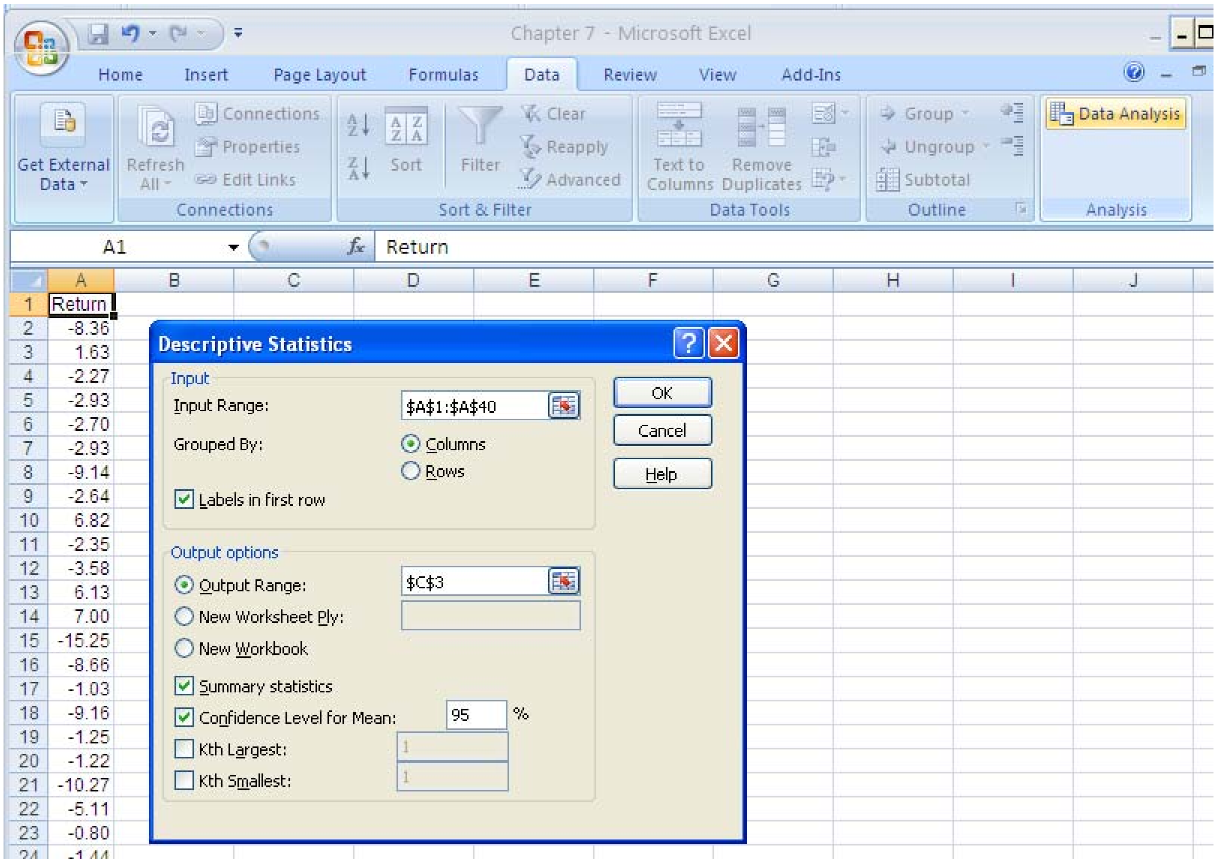
Inference for Distributions

One-Sample *t* Procedures

An investor was concerned about the poor performance of his portfolio. The data gives the rates of returns for 39 months that the account was managed by a particular broker. Consider the 39 monthly returns as a random sample from the population of monthly returns the broker would generate if he managed the account forever.

-8.36	1.63	-2.27	-2.93	-2.70	-2.93	-9.14	-2.64	6.82	-2.35	-3.58	6.13	7.00
-15.25	-8.66	-1.03	-9.16	-1.25	-1.22	-10.27	-5.11	-0.80	-1.44	1.28	-0.65	4.34
12.22	-7.21	-0.09	7.34	5.04	-7.24	-2.14	-1.01	-1.41	12.03	-2.56	4.33	2.35

To find a 95% confidence interval for the mean choose **Data ➤ Data Analysis ➤ Descriptive Statistics** from the menu. In the dialog box, supply the Input range for the data being analyzed. If there are labels in the first row, check the appropriate box. Indicate where you'd like the output and check the Summary Statistics and Confidence Level for Mean boxes along with the confidence level. Click OK to calculate the estimate and margin of error for a 95% confidence interval.



A1 f_x Return

	A
1	Return
2	-8.36
3	1.63
4	-2.27
5	-2.93
6	-2.70
7	-2.93
8	-9.14
9	-2.64
10	6.82
11	-2.35
12	-3.58
13	6.13
14	7.00
15	-15.25
16	-8.66
17	-1.03
18	-9.16
19	-1.25
20	-1.22
21	-10.27
22	-5.11
23	-0.80
24	-1.44

Descriptive Statistics

Input

Input Range: $A$1:$A$40

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $C$3

○ New Worksheet Ply:

○ New Workbook

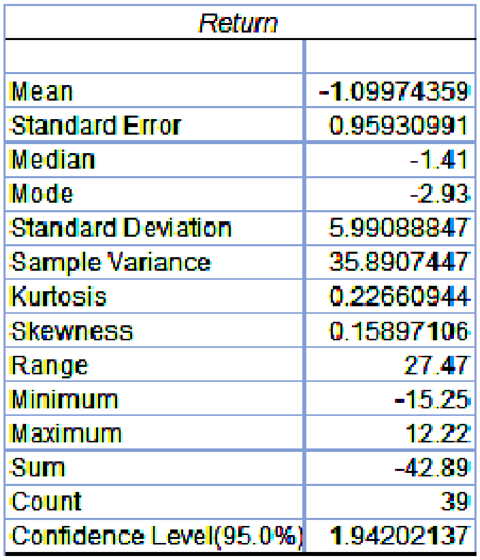
☑ Summary statistics

☑ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Excel gives the following output indicating that the mean is equal to -1.10 and the margin of er ror is equal to 1.94. This means that the 95% confidence interval for the mean rate of return is -1.010 ± 1.94, or (-3.04, -0.84).



Return	
Mean	-1.09974359
Standard Error	0.95930991
Median	-1.41
Mode	-2.93
Standard Deviation	5.99088847
Sample Variance	35.8907447
Kurtosis	0.22660944
Skewness	0.15897106
Range	27.47
Minimum	-15.25
Maximum	12.22
Sum	-42.89
Count	39
Confidence Level(95.0%)	1.94202137

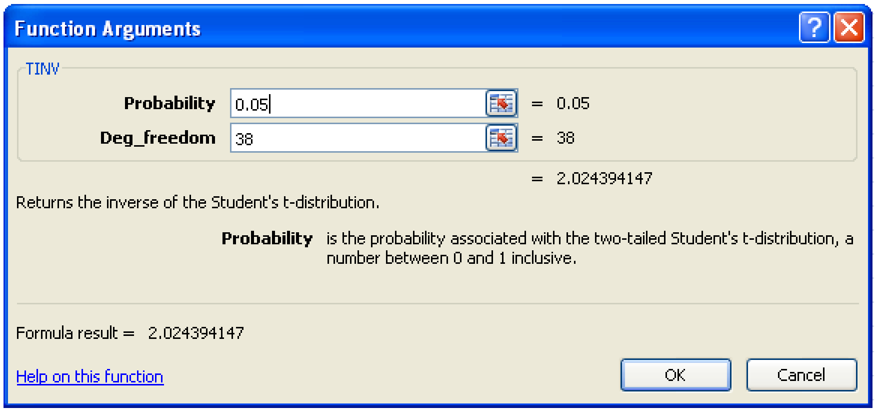
Excel calculated the confidence interval as

$$\overline{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \text{ to } \overline{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $\overline{x}$ is the mean of the data, s is the sample standard deviation, n is the sample size, and $t_{\alpha/2}$ is the critical value from a t-distribution with n–1 degrees of freedom. Since the sample size is equal to 39, $t_{\alpha/2}$ is the critical value from a t-distribution with 38 degrees of freedom.

You can also compute a confidence interval for the mean improvement by first computing the improvement for each subject and then selecting **Data ➤ Data Analysis ➤ Descriptive Statistics** from the menu**.**

The **TINV** function can be used to find the critical value you would use for a 95% confidence interval based on the t(38) distribution. The function arguments are the Probability, which is 0.05 for a 95% confidence interval, and the Degrees of Freedom. As shown below, the critical value from the t(38) distribution is equal to 2.024.



Probability 0.05 = 0.05
Deg_freedom 38 = 38
= 2.024394147

Returns the inverse of the Student's t-distribution.

Probability is the probability associated with the two-tailed Student's t-distribution, a number between 0 and 1 inclusive.

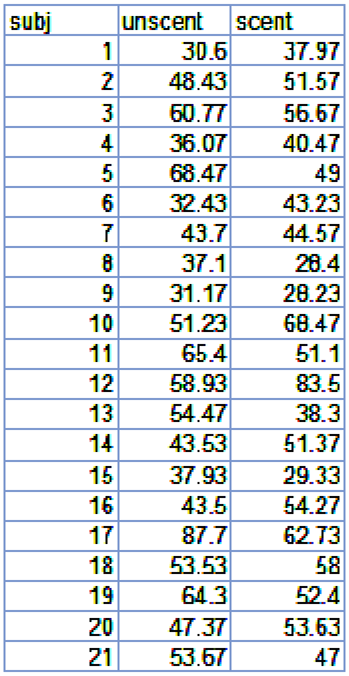
Formula result = 2.024394147

Matched Pairs

In a matched pairs study, subjects are matched in pairs and the outcomes are compared within each matched pair. In this example, subjects worked a paper-and-pencil maze while wearing masks. Each mask was either unscented or carried a floral scent. The response variable is their mean time on three trials. Each subject worked the maze with both types of mask. The data gives the subjects average times. To assess whether the floral scent significantly improved performance, we test

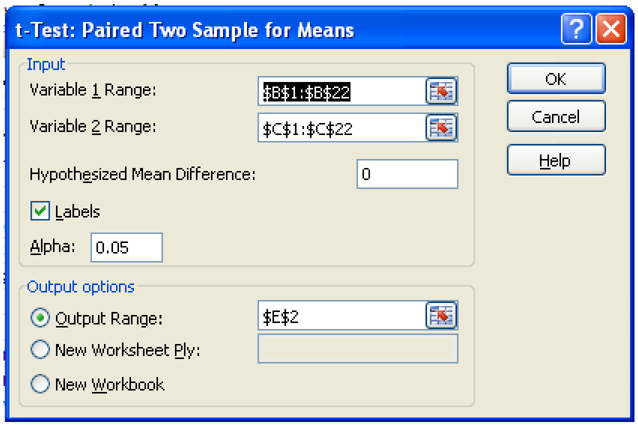
$$H_0: \mu = 0$$
$$H_a: \mu > 0,$$

where μ is the mean improvement if all subjects received similar instruction.



subj	unscent	scent
1	30.6	37.97
2	48.43	51.57
3	60.77	56.67
4	36.07	40.47
5	68.47	49
6	32.43	43.23
7	43.7	44.57
8	37.1	28.4
9	31.17	28.23
10	51.23	68.47
11	65.4	51.1
12	58.93	83.5
13	54.47	38.3
14	43.53	51.37
15	37.93	29.33
16	43.5	54.27
17	87.7	62.73
18	53.53	58
19	64.3	52.4
20	47.37	53.63
21	53.67	47

To do a *t*-test for matched pairs using Excel, select **Data ➤ Data Analysis ➤ t-Test: Paired Two Sample for Means** from the Excel menu. Since Paired *t* evaluates the first sample minus the second sample, we select the Unscented values for Variable 1 and the Scented for Variable 2. The Hypothesized Mean Difference is 0.



t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $B$1:$B$22

Variable 2 Range: $C$1:$C$22

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

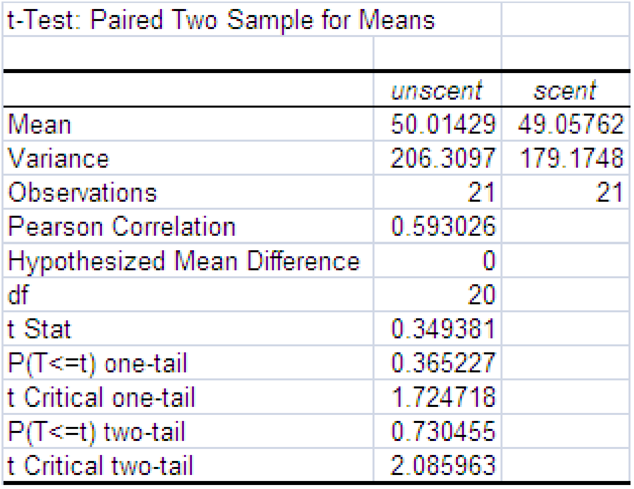
Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

The large value (0.349) given for the *P*-value shows that the data do not support the claim that floral scents improve performance.



t-Test: Paired Two Sample for Means

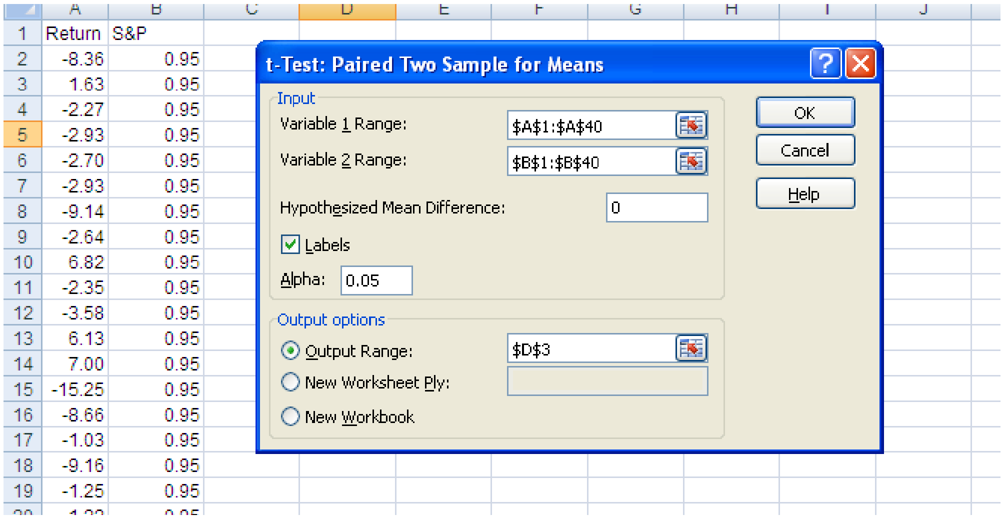
	unscent	scent
Mean	50.01429	49.05762
Variance	206.3097	179.1748
Observations	21	21
Pearson Correlation	0.593026	
Hypothesized Mean Difference	0	
df	20	
t Stat	0.349381	
P(T<=t) one-tail	0.365227	
t Critical one-tail	1.724718	
P(T<=t) two-tail	0.730455	
t Critical two-tail	2.085963	

Excel does not list a procedure for 1-Sample tests, but it is easy to use the t-Test: Paired Two Sample for Means for 1-Sample. Consider the investment example from earlier in this chapter. During the same 39 months, the S&P 500 had an average return of μ = 0.95%. To see if the investors returns are compatible with the S&P 500 average for those same months, we test

$$H_0: \mu = 0.95$$
$$H_a: \mu \neq 0.95.$$

To do the hypothesis test, enter the investment data in one column and the values from the null hypothesis in another column of the same length. As with paired comparisons, select **Data ➤ Data Analysis ➤ t-Test: Paired Two Sample for Means** from the Excel menu. Although not intended for this purpose, this works because the calculations are the same for 1-Sample and paired data.

Enter the actual data for Variable 1 and the constant values for Variable 2. Check the box for labels, if appropriate, specify the location of the output, and click OK.



	A	B
1	Return	S&P
2	-8.36	0.95
3	1.63	0.95
4	-2.27	0.95
5	-2.93	0.95
6	-2.70	0.95
7	-2.93	0.95
8	-9.14	0.95
9	-2.64	0.95
10	6.82	0.95
11	-2.35	0.95
12	-3.58	0.95
13	6.13	0.95
14	7.00	0.95
15	-15.25	0.95
16	-8.66	0.95
17	-1.03	0.95
18	-9.16	0.95
19	-1.25	0.95

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $A$1:$A$40

Variable 2 Range: $B$1:$B$40

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

● Output Range: $D$3

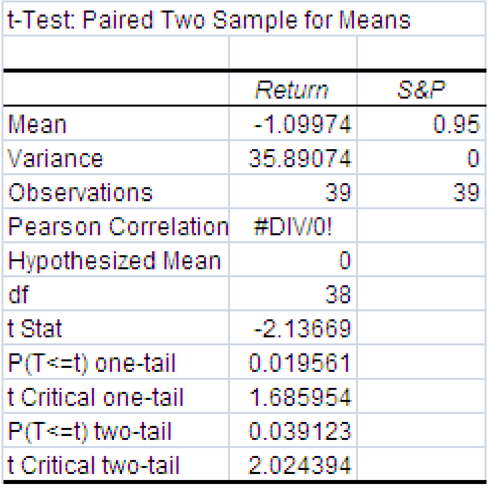
○ New Worksheet Ply:

○ New Workbook

Excel calculates the test statistic as

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ is the mean of the data, s is the sample standard deviation, n is the sample size, and μ_0 is the hypothesized population mean. As shown below, the test statistic t is calculated as -2.14. The *P*-value of this test is 0.39. There is strong evidence that the investor's returns were different from the S&P 500 average. We can safely reject H_0.



t-Test: Paired Two Sample for Means		
	Return	*S&P*
Mean	-1.09974	0.95
Variance	35.89074	0
Observations	39	39
Pearson Correlation	#DIV/0!	
Hypothesized Mean	0	
df	38	
t Stat	-2.13669	
P(T<=t) one-tail	0.019561	
t Critical one-tail	1.685954	
P(T<=t) two-tail	0.039123	
t Critical two-tail	2.024394	

Two-Sample *t* Procedures

To perform a hypothesis test and of the difference between two population means, select

Data ➤ Data Analysis ➤ t-Test: Two Sample Assuming Unequal Variances

from the Excel menu.

The following example fits the two-sample setting. A researcher buried polyester strips in the soil to see how quickly they decay. Five of the strips, chosen at random, were dug up after two weeks. Another five were dug up after 16 weeks. The breaking strength (in pounds) of all 10 strips was measured and entered into an Excel worksheet. The dialog box looks exactly the same as the one for t-Test: Paired Sample for Means, however the calculations are very different.

Excel calculates the Two-Sample *t*-test statistic as

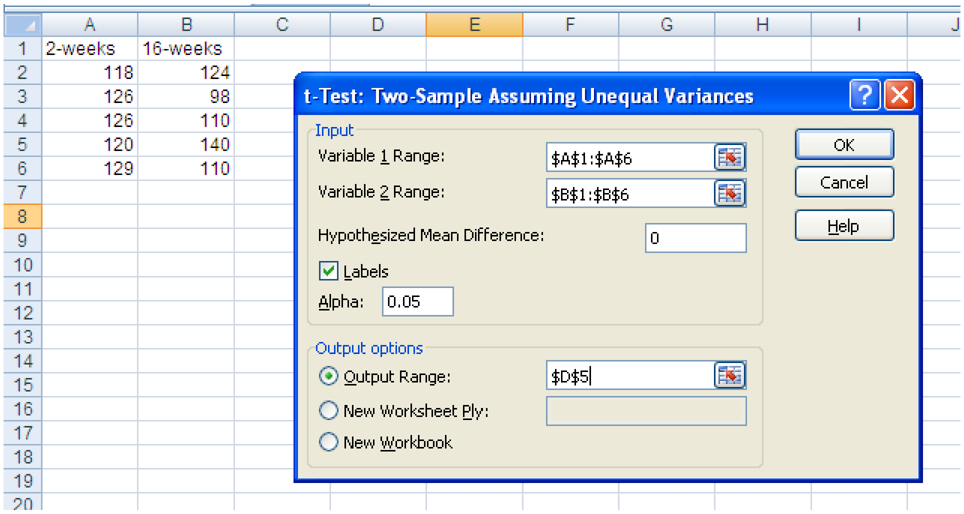
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This statistic has an approximate t distribution with degrees of freedom given by the Scatterthwaite approximation.

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Excel rounds the number to an integer, if necessary.

Enter the actual data for Variable 1 and the constant values for Variable 2 and the Hypothsized Mean Difference equal to 0. Check the box for labels, if appropriate, specify the location of the output, and click OK.



	A	B
1	2-weeks	16-weeks
2	118	124
3	126	98
4	126	110
5	120	140
6	129	110

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $A$1:$A$6

Variable 2 Range: $B$1:$B$6

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

● Output Range: $D$5

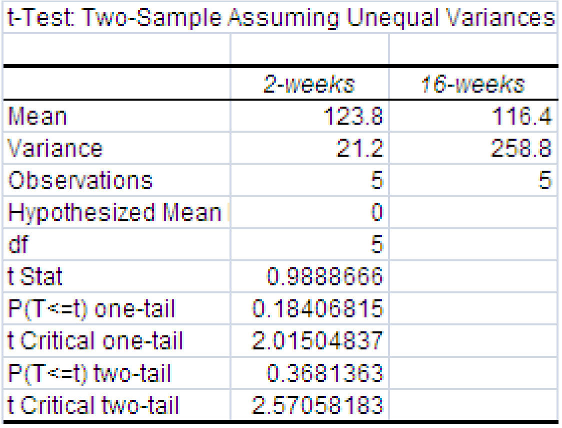
○ New Worksheet Ply:

○ New Workbook

We wish to test

$$H_0 : \mu_2 = \mu_{16}$$
$$H_a : \mu_2 > \mu_{16}$$

so we note that the mean for 2 weeks is larger than the mean for 16 weeks, and we use the 1-tail P-value.



t-Test: Two-Sample Assuming Unequal Variances

	2-weeks	16-weeks
Mean	123.8	116.4
Variance	21.2	258.8
Observations	5	5
Hypothesized Mean	0	
df	5	
t Stat	0.9888666	
P(T<=t) one-tail	0.18406815	
t Critical one-tail	2.01504837	
P(T<=t) two-tail	0.3681363	
t Critical two-tail	2.57058183	

The results show that the *P*-value was calculated to be 0.184. The experiment did not find convincing evidence that polyester decays more in 16 weeks than in 2 weeks. If a confidence interval is needed, the "not equal" alternative must be selected on the Options subdialog box.

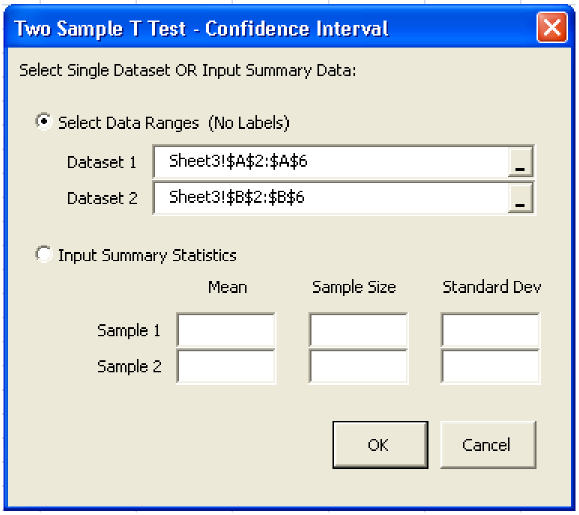
If a confidence interval is needed, we must calculate

$$\overline{X}_{.1} - \overline{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Fortunately, this is easily down with the previous Excel output. In our example, $\bar{X}_1 = 123.8$, , $\bar{X}_2 = 116.5$, , $s_1^2 = 21.2$, $s_2^2 = 258.8$, $n_1 = n_2 = 5$, so the confidence interval is equal to 7.4±2.777 or (-13.377, 28.177).

If we select *t*-Test: Two-Sample Assuming Equal Variances instead of Unequal Variances, Excel uses a pooled procedure, which assumes that the two populations have equal variances and "pools" the two sample variances to estimate the common population variance. The test statistic has a *t* distribution with exactly $n_1 + n_2 - 2$ degrees of freedom. The pooled procedure can be seriously in error if the variances are not equal. It is recommended that the Equal Variances procedure never be used.

Alternatively, we can select **Add-Ins ➤ WHFStat ➤ Estimating and Testing Means ➤ 2-Sample t-Test – Unequal Variances** from the Excel menu. In the dialog box, enter either the data range without labels or the summary statistics. Click on OK to find the 95% confidence interval for the difference and results from the hypothesis test described above.



Two Sample T Test - Confidence Interval

Select Single Dataset OR Input Summary Data:

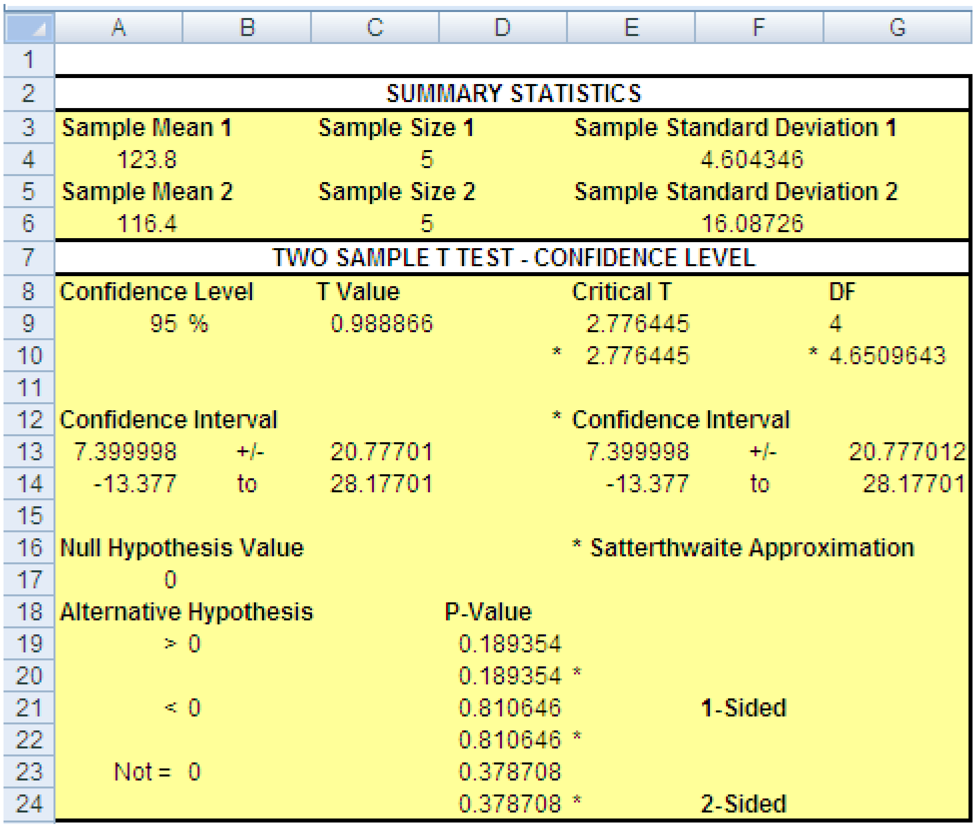
● Select Data Ranges (No Labels)

Dataset 1: Sheet3!$A$2:$A$6

Dataset 2: Sheet3!$B$2:$B$6

○ Input Summary Statistics

	Mean	Sample Size	Standard Dev
Sample 1			
Sample 2			



	A	B	C	D	E	F	G
1							
2	SUMMARY STATISTICS						
3	Sample Mean 1		Sample Size 1		Sample Standard Deviation 1		
4	123.8		5		4.604346		
5	Sample Mean 2		Sample Size 2		Sample Standard Deviation 2		
6	116.4		5		16.08726		
7	TWO SAMPLE T TEST - CONFIDENCE LEVEL						
8	Confidence Level		T Value		Critical T		DF
9	95 %		0.988866		2.776445		4
10				*	2.776445	*	4.6509643
11							
12	Confidence Interval			*	Confidence Interval		
13	7.399998	+/-	20.77701		7.399998	+/-	20.777012
14	-13.377	to	28.17701		-13.377	to	28.17701
15							
16	Null Hypothesis Value				* Satterthwaite Approximation		
17	0						
18	Alternative Hypothesis			P-Value			
19	> 0			0.189354			
20				0.189354	*		
21	< 0			0.810646		1-Sided	
22				0.810646	*		
23	Not = 0			0.378708			
24				0.378708	*	2-Sided	

The *F*-Test for Equality of Variance

Consider the experiment to compare the mean breaking strengths of polyester fabric after being buried for two weeks and for 16 weeks. We might also compare the standard deviations to see whether strength loss is more or less variable after 16 weeks. We want to test

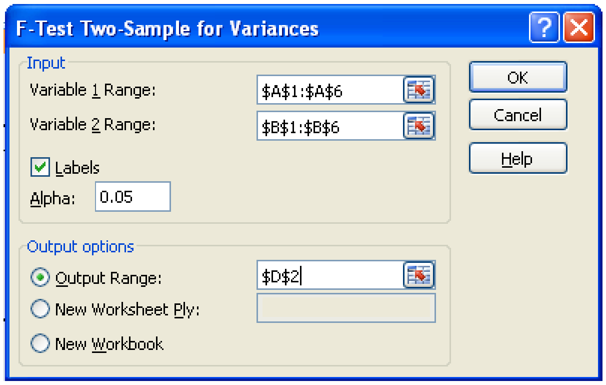
$$H_0 : \sigma_1 = \sigma_2$$
$$H_a : \sigma_1 \neq \sigma_2$$

The hypothesis of equal spread can be tested in Excel using an *F*-test by selecting

Data ➤ Data Analysis ➤ F-Test Two-Sample for Variances

from the menu. The *F*-test is not recommended for distributions that are not normal. Before we calculate the *F*-statistic, it is important to verify that the distributions are normal. This is done graphically by selecting **Data ➤ Data Analysis ➤ Histogram** from the menu.

To test the equality of two variances, select **Data ➤ Data Analysis ➤ F-Test Two-Sample for Variances** and fill out the dialog box. Check Samples in one column, Samples in different columns, or Summarized data depending on the format of your data.



F-Test Two-Sample for Variances

Input

Variable 1 Range: $A$1:$A$6

Variable 2 Range: $B$1:$B$6

☑ Labels

Alpha: 0.05

Output options

● Output Range: $D$2

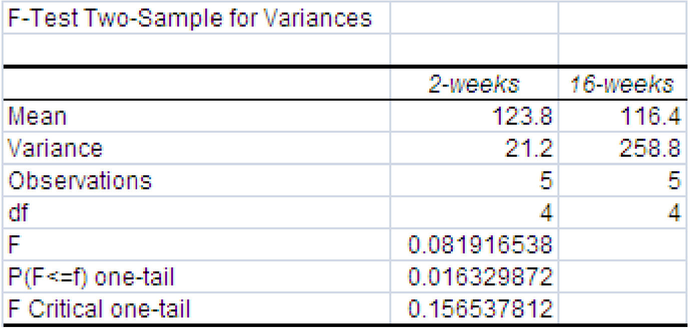
○ New Worksheet Ply:

○ New Workbook

The *F* statistic is the ratio of the sample variances,

$$F = \frac{s_1^2}{s_2^2}$$

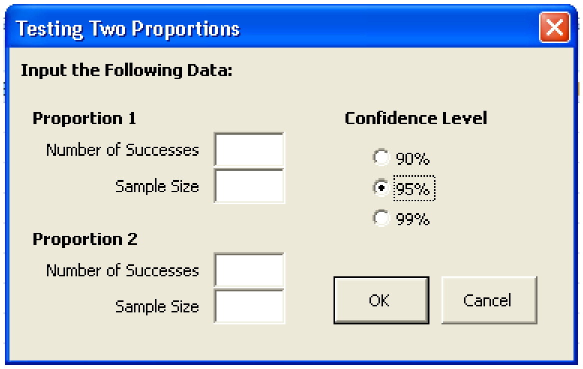
The one tail *P*-value is listed as .0163, but since we are doing a two-tail test, the correct *P*-value is equal to 0.033, so the difference between the spread on the two tests is statistically significant at the 5% level. The results from the test follow.



F-Test Two-Sample for Variances		
	2-weeks	*16-weeks*
Mean	123.8	116.4
Variance	21.2	258.8
Observations	5	5
df	4	4
F	0.081916538	
P(F<=f) one-tail	0.016329872	
F Critical one-tail	0.156537812	

CHAPTER

8


Testing Two Proportions

Input the Following Data:

Proportion 1

Number of Successes

Sample Size

Proportion 2

Number of Successes

Sample Size

Confidence Level

○ 90%

● 95%

○ 99%

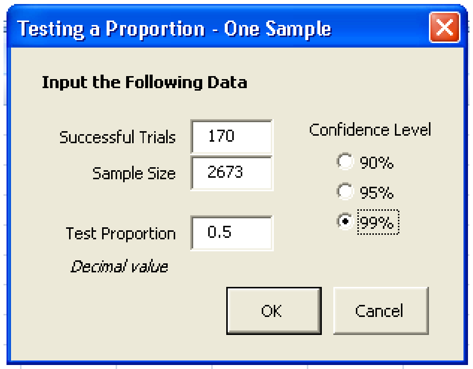
Inference for Proportions

Confidence Intervals for a Single Proportion

To compute a confidence interval and perform a hypothesis test of the proportion, select

Add-Ins ➤ WHFStat ➤ Proportion Testing ➤ One Sample

from the menu. In the dialog box, enter the number of successful trials, the sample size, the test proportion from the null hypothesis, and the desired level of confidence. A sample survey found that 170 of a sample of 2673 adult heterosexuals had multiple partners. The sample size is $n = 2673$ and the count of successes is $X = 170$. Fill in the data in the dialog box to make a 99% confidence interval for the proportion p of all adult heterosexuals with multiple partners. Although we are not now interested in a hypothesis test, it is required to select a test proportion.


Testing a Proportion - One Sample

Input the Following Data

Successful Trials 170

Sample Size 2673

Test Proportion 0.5

Decimal value

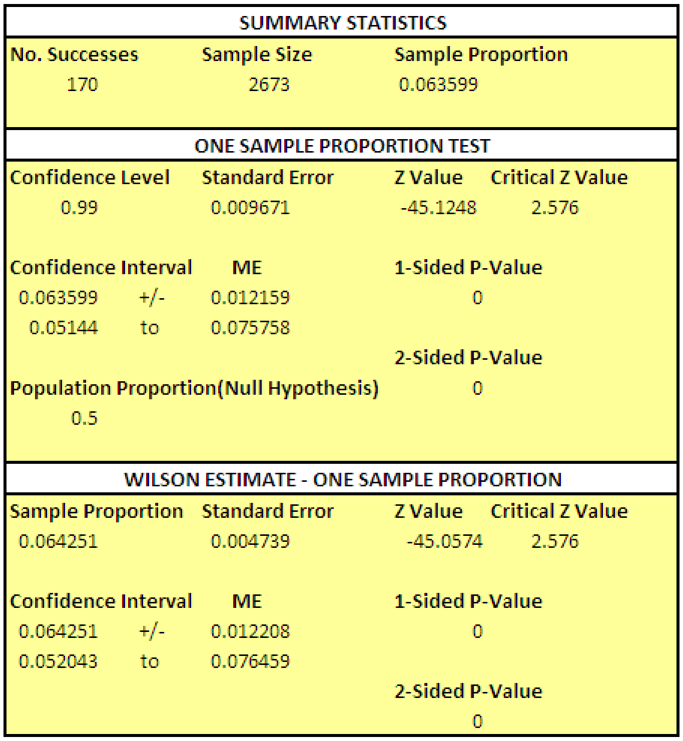
Confidence Level

○ 90%

○ 95%

● 99%

The output that follows shows that the 99% confidence interval for the proportion of adult heterosexuals with multiple partners is 0.0636 ± 0.0125 or (0.0514, 0.0758).



```
                         SUMMARY STATISTICS
No. Successes        Sample Size           Sample Proportion
     170                 2673               0.063599

                     ONE SAMPLE PROPORTION TEST
Confidence Level     Standard Error        Z Value    Critical Z Value
     0.99              0.009671             -45.1248       2.576

Confidence Interval     ME                 1-Sided P-Value
 0.063599     +/-      0.012159                    0
  0.05144     to       0.075758
                                           2-Sided P-Value
Population Proportion(Null Hypothesis)             0
      0.5

              WILSON ESTIMATE - ONE SAMPLE PROPORTION
Sample Proportion    Standard Error        Z Value    Critical Z Value
 0.064251              0.004739             -45.0574       2.576

Confidence Interval     ME                 1-Sided P-Value
 0.064251     +/-      0.012208                    0
 0.052043     to       0.076459
                                           2-Sided P-Value
                                                   0
```

A more accurate confidence interval for the population proportion p can be obtained by using the "plus four" method.

$$\tilde{p} = \frac{X+2}{n+4}$$

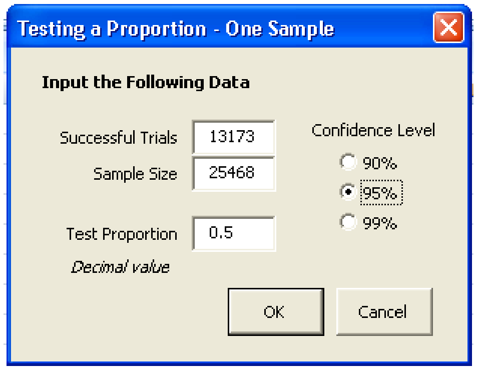
This is equivalent to adding two successes and two failures to the actual data. Calculations based on the "plus four" method are given in the Wilson Estimate section. The confidence interval using this method is 0.643 ± 0.0122 or (0.52, .076). The "plus four" method is always recommended and is particularly important to do if sample sizes are not large.

Significance Tests for a Proportion

Consider whether newborn babies are more likely to be boys than girls, presumably to compensate for higher mortality among boys in early life. A random sample found 13,173 boys among 25,468 first-born children. The sample proportion of boys is $\hat{p} = 13{,}173/25{,}468 = 0.5172$. Is this sample evidence that boys are more common than girls in the entire population? To answer this question we test

$$H_0: p = 0.5$$
$$H_a: p > 0.5$$

In the Testing a Proportion – One Sample dialog box, enter the data and the hypothesized proportion. Although we aren't interested in the confidence interval, it is also necessary to select a Confidence Level.



Testing a Proportion - One Sample

Input the Following Data

Successful Trials: 13173

Sample Size: 25468

Test Proportion: 0.5

Decimal value

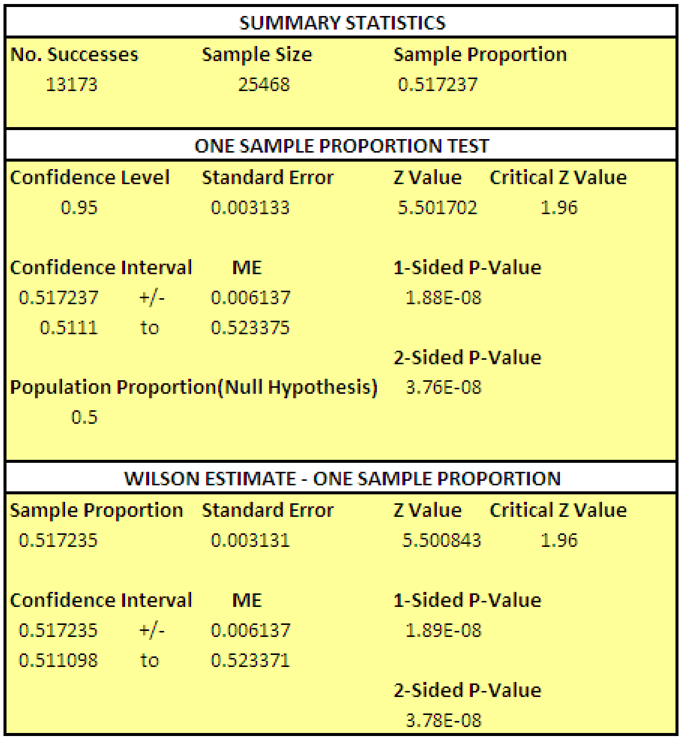
Confidence Level

○ 90%
● 95%
○ 99%

Excel calculates the test statistic as

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

where $\hat{p}$ is the observed probability equal to X/n, X is the observed number of successes in n trials, and p_0 is the hypothesized probability. The probabilities are obtained from the standard normal distribution. In other words, Excel uses tests and intervals based on the normal approximation.



```
                    SUMMARY STATISTICS
No. Successes      Sample Size        Sample Proportion
   13173              25468            0.517237

                 ONE SAMPLE PROPORTION TEST
Confidence Level   Standard Error     Z Value    Critical Z Value
     0.95             0.003133         5.501702       1.96

Confidence Interval     ME            1-Sided P-Value
 0.517237    +/-      0.006137          1.88E-08
   0.5111    to       0.523375
                                      2-Sided P-Value
Population Proportion(Null Hypothesis)  3.76E-08
      0.5

            WILSON ESTIMATE - ONE SAMPLE PROPORTION
Sample Proportion  Standard Error     Z Value    Critical Z Value
 0.517235             0.003131         5.500843       1.96

Confidence Interval     ME            1-Sided P-Value
 0.517235    +/-      0.006137          1.89E-08
 0.511098    to       0.523371
                                      2-Sided P-Value
                                        3.78E-08
```

Excel gives the one sided *P*-value as 1.88×10^{-8}. This means that there is very strong evidence that more than half of newborns are boys.

Confidence Interval for Comparing Proportions

To compute a confidence interval and perform a hypothesis test of the difference between two proportions, select

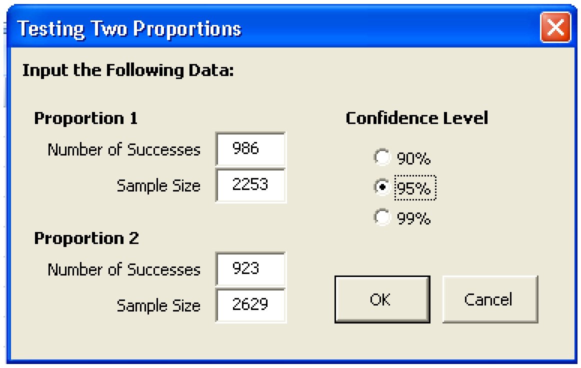
Add-Ins ➤ WHFStat ➤ Proportion Testing ➤ Two Samples

from the menu. Enter the summary data, and the confidence level.

A random sample by the National Institutes of Health concerns the number of young adults (ages 19 to 25) that still live in their parents home. Included in the sample are 2253 men and 2629 women in this age group. The survey found that 986 of the men and 923 of the women lived with their parents. Here is the data summary:

Population	n	X	$\hat{p} = X/n$
1 (men)	2253	986	0.4376
2 (women)	2629	923	0.3511

The difference $p_1 - p_2$ allows us to see how large the difference is between the proportions of young men and young women who live with their parents. To compute a 95% confidence interval for $p_1 - p_2$, select **Add-Ins ➤ WHFStat ➤ ProportionTesting ➤ Two Samples** from the Excel menu. Enter the summarized data in the dialog box, select the 95% Confidence Level, and click OK.



Testing Two Proportions

Input the Following Data:

Proportion 1

Number of Successes 986

Sample Size 2253

Proportion 2

Number of Successes 923

Sample Size 2629

Confidence Level

○ 90%

● 95%

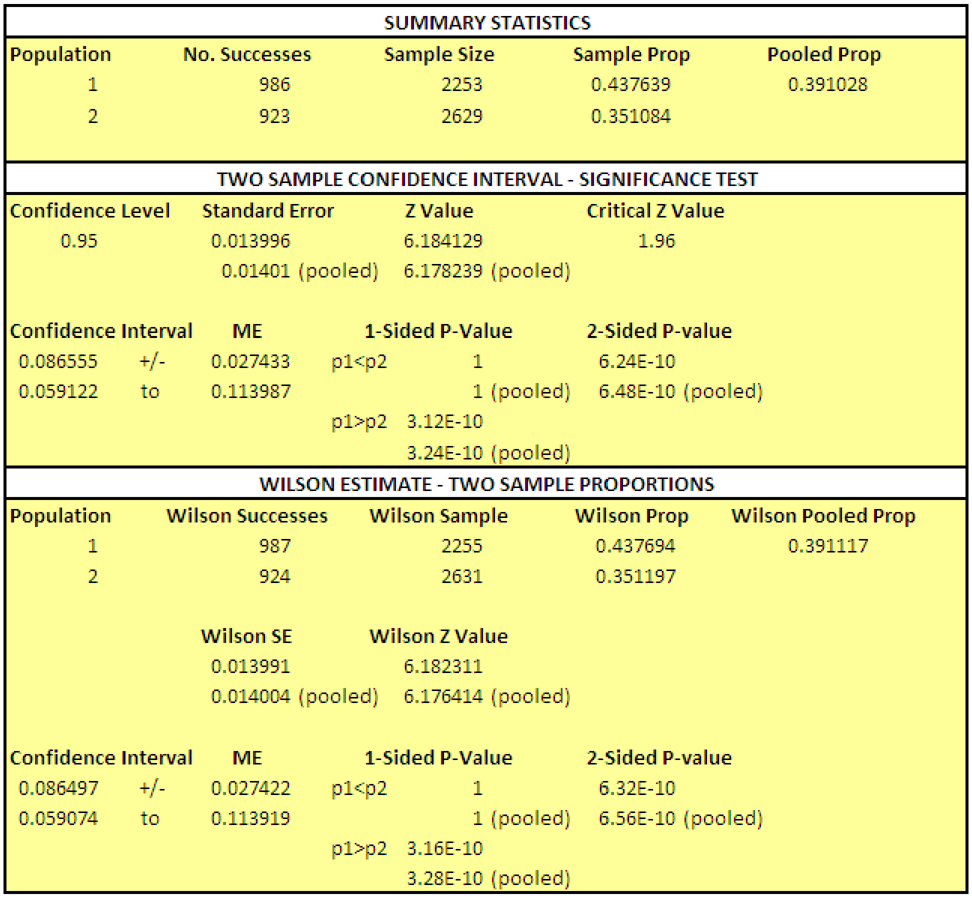
○ 99%

Excel calculates the confidence interval as

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the observed probabilities of sample one and sample two, respectively, and $\hat{p} = X/n$, where X is the observed success in n trials.

The following results show that we are 95% confident that the percent of women living at home is somewhere between 5.9 and 11.4 percentage points lower among women than among men.



```
                                  SUMMARY STATISTICS
Population        No. Successes        Sample Size        Sample Prop        Pooled Prop
     1                 986                 2253             0.437639           0.391028
     2                 923                 2629             0.351084

                  TWO SAMPLE CONFIDENCE INTERVAL - SIGNIFICANCE TEST
Confidence Level   Standard Error       Z Value            Critical Z Value
    0.95             0.013996            6.184129               1.96
                       0.01401 (pooled)  6.178239 (pooled)

Confidence Interval    ME             1-Sided P-Value      2-Sided P-value
 0.086555     +/-    0.027433     p1<p2        1            6.24E-10
 0.059122     to     0.113987                  1 (pooled)   6.48E-10 (pooled)
                                  p1>p2   3.12E-10
                                          3.24E-10 (pooled)

                     WILSON ESTIMATE - TWO SAMPLE PROPORTIONS
Population      Wilson Successes    Wilson Sample      Wilson Prop     Wilson Pooled Prop
     1                 987               2255            0.437694           0.391117
     2                 924               2631            0.351197

                   Wilson SE           Wilson Z Value
                     0.013991            6.182311
                     0.014004 (pooled)   6.176414 (pooled)

Confidence Interval    ME             1-Sided P-Value      2-Sided P-value
 0.086497     +/-    0.027422     p1<p2        1            6.32E-10
 0.059074     to     0.113919                  1 (pooled)   6.56E-10 (pooled)
                                  p1>p2   3.16E-10
                                          3.28E-10 (pooled)
```

More Accurate Confidence Intervals

A simple modification improves the accuracy of confidence interval for comparing propositions. As with a single proportion, the interval is called the "plus four" interval because you add four imaginary observations, one success and one failure in each of the two samples. That is, we let

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \text{ and } \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

The confidence interval based on this modification is given in the output in the Wilson output section. For this example, the results are about the same as without the modification. The "plus four" interval is generally much more accurate than the large-sample interval when the samples are small.

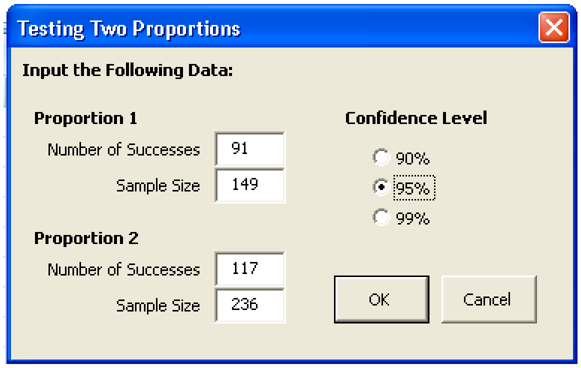
Significance Tests for Comparing Proportions

We also can use Excel to do significance tests to help us decide if the effect we see in the samples is really there in the populations. The null hypothesis says that there is no difference between the two populations: H_0: $p_1 = p_2$.

In the next example, researchers ask the question, "Would you marry a person from a lower social class than your own?" Of the 149 men in the sample, 91 said "Yes." Among the 236 women, 117 said "Yes." To see if there is a statistically significant difference we have a two-sided alternative:

$$H_0: p_1 = p_2$$
$$H_a: p_1 \neq p_2$$

Select **Add-Ins ➤ WHFStat ➤ Proportion Testing ➤ Two Samples** to perform the significance test. Enter the summary information in the dialog box and select a Confidence Level before clicking on OK.



Testing Two Proportions

Input the Following Data:

Proportion 1

Number of Successes 91

Sample Size 149

Proportion 2

Number of Successes 117

Sample Size 236

Confidence Level

○ 90%

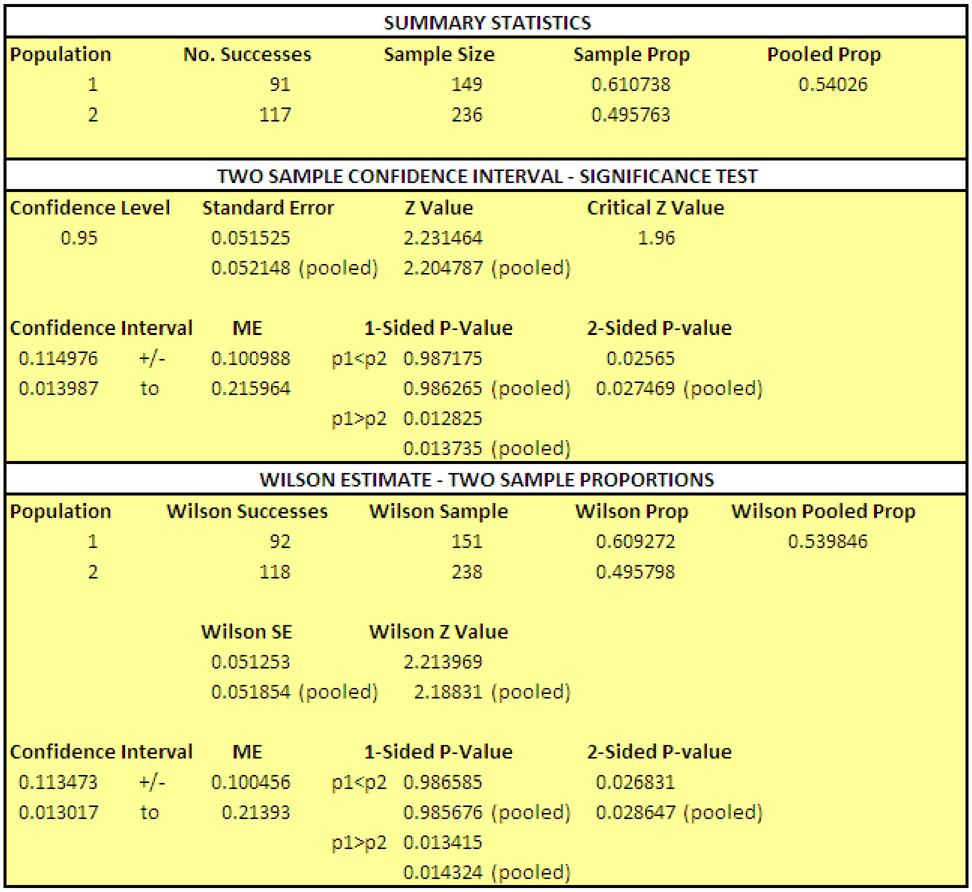
● 95%

○ 99%

Excel uses a pooled estimate of p for the hypothesis test and calculates z as

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Because the P-value shown on the following page is equal to 0.027, the results are statistically significant at the $\alpha = 0.05$ level. There is good evidence that men are more likely than women to say they will marry someone from a lower social class.



SUMMARY STATISTICS

Population	No. Successes	Sample Size	Sample Prop	Pooled Prop
1	91	149	0.610738	0.54026
2	117	236	0.495763	

TWO SAMPLE CONFIDENCE INTERVAL - SIGNIFICANCE TEST

Confidence Level	Standard Error	Z Value	Critical Z Value
0.95	0.051525	2.231464	1.96
	0.052148 (pooled)	2.204787 (pooled)	

Confidence Interval		ME	1-Sided P-Value		2-Sided P-value
0.114976	+/-	0.100988	p1<p2	0.987175	0.02565
0.013987	to	0.215964		0.986265 (pooled)	0.027469 (pooled)
			p1>p2	0.012825	
				0.013735 (pooled)	

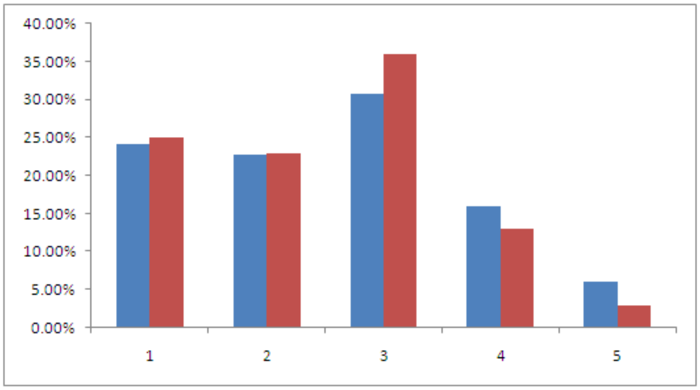
WILSON ESTIMATE - TWO SAMPLE PROPORTIONS

Population	Wilson Successes	Wilson Sample	Wilson Prop	Wilson Pooled Prop
1	92	151	0.609272	0.539846
2	118	238	0.495798	

Wilson SE	Wilson Z Value
0.051253	2.213969
0.051854 (pooled)	2.18831 (pooled)

Confidence Interval		ME	1-Sided P-Value		2-Sided P-value
0.113473	+/-	0.100456	p1<p2	0.986585	0.026831
0.013017	to	0.21393		0.985676 (pooled)	0.028647 (pooled)
			p1>p2	0.013415	
				0.014324 (pooled)	

CHAPTER 9



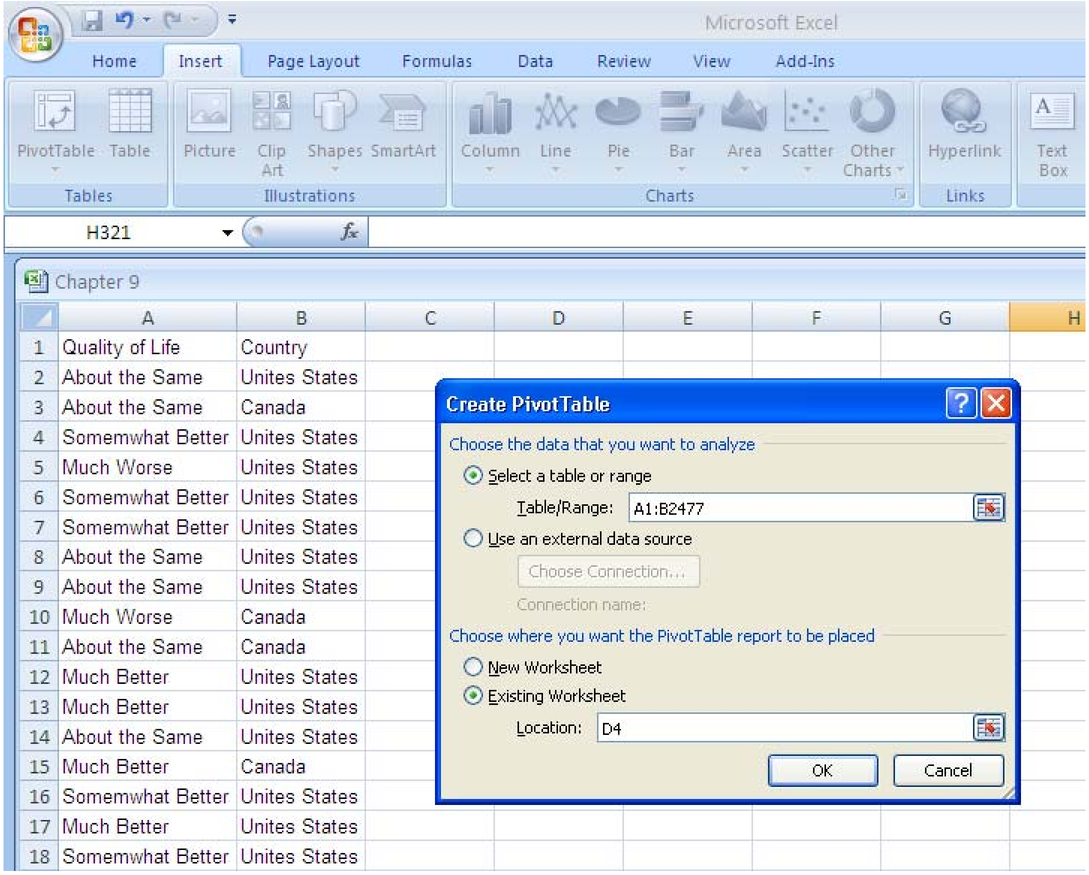
Inference for Two-Way Tables

The Chi-Square Test

We can use Excel to do a χ^2 test of the null hypothesis that there is "no relationship" between the column variable and the row variable in a two-way table. Our example looks at the health care system in the United States and Canada. The study looked at outcomes a year after a heart attack. One outcome was the patients' own assessment of their quality of life relative to what it had been before the heart attack. The data for the patients who survived a year are in an Excel worksheet with the columns Quality of Life and Country. To obtain tables of counts from this data, select

Insert ➤ Pivot Table

from the Excel menu. In the first dialog box enter the table range including the column titles for the variables containing the categories that define the rows and column, of the table, as shown on the following page. Also, choose where the Pivot Table will be placed and then click OK.



	A	B
1	Quality of Life	Country
2	About the Same	Unites States
3	About the Same	Canada
4	Somemwhat Better	Unites States
5	Much Worse	Unites States
6	Somemwhat Better	Unites States
7	Somemwhat Better	Unites States
8	About the Same	Unites States
9	About the Same	Unites States
10	Much Worse	Canada
11	About the Same	Canada
12	Much Better	Unites States
13	Much Better	Unites States
14	About the Same	Unites States
15	Much Better	Canada
16	Somemwhat Better	Unites States
17	Much Better	Unites States
18	Somemwhat Better	Unites States

Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: A1:B2477

○ Use an external data source

Connection name:

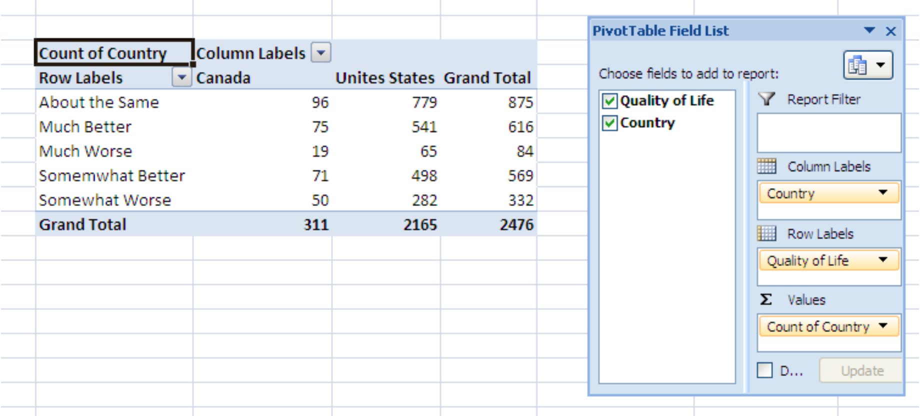
Choose where you want the PivotTable report to be placed

○ New Worksheet

● Existing Worksheet

Location: D4

An empty Pivot Table will appear along with the Pivot Table Field List. Check variables and move them into the Row Labels and Column Labels fields. In addition, you must move one variable into the Σ Values field as shown. Click the Update button to create a table with counts.



Count of Country	Column Labels		
Row Labels	Canada	Unites States	Grand Total
About the Same	96	779	875
Much Better	75	541	616
Much Worse	19	65	84
Somemwhat Better	71	498	569
Somewhat Worse	50	282	332
Grand Total	311	2165	2476

PivotTable Field List

Choose fields to add to report:

☑ Quality of Life

☑ Country

Report Filter

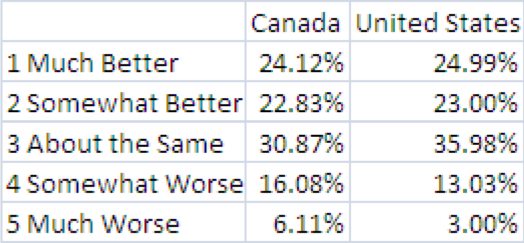
Column Labels: Country

Row Labels: Quality of Life

Σ Values: Count of Country

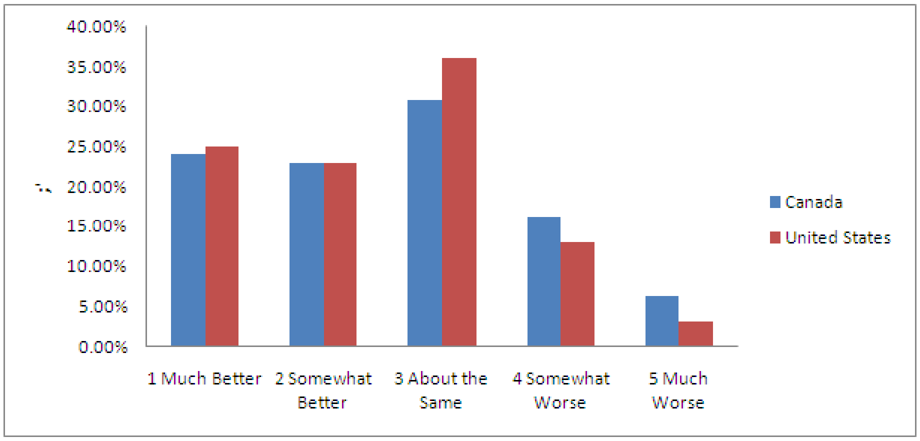
Once the data has been summarized, it is helpful to graph the data so that the reported outcomes can be compared for each country. The pivot table automatically orders the variables, in this case alphabetically. Since alphabetical ordering doesn't make sense for the quality of life outcomes, these were coded in a logical order from "Much

Better" to "Much Worse." Additionally, the counts were each divided by their column total so that the percentages in each category could be compared.



	Canada	United States
1 Much Better	24.12%	24.99%
2 Somewhat Better	22.83%	23.00%
3 About the Same	30.87%	35.98%
4 Somewhat Worse	16.08%	13.03%
5 Much Worse	6.11%	3.00%

As seen in the table above and the following graph, the reported outcomes look similar in Canada and the United States. The outcome "About the Same" is selected most frequently and "Much Worse" is selected the least frequently in both countries. However, there are also differences between the countries as well. In the United States, the outcomes "Somewhat Worse" and "Much Worse" are selected less often than in Canada.

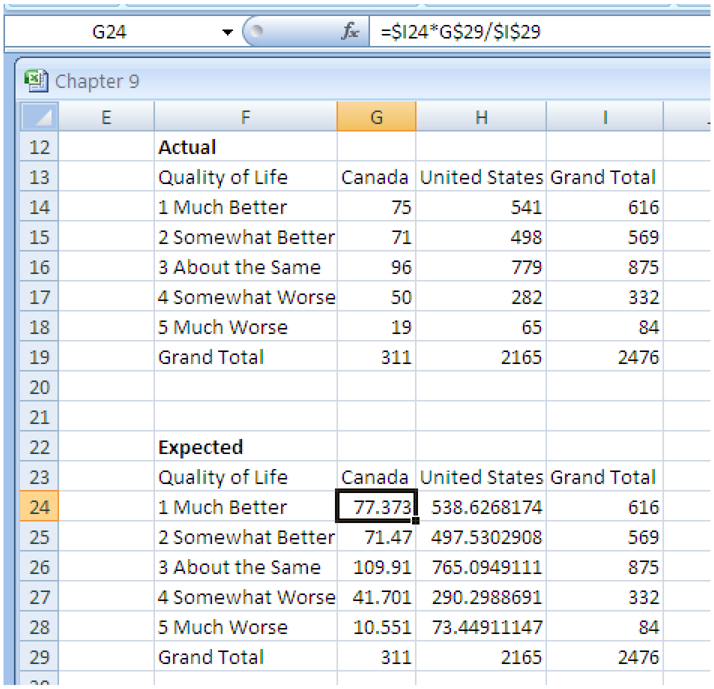


The chi-square test will help us see whether the differences between the two countries are statistically significant. The null hypothesis, H_0 for this test is that there is no association between the row variable and the column variable. H_a is that there is an association. In this example, H_0 is that there is no difference between the reported outcomes in the United States and Canada.

To perform a chi-square test of association between variables, expected cell counts are required. These can be calculated by first copying the row and column totals from the Pivot Table and then calculating the expected counts for the interior cells on the table. The expected count for each outcome/country combination is calculated as

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Overall total}}$$

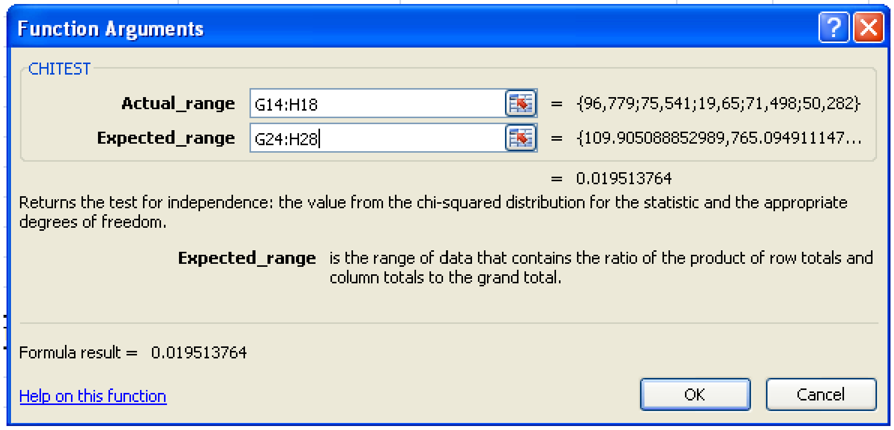
as shown in the spreadsheet below.


G24 =$I24*G$29/$I$29

Chapter 9

	E	F	G	H	I
12		**Actual**			
13		Quality of Life	Canada	United States	Grand Total
14		1 Much Better	75	541	616
15		2 Somewhat Better	71	498	569
16		3 About the Same	96	779	875
17		4 Somewhat Worse	50	282	332
18		5 Much Worse	19	65	84
19		Grand Total	311	2165	2476
20					
21					
22		**Expected**			
23		Quality of Life	Canada	United States	Grand Total
24		1 Much Better	77.373	538.6268174	616
25		2 Somewhat Better	71.47	497.5302908	569
26		3 About the Same	109.91	765.0949111	875
27		4 Somewhat Worse	41.701	290.2988691	332
28		5 Much Worse	10.551	73.44911147	84
29		Grand Total	311	2165	2476

Excel's **CHITEST** function provides the *P*-value for the chi-squared test of association between the row and column variables. The function arguments are the actual counts and the expected counts (interior cells) on the tables. The *P*-value for this example is .0195, a small value. This means that there is a statistically significant relationship between patients' assessment of their quality of life and the country where they are treated for a heart attack.



CHITEST

Actual_range G14:H18 = {96,779;75,541;19,65;71,498;50,282}

Expected_range G24:H28 = {109.905088852989,765.094911147...

= 0.019513764

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

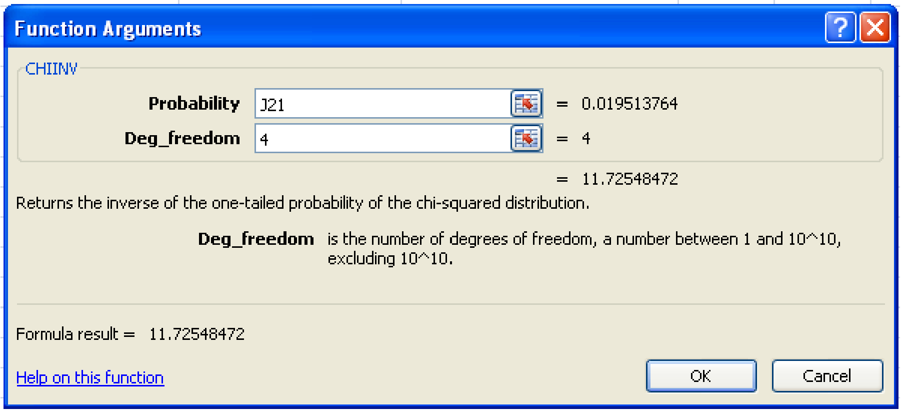
Expected_range is the range of data that contains the ratio of the product of row totals and column totals to the grand total.

Formula result = 0.019513764

The number of degrees of freedom for the χ^2 statistic is equal to $(\text{rows} - 1) \times (\text{columns} - 1)$. For our example the degrees of freedom is equal to 4. The χ^2 statistic compares the table of observed counts with the table of expected counts.

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

Excel does not provide the χ^2 statistic, but since we have the *P*-value, we can work backward to obtain the value using the **CHIINV** function. The function arguments are the probability, i.e., the *P*-value entered directly from the spreadsheet, and the degrees of freedom.



CHIINV

Probability J21 = 0.019513764

Deg_freedom 4 = 4

= 11.72548472

Returns the inverse of the one-tailed probability of the chi-squared distribution.

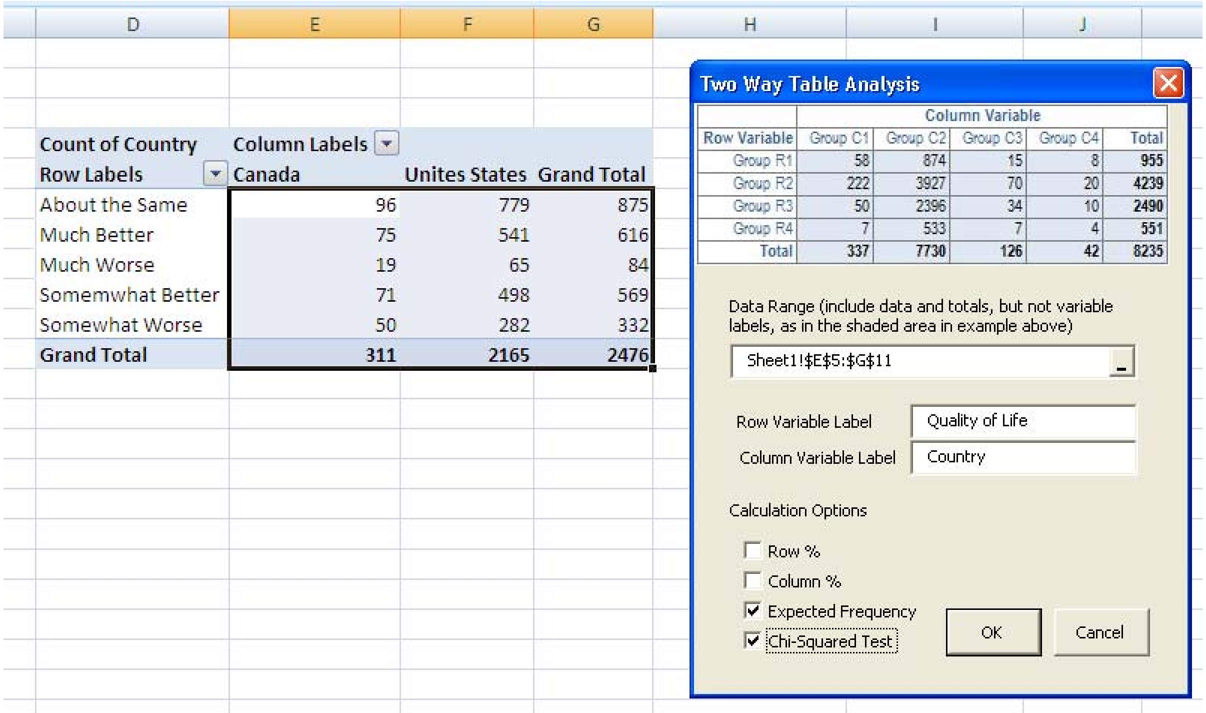
Deg_freedom is the number of degrees of freedom, a number between 1 and 10^10, excluding 10^10.

Formula result = 11.72548472

Help on this function

The χ^2 value calculated by Excel is equal to 11.725.

Alternatively, you may select **Add-Ins ➤ WHFStat ➤ Two-Way Table / Chi-Squared Test** from the Excel menu and fill in the dialog box as shown below. The results are identical to those described above.



Count of Country	Column Labels		
Row Labels	Canada	Unites States	Grand Total
About the Same	96	779	875
Much Better	75	541	616
Much Worse	19	65	84
Somemwhat Better	71	498	569
Somewhat Worse	50	282	332
Grand Total	311	2165	2476

	Column Variable				
Row Variable	Group C1	Group C2	Group C3	Group C4	Total
Group R1	58	874	15	8	955
Group R2	222	3927	70	20	4239
Group R3	50	2396	34	10	2490
Group R4	7	533	7	4	551
Total	337	7730	126	42	8235

Data Range (include data and totals, but not variable labels, as in the shaded area in example above)

Sheet1!$E$5:$G$11

Row Variable Label: Quality of Life

Column Variable Label: Country

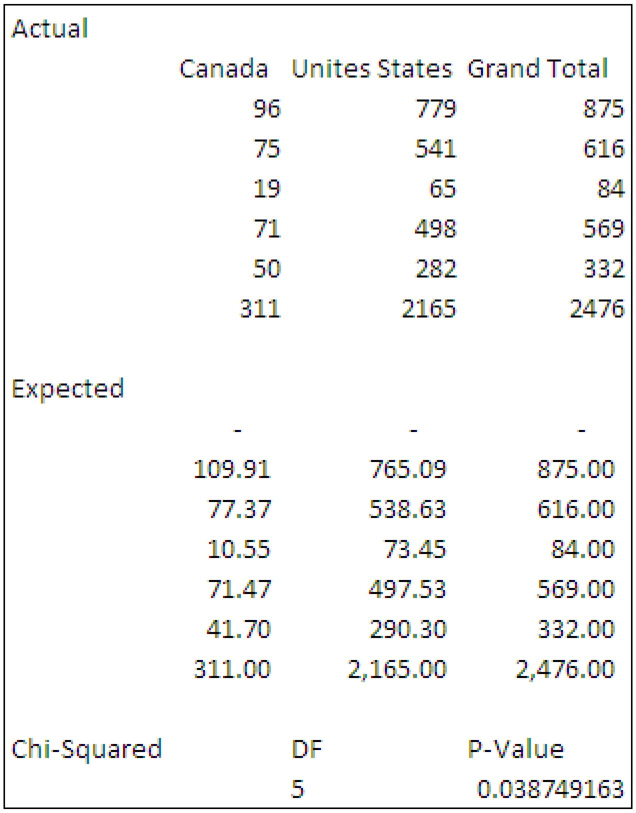
Calculation Options

□ Row %

□ Column %

☑ Expected Frequency

☑ Chi-Squared Test



Actual

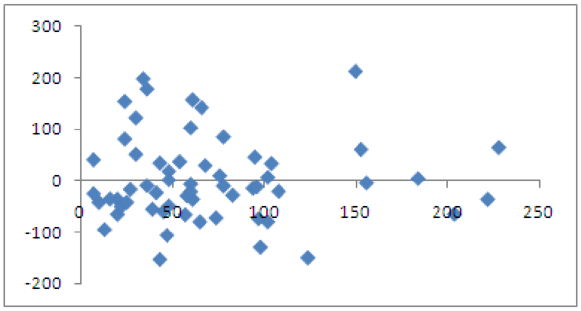
	Canada	Unites States	Grand Total
	96	779	875
	75	541	616
	19	65	84
	71	498	569
	50	282	332
	311	2165	2476

Expected

	-	-	-
	109.91	765.09	875.00
	77.37	538.63	616.00
	10.55	73.45	84.00
	71.47	497.53	569.00
	41.70	290.30	332.00
	311.00	2,165.00	2,476.00

Chi-Squared	DF	P-Value
	5	0.038749163

CHAPTER 10



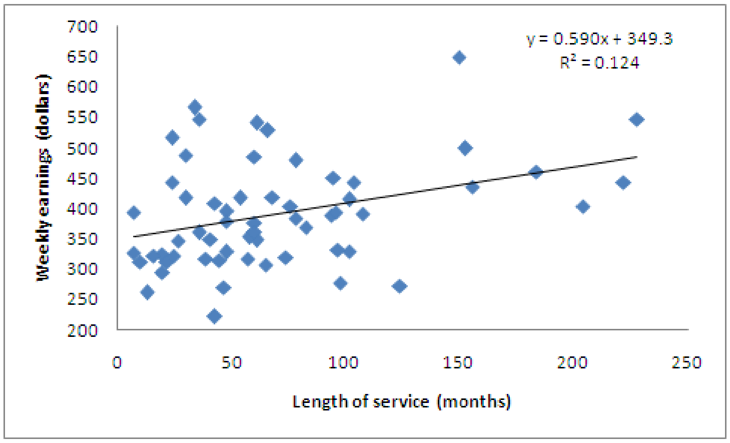
Inference for Regression

Estimating the Regression Parameters

In the following example we will examine the relationship between bank wages and length of service for 59 married women who hold customer service jobs in Indiana banks. The data are below.

WAGES	LOS	WAGES	LOS	WAGES	LOS	WAGES	LOS
389	94	541	61	486	60	404	204
395	48	312	10	393	7	443	24
329	102	418	68	311	22	261	13
295	20	417	54	316	57	417	30
377	60	516	24	384	78	450	95
479	78	443	222	360	36	443	104
315	45	353	58	369	83	566	34
316	39	349	41	529	66	461	184
324	20	499	153	270	47	436	156
307	65	322	16	332	97	321	25
403	76	408	43	547	228	221	43
378	48	393	96	347	27	547	36
348	61	277	98	328	48	362	60
488	30	649	150	327	7	415	102
391	108	272	124	320	74		

Before attempting inference, examine the data by (1) making a scatterplot; (2) fitting the least-squares regression, $\hat{y} = b_0 + b_1 x$; (3) checking for outliers and influential observations; and (4) computing the value of r^2. These can all be done at once by making a fitted line plot. First make a scatterplot. Begin by highlighting your data with the explanatory variable to the left of the response variable. Selecting **Insert ➤ Scatter ➤ Scatter with only Markers,** then right click on an observation, right click on the scatterplot, select Add Trendline from the list, select Linear on the Trendline Options, check the box next to Display Equation on the chart, and Display R-squared value on the chart if desired.

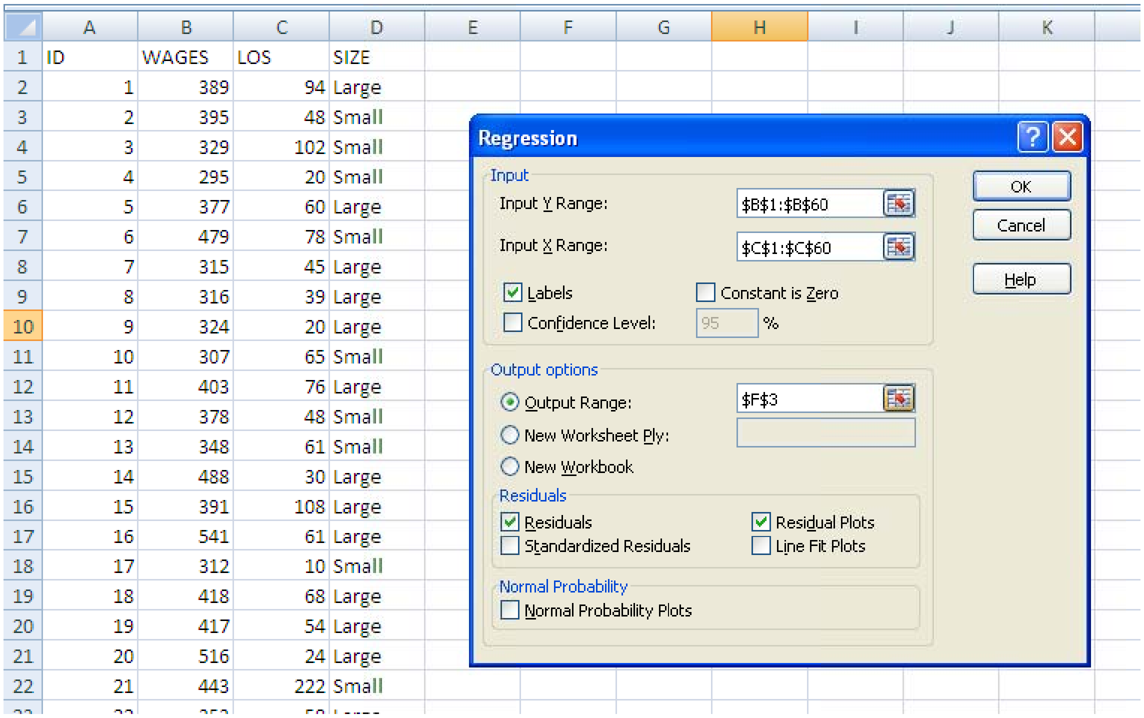


The scatterplot shows a moderate linear relationship with no extreme outliers. The least-squares line is given to be

$$y = 349.4 + 0.5905x,$$

and $r^2 = 0.124$ The change in wages along the regression line as length of service increases explains only about 12% of the variation. The change in wages along the regression line as length of service increases explains only about 12% of the variation.

Additional Regression information can be obtained by selecting **Data ➤ Data Analysis ➤ Regression** from the Excel menu. The response variable (WAGES) and the explanatory variable (LOS) are entered in the dialog box in the windows labeled Input Y Range and Input X Range, respectively.



	A	B	C	D
1	ID	WAGES	LOS	SIZE
2	1	389	94	Large
3	2	395	48	Small
4	3	329	102	Small
5	4	295	20	Small
6	5	377	60	Large
7	6	479	78	Small
8	7	315	45	Large
9	8	316	39	Large
10	9	324	20	Large
11	10	307	65	Small
12	11	403	76	Large
13	12	378	48	Small
14	13	348	61	Small
15	14	488	30	Large
16	15	391	108	Large
17	16	541	61	Large
18	17	312	10	Small
19	18	418	68	Large
20	19	417	54	Large
21	20	516	24	Large
22	21	443	222	Small

Regression

Input

Input Y Range: $B$1:$B$60

Input X Range: $C$1:$C$60

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $F$3

○ New Worksheet Ply:

○ New Workbook

Residuals

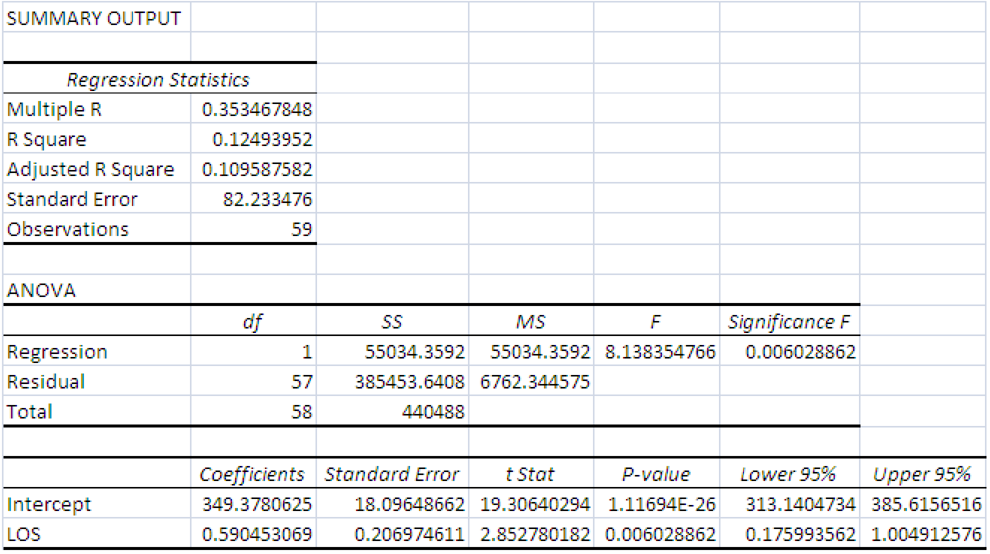
☑ Residuals ☑ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

E
F
G
H
I
J
K
OK
Cancel
Help



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.353467848
R Square	0.12493952
Adjusted R Square	0.109587582
Standard Error	82.233476
Observations	59

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	55034.3592	55034.3592	8.138354766	0.006028862
Residual	57	385453.6408	6762.344575		
Total	58	440488			

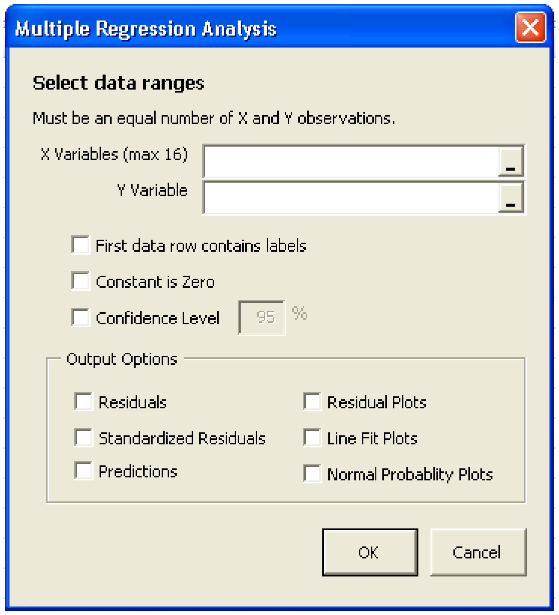
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	349.3780625	18.09648662	19.30640294	1.11694E-26	313.1404734	385.6156516
LOS	0.590453069	0.206974611	2.852780182	0.006028862	0.175993562	1.004912576

The values of b_0 and b_1 are given in the bottom section in the column labeled Coefficients. The first entry in that column is b_0, the intercept, and the second is b_1, the slope. We see that b_0 = 349.38 and b_1 = .5905. These are the estimates of β_0 and β_1. Therefore, the regression equation is

$$\text{WAGES} = 349 + 0.590\ \text{LOS}$$

The standard error, s = 82.23, found in the top section, is used to estimate σ, the standard deviation of responses about the true regression line.

Alternatively, the same results can be obtained by selecting **Add-Ins ➤ WHFStat ➤ Correlation and Regression ➤ Regression** from the Excel menu and filling in the following dialog box.



Multiple Regression Analysis

Select data ranges

Must be an equal number of X and Y observations.

X Variables (max 16)

Y Variable

□ First data row contains labels

□ Constant is Zero

□ Confidence Level 95 %

Output Options

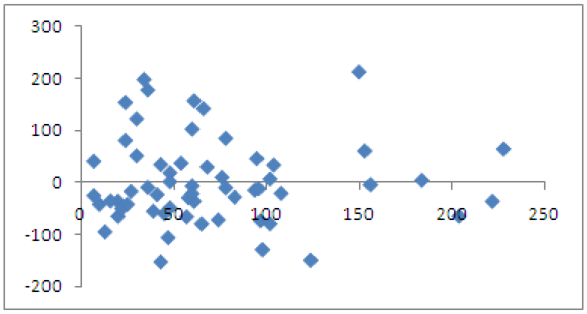
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

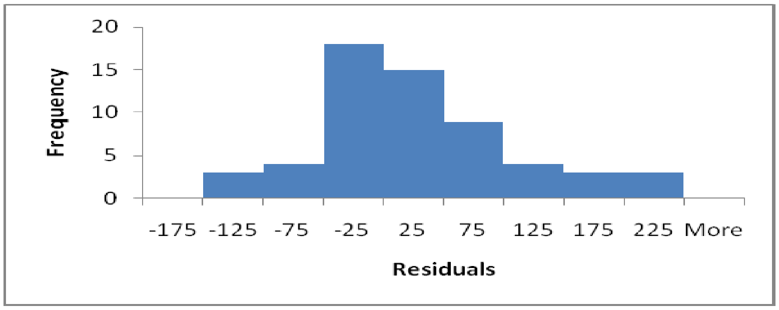
□ Predictions □ Normal Probablity Plots

Remember that before using regression inference, the data must satisfy the regression model assumptions. Use a scatterplot to check that the true relationship is linear. The scatter of the data points about the line should be roughly the same over the entire range of the data. A plot of the residuals against *x* should not show any pattern. A histogram or stemplot of the residuals should not show any major departures from normality.

We have fitted a regression line and we should now examine the residuals. To obtain residual plot, make sure that the Residual plots box is checked as shown on previous page. The residual plots for this data looks satisfactory.



You can also check the Residuals box to obtain a list of the residuals. These can be used to make a histogram of the residuals.



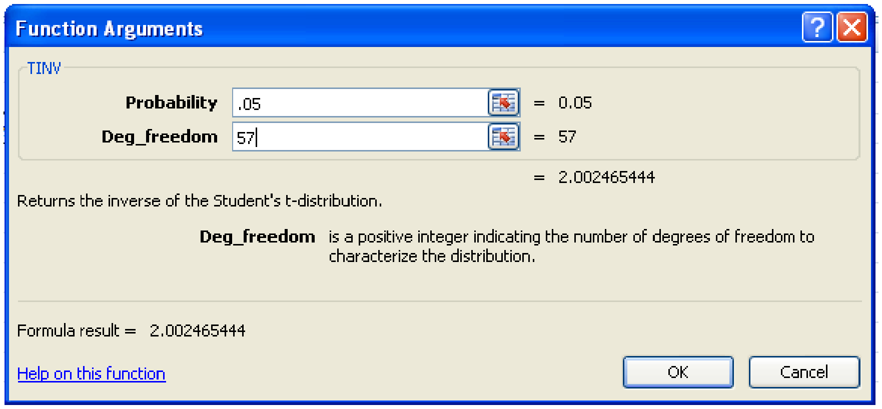
The assumption of normally distributed residuals also appears to be reasonable. There are no serious deviations from a Normal Distribution. This is important for the inference that follows.

Confidence Intervals and Hypothesis Tests for β_0 and β_1

Confidence intervals and tests for the slope and intercept are based on the normal sampling distributions of the estimates b_0 and b_1. Since the standard deviations are not known, a t distribution is used. The value of SE_{b_1} is 0.207. It appears in the output from the regression to the right of the estimated slope, $b_1 = 0.5905$. Similarly, the value of SE_{b_0} is 18.1. It appears to the right of the estimated constant, $b_0 = 349.38$. Confidence intervals for β_0 and β_1 have the form

$$\text{estimate} \pm t^{*}\text{SE}_{\text{estimate}}$$

The **TINV** function can be used to find the critical value you would use for a 95% confidence interval based on the $t(57)$ distribution. The t distributions for this problem have $n–2 = 57$ degrees of freedom. The function arguments are the Probability, which is 0.05 for a 95% confidence interval, and the Degrees of Freedom. As shown below, the critical value from the $t(57)$ distribution is equal to 2.002.



TINV

Probability .05 = 0.05

Deg_freedom 57 = 57

= 2.002465444

Returns the inverse of the Student's t-distribution.

Deg_freedom is a positive integer indicating the number of degrees of freedom to characterize the distribution.

Formula result = 2.002465444

Help on this function

The upper and lower bounds for the confidence interval can be calculated with Excel's calculator functions. A 95% confidence interval for β_1 is (.176, 1.005). Fortunately, this calculation isn't needed as the confidence interval is listed in the Excel output on the right side of the bottom section on the same row as the slope.

The *t* statistic and *P*-value for the test of

$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

appear in the columns labeled *t* Stat and *P*-value. The *t* ratio can also be obtained from the formula

$$t = \frac{b_1}{\mathrm{SE}_{b_1}} = \frac{.5905}{.207} = 2.85$$

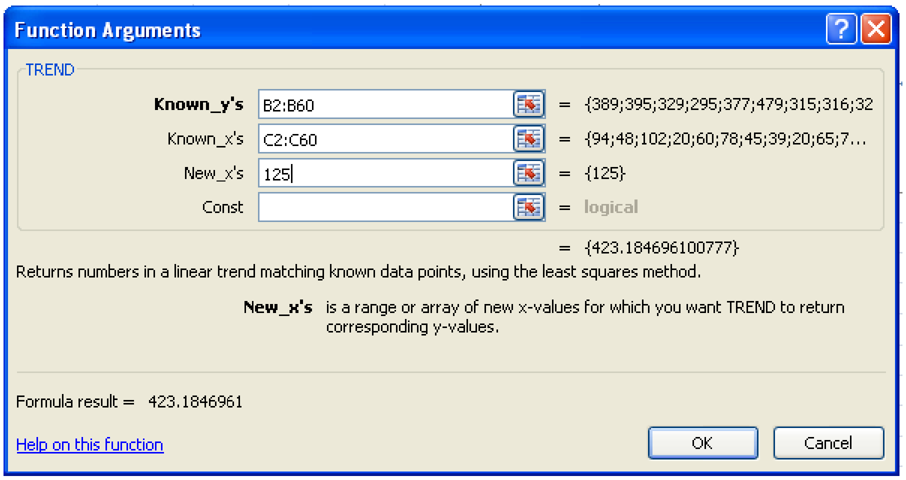
The *P*-value is listed as 0.006. We expect that wages will rise with length of service, so our alternative is one-sided, $H_a\colon \beta_1 > 0$. The *P*-value for this alternative is one-half the two-sided value; that is, the *P*-value is 0.003. There is strong evidence that mean wages increase as length of service increases. Confidence intervals and hypothesis tests for β_0 can be obtained similarly.

Inference about Prediction

We found that the least-squares line for predicting WAGES from LOS is

$$\hat{y} = 349.4 + 0.5905x.$$

Excel can be used to predict WAGES for a worker who has been with the bank 125 months either by plugging 125 into the least-squares equation or using the **TREND** function. The function arguments are the data, i.e., the Known y's and the Known x's, and the New x, which is equal to 125 in this example. The last argument is left blank.



Function Arguments

TREND

Known_y's B2:B60 = {389;395;329;295;377;479;315;316;32

Known_x's C2:C60 = {94;48;102;20;60;78;45;39;20;65;7...

New_x's 125 = {125}

Const = logical

= {423.184696100777}

Returns numbers in a linear trend matching known data points, using the least squares method.

New_x's is a range or array of new x-values for which you want TREND to return corresponding y-values.

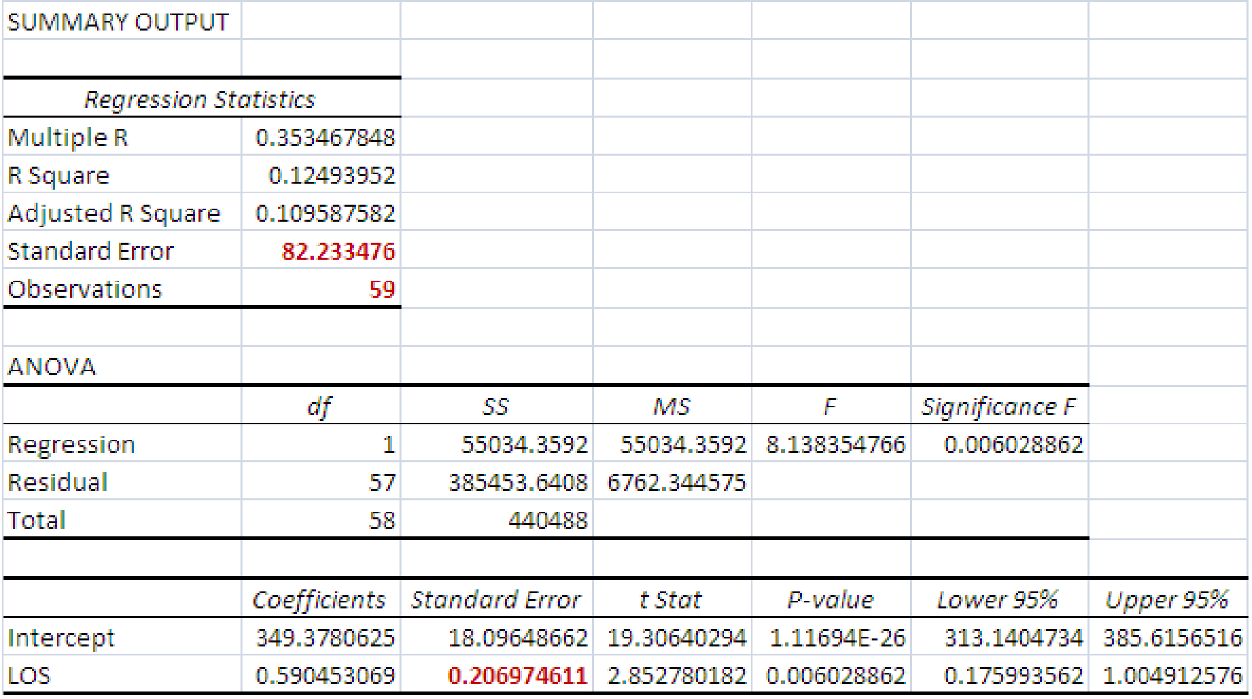
Formula result = 423.1846961

We may be interested in predicting the *mean response*, the average earnings of all workers in the subpopulation with 125 months on the job, or we may be interested in predicting the earnings of *one individual worker* with 125 months of service. The prediction is the same for both, $\hat{y} = 423.2$ dollars per week. However, the margin of error is different for the two kinds of prediction.

Individual workers with 125 months of service do not all earn the same amount. So we need a larger margin of error to pin down one worker's earnings than to estimate the mean earnings of all workers who have been with their employer 125 months. The confidence interval for the *mea response* is $\hat{y} \pm t^* \mathrm{SE}_{\hat{\mu}}$. The value for $\mathrm{SE}_{\hat{\mu}}$ is not listed and must be calculated. The value is most easily calculated using the formula

$$\mathrm{SE}_{\hat{\mu}} = \sqrt{\frac{s^2}{n} + SE_{b_1}^2 (x^* - \bar{x})^2}$$

In this example, $s^*= 82.23$, $SE_{b_1}^2 = 0.207$, $n = 59$, $x^* = 125$, and $\bar{x} = 70.49$, is calculated from the values of the explanatory variable (LOS.) Based on the data given below, $\mathrm{SE}_{\hat{\mu}}$=15.55. The value for *t** for 57 degrees of freedom was found previously to be 2.002. Therefore, the 95% confidence interval for the mean response is equal to 423.2 ± 31.14 or (392.0, 454.3).



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.353467848					
R Square	0.12493952					
Adjusted R Square	0.109587582					
Standard Error	82.233476					
Observations	59					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	55034.3592	55034.3592	8.138354766	0.006028862	
Residual	57	385453.6408	6762.344575			
Total	58	440488				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	349.3780625	18.09648662	19.30640294	1.11694E-26	313.1404734	385.6156516
LOS	0.590453069	0.206974611	2.852780182	0.006028862	0.175993562	1.004912576

The *individual* prediction interval will be wider than the confidence interval for the mean response. This interval is $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$. The value of $\mathrm{SE}_{\hat{y}}$ is also not given on the Excel output, but it is easily obtained from the following formula

$$\mathrm{SE}_{\hat{y}} = \sqrt{s^2 + \left(\mathrm{SE}_{\hat{\mu}}\right)^2}.$$

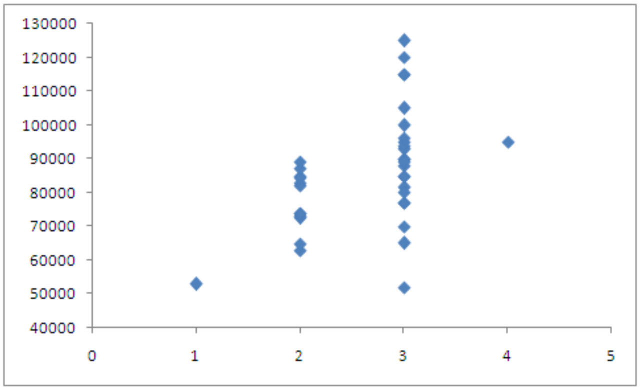
The value from the formula is 83.7, giving a 95% prediction interval of 423.2 ± 167.6 or (255.6, 590.8). Alternatively, the value of $\mathrm{SE}_{\hat{y}}$ is easily approximated by

$$\mathrm{SE}_{\hat{y}} \cong \sqrt{s^2 + \frac{1}{n}}.$$

In this example, the approximation gives a value of 82.23.

CHAPTER

11



Multiple Regression

Multiple Regression

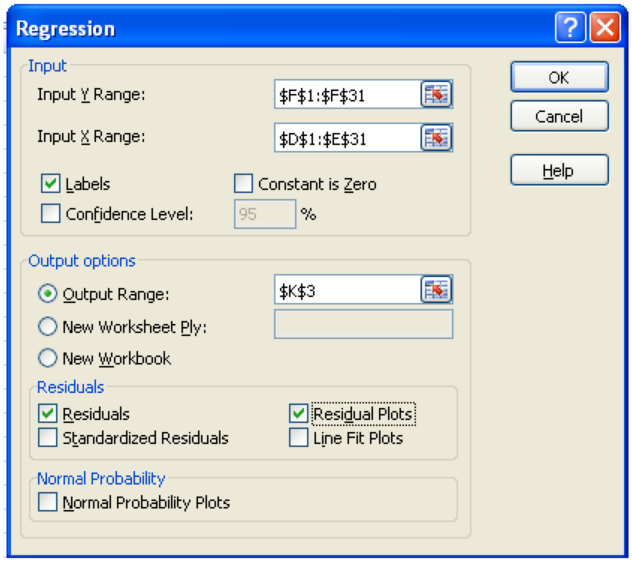
Multiple regression fits the regression equation

$$\mu_y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

to data in selected response and predictor variables. We will illustrate multiple regression with the data showing characteristics of the 30 stocks in the Dow Jones Industrial Average (DJIA). We'll examine how profits are related to sales and assets. To run a multiple regression analysis using Excel, select

Data ➤ Data Analysis ➤ Regression

from the menu. In the dialog box, enter the response variable in the Input Y Range window and as many explanatory variables as you like in the Input X Range window. Since you need a single cell range for the *x* values, all of the explanatory variables must be adjacent.



Regression

Input

Input Y Range: $F$1:$F$31

Input X Range: $D$1:$E$31

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $K$3

○ New Worksheet Ply:

○ New Workbook

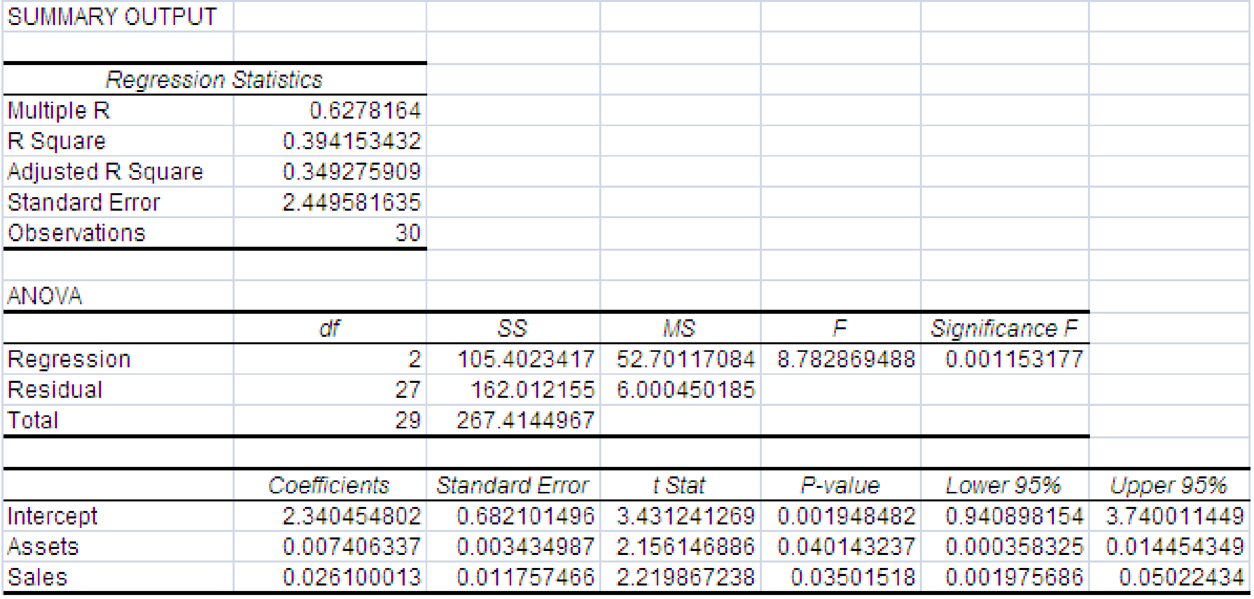
Residuals

☑ Residuals ☑ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6278164
R Square	0.394153432
Adjusted R Square	0.349275909
Standard Error	2.449581635
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	105.4023417	52.70117084	8.782869488	0.001153177
Residual	27	162.012155	6.000450185		
Total	29	267.4144967			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.340454802	0.682101496	3.431241269	0.001948482	0.940898154	3.740011449
Assets	0.007406337	0.003434987	2.156146886	0.040143237	0.000358325	0.014454349
Sales	0.026100013	0.011757466	2.219867238	0.03501518	0.001975686	0.05022434

The above output provides the estimated regression coefficients; the intercept, b_0 = 2.34045, the slope for Assets, b_1 = 0.007406, and the slope for Sales, b_2 = 0.02610. Therefore, the regression equation is

$$\text{Profits} = 2.34 + 0.00741 \text{ Assets} + 0.0261 \text{ Sales}$$

The estimate of σ is given as s = 2.450. The estimate is calculated as

$$s = \sqrt{\frac{\Sigma e_i^2}{n-p-1}}$$

where the e_i's are the residuals and p is the number of predictor variables. In the bottom section, the column marked Standard Error gives the estimated standard errors: $s_{b_0} = 0.6821$, $s_{b_1} = 0.007406$, and $s_{b_2} = 0.01176$. A level C confidence interval for β_j can be computed as

$$b_j \pm t^* s_{b_j}$$

where t^* is the upper $(1-C)/2$ critical value for the $t(n-p-1)$ distribution. This is exactly the same as for simple linear regression. In that case, $p = 1$, so the number of degrees of freedom is $n-2$.

To test the hypothesis $H_0 : \beta_j = 0$, the value of t is computed as b_j / s_{b_j}. For each coefficient, the value appears in the column marked t-ratio. The values are given as 3.43, 2.16, and 2.22. The P-values for a test against $H_a : \beta_j \neq 0$ are provided in the column marked p and are from the $t(n-p-1)$ distribution.

The analysis of variance table for multiple regression is illustrated below. It has the same format as for simple linear regression. The only difference is that the number of degrees of freedom for the model increases from 1 to p, reflecting the fact that there are p explanatory variables. Similarly, the number of degrees of freedom for the error decreases from $n-2$ to $n-p-1$.

Analysis of Variance

SOURCE	DF	SS	MS	F	SIGNIFICANCE F
Regression	p	$\Sigma(\hat{y}_i - \bar{y})^2$	MSM=SSM/DFM	MSM/MSE	
Error	$n-p-1$	$\Sigma(y_i - \hat{y}_i)^2$	MSE=SSE/DFE		
Total	$n-1$	$\Sigma(y_i - \bar{y})^2$			

The value of MSE is the estimate of σ^2. In the example above, it is given as 6.000. This value could also be obtained by squaring the estimate of σ ($s = 2.450$). The ratio MSM/MSE is an F statistic for testing the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

against the alternative hypothesis

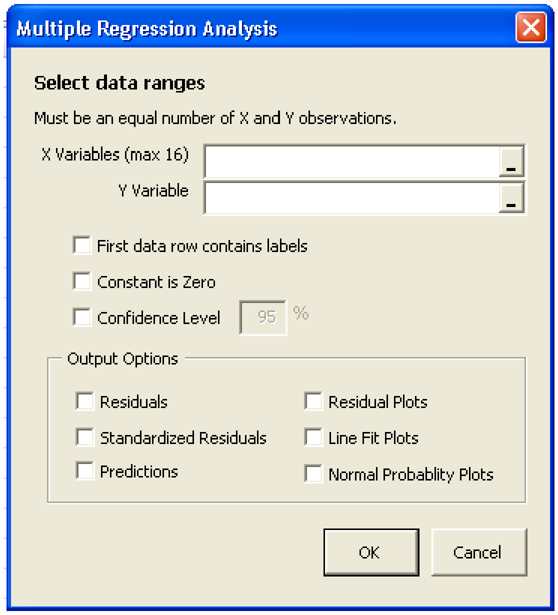
$$H_a : \beta_j \neq 0 \quad \text{for at least one } j = 1, 2, \ldots, p$$

The test statistic has the $F(p, n-p-1)$ distribution. In the example above, $F = 8.78$. The P-value listed under the column marked p is given as 0.001. This means that there is strong evidence that at least one $\beta_j \neq 0$.

The value of R-sq is listed above as 0.349. This means that the proportion of the variation in profits that is explained by assets and sales is

$$R^2 = \frac{\text{SSM}}{\text{SST}} = 0.349$$

Alternatively, the same results can be obtained by selecting **Add-Ins ➤ WHFStat ➤ Correlation and Regression ➤ Regression** from the Excel menu and filling in the following dialog box.



Multiple Regression Analysis

Select data ranges

Must be an equal number of X and Y observations.

X Variables (max 16)

Y Variable

□ First data row contains labels

□ Constant is Zero

□ Confidence Level 95 %

Output Options

□ Residuals □ Residual Plots

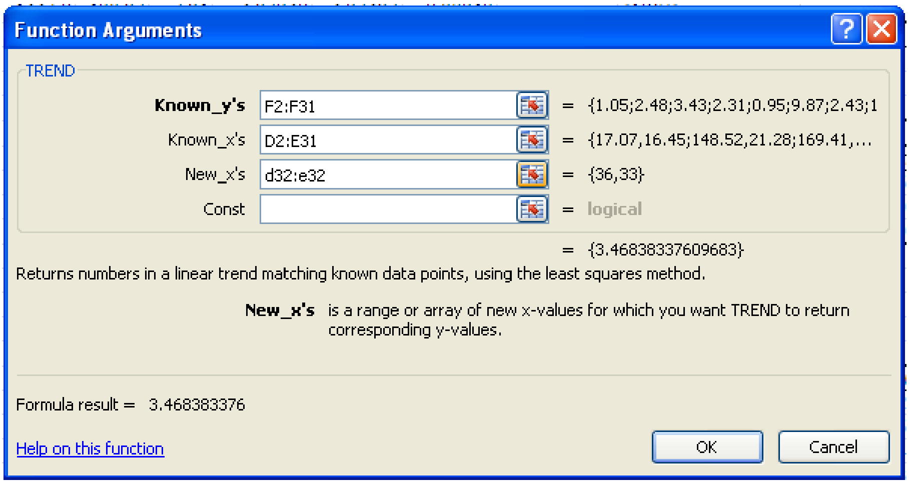
□ Standardized Residuals □ Line Fit Plots

□ Predictions □ Normal Probablity Plots

Remember that before using regression inference, the data must satisfy the regression model assumptions. The scatter of the data points about the line should be roughly the same over the entire range of the data. A plot of the residuals against each explanatory variable should not show any pattern. A histogram of the residuals should not show any major departures from normality.

The residual plots can be obtained by checking the Residual Plots box in the dialog box above or as shown in the earlier dialog box. A list of the residuals is obtained by checking the Residuals box in the same dialog box. A histogram of the residuals can be obtained by selecting **Data ➤ Data Analysis ➤ Histogram** or **Add-Ins ➤WHFStat ➤ Graphs ➤ Histogram** from the Excel menu.

To obtain prediction intervals for a new observation, we can plug the values of the explanatory variable into the least squares regression equation. Alternatively, we use Excel's **TREND** function to find the predicted value. The function arguments are the values of the response variable, the Known y's, the values of the explanatory variables, the Known x's, and the New x's. The values must be entered in the same order as in the regression equation. The values entered below correspond to Assets = 36 and Sales = 33. As shown below, the predicted value is 3.468.



TREND

Known_y's F2:F31 = {1.05;2.48;3.43;2.31;0.95;9.87;2.43;1

Known_x's D2:E31 = {17.07,16.45;148.52,21.28;169.41,...

New_x's d32:e32 = {36,33}

Const = logical

= {3.46838337609683}

Returns numbers in a linear trend matching known data points, using the least squares method.

New_x's is a range or array of new x-values for which you want TREND to return corresponding y-values.

Formula result = 3.468383376

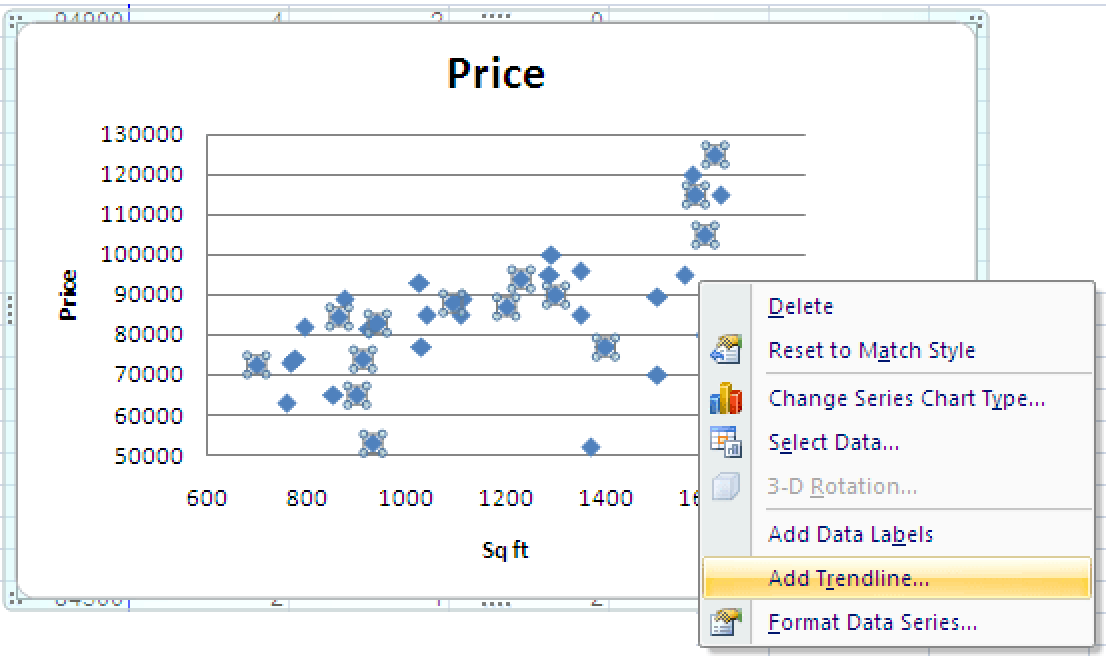
The individual prediction interval for an individual response is $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$. The value of $\mathrm{SE}_{\hat{y}}$ is also not given on the Excel output, but it is easily approximated by

$$\mathrm{SE}_{\hat{y}} \cong \sqrt{s^2 + \frac{1}{n}}.$$

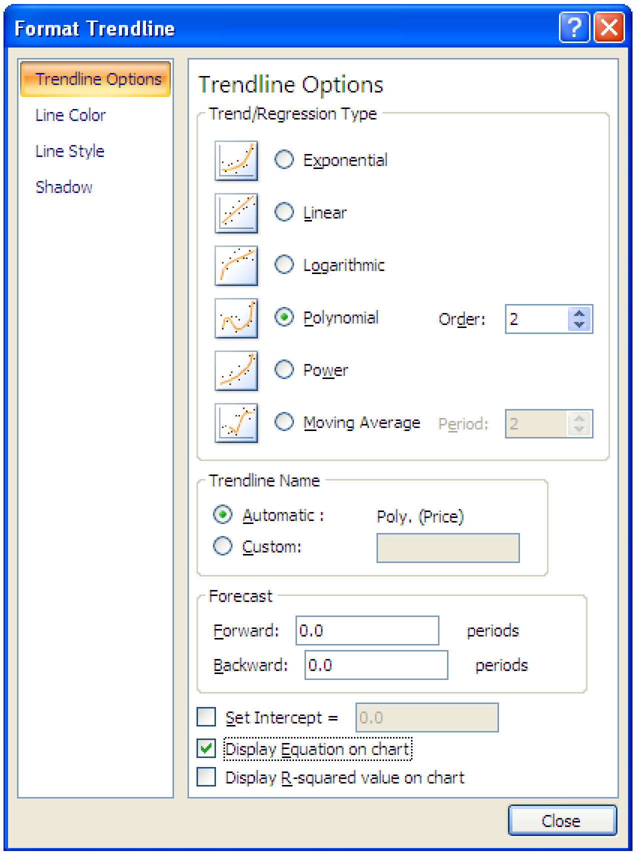
In this example, the approximation gives a value of 2.49453. The critical value from the *t* distribution with 27 degrees of freedom is obtained from the TINV function with the arguments, probability = 0.05, and degrees of freedom = 27, giving *t** = 2.052. Therefore, the approximate 95% prediction interval is 3.468 ± 5.118 or (–1.65, 8.59).

Model Building

When the relationship between a response and an explanatory variable is curved, a quadratic function may be an appropriate model. The data for the following example examines the relationship between the price and size of homes.



A scatterplot of the data suggests that a quadratic model may be reasonable for this data. A fitted-line plot of the quadratic model is a good way to see if a quadratic relationship is reasonable. To make a fitted-line plot, right click on any data point and select Add Trendline from the menu and select a polynomial, order 2 model as shown on the next page. This fits a model of the form $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$.



Trendline Options

Trend/Regression Type

○ Exponential

○ Linear

○ Logarithmic

● Polynomial Order: 2

○ Power

○ Moving Average Period: 2

Trendline Name

● Automatic : Poly. (Price)

○ Custom:

Forecast

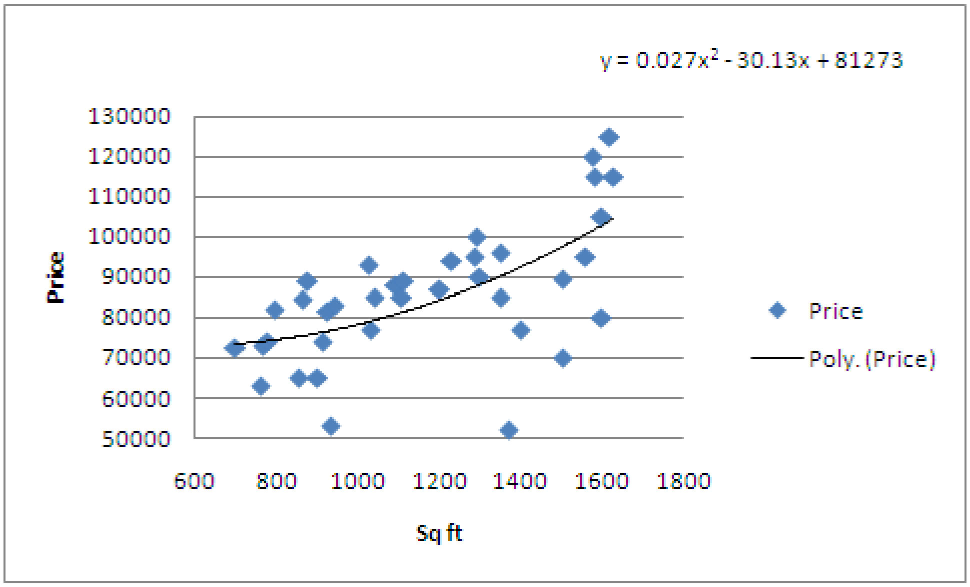
Forward: 0.0 periods

Backward: 0.0 periods

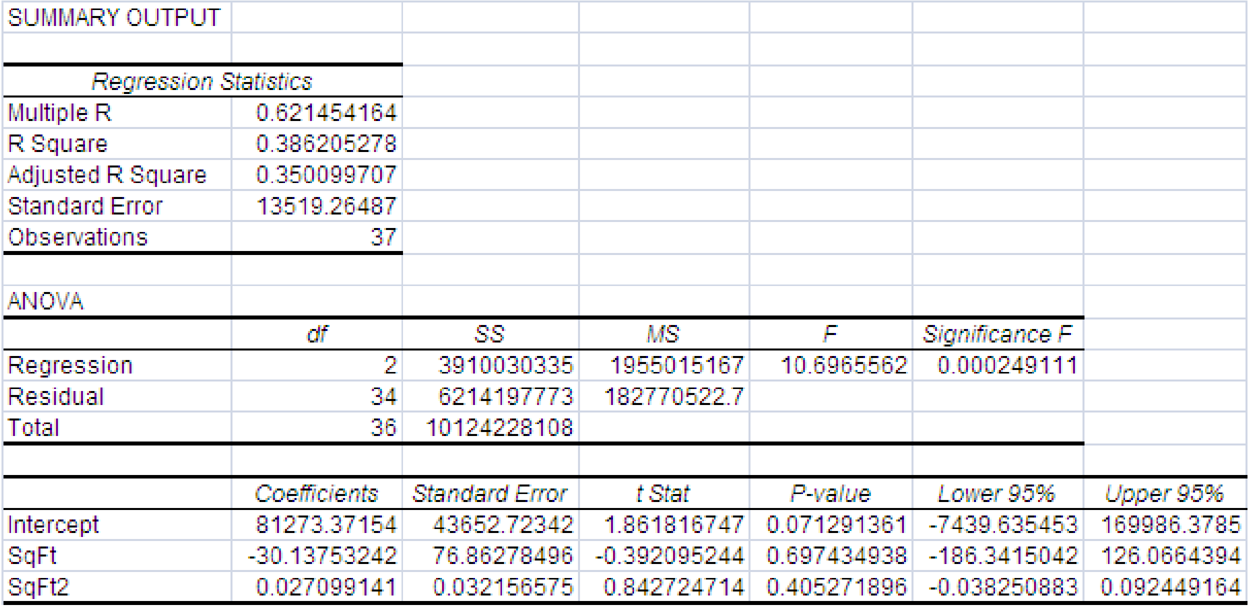
□ Set Intercept = 0.0

☑ Display Equation on chart

□ Display R-squared value on chart



To fully evaluate the model, make a variable for sqft squared. Then select **Data ➤ Data Analysis ➤ Regression** to run a regression using sqft and sqft^2.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.621454164
R Square	0.386205278
Adjusted R Square	0.350099707
Standard Error	13519.26487
Observations	37

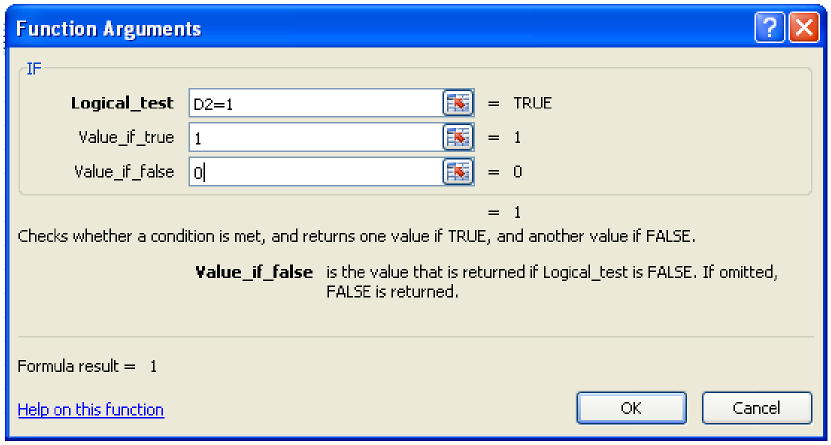
ANOVA

	df	SS	MS	F	Significance F
Regression	2	3910030335	1955015167	10.6965562	0.000249111
Residual	34	6214197773	182770522.7		
Total	36	10124228108			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	81273.37154	43652.72342	1.861816747	0.071291361	-7439.635453	169986.3785
SqFt	-30.13753242	76.86278496	-0.392095244	0.697434938	-186.3415042	126.0664394
SqFt2	0.027099141	0.032156575	0.842724714	0.405271896	-0.038250883	0.092449164

Notice that the individual *t* tests for both SqFt and SqFt2 are not significant, but the overall *F* test for the null hypothesis that both coefficients are zero is significant (*P*-value = 0.00025). This is due to a high degree of correlation between SqFt and SqFt2. As we will keep one of SqFt and SqFt2, it is natural to keep SqFt and drop its square.

When a categorical variable is used for prediction, it is usually best to made variables to indicate whether or not the value is in a particular category. In our example, the data has the number of bedrooms available to predict house prices. We define Bed1, an indicator variable that is equal to 1 if the variable bedrooms is equal to 1. Bed1 is equal to 0, otherwise. The **IF** function can be used to create indicator variables in Excel. The function arguments are the Logical test, in this case whether or not the variable Bedrooms is equal to 1, the Value if true, in this case, 1, and the Value if false, in this case 0. Copying the formula down the entire column will create the indicator variable.



Logical_test	D2=1	= TRUE
Value_if_true	1	= 1
Value_if_false	0	= 0

= 1

Checks whether a condition is met, and returns one value if TRUE, and another value if FALSE.

Value_if_false is the value that is returned if Logical_test is FALSE. If omitted, FALSE is returned.

Formula result = 1

We can create four indicator variables, Bed1, Bed2, Bed3, and Bed4, because there are four possible bedroom sizes. Alternative variables are possible. For example, we can create a variable that is equal to 1 for homes with 3 or more bedrooms and equal to 0 otherwise. Interaction effects are easily modeled by multiplying two variables together.

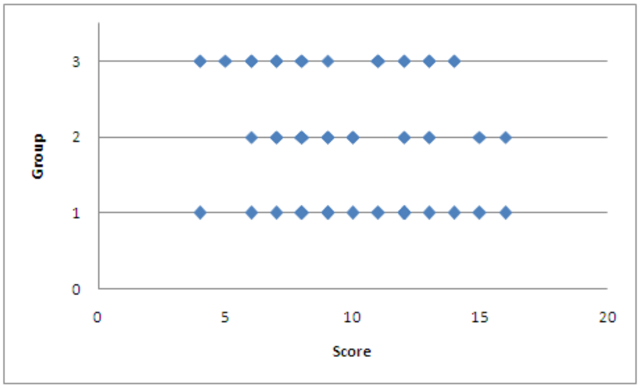
Variable Selection

Excel can help you to select a satisfactory model. Generally, you want a model for which the overall *F* test for the null hypothesis that all coefficients are zero is significant and all the individual *t* tests for the explanatory variable are significant. In choosing between different models, you want the model with the highest Adjusted R Squared value.

In Excel, the easiest way to select a model is by backward elimination. Select **Data ➤ Data Analysis ➤ Regression** or **Add-Ins ➤ WHFStat ➤ Correlation and Regression ➤ Regression** from the Excel menu. Begin with a model containing all the explanatory variables of interest. Then, at each step the variable with highest *P*-value is deleted. Since you need a single cell range for the *x* values, all of the explanatory variables must continue to be adjacent as variables are eliminated. Continue the procedure until the overall *F*-test is significant and all the individual *t*-tests for the remaining explanatory variable are significant.

CHAPTER

12



One-Way Analysis of Variance

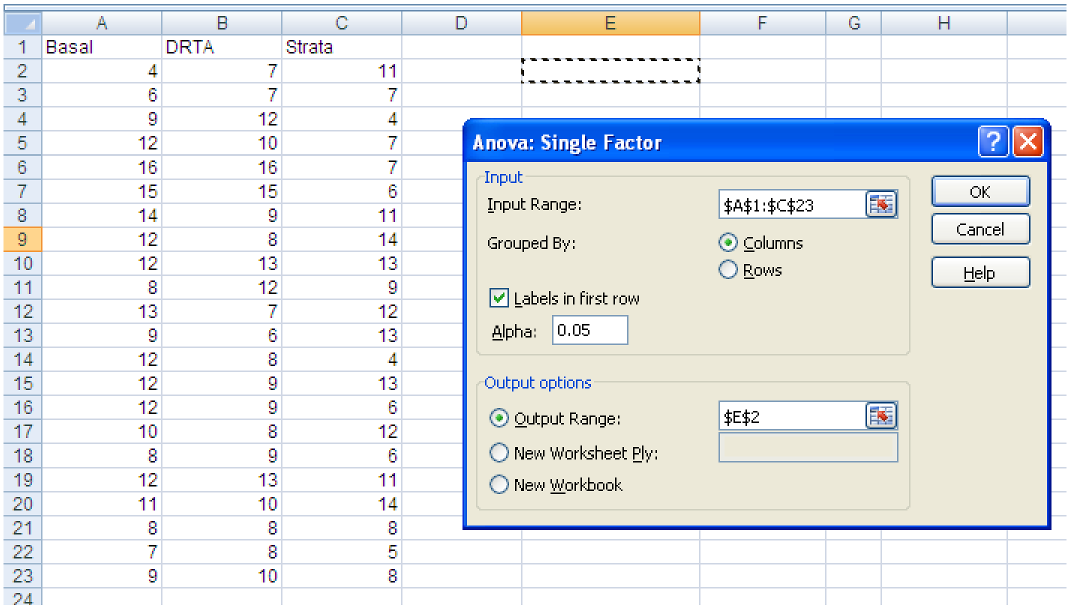
One-Way Analysis of Variance

Excel can perform a one-way analysis of variance test to compare means of different populations. The response variable must be numeric. The data should be entered with each population in separate columns (or rows) on a worksheet. To perform a one-way analysis of variance with stacked data, choose

Data ➤ Data Analysis ➤ ANOVA: Single Factor

from the Excel menu.

Our example comes from a study conducted to compare three educational approaches (basal, DRTA, and strategies) to improve children's reading comprehension. We want to test the null hypothesis that the three groups represent three populations that all have the same mean score on the pretest. The data are given in CA14_001.MTW. The data are arranged with one row for each student. The teaching method is given in one column and the student's pretest score given in a second column. To analyze these data, select **Stat ➤ ANOVA ➤ One-way** from the menu. In the dialog box enter the columns containing the test scores in the Input Range box, check the Chart Titles box, indicate where the output should go, click OK.



	A	B	C	D	E	F	G	H
1	Basal	DRTA	Strata					
2	4	7	11					
3	6	7	7					
4	9	12	4					
5	12	10	7					
6	16	16	7					
7	15	15	6					
8	14	9	11					
9	12	8	14					
10	12	13	13					
11	8	12	9					
12	13	7	12					
13	9	6	13					
14	12	8	4					
15	12	9	13					
16	12	9	6					
17	10	8	12					
18	8	9	6					
19	12	13	11					
20	11	10	14					
21	8	8	8					
22	7	8	5					
23	9	10	8					
24								

Anova: Single Factor

Input

Input Range: $A$1:$C$23

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

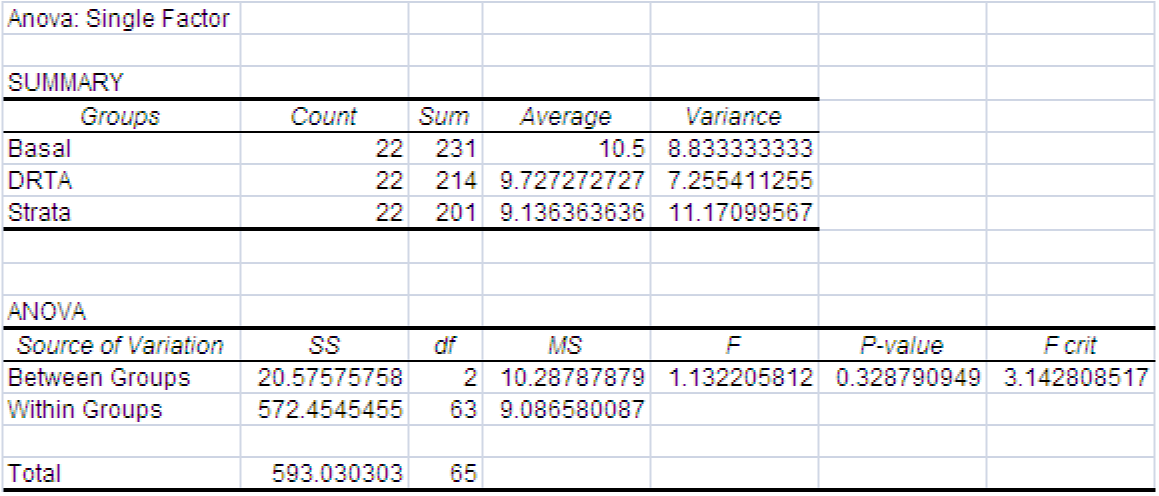
It is important to check that the assumptions of one-way analysis of variance are satisfied. Specifically, the populations are normal with possibly different means and the same variance. Histogram of the variables serve to detect outliers or extreme deviations from Normality. Compute the ratio of the largest to the smallest sample standard deviation. If this ratio is less than 2 and the histograms are satisfactory, the assumptions of ANOVA are satisfied.

Call the mean test scores for the three educational approaches μ_1, μ_2, and μ_3. We want to test the null hypothesis that there are *no differences* among the test scores for the three groups:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$

The alternative is that there is some difference.

$$H_a\text{: not all of } \mu_1, \mu_2, \text{and } \mu_3 \text{ are equal}$$



Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Basal	22	231	10.5	8.833333333		
DRTA	22	214	9.727272727	7.255411255		
Strata	22	201	9.136363636	11.17099567		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	20.57575758	2	10.28787879	1.132205812	0.328790949	3.142808517
Within Groups	572.4545455	63	9.086580087			
Total	593.030303	65				

The output provides the ANOVA table. The columns in this table are labeled Source of Variation, SS (sum of squares), df (degrees of freedom), MS (mean square), *F*, *P*-value, and *F* crit. The rows in the table are labeled Between Groups, Within Groups, and Total. Consider our model

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The Between Groups row corresponds to the FIT term, the Within Groups row corresponds to the RESIDUAL term, and the Total row corresponds to the DATA term. Notice that both the degrees of freedom and the sum of squares add to the value in the Total row.

The pooled standard deviation can be computed from the ANOVA table using the sum of squares and degrees of freedom for the Within Groups row. That is,

$$s_p^2 = \frac{\text{SS}}{\text{DF}} = \frac{572.45}{63} = 9.09$$

which implies that $s_p = 3.014$.

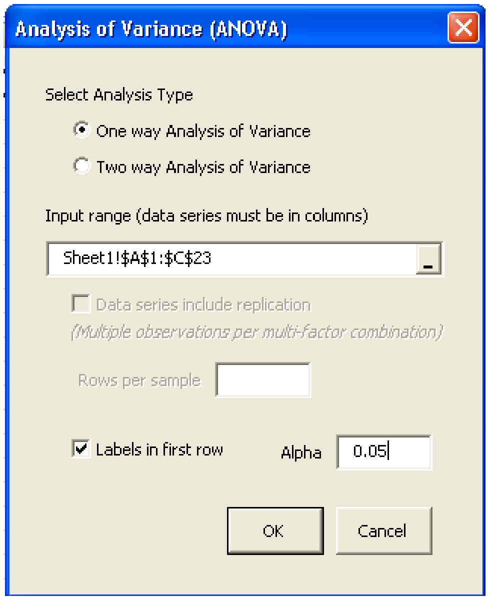
The *F* statistic is given in the ANOVA table. If H_0 is true, the *F* statistic has an $F(\text{DFG}, \text{DFE})$ distribution, where DFG stands for degrees of freedom for groups and DFE stands for degrees of freedom for error. $\text{DFG} = I - 1$, the number of groups minus 1. $\text{DFE} = N - I$, the number of observations minus the number of groups. The *P*-value for this distribution is also given above. In this example, the *P*-value is given as 0.329. A *P*-value this large does not give us evidence against the hypothesis that the means are all equal.

For information purposes, the output from one-way analysis of variance provides the mean and variance for each group. Individual 95% confidence intervals for the means are of the form

$$\left(\bar{x}_i - t^* \frac{s_p}{\sqrt{n_i}}, \bar{x}_i + t^* \frac{s_p}{\sqrt{n_i}} \right)$$

where $\bar{x}_i$ and n_i are the sample mean and sample size for level *i*, s_p= Pooled StDev is the pooled estimate of the common standard deviation, and *t** is the value from a *t* table corresponding to 95% confidence and the degrees of freedom for within groups.

Alternatively, the exact same results can be obtained by selecting **Add-Ins ➤WHFStat ➤ Analysis of Variance - ANOVA** from the Excel menu. Select the radio button for One way Analysis of Variance and fill in the dialog box as shown below.



Analysis of Variance (ANOVA)

Select Analysis Type

● One way Analysis of Variance

○ Two way Analysis of Variance

Input range (data series must be in columns)

Sheet1!$A$1:$C$23

□ Data series include replication

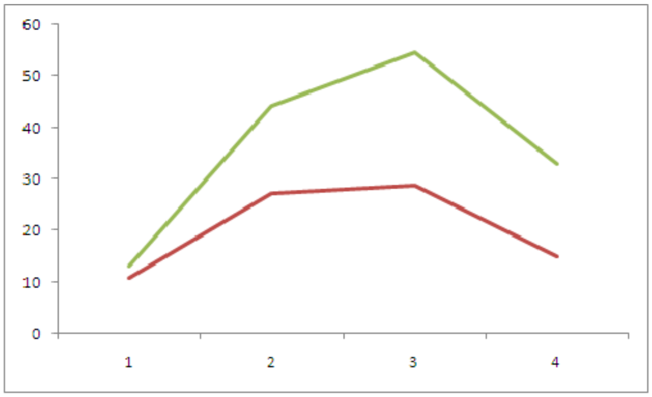
(Multiple observations per multi-factor combination)

Rows per sample

☑ Labels in first row

Alpha 0.05

CHAPTER 13



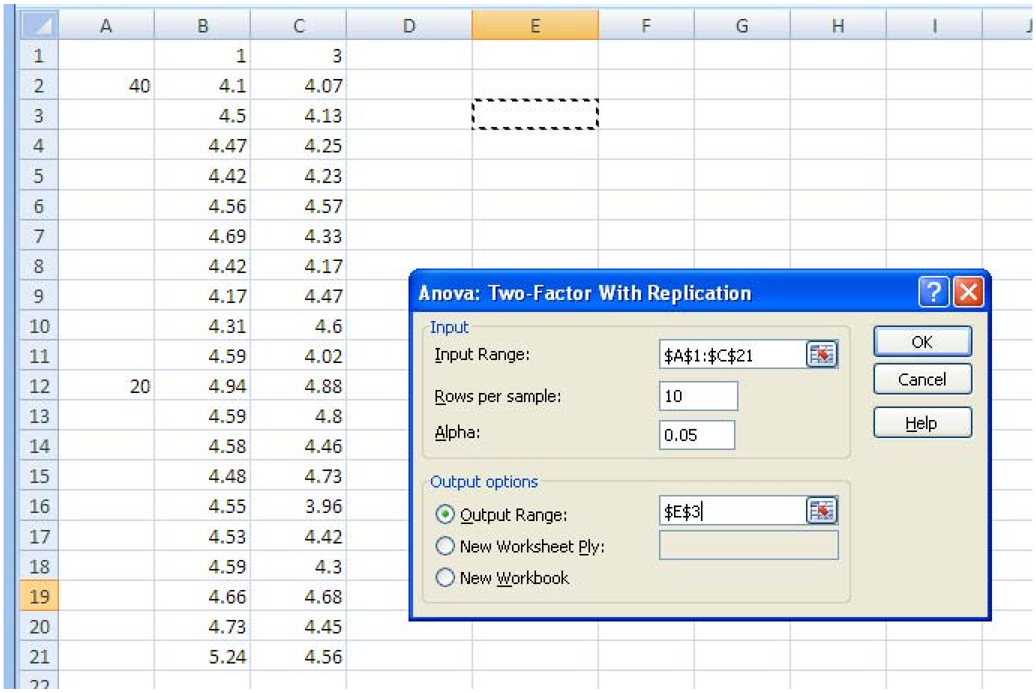
Two-Way Analysis of Variance

Cross Tabulation

The following data is from a study by researchers conducted on students enrolled in an introductory management course at a large midwestern university. The purpose of the study is to examine if the frequency with which a supermarket product is offered at a discount and the percent reduction affect the price that customers expect to pay for the product. For 10 weeks 160 subjects received information about the products. The treatment conditions corresponded to the number of promotions during this 10-week period and the percent that the product was discounted. Ten students were randomly assigned to each treatment. The case study examines the price customers expect to pay for two levels of promotions (1 and 3) and two levels of discount (40% and 20%). Thus, we have a two-way analysis of variance with each of the factors having two levels and 10 observations in each of the four treatment combinations. The data are on the following page.

Number of promotions	Percent discount	Expected price									
1	40	4.10	4.50	4.47	4.42	4.56	4.69	4.42	4.17	4.31	4.59
1	20	4.94	4.59	4.58	4.48	4.55	4.53	4.59	4.66	4.73	5.24
3	40	4.07	4.13	4.25	4.23	4.57	4.33	4.17	4.47	4.60	4.02
3	20	4.88	4.80	4.46	4.73	3.96	4.42	4.30	4.68	4.45	4.56

The data must be entered with one variable for rows and the other variable for columns as shown below. The input range includes the variable titles in addition to the sample results. Since replication is allowed, you must also specify the number of rows per sample, in this case 10. Select **Data ➤ Data Analysis ➤ Anova: Two Factor With Replication**, fill out the Input and Output Range, the Rows per sample, and click OK.



	A	B	C
1		1	3
2	40	4.1	4.07
3		4.5	4.13
4		4.47	4.25
5		4.42	4.23
6		4.56	4.57
7		4.69	4.33
8		4.42	4.17
9		4.17	4.47
10		4.31	4.6
11		4.59	4.02
12	20	4.94	4.88
13		4.59	4.8
14		4.58	4.46
15		4.48	4.73
16		4.55	3.96
17		4.53	4.42
18		4.59	4.3
19		4.66	4.68
20		4.73	4.45
21		5.24	4.56

Anova: Two-Factor With Replication

Input

Input Range: $A$1:$C$21

Rows per sample: 10

Alpha: 0.05

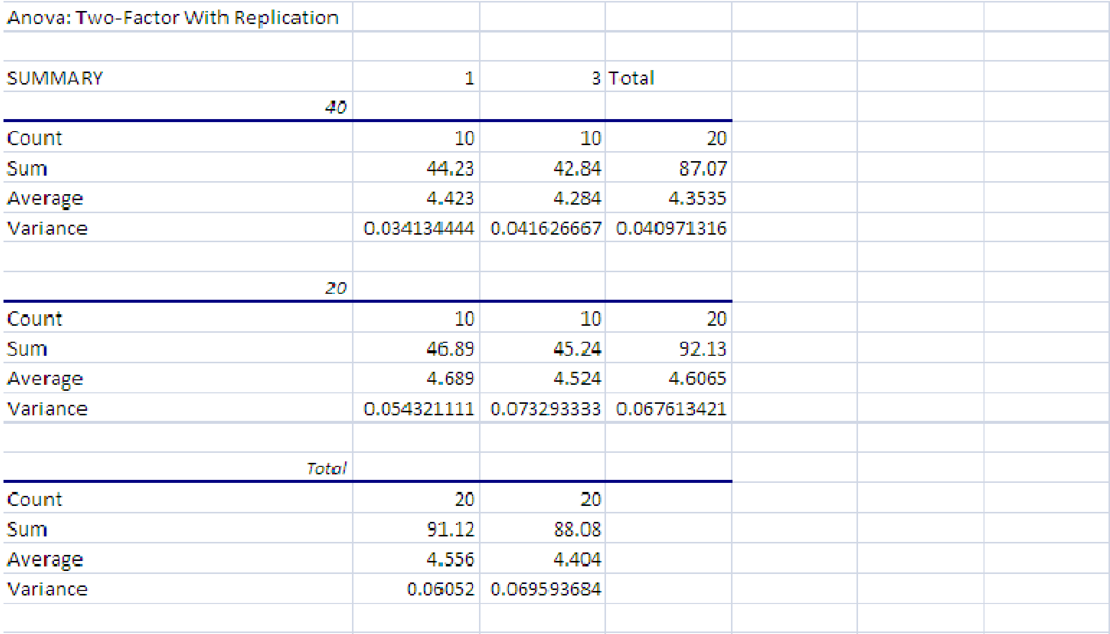
Output options

● Output Range: $E$3

○ New Worksheet Ply:

○ New Workbook

Excel provides numerical summaries of the data. The first section of the data gives the Count, Sum, and Variance of the different groups of data. The variance is the square of the sample standard deviation.

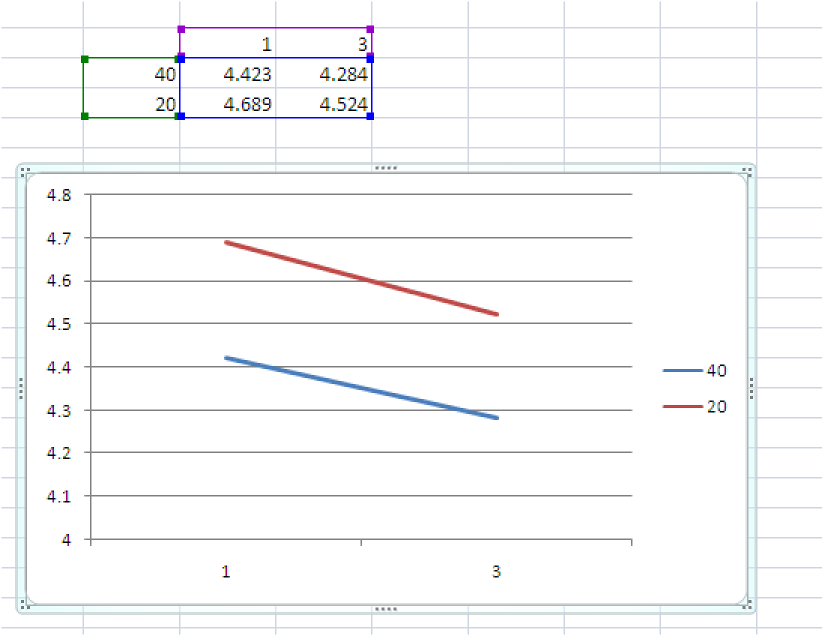


Anova: Two-Factor With Replication

SUMMARY	1	3	Total
40			
Count	10	10	20
Sum	44.23	42.84	87.07
Average	4.423	4.284	4.3535
Variance	0.034134444	0.041626667	0.040971316
20			
Count	10	10	20
Sum	46.89	45.24	92.13
Average	4.689	4.524	4.6065
Variance	0.054321111	0.073293333	0.067613421
Total			
Count	20	20	
Sum	91.12	88.08	
Average	4.556	4.404	
Variance	0.06052	0.069593684	

Interactions Plots

Interactions plots can be used to describe the interaction effects in a two-way analysis of variance. The averages for the promotions and discount data can be plotted in an interactions plot. Enter the appropriate summary data in Excel, highlight the data and select **Insert ➤ Line ➤ 2-D Line** from the menu.



	1	3
40	4.423	4.284
20	4.689	4.524

The two lines are approximately parallel, suggesting that there is little or no interaction between promotion and discount.

The ANOVA Table

The results of a two-way analysis of variance are given in the ANOVA table. The total variation (SS) is split among the two main effects, the interaction, and the error. The degrees of freedom (df) is split the same way. If the sample size is the same for all groups, as in our example, then

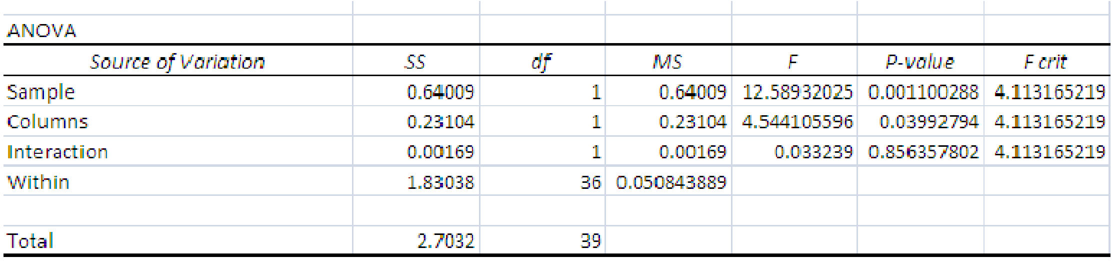
$$SST = SSA + SSB + SSAB + SSE$$
$$DFT = DFA + DFB + DBAB + DFE$$

Where A and B are the main effects and AB is the interaction. Here is the form of the ANOVA table.

Source	Degrees of freedom	Sum of squares	Mean square	F
A	$I - 1$	SSA	SSA/DFA	MSA/MSE
B	$J - 1$	SSB	SSB/DFB	MSB/MSE
AB		SSAB	SSAB/DFAB	MSAB/MSE
Error		SSE	SSE/DBE	
Total	$N - 1$	SST	SST/DFT	

There are three null hypotheses in two-way analysis of variances, with an *F* test for each. We test for significance of the two main effects and the interaction. The bottom section of the Excel output provides the ANOVA output. To the right of the *F* column are columns labeled *P*-value and *F* crit. The *P*-value is the result of the significance test and the *F* crit value is the critical value from the *F* distribution for the selected significance level, in this case 0.05.

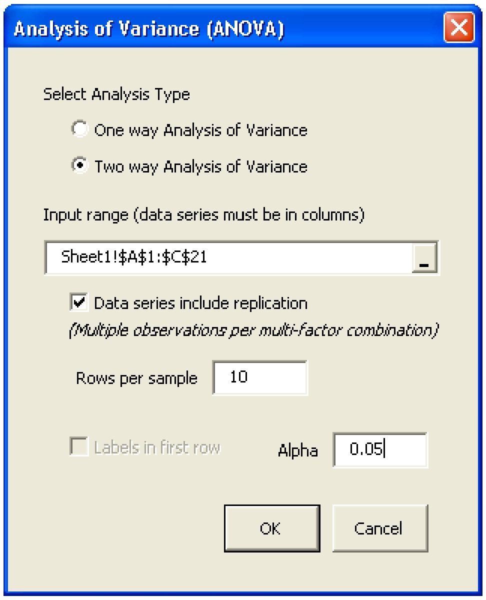
As expected, in our example, the interaction is not statistically significant (*P*-value = 0.856.) On the other hand, the main effects of discount (sample) and promotion (columns) are significant with *P*-values 0.001 and 0.040, respectively.



ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.64009	1	0.64009	12.58932025	0.001100288	4.113165219
Columns	0.23104	1	0.23104	4.544105596	0.03992794	4.113165219
Interaction	0.00169	1	0.00169	0.033239	0.856357802	4.113165219
Within	1.83038	36	0.050843889			
Total	2.7032	39				

Alternatively, the exact same results can be obtained by selecting **Add-Ins ➤WHFStat ➤ Analysis of Variance - ANOVA** from the Excel menu. Select the radio button for Two way Analysis of Variance and fill in the dialog box as shown below.



Analysis of Variance (ANOVA)

Select Analysis Type

○ One way Analysis of Variance

● Two way Analysis of Variance

Input range (data series must be in columns)

Sheet1!$A$1:$C$21

☑ Data series include replication
(Multiple observations per multi-factor combination)

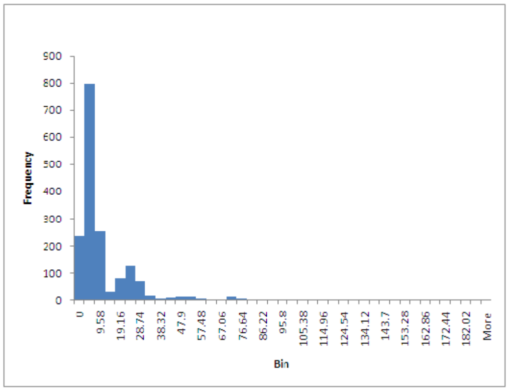
Rows per sample: 10

□ Labels in first row

Alpha: 0.05

CHAPTER

14



Bootstrap Methods and Permutation Tests

Bootstrap methods are based on resampling from data and were first introduced in 1979 for estimating the standard error of the estimate of a parameter. Resampling methods allow us to quantify uncertainty by calculating standard errors and confidence intervals and performing significance tests. They require fewer assumptions than traditional methods and generally give more accurate answers (sometimes very much more accurate). Moreover, resampling lets us tackle new inference settings easily. Resampling also helps us understand the concepts of statistical inference.

The bootstrap is best carried out with specialized software that does simulation well and quickly. Excel is not well suited to large-scale simulations. However, we will use Excel to demonstrate the principles of the bootstrap with small-scale simulations.

The Bootstrap Procedure

Step 1 - Resample. Create hundreds of new samples, called bootstrap samples or bootstrap samples resamples, by sampling with replacement from the original random sample. Each resample is the same size as the original random sample.

Step 2 - Calculate the bootstrap distribution. Calculate the statistic for each resample. The distribution of these resample statistics is called a bootstrap distribution. If we want to estimate the population mean μ, the statistic is the sample mean $\bar{x}$.

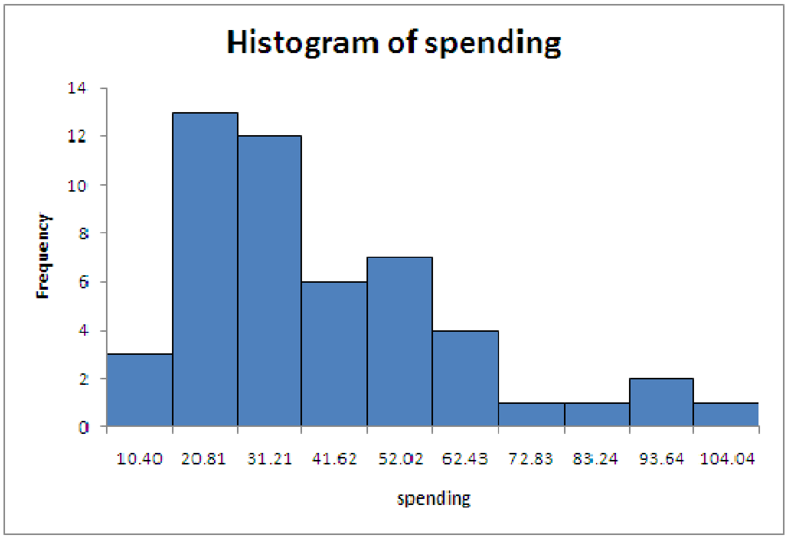
Step 3 - Use the bootstrap distribution. The bootstrap distribution gives information about the shape, center, and spread of the sampling distribution of the statistic. The bootstrap standard error of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Bootstrap Distribution

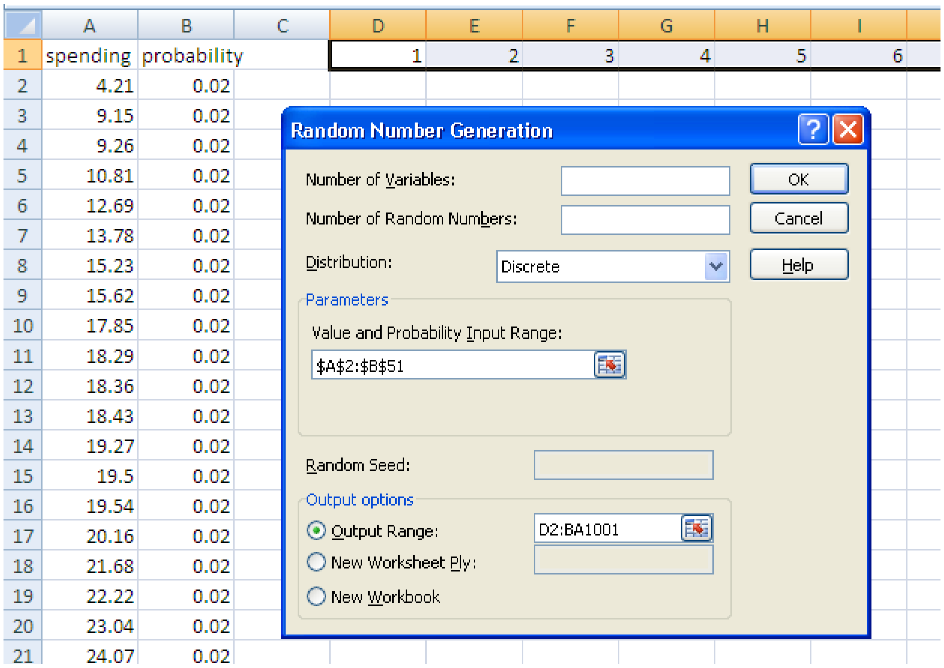
We show how to create a bootstrap distribution for the spending of a small sample of shoppers. Here are the dollar amounts spent by 50 consecutive shoppers at a supermarket. We are willing to regard this as an SRS of all shoppers at this market.

4.21	9.15	9.26	10.81	12.69	13.78	15.23	15.62	17.85	18.29
18.36	18.43	19.27	19.5	19.54	20.16	21.68	22.22	23.04	24.07
24.58	25.13	26.24	26.26	27.04	28.06	28.08	28.38	32.03	35.28
36.37	38.64	38.99	41.02	42.97	44.08	44.67	45.4	46.69	49.64
51.3	52.75	54.8	59.07	61.22	70.32	82.7	85.76	88.77	97.85

To begin with, we will select **Add-Ins ➤ WHFStat ➤ Graphs ➤ Histogram** from the Excel menu. Alternatively, we can select **Data ➤ Data Analysis ➤ Histogram.** Either way, we fill in the dialog box to examine the shape of the shopping distribution. The histogram on the following page shows that the data is skewed to right.



To create the bootstrap distribution, we will sample with replacement 1000 times and create 1000 samples each of size 50. To sample, enter the observations into a worksheet and enter the probability of each, i.e., 0.02. Select **Data ➤ Data Analysis ➤ Random Number Generation** from the Excel menu. Fill out the dialog box by entering the data and probabilities (without column titles) and the location of the samples. In the dialog box shown below, we choose the Output Range D2:BA1001, a range that will give us 1000 rows each of sample size 50.



	A	B	C	D	E	F	G	H	I
1	spending	probability		1	2	3	4	5	6
2	4.21	0.02							
3	9.15	0.02							
4	9.26	0.02							
5	10.81	0.02							
6	12.69	0.02							
7	13.78	0.02							
8	15.23	0.02							
9	15.62	0.02							
10	17.85	0.02							
11	18.29	0.02							
12	18.36	0.02							
13	18.43	0.02							
14	19.27	0.02							
15	19.5	0.02							
16	19.54	0.02							
17	20.16	0.02							
18	21.68	0.02							
19	22.22	0.02							
20	23.04	0.02							
21	24.07	0.02							

Random Number Generation

Number of Variables:

Number of Random Numbers:

Distribution: Discrete

Parameters

Value and Probability Input Range: $A$2:$B$51

Random Seed:

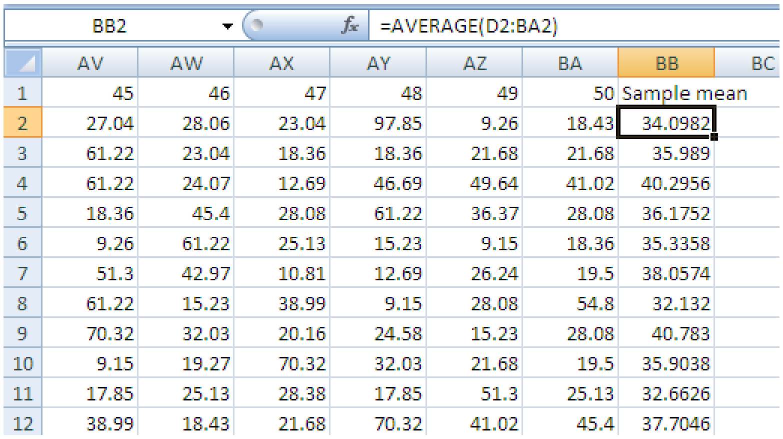
Output options

● Output Range: D2:BA1001

○ New Worksheet Ply:

○ New Workbook

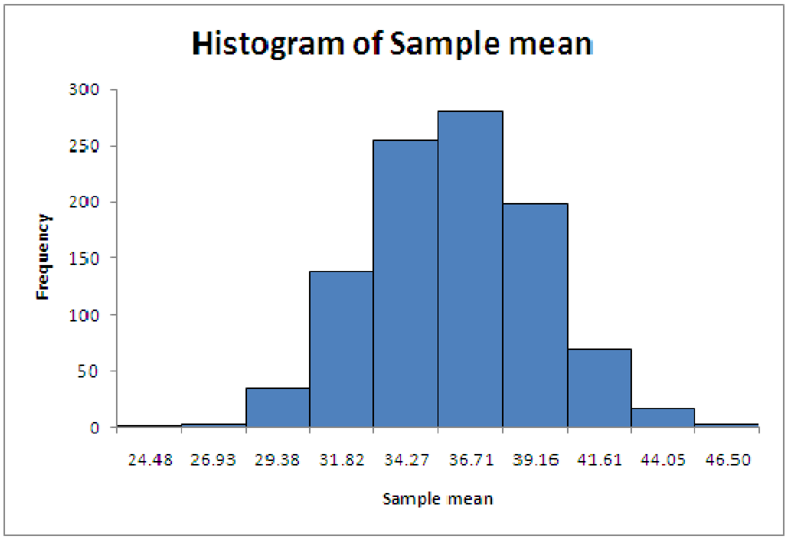
We calculate the sample mean for each row using Excel's **AVERAGE** function as shown on the following page.



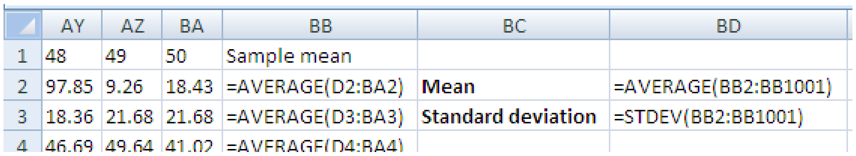
BB2 | =AVERAGE(D2:BA2)

	AV	AW	AX	AY	AZ	BA	BB	BC
1	45	46	47	48	49	50	Sample mean	
2	27.04	28.06	23.04	97.85	9.26	18.43	34.0982	
3	61.22	23.04	18.36	18.36	21.68	21.68	35.989	
4	61.22	24.07	12.69	46.69	49.64	41.02	40.2956	
5	18.36	45.4	28.08	61.22	36.37	28.08	36.1752	
6	9.26	61.22	25.13	15.23	9.15	18.36	35.3358	
7	51.3	42.97	10.81	12.69	26.24	19.5	38.0574	
8	61.22	15.23	38.99	9.15	28.08	54.8	32.132	
9	70.32	32.03	20.16	24.58	15.23	28.08	40.783	
10	9.15	19.27	70.32	32.03	21.68	19.5	35.9038	
11	17.85	25.13	28.38	17.85	51.3	25.13	32.6626	
12	38.99	18.43	21.68	70.32	41.02	45.4	37.7046	

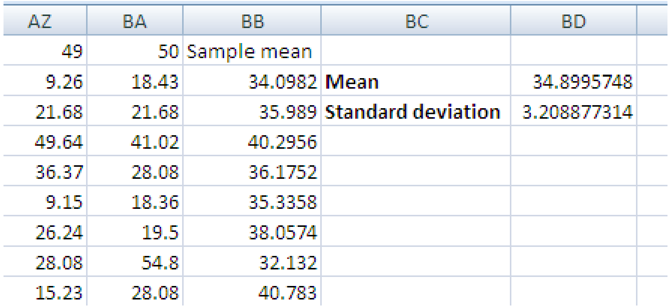
The central limit theorem says that the sampling distribution of the sample mean $\bar{x}$ becomes Normal as the sample size increases. To find out if the sampling distribution is Normal for $n = 50$, we make a histogram of the bootstrap distribution of the mean. As we can see from the following histogram, the bootstrap distribution is approximately Normal.



We can calculate the mean and standard error for the bootstrap distribution using Excel's AVERAGE and STDEV function as shown below.



	AY	AZ	BA	BB	BC	BD
1	48	49	50	Sample mean		
2	97.85	9.26	18.43	=AVERAGE(D2:BA2)	**Mean**	=AVERAGE(BB2:BB1001)
3	18.36	21.68	21.68	=AVERAGE(D3:BA3)	**Standard deviation**	=STDEV(BB2:BB1001)
4	46.69	49.64	41.02	=AVERAGE(D4:BA4)		



AZ	BA	BB	BC	BD
49	50	Sample mean		
9.26	18.43	34.0982	**Mean**	34.8995748
21.68	21.68	35.989	**Standard deviation**	3.208877314
49.64	41.02	40.2956		
36.37	28.08	36.1752		
9.15	18.36	35.3358		
26.24	19.5	38.0574		
28.08	54.8	32.132		
15.23	28.08	40.783		

The mean of the bootstrap distribution is 34.90 and the bootstrap standard error is 3.209. All of these values will differ if you repeat 1000 resamples, because resamples are drawn at random.

The bootstrap estimate of bias is the mean of the bootstrap distribution minus the statistic for the original data. For the shopping data the mean of the original sample is 34.945. Therefore, the bootstrap estimate of bias for these resamples is –0.045.

When a bootstrap distribution is approximately Normal and has small bias, an approximate level *C* confidence interval is

$$\text{statistic} \pm t^{*}\text{SE}_{\textit{boot, statistic}}$$

A 95% confidence interval for the population mean for the shopping data is therefore 34.90 ± 1.96(3.209) = (18.61, 41.19).

Hypothesis Tests

The following example considers a test to determine whether new "directed reading activities" improved the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) score. The study assigned students at random to either the new method (treatment group, 21 students) or traditional teaching methods (control group, 23 students). Their DRP scores are given below.

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

The statistic that measures the success of the new method is the difference in mean DRP scores,

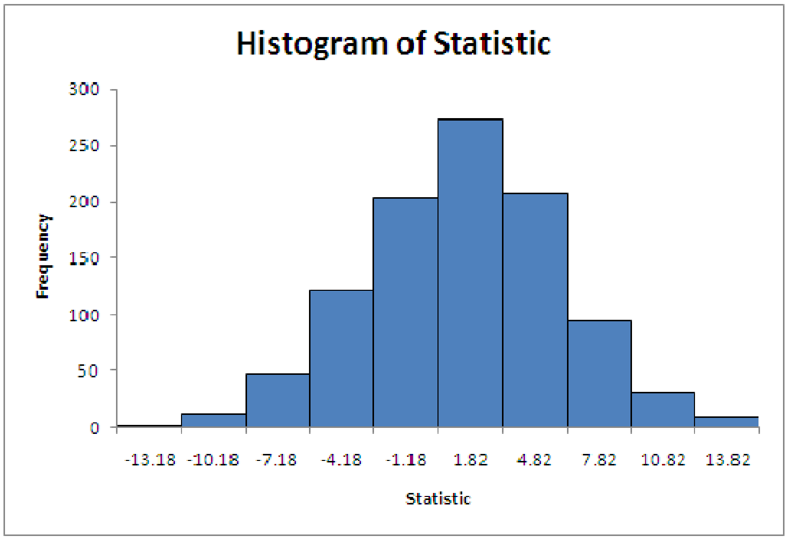
$$\bar{x}_{treatment} - \bar{x}_{control}.$$

The null hypothesis is "no difference" between the two methods. We will use the bootstrap distribution to do a hypothesis test using Excel.

To create the bootstrap distribution for the difference between the two methods, we create a discrete probability distribution using the scores for all 44 students. The simplest way is to list all 44 scores and an associated probability of 1/44. . Select **Data ➤ Data Analysis ➤ Random Number Generation** from the Excel menu. Fill out the dialog box by entering the data and probabilities (without column titles) and the location of the samples to create 999 re-samples each of size 44 into 999 rows. The output range for our samples was D2:AV1000.

Since the null hypothesis says that there is no difference between the control group and the treatment group, we arbitrarily select the first 21 columns to be the control group and the remaining 23 columns to be the treatment group. We then calculate the row mean for each group and calculate the difference for each row. That is, we calculate =AVERAGE(E2:Y2)–AVERAGE(Z2:AV2) for the first sample. The formula can be copied to calculate the statistic for the remaining samples. This is the bootstrap distribution of the statistic $\bar{x}_{treatment} - \bar{x}_{control}$ under the condition that the null hypothesis is true.

Select **Add-Ins ➤ WHFStat ➤ Graph ➤ Histogram** or **Data ➤ Data Analysis ➤ Histogram** from the Excel menu and make a histogram of the values of the statistic. We see that the bootstrap distribution is nearly Normal.



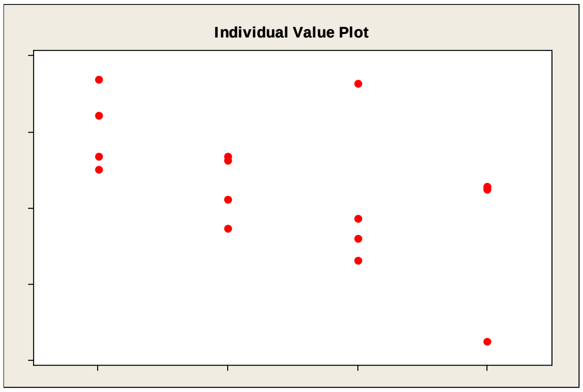
The value of the statistic actually observed in the study was

$$\bar{x}_{treatment} - \bar{x}_{control} = 51.476 - 41.522 = 9.954$$

Count the number of values that exceed 9.954. If your data is in column AW, this is easily done by entering =COUNTIF(AW2:AW1000,">9.954") into a cell. For these resamples, 17 of the 1000 resamples gave a value 9.954 or larger. This value will differ if you repeat 1000 resamples, because resamples are drawn at random. The proportion of samples that exceed the observed value 9.954 is 17/1000 or 0.017. Recall from Chapter 8 that we can improve the estimate of a population proportion by adding two successes and two failures to the sample. We can similarly improve the estimate of the *P*-value by adding one sample result above the observed statistic. The final bootstrap test estimate of the *P*-value is $\frac{17+1}{1000+1} = \frac{18}{1000}$ = 0.018. This is a one-sided *P*-value. The data give good evidence that the new method beats the standard method.

CHAPTER

15



Nonparametric Tests

The investigations in previous chapters into various methods of inference have proven to be robust and not very sensitive to a moderate lack of normality, especially when the sample size is fairly large. There are several options for dealing with nonnormal distributions. This chapter deals with nonparametric methods, defined as inference procedures that do not require any specific type of population distribution. The analysis is done using existing Excel tools only.

The Wilcoxon Rank Sum Test

One type of nonparametric test is based on the rank (ordered position) of each observation in the dataset. The Wilcoxon rank sum test addresses the common two sample problem. The rank tests studied in this chapter focus on the center of a population. If a population has a normal distribution, the center is the mean. For a skewed distribution, the center is best represented by the median. The hypotheses for the rank tests will replace the mean with the median as the measure of the center.

Observations are ranked by sorting them in order from smallest to largest, combining observations from both datasets (for a two-sample problem). The rank of each observation is its position in this ordered list, starting with rank 1 for the smallest observation. If an SRS is taken from each population, there are a total of N observations in all, where $N = n_1 + n_2$. The sum W of the ranks for the first sample is the Wilcoxon rank sum statistic. The mean of W is defined by:

$$\mu_W = \frac{n_1(N+1)}{2}$$

The standard deviation is defined by:

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

The Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions when the rank sum *W* is far from its mean. The hypotheses tested are:

H_0: No difference in the distribution of yields against the one sided alternative

H_a: Yields are systematically higher in weed free plots.

The rank sum statistic W becomes approximately Normal as the two sample sizes increase. We can form the test statistic by standardizing W.

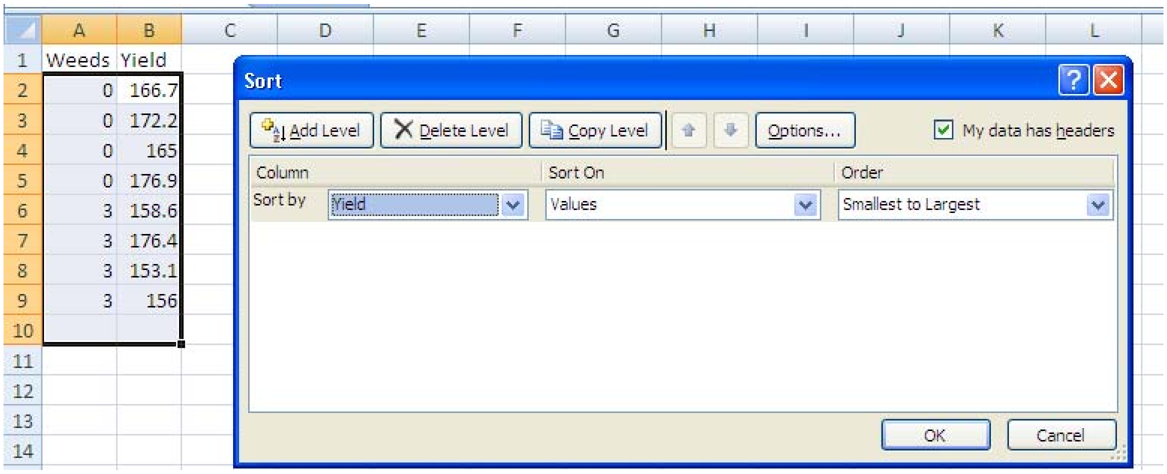
$$z = \frac{W - \mu_W}{\sigma_W}$$

Use standard Normal probability calculations to find *P*-values for this statistic. Because *W* takes only whole number values, the continuity *correction* improves the accuracy of the approximation.

The following example follows the setting for the Wilcoxon rank sum test. A researcher planted corn in eight plots of ground, then weeded the corn to allow no weeds in four plots and exactly three weeds per meter in the other four plots. The table here shows yields of corn (bushels per acre) in each of the plots.

Weeds per meter	Yield (bu/acre)			
0	166.7	172.2	165.0	176.9
3	158.6	176.4	153.1	156.0

The Wilcoxon rank sum test assumes that the data are independent random samples from two populations that have the same shape (hence the same variance) and a scale that is at least ordinal. The data need not be from normal populations. The first step in the calculation is to sort the data as shown and provide the ranks.



	A	B
1	Weeds	Yield
2	0	166.7
3	0	172.2
4	0	165
5	0	176.9
6	3	158.6
7	3	176.4
8	3	153.1
9	3	156
10		

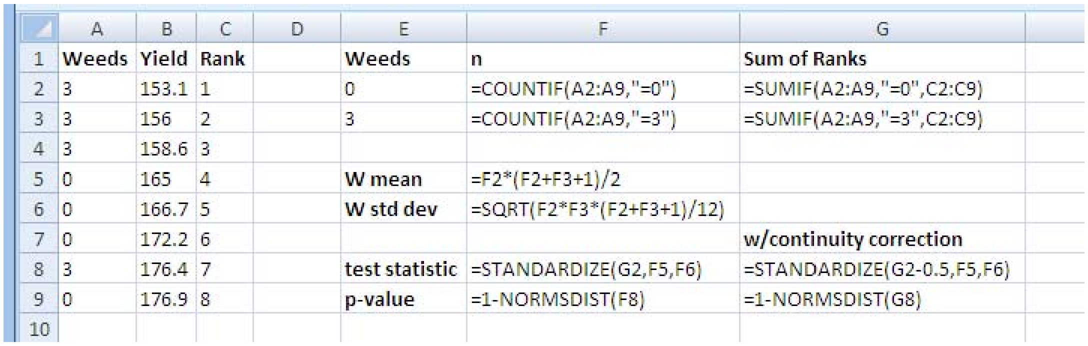
Sort

Add Level | Delete Level | Copy Level | Options... | ☑ My data has headers

	Column	Sort On	Order
Sort by	Yield	Values	Smallest to Largest

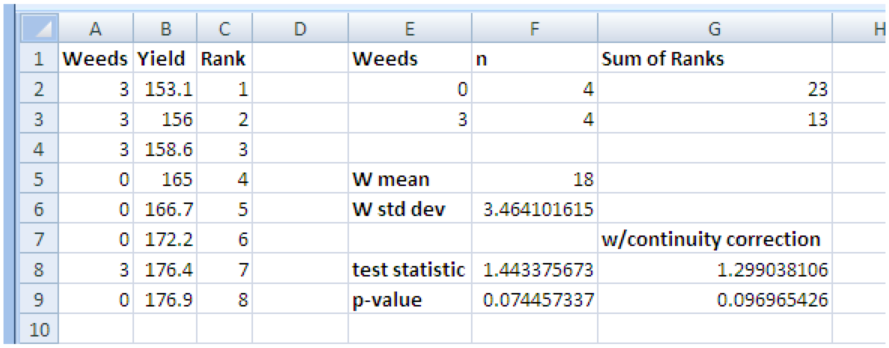
OK | Cancel

Formulas for the calculations are shown in the following spreadsheet. Since the ranks are a simple series, 1, 2, 3, 4,..., type **1** and **2** in the first two cells. Select the cells that contain the starting values. Drag the fill handle down to fill in the remaining ranks. The **COUNTIF** function counts the number of cells with 0 or 3 weeds. The **ADDIF** function adds the values of the ranks with 0 or 3 weeds. The test statistic is calculated using the **STANDARDIZE** function, and the *P*-value is calculated using the **NORMSDIST** function. Since we have an upper tail test, we must subtract from one as shown below.



	A	B	C	D	E	F	G
1	**Weeds**	**Yield**	**Rank**		**Weeds**	**n**	**Sum of Ranks**
2	3	153.1	1		0	=COUNTIF(A2:A9,"=0")	=SUMIF(A2:A9,"=0",C2:C9)
3	3	156	2		3	=COUNTIF(A2:A9,"=3")	=SUMIF(A2:A9,"=3",C2:C9)
4	3	158.6	3				
5	0	165	4		**W mean**	=F2*(F2+F3+1)/2	
6	0	166.7	5		**W std dev**	=SQRT(F2*F3*(F2+F3+1)/12)	
7	0	172.2	6				**w/continuity correction**
8	3	176.4	7		**test statistic**	=STANDARDIZE(G2,F5,F6)	=STANDARDIZE(G2-0.5,F5,F6)
9	0	176.9	8		**p-value**	=1-NORMSDIST(F8)	=1-NORMSDIST(G8)
10							

The results determine the attained significance level of the test using a normal approximation with and without a continuity correction factor. We see from the output that follows that the sum of the ranks in the first group (0 weeds) is $W = 23$. The approximate *P*-value using the continuity correction is 0.0970. The effect of weeds on yield is not statistically significant at the 0.05 level.



	A	B	C	D	E	F	G	H
1	Weeds	Yield	Rank		Weeds	n	Sum of Ranks	
2	3	153.1	1		0	4	23	
3	3	156	2		3	4	13	
4	3	158.6	3					
5	0	165	4		W mean	18		
6	0	166.7	5		W std dev	3.464101615		
7	0	172.2	6				w/continuity correction	
8	3	176.4	7		test statistic	1.443375673	1.299038106	
9	0	176.9	8		p-value	0.074457337	0.096965426	
10								

Wilcoxon Signed Rank Test

We can use Excel to perform a one-sample Wilcoxon signed rank test of the median for single samples or matched pairs. The Wilcoxon test assumes that the data are a random sample from a symmetric population that is not necessarily normal.

Consider a study of early childhood education. Kindergarten students were asked to tell a fairy tale that had been read to them earlier in the week. Each child told two stories. The first had been read to them and the second had been read and also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here and in EG26-06.MTW are the data for five "low progress" readers in a pilot study:

Child	1	2	3	4	5
Story 1	0.77	0.49	0.66	0.28	0.38
Story 2	0.40	0.72	0.00	0.36	0.55
Difference	0.37	−0.23	0.66	−0.08	−0.17

We will test the hypotheses

H_0: scores have the same distribution for both stories

H_a: scores are systematically higher for Story 2

Because these are matched pairs data, we base our inference on the differences. Enter the differences into an Excel worksheet. Calculate the absolute value of the differences as well. If there are any zero differences, remove them. Sort and then rank the absolute value of the differences. The sum W^+ of the ranks for the positive differences is the Wilcoxon signed rank statistic. If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has mean

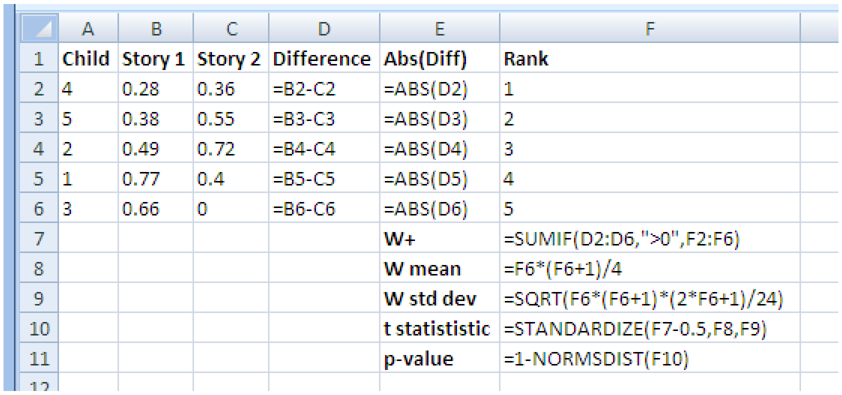
$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

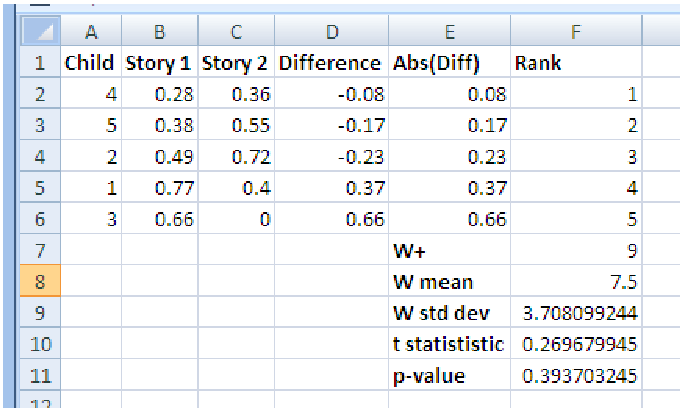
The Wilcoxon signed rank test rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

Formulas for the calculations are shown in the following spreadsheet. Since the ranks are a simple series, 1, 2, 3, 4,..., type **1** and **2** in the first two cells. Select the cells that contain the starting values. Drag the fill handle down to fill in the remaining ranks. The **ADDIF** function adds the values of the ranks with positive differences. The test statistic with the continuity correction is calculated using the **STANDARDIZE** function and the *P*-value is calculated using the **NORMSDIST** function. Since we have an upper tail test, we must subtract from one as shown below.



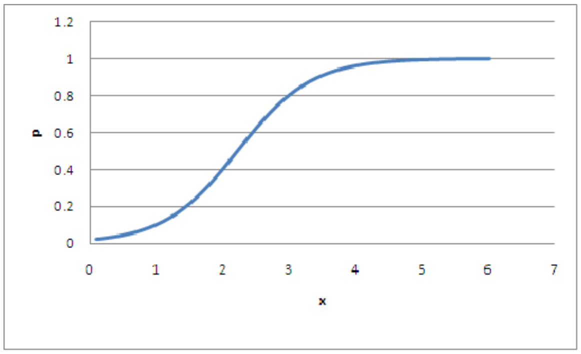
	A	B	C	D	E	F
1	Child	Story 1	Story 2	Difference	Abs(Diff)	Rank
2	4	0.28	0.36	=B2-C2	=ABS(D2)	1
3	5	0.38	0.55	=B3-C3	=ABS(D3)	2
4	2	0.49	0.72	=B4-C4	=ABS(D4)	3
5	1	0.77	0.4	=B5-C5	=ABS(D5)	4
6	3	0.66	0	=B6-C6	=ABS(D6)	5
7					W+	=SUMIF(D2:D6,">0",F2:F6)
8					W mean	=F6*(F6+1)/4
9					W std dev	=SQRT(F6*(F6+1)*(2*F6+1)/24)
10					t statistic	=STANDARDIZE(F7-0.5,F8,F9)
11					p-value	=1-NORMSDIST(F10)

The following output shows that the observed value $W^+ = 9$. The Normal approximation with the continuity correction gives the approximate *P*-value of 0.394. This small sample is not statistically significant.



	A	B	C	D	E	F
1	Child	Story 1	Story 2	Difference	Abs(Diff)	Rank
2	4	0.28	0.36	-0.08	0.08	1
3	5	0.38	0.55	-0.17	0.17	2
4	2	0.49	0.72	-0.23	0.23	3
5	1	0.77	0.4	0.37	0.37	4
6	3	0.66	0	0.66	0.66	5
7					W+	9
8					W mean	7.5
9					W std dev	3.708099244
10					t statistic	0.269679945
11					p-value	0.393703245

CHAPTER 16

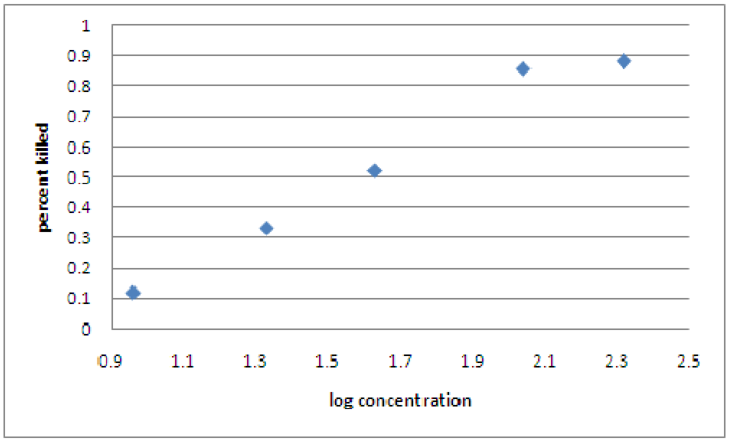


Logistic Regression

The data for logistic regression are n independent observations each consisting of a value of the explanatory variable x and either a success or failure for each trial. Our example concerns an experiment that was designed to examine how well the insecticide rotenone kills an aphid, called *Macrosiphoniella sanborni*, that feeds on the chrysanthemum plant. The explanatory variable is the concentration (in log of milligrams per liter) of the insecticide. About 50 aphids were exposed to each of five concentrations. Each insect was either killed or not killed. Here are the data, along with the results of some calculations:

Concentration x (log scale)	Number of insects	Number killed	Propoertion killed
0.96	50	6	0.1200
1.33	48	16	0.3333
1.63	46	24	0.5217
2.04	49	42	0.8571
2.32	50	44	0.8800

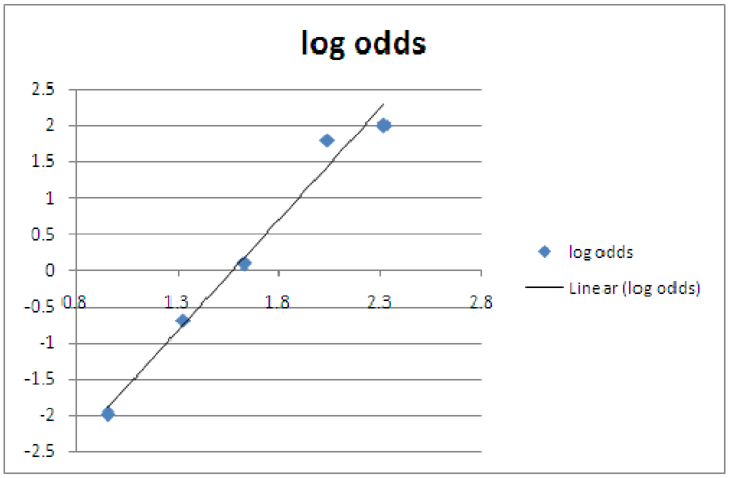
A plot of the proportion killed versus x illustrates the need for logistic regression. If a line is fit to the data, values of x above 2.5 will predict proportions above 1. Similarly, values of x below 0.7 will predict proportions below 0.



The logistic regression model removes this difficulty by working with the natural logarithm of the odds, $p/(1\text{-}p)$. We use the term log odds for this transformation. As p moves from 0 to 1, the log odds moves through all negative and positive numerical values. We model the log odds as a linear function of the explanatory variable:

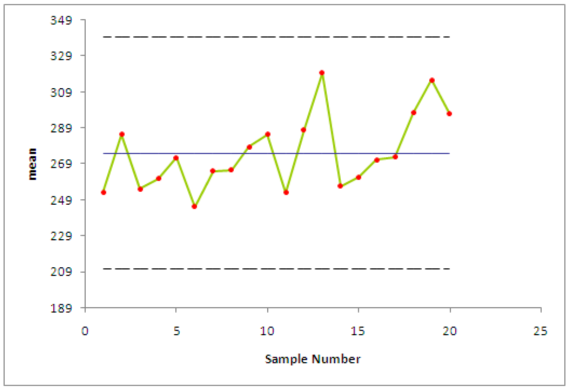
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

The plot of log odds versus log concentration is close to linear. The graph strongly suggests that insecticide concentration affects the kill rate in a way that fits the logistic regression model.



Unfortunately, we cannot perform logistic regression analysis in Excel. Complete analysis should be done in a statistical program such as SPSS, SAS, or Minitab.

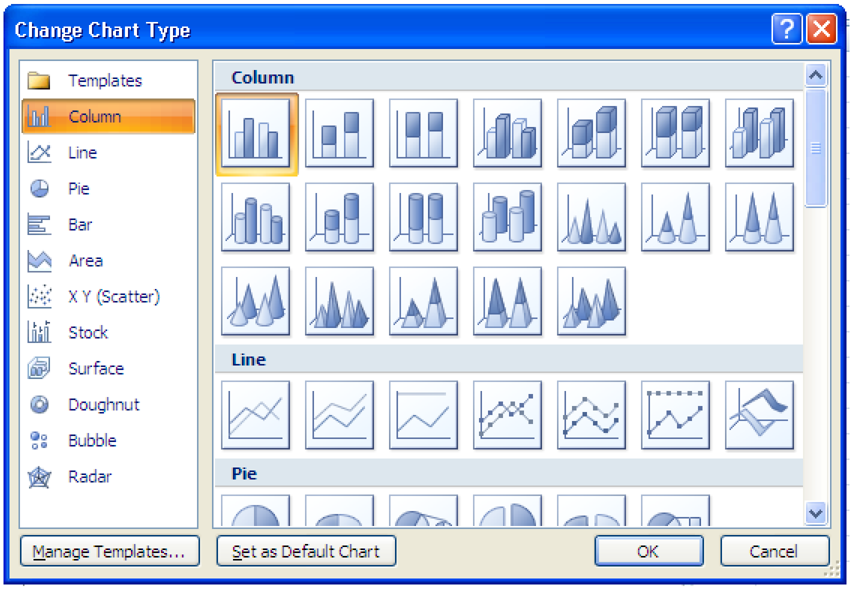
CHAPTER 17



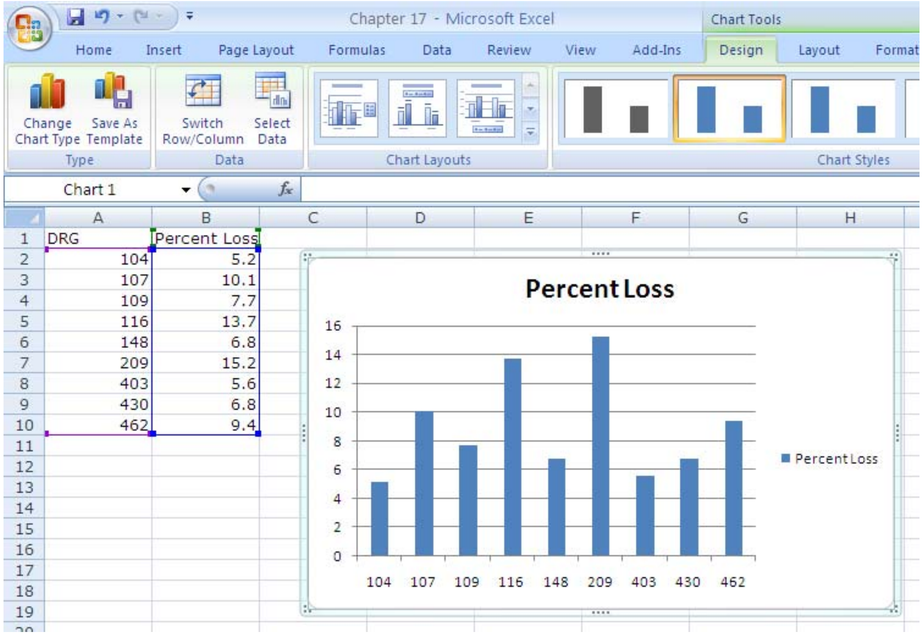
Statistics for Quality

Pareto Charts

Pareto charts are bar graphs with the bars ordered by height. They are often used to isolate the "vital few" categories on which we should focus our attention. Consider the following example: A large medical center, financially pressed by restrictions on reimbursement by insurers and the government, looked at losses broken down by diagnosis. Government standards place cases into diagnostic related groups (DRGs). For example, major joint replacements (mostly hip and knee) are DRG 209. The data list the nine DRGs with the most losses along with the percent of losses. Since the percents are given, a Pareto chart can be constructed by highlighting the data and selecting **Insert ➤ Scatter** from the menu and then select a scatterplot type. Next, click on the graph and select **Design ➤ Change Chart Type**. In the dialog box, select the Column type. Surprisingly, this gives a different graph from the one that we obtain if we simply select the Column type to begin with.

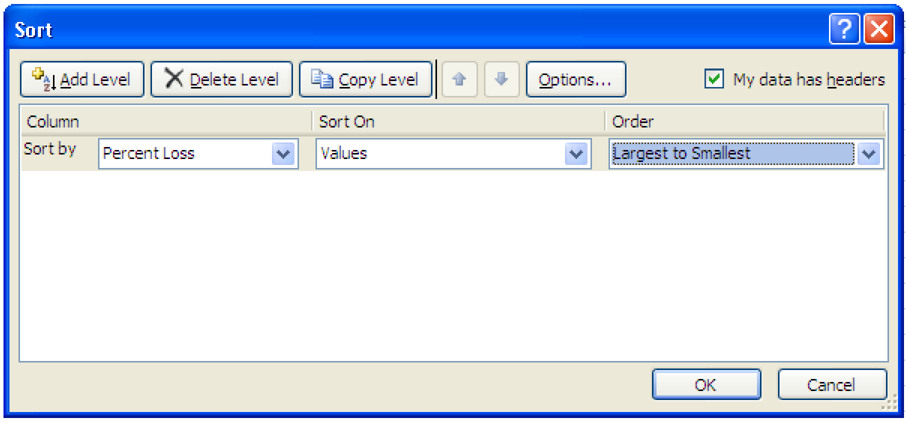


The advantage of this approach is that we obtain a bar chart with the DRGs on the *x*-axis and the Percent Loss on the *y*-axis as shown below.



DRG	Percent Loss
104	5.2
107	10.1
109	7.7
116	13.7
148	6.8
209	15.2
403	5.6
430	6.8
462	9.4

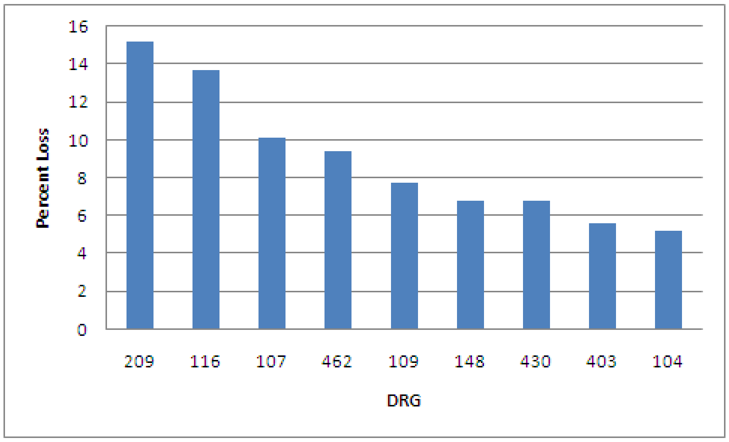
The categories in the bar chart above are ordered numerically instead of from most frequent to least frequent as required for a Pareto chart. To change this, highlight the data and select **Data ➤ Sort** from the menu. In the dialog box, specify that you wish to sort Percent Loss values from largest to smallest.



☑ My data has headers

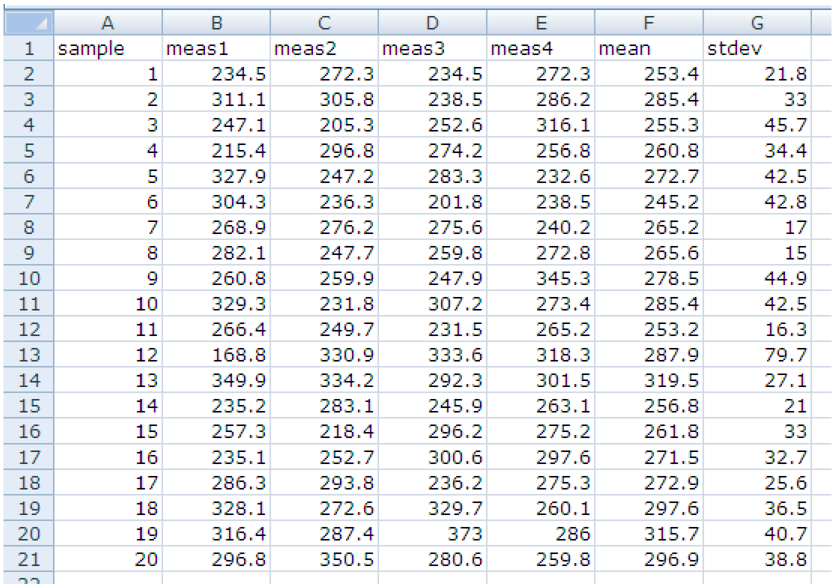
Column	Sort On	Order
Sort by Percent Loss	Values	Largest to Smallest

The axis titles on the Pareto Chart can be obtained by clicking on the graph and then selecting **Layout ➤ Axis Titles** from the Excel menu. The chart below allows us to identify the "vital few" categories that contain most of the observations.



Control Charts for Sample Means

Our next example considers a manufacturer of computer monitors. The manufacturer measures the tension of fine wires behind the viewing screen. Tension is measured by an electrical device with output readings in millivolts (mV). The proper tension is 275 mV. Some variation is always present in the production process. When the process is operating properly, the standard deviation of the tension readings is $\sigma = 43$ mV. Four measurements are made every hour. The data contain the measurements for 20 hours. The first row of observations is from the first hour, the next row is from the second hour, and so on. There are a total of 80 observations.

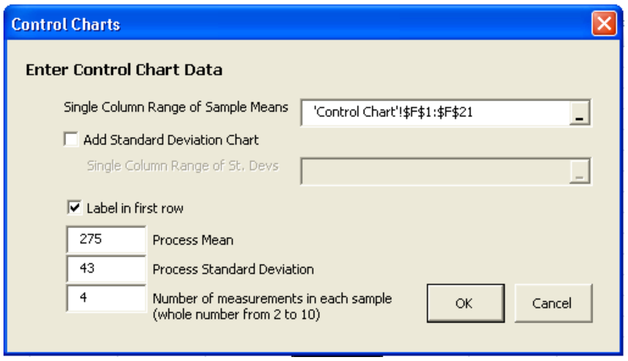


	A	B	C	D	E	F	G
1	sample	meas1	meas2	meas3	meas4	mean	stdev
2	1	234.5	272.3	234.5	272.3	253.4	21.8
3	2	311.1	305.8	238.5	286.2	285.4	33
4	3	247.1	205.3	252.6	316.1	255.3	45.7
5	4	215.4	296.8	274.2	256.8	260.8	34.4
6	5	327.9	247.2	283.3	232.6	272.7	42.5
7	6	304.3	236.3	201.8	238.5	245.2	42.8
8	7	268.9	276.2	275.6	240.2	265.2	17
9	8	282.1	247.7	259.8	272.8	265.6	15
10	9	260.8	259.9	247.9	345.3	278.5	44.9
11	10	329.3	231.8	307.2	273.4	285.4	42.5
12	11	266.4	249.7	231.5	265.2	253.2	16.3
13	12	168.8	330.9	333.6	318.3	287.9	79.7
14	13	349.9	334.2	292.3	301.5	319.5	27.1
15	14	235.2	283.1	245.9	263.1	256.8	21
16	15	257.3	218.4	296.2	275.2	261.8	33
17	16	235.1	252.7	300.6	297.6	271.5	32.7
18	17	286.3	293.8	236.2	275.3	272.9	25.6
19	18	328.1	272.6	329.7	260.1	297.6	36.5
20	19	316.4	287.4	373	286	315.7	40.7
21	20	296.8	350.5	280.6	259.8	296.9	38.8

Excel can be used to produce control charts for sample means by selecting

Add-Ins ➤ WHFStat ➤ Graphs ➤ Control Chart

from the menu. In the dialog box, enter the $\overline{x}$ values. These values will be plotted on the chart. In addition, a center line, an upper control limit (UCL) at 3σ above the center line, and a lower control limit (LCL) at 3σ below the center line are drawn on the chart. The parameters μ and σ must be specified from historical data by filling in the Process Mean and Process Standard Deviation in the dialog box. In the following dialog box we specify that the historical mean is equal to 275 and the historical standard deviation is equal to 43 and the sample size of 4.



Control Charts

Enter Control Chart Data

Single Column Range of Sample Means: 'Control Chart'!$F$1:$F$21

□ Add Standard Deviation Chart

Single Column Range of St. Devs

☑ Label in first row

275 Process Mean

43 Process Standard Deviation

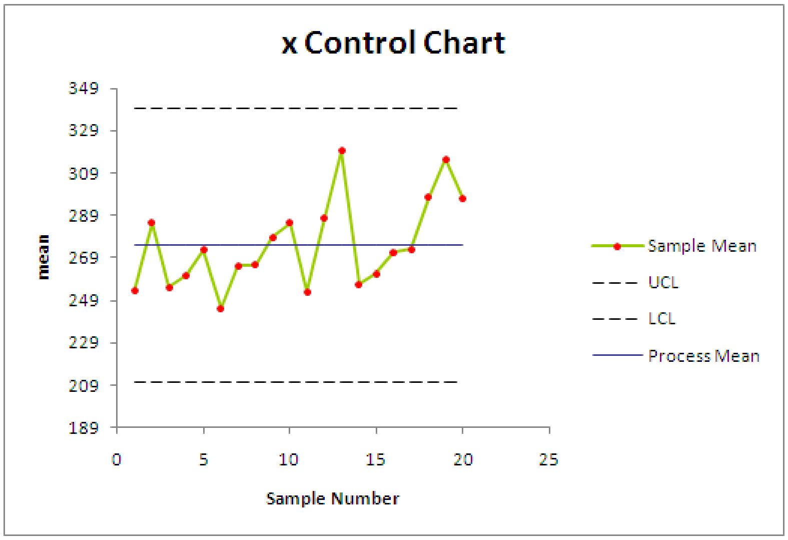
4 Number of measurements in each sample (whole number from 2 to 10)

The center line is at m = 275 mV. The upper and lower control limits are

$$\mu + 3\frac{\sigma}{\sqrt{n}} = 275 + 3\frac{43}{\sqrt{4}} = 339.5 \text{ mV}$$

$$\mu - 3\frac{\sigma}{\sqrt{n}} = 275 - 3\frac{43}{\sqrt{4}} = 210.5 \text{ mV}$$

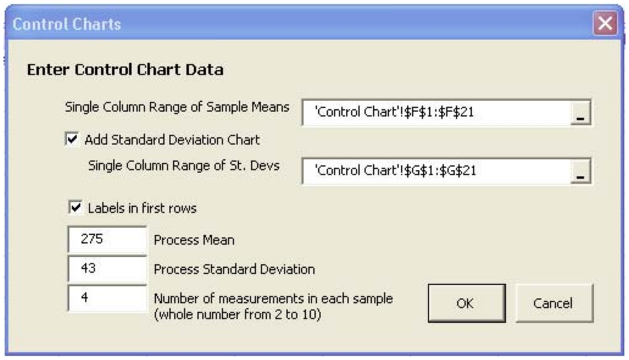
The $\overline{x}$ chart for the mesh tension data show that no points lie outside the control limits.



In practice, we must monitor both the process center, using an $\overline{x}$ chart, and the process spread, using a control chart for the sample standard deviation s. This is commonly done with an s chart, a chart of standard deviations against time. Usually, the $\overline{x}$ chart and the s chart will be looked at together. The $\overline{x}$ chart and s chart can be produced by selecting

Add-Ins ➤ WHFStat ➤ Graphs ➤ Control Chart

from the menu. Check the box next to Add Standard Deviation Chart and enter the data into the dialog box as shown.


Control Charts

Enter Control Chart Data

Single Column Range of Sample Means: 'Control Chart'!$F$1:$F$21

☑ Add Standard Deviation Chart

Single Column Range of St. Devs: 'Control Chart'!$G$1:$G$21

☑ Labels in first rows

275 Process Mean

43 Process Standard Deviation

4 Number of measurements in each sample (whole number from 2 to 10)

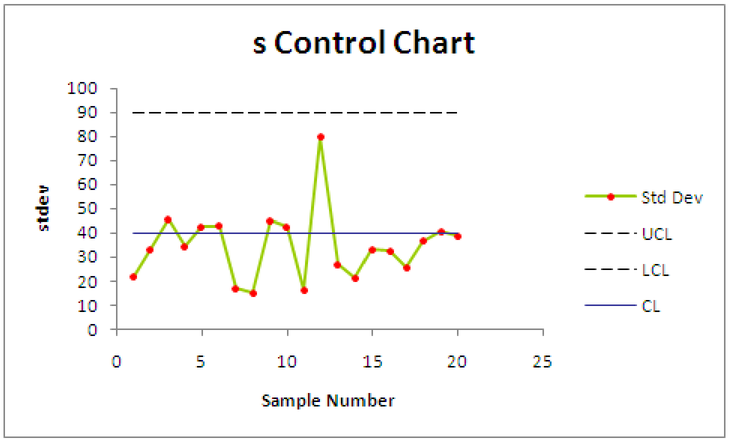
The samples are of size n = 4 and the process standard deviation in control is σ = 43 mV. The centerline is therefore

$$CL = c_4\sigma = (0.9213)(43) = 39.6 \text{ mV}$$

The control limits are

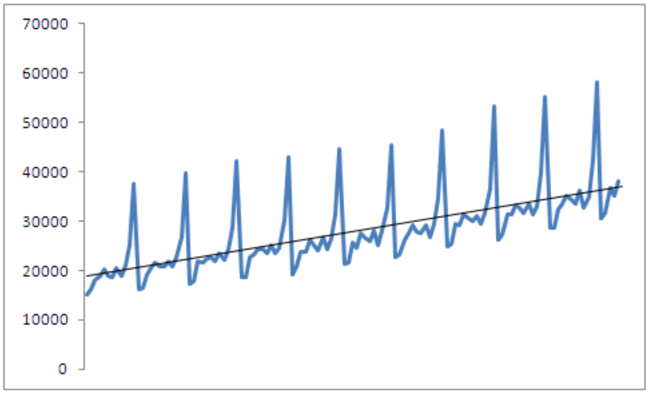
$$UCL = B_6\sigma = (2.088)(43) = 89.9$$
$$LCL = B_5\sigma = (0)(43) = 0$$

as described in your text. The *s* chart for the mesh tension data is also in control.



CHAPTER

18



Time Series Forecasting

Time Series Plots

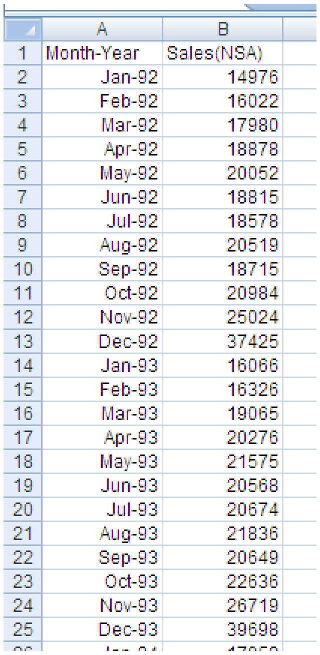
Consider the monthly retail sales data beginning January 1992 and ending May 2002 (125 months). Excel can be used to plot the monthly retail sales by highlighting the data and selecting

Insert ➤ Scatter ➤ Scatter with Straight Lines

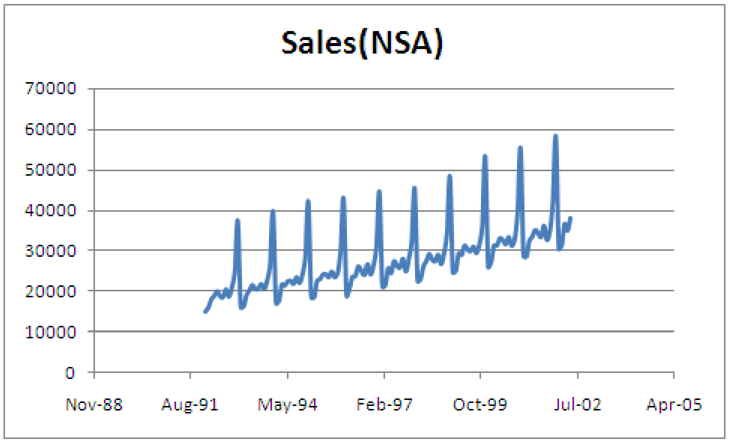
from the menu. This command plots measurement data on the *y*-axis versus time data on the *x*-axis. If you only have the *y*-axis data in the order that the values appear in the column, in equally spaced time intervals, you may want to use

Insert ➤ Line

to plot the *y*-axis data versus consecutive numbers on the *x*-axis.

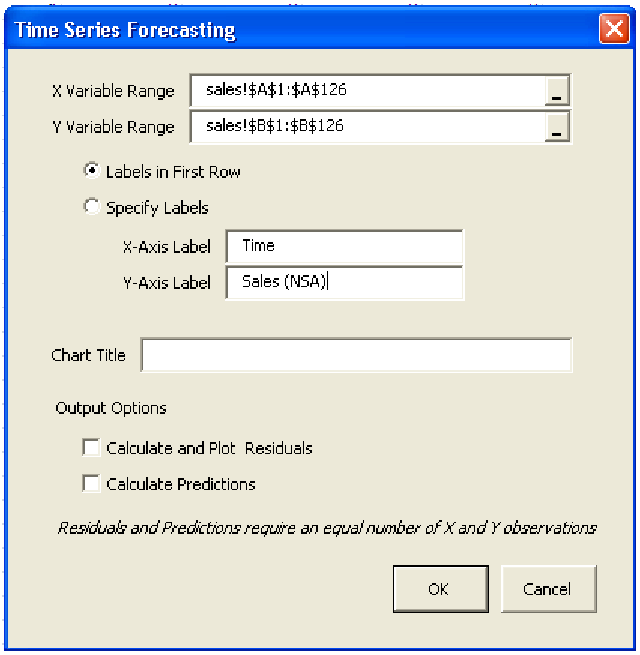


	A	B
1	Month-Year	Sales(NSA)
2	Jan-92	14976
3	Feb-92	16022
4	Mar-92	17980
5	Apr-92	18878
6	May-92	20052
7	Jun-92	18815
8	Jul-92	18578
9	Aug-92	20519
10	Sep-92	18715
11	Oct-92	20984
12	Nov-92	25024
13	Dec-92	37425
14	Jan-93	16066
15	Feb-93	16326
16	Mar-93	19065
17	Apr-93	20276
18	May-93	21575
19	Jun-93	20568
20	Jul-93	20674
21	Aug-93	21836
22	Sep-93	20649
23	Oct-93	22636
24	Nov-93	26719
25	Dec-93	39698



Since January 1992, overall sales have gradually increased and a distinct pattern repeats itself approximately every 12 months.

Alternatively, the time series plot can be obtained by selecting **Add-Ins ➤ WHFStat ➤ Times Series Forecasting ➤ Times Series Scatterplot** and filling out the dialog box as shown.



Time Series Forecasting

X Variable Range: sales!$A$1:$A$126

Y Variable Range: sales!$B$1:$B$126

● Labels in First Row

○ Specify Labels

X-Axis Label: Time

Y-Axis Label: Sales (NSA)

Chart Title:

Output Options

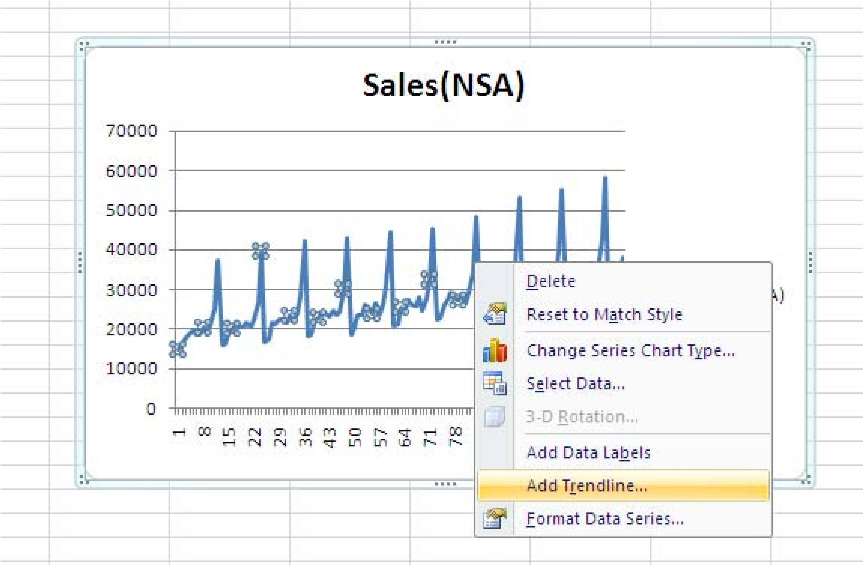
□ Calculate and Plot Residuals

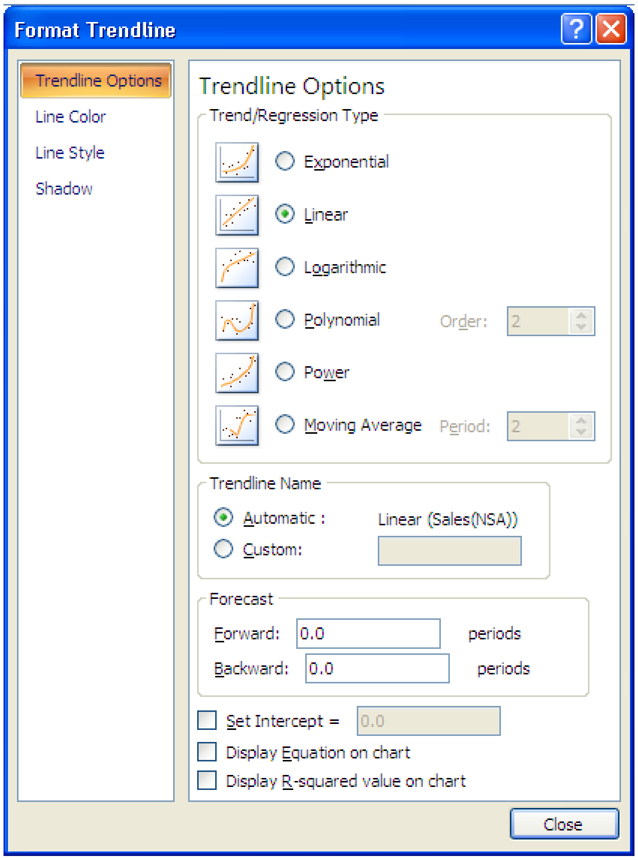
□ Calculate Predictions

Residuals and Predictions require an equal number of X and Y observations

Trend Analysis

Our example discusses monthly retail sales of General Merchandise Stores beginning January 1992 and ending May 2002 (125 months). The pattern of increasing growth in the time series plot of the retail sales data is an example of a *linear trend*. Excel can be used to estimate the linear trend. Right click on any point in the graph and select Add Trendline from the menu as shown to fit a general trend model to time series data and provide forecasts.





Trendline Options

Trend/Regression Type

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order: 2

○ Power

○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Sales(NSA))

○ Custom:

Forecast

Forward: 0.0 periods

Backward: 0.0 periods

□ Set Intercept = 0.0

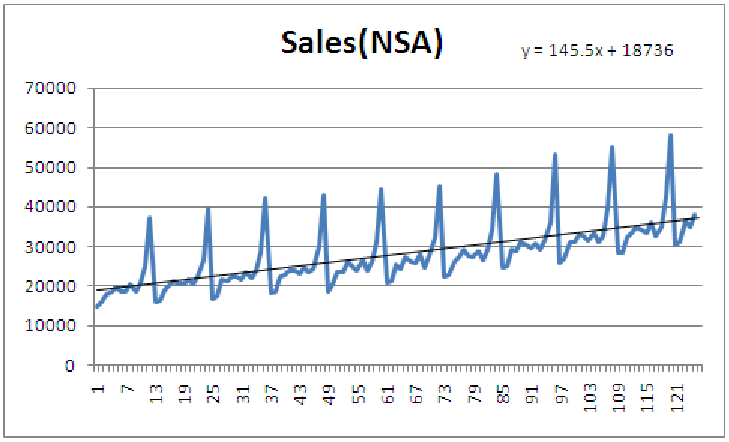
□ Display Equation on chart

□ Display R-squared value on chart

You can choose from among a variety of time series models. Here we select Linear and check the appropriate box to display the equation on the chart. This is best done on a graph with consecutive numerical values on the *x*-axis. Recall that this graph is obtained by highlighting only the sales data and selecting **Insert ➤ Line** from the menu. Excel estimates the linear trend to be

$$y_t = 18736 + 145.5t$$

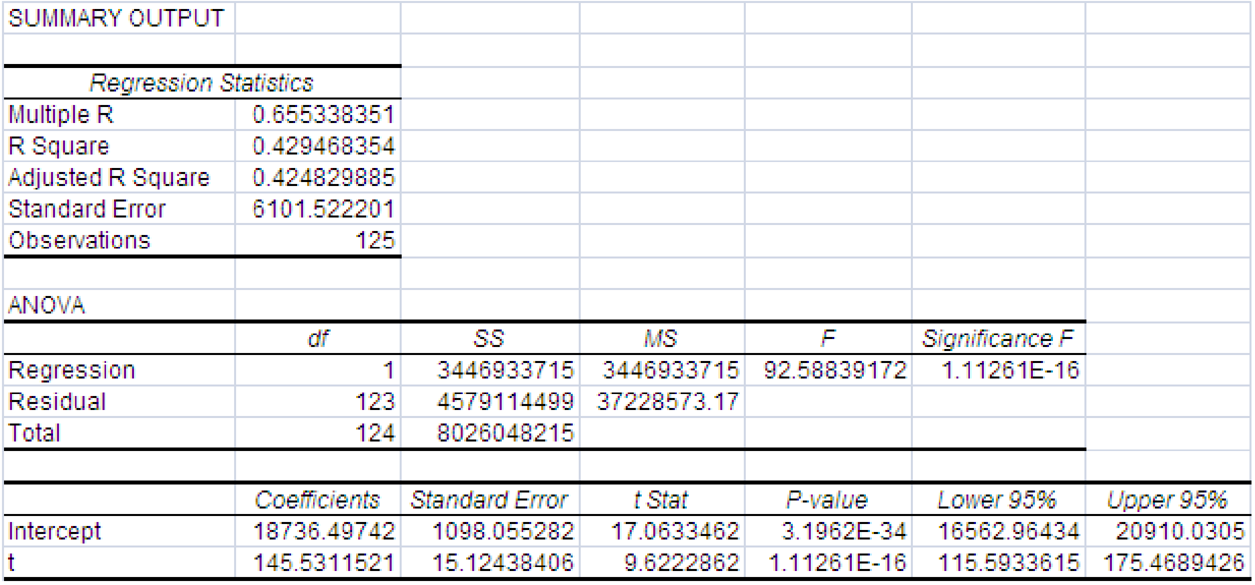
where *t* is the number of months elapsed beginning with the first month of the time series.



We can also use regression techniques to fit a linear model to the above data. This gives additional output. First, create a variable *x* (or *t*) where *x* is the number of months elapsed beginning with the first month of the time series. That is, *x* = 1 corresponds to January 1992, *x* = 2 corresponds to February 1992, etc. Next, select

Data ➤ Data Analysis ➤ Regression

from the menu. Enter Sales in the Input Y Range and *t* as the Input X Range. Specify that there are titles in the first row, where you want your output, and click OK.

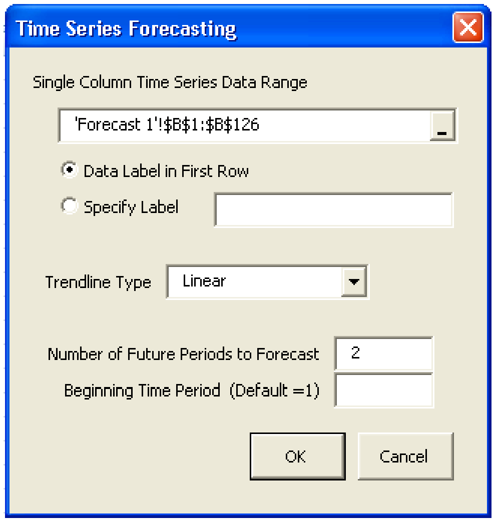


SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.655338351					
R Square	0.429468354					
Adjusted R Square	0.424829885					
Standard Error	6101.522201					
Observations	125					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	3446933715	3446933715	92.58839172	1.11261E-16	
Residual	123	4579114499	37228573.17			
Total	124	8026048215				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	18736.49742	1098.055282	17.0633462	3.1962E-34	16562.96434	20910.0305
t	145.5311521	15.12438406	9.6222862	1.11261E-16	115.5933615	175.4689426

The trend-only model ignores the seasonal variation in the retail sales time series. Notice that the R^2 value for the trend-only model is 42.9% or 0.429.

To forecast sales, we can either plug in a specific value for x into the regression equation, $Y_t = 18736 + 145.5\, t$, or we can use Excel's **TREND** function. For example to forecast the sales for January, we could enter =TREND(C2:C126,B2:B126,133) into a cell. The new value of x is 133, corresponding to the 133rd month, where January 2002 is considered the first month. The resulting forecast is 38,092.

Selecting **Add-Ins ➤ WHFStat ➤ Times Series Forecasting ➤ Times Series Scatterplot** and filling out the dialog box as shown gives us another way to obtain the same results.



Time Series Forecasting

Single Column Time Series Data Range: 'Forecast 1'!$B$1:$B$126

● Data Label in First Row

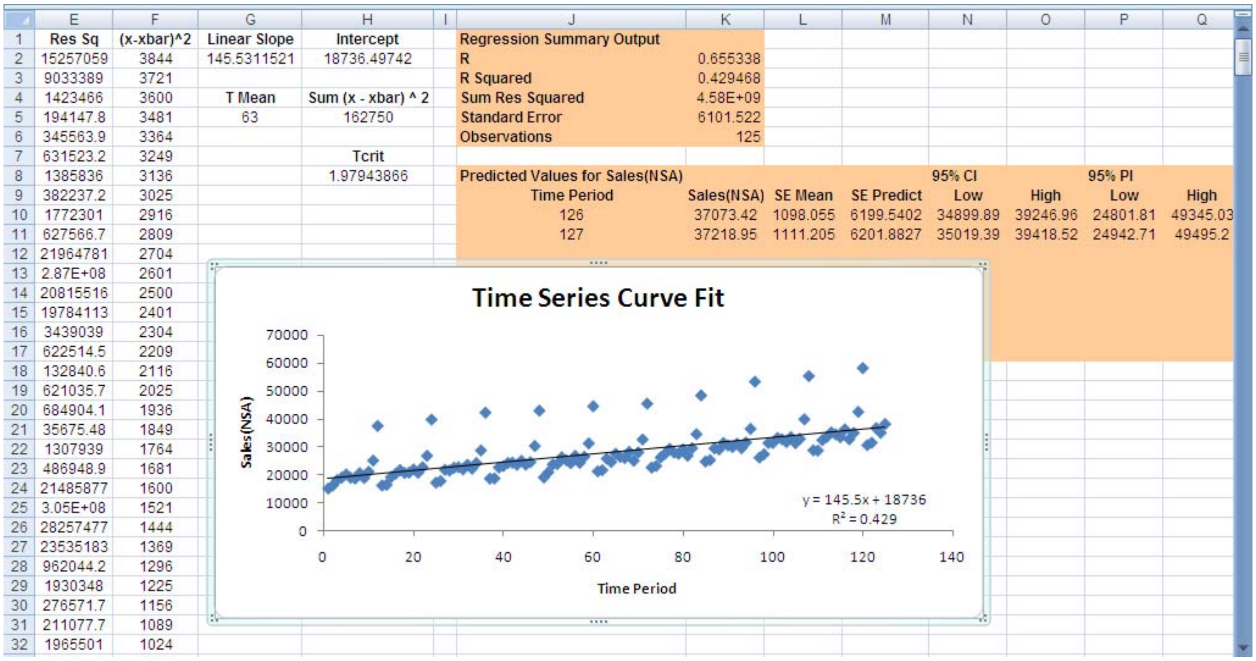
○ Specify Label

Trendline Type: Linear

Number of Future Periods to Forecast: 2

Beginning Time Period (Default =1):

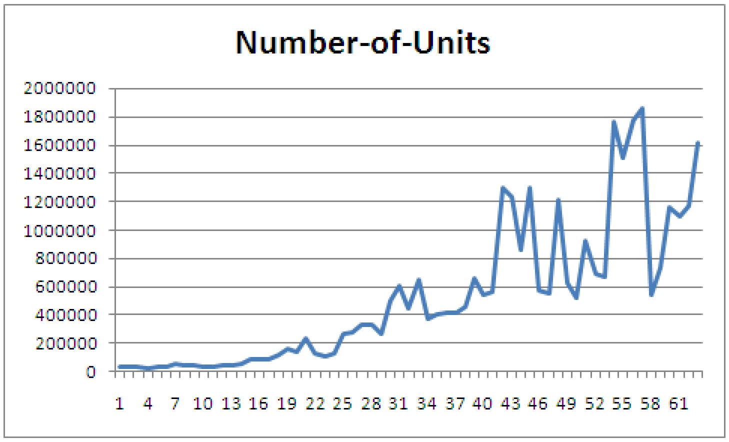
The WHFStat Add-In does not require that the values of the explanatory variable be entered. As a default, these values are assumed to be 1, 2, 3, …. The results are the same as shown previously, but the output is substantially different as shown below.



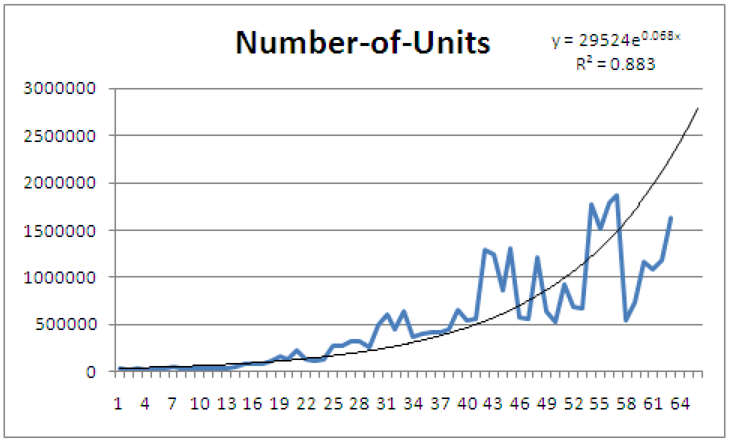
	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Res Sq	(x-xbar)^2	Linear Slope	Intercept		Regression Summary Output							
2	15257059	3844	145.5311521	18736.49742		R	0.655338						
3	9033389	3721				R Squared	0.429468						
4	1423466	3600	T Mean	Sum (x - xbar) ^ 2		Sum Res Squared	4.58E+09						
5	194147.8	3481	63	162750		Standard Error	6101.522						
6	345563.9	3364				Observations	125						
7	631523.2	3249		Tcrit									
8	1385836	3136		1.97943866		Predicted Values for Sales(NSA)				95% CI		95% PI	
9	382237.2	3025				Time Period	Sales(NSA)	SE Mean	SE Predict	Low	High	Low	High
10	1772301	2916				126	37073.42	1098.055	6199.5402	34899.89	39246.96	24801.81	49345.03
11	627566.7	2809				127	37218.95	1111.205	6201.8827	35019.39	39418.52	24942.71	49495.2
12	21964781	2704											
13	2.87E+08	2601											
14	20815516	2500											
15	19784113	2401											
16	3439039	2304											
17	622514.5	2209											
18	132840.6	2116											
19	621035.7	2025											
20	684904.1	1936											
21	35675.48	1849											
22	1307939	1764											
23	486948.9	1681											
24	21485877	1600											
25	3.05E+08	1521											
26	28257477	1444											
27	23535183	1369											
28	962044.2	1296											
29	1930348	1225											
30	276571.7	1156											
31	211077.7	1089											
32	1965501	1024											

Exponential Growth

Consider the sales of DVD players since the introduction of the DVD format in March 1997. At the end of June 2002, nearly 33 million DVD players had been sold in the United States with over 18,000 titles available in the DVD format. The Consumers Electronic Association tracks monthly sales of DVD players. We can highlight the dates and units sold and select **Insert ➤ Scatter ➤ Scatter with Lines** from the menu to plot the DVD sales data. Alternatively, we can highlight on the units sold and select **Insert ➤ Line** from the menu to obtain the graph shown.



The pattern of increasing growth in this plot is an example of an *exponential trend.* Excel can be used to estimate the exponential trend. Right click on a point in the graph and select **Add Trendline** from the menu. In the dialog box, under Trend/Regression Type, choose Exponential.

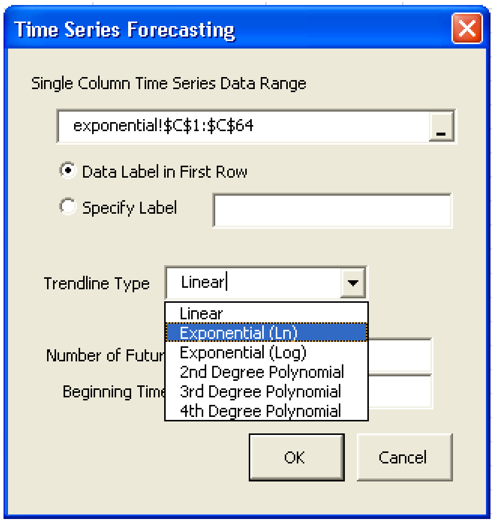


As shown above in the trend analysis, Excel estimates the exponential trend to be

$$y_t = 29524\mathrm{e}^{\,0.068t}$$

where *t* is the number of months elapsed beginning with the first month of the time series.

As with the linear trend, we can select **Add-Ins ➤ WHFStat ➤ Times Series Forecasting ➤ Forecast** and select Exponential (Ln) for the Trendline Type to get the same results as above.



Time Series Forecasting

Single Column Time Series Data Range: exponential!$C$1:$C$64

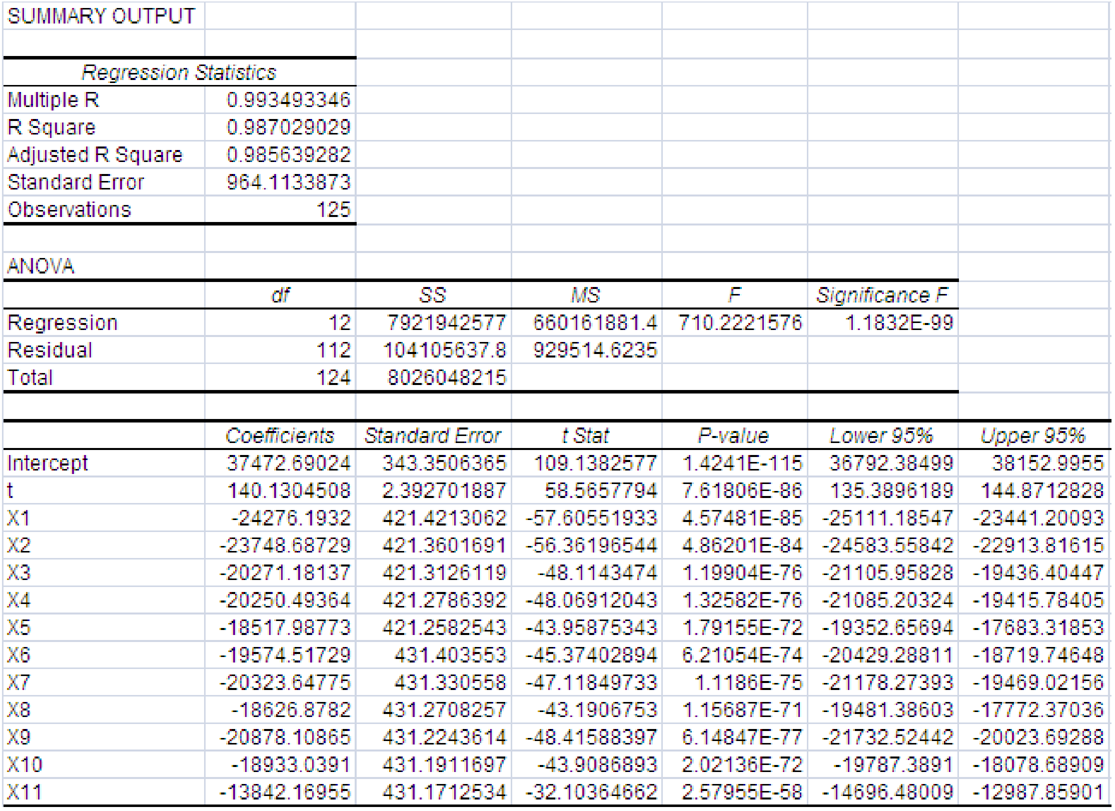
● Data Label in First Row

○ Specify Label

Trendline Type: Linear

Seasonal Models

A trend equation may be a good description of the long run behavior of the data, but we need to account for short run phenomena like seasonal variation to improve the accuracy of our forecasts. In this example, we can use indicator variables to add the seasonal pattern to the trend model for the monthly retail sales data. Name the indicator variables X1, X2,….X11. Enter data such that the value of X1 is 1 for January and 0 otherwise. Similarly, the value for X2 is 1 for February and 0 otherwise, etc. We do not need an X12 variable as it would provide the same information as X1, X2,…,X11. Select **Data ➤ Data Analysis ➤ Regression** from the menu and enter Sales as the Input Y variable. Enter *t* and all 12 indicator variables as the Input X Variables and click OK. (Recall that we defined the variable *x* in the previous section as the number of months elapsed beginning with the first month of the time series.) We get the following output from Excel.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.993493346
R Square	0.987029029
Adjusted R Square	0.985639282
Standard Error	964.1133873
Observations	125

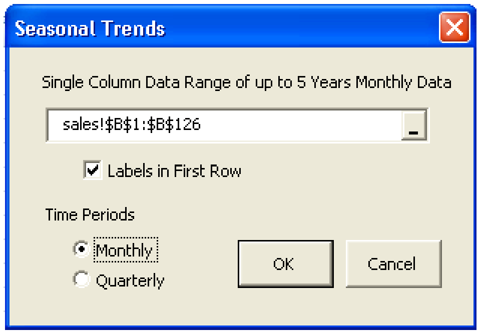
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	12	7921942577	660161881.4	710.2221576	1.1832E-99
Residual	112	104105637.8	929514.6235		
Total	124	8026048215			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	37472.69024	343.3506365	109.1382577	1.4241E-115	36792.38499	38152.9955
t	140.1304508	2.392701887	58.5657794	7.61806E-86	135.3896189	144.8712828
X1	-24276.1932	421.4213062	-57.60551933	4.57481E-85	-25111.18547	-23441.20093
X2	-23748.68729	421.3601691	-56.36196544	4.86201E-84	-24583.55842	-22913.81615
X3	-20271.18137	421.3126119	-48.1143474	1.19904E-76	-21105.95828	-19436.40447
X4	-20250.49364	421.2786392	-48.06912043	1.32582E-76	-21085.20324	-19415.78405
X5	-18517.98773	421.2582543	-43.95875343	1.79155E-72	-19352.65694	-17683.31853
X6	-19574.51729	431.403553	-45.37402894	6.21054E-74	-20429.28811	-18719.74648
X7	-20323.64775	431.330558	-47.11849733	1.1186E-75	-21178.27393	-19469.02156
X8	-18626.8782	431.2708257	-43.1906753	1.15687E-71	-19481.38603	-17772.37036
X9	-20878.10865	431.2243614	-48.41588397	6.14847E-77	-21732.52442	-20023.69288
X10	-18933.0391	431.1911697	-43.9086893	2.02136E-72	-19787.3891	-18078.68909
X11	-13842.16955	431.1712534	-32.10364662	2.57955E-58	-14696.48009	-12987.85901

The R^2 value for the trend-and-season model is 98.7%, which is a dramatic improvement over the trend-only model. Recall that the R^2 value for the trend-only model was 42.9%. The forecast value for January 2002 would be 37472.69 + 140.1305(133) – 24276.19 = 31833.85, since in January X1 is equal to 1 and the other indicator variables are equal to 0.

A much easier alternative is available by selecting **Add-Ins ➤ WHFStat ➤ Times Series Forecasting ➤ Forecast** and selecting Monthly under Time Periods. Unfortunately, this capability only works if you are using at most five years of data.



Seasonal Trends

Single Column Data Range of up to 5 Years Monthly Data

sales!$B$1:$B$126

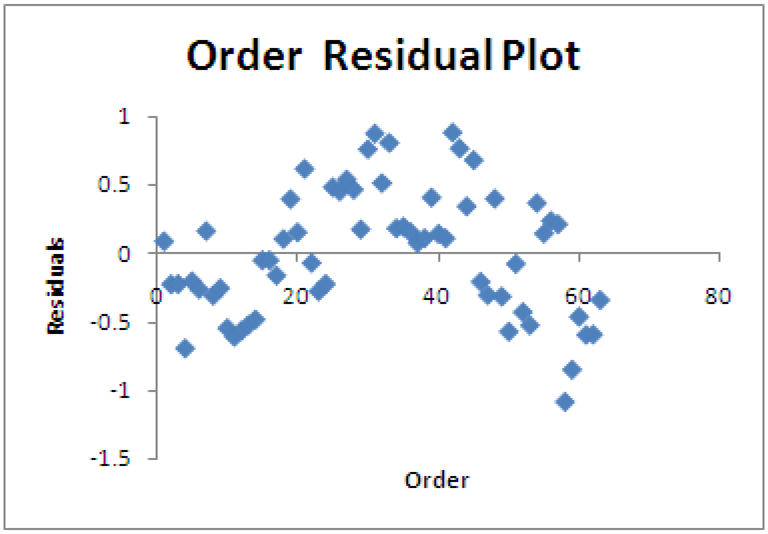
☑ Labels in First Row

Time Periods

● Monthly
○ Quarterly

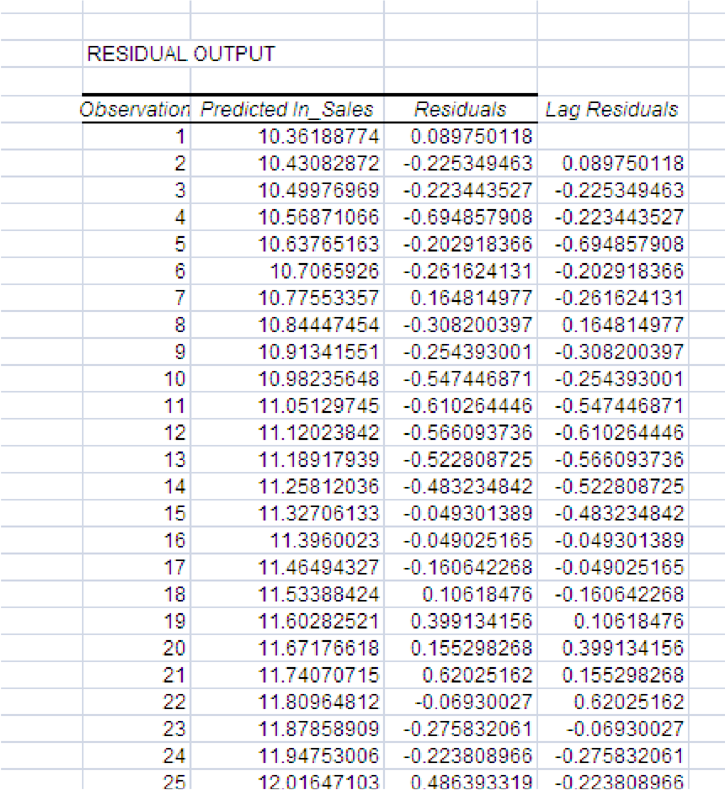
Autocorrelation

The residuals from a regression model that uses time as an explanatory variable should be examined for signs of autocorrelation. Examine the residuals that result from fitting an exponential trend to the DVD player sales data. First, calculate $\log_e$(Sales). Name this column lnSales and calculate the values using Excel's **LN** function. Next, select **Data ➤ Data Analysis ➤ Regression** from the menu and regress lnSales on the predictor variable x, where x is the number of months elapsed beginning with the first month of the time series. In the dialog box, check Residual Plots box to obtain the residual plot from the exponential trend model.



The pattern in the plot indicates positive autocorrelation among the residuals.

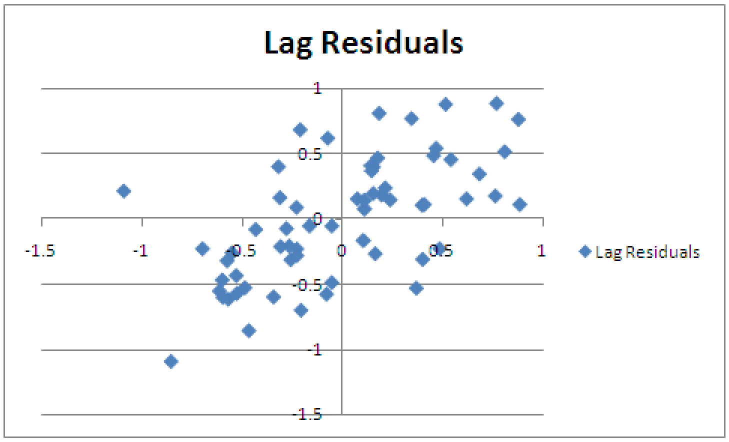
An alternative plot for detecting autocorrelation is a lagged residual plot. Create a column of lagged residuals by copying the values one row down as shown. There will be one missing at the top of the output column. The lagged column should have the same number of rows as the Residual column, so the last value from the Residual column does need to be lagged.



RESIDUAL OUTPUT

Observation	Predicted ln_Sales	Residuals	Lag Residuals
1	10.36188774	0.089750118	
2	10.43082872	-0.225349463	0.089750118
3	10.49976969	-0.223443527	-0.225349463
4	10.56871066	-0.694857908	-0.223443527
5	10.63765163	-0.202918366	-0.694857908
6	10.7065926	-0.261624131	-0.202918366
7	10.77553357	0.164814977	-0.261624131
8	10.84447454	-0.308200397	0.164814977
9	10.91341551	-0.254393001	-0.308200397
10	10.98235648	-0.547446871	-0.254393001
11	11.05129745	-0.610264446	-0.547446871
12	11.12023842	-0.566093736	-0.610264446
13	11.18917939	-0.522808725	-0.566093736
14	11.25812036	-0.483234842	-0.522808725
15	11.32706133	-0.049301389	-0.483234842
16	11.3960023	-0.049025165	-0.049301389
17	11.46494327	-0.160642268	-0.049025165
18	11.53388424	0.10618476	-0.160642268
19	11.60282521	0.399134156	0.10618476
20	11.67176618	0.155298268	0.399134156
21	11.74070715	0.62025162	0.155298268
22	11.80964812	-0.06930027	0.62025162
23	11.87858909	-0.275832061	-0.06930027
24	11.94753006	-0.223808966	-0.275832061
25	12.01647103	0.486393319	-0.223808966

Next, highlight the residuals and lagged residuals and select **Insert ➤ Scatter** from the menu and plot the residuals versus the lagged residuals. The graph on the following page shows a linear pattern with positive slope. This indicates that the residuals may have positive autocorrelation.

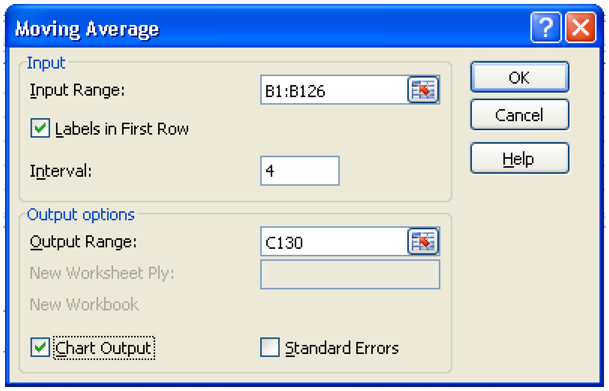


Excel can be used to calculate the autocorrelation by calculating the correlation between the residuals and the lagged residuals. This is easily done with the **CORREL** function. The function arguments are the two arrays of data. Make sure to enter the data from the same rows. In our example, =CORREL(H56:H117,I56:I117) results in a value of 0.616, indicating a strong autocorrelation.

When autocorrelation is present, an autoregressive model can be used. The first-order autoregressive model specifies a linear relationship between the response variable and the lagged values of the response variable. The shorthand for this model is AR(1). The model can be found with Excel either by making a scatterplot and adding a trendline or by selecting **Data ➤ Data Analysis ➤ Regression** from the menu and letting the Input Y Variable be the response variable and the Input X Variable be the lagged values of the response variable.

Moving average models

With some time series, our forecasts can be obtained by using the average of several past time periods. Moving average models use the average of the last *k* values of the time series as the forecast for period *t*. To find the forecasts using Excel, select **Data ➤ Data Analysis ➤ Moving Average** from the Excel menu. Specify the values of the time series and the value for *k* in the dialog box as shown.



Moving Average

Input

Input Range: B1:B126

☑ Labels in First Row

Interval: 4

Output options

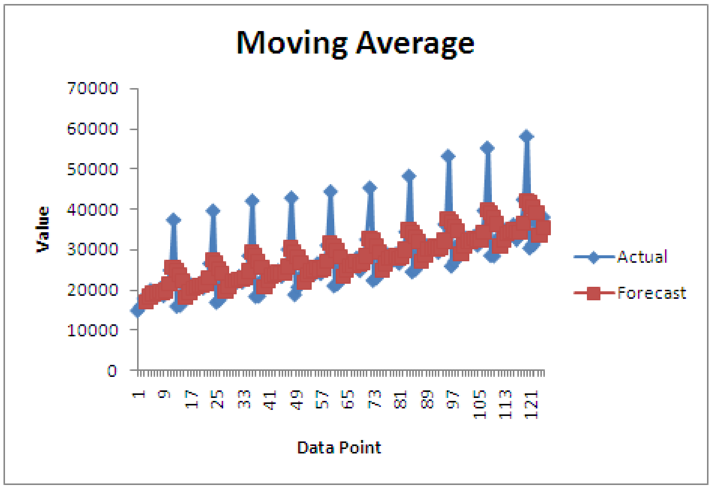
Output Range: C130

New Worksheet Ply:

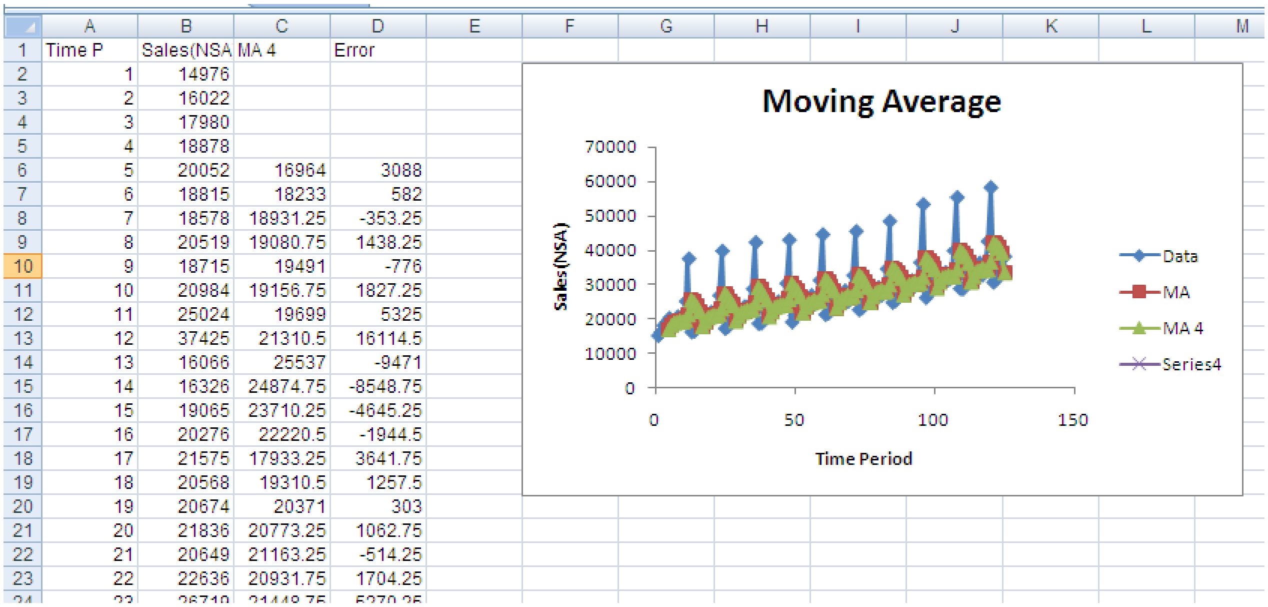
New Workbook

☑ Chart Output ☐ Standard Errors

If you check the Chart Output box, a graph of the time series and the moving average values is provided in addition to the calculated values. The graph for the sales data considered earlier is provided below.



The moving average analysis can also be obtained by selecting **Add-Ins ➤ WHFStat ➤ Times Series Forecasting ➤ Moving Average** from the Excel menu. In this case, the output includes the moving average for each point and the error = actual – forecast for each point.



	A	B	C	D
1	Time P	Sales(NSA	MA 4	Error
2	1	14976		
3	2	16022		
4	3	17980		
5	4	18878		
6	5	20052	16964	3088
7	6	18815	18233	582
8	7	18578	18931.25	-353.25
9	8	20519	19080.75	1438.25
10	9	18715	19491	-776
11	10	20984	19156.75	1827.25
12	11	25024	19699	5325
13	12	37425	21310.5	16114.5
14	13	16066	25537	-9471
15	14	16326	24874.75	-8548.75
16	15	19065	23710.25	-4645.25
17	16	20276	22220.5	-1944.5
18	17	21575	17933.25	3641.75
19	18	20568	19310.5	1257.5
20	19	20674	20371	303
21	20	21836	20773.25	1062.75
22	21	20649	21163.25	-514.25
23	22	22636	20931.75	1704.25
24	23	26719	21448.75	5270.25

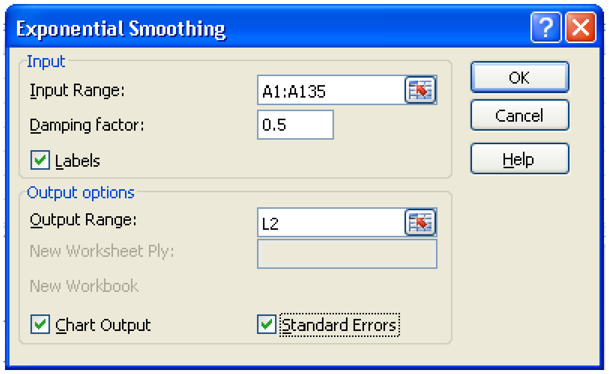
Exponential smoothing models

The exponential smoothing model uses a weighted average of the observed value from and the forecast value for time *t*-1 to calculate the forecast for time *t*. The weight, *w*, is called the smoothing constant and is a value between 0 and 1. The forecasting equation is

$$\hat{y}_t = w y_{t-1} + (1 - w)\hat{y}_{t-1}$$

Choosing *w* close to 1 puts more weight on the most recent value.

To find the forecasts using Excel, select **Data ➤ Data Analysis ➤ Exponential Smoothing** from the menu. Specify the values of the time series and the value for the damping factor (1–w) in the dialog box as shown.



Exponential Smoothing

Input

Input Range: A1:A135

Damping factor: 0.5

☑ Labels

Output options

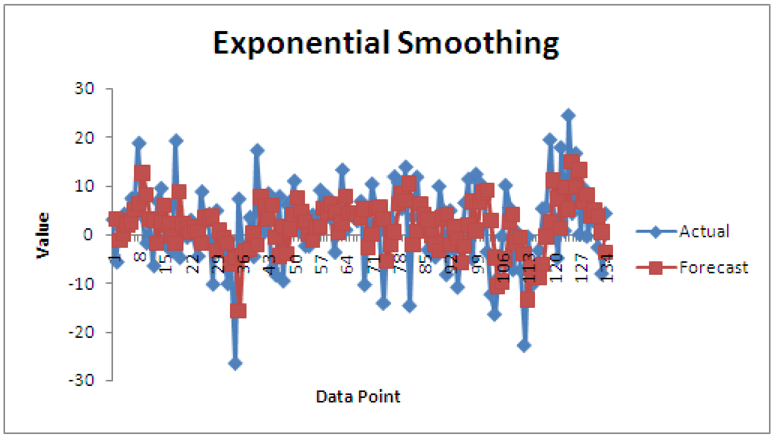
Output Range: L2

New Worksheet Ply:

New Workbook

☑ Chart Output ☑ Standard Errors

If you check the Chart Output box, a graph of the time series and the forecast values is provided in addition to the calculated values. A sample graph is provided below.



The exponential smoothing analysis can also be obtained by selecting **Add-Ins ➤ WHFStat ➤ Time Series Forecasting ➤ Exponential Smoothing** from the Excel menu. In this case, the smoothed values are listed and plotted.

Chapter 1 Exercises

1.7 Refer to the first exam scores from Exercise 1.5 (reproduced below) and this histogram you produced in Exercise 1.6. Now make a histogram for these data using classes 40-59, 60-79, and 80-100. Compare this histogram with the one that you produced in Exercise 1.6.

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

1.19 Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetical and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest is sometimes called a **Pareto chart**, after the Italian economist who recommended this procedure.

1.31 Table 1.7 (reproduced below) contains data on the mean annual temperatures (degrees Fahrenheit) for the years 1941 to 2000 at two locations in California: Pasadena and Redding. Make time plots of both time series and compare their main features. You can see why discussions of climate change often bring disagreement.

Year	Pasadena	Redding	Year	Pasadena	Redding
1951	62.27	62.02	1976	64.23	63.51
1952	61.59	62.27	1977	64.47	63.89
1953	62.64	62.06	1978	64.21	64.05
1954	62.88	61.65	1979	63.76	60.38
1955	61.75	62.48	1980	65.02	60.04
1956	62.93	63.17	1981	65.80	61.95
1957	63.72	62.42	1982	63.50	59.14
1958	65.02	64.42	1983	64.19	60.66
1959	65.69	65.04	1984	66.06	61.72
1960	64.48	63.07	1985	64.44	60.51
1961	64.12	63.50	1986	65.31	61.76

1962	62.82	63.97	1987	64.58	62.94
1963	63.71	62.42	1988	65.22	63.70
1964	62.76	63.29	1989	64.53	61.50
1965	63.03	63.32	1990	64.96	62.22
1966	64.25	64.51	1991	65.60	62.73
1967	64.36	64.21	1992	66.07	63.59
1968	64.15	63.40	1993	65.16	61.55
1969	63.51	63.77	1994	64.63	61.63
1970	64.08	64.30	1995	65.43	62.62
1971	63.59	62.23	1996	65.76	62.93
1972	64.53	63.06	1997	66.72	62.48
1973	63.46	63.75	1998	64.12	60.23
1974	63.93	63.80	1999	64.85	61.88
1975	62.36	62.66	2000	66.25	61.58

1.47 Here are the scores on the first exam in an introductory statistics course for 10 students. Find the mean first exam score for these students.

80	73	92	85	75	98	93	55	80	90

1.49 Here are the scores on the first exam in an introductory statistics course for 10 students. Find the quartiles for these first-exam scores.

80	73	92	85	75	98	93	55	80	90

1.57 C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and reach a peak within 24 to 48 hours after. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of appar3ently healthy children aged 6 to 60 months in Papua, New Guinea, CRP was measured in 90 children. The units are milligrams per liger (mg/l). Here are the data from a random sample of 40 of these children:

0	0	30.61	46.7	22.82	0	5.36	59.76	0	20.78
3.9	5.62	0	0	0	0	0	12.38	0	7.1
5.64	3.92	73.2	0	0	4.81	5.66	15.74	0	7.89
8.22	6.81	0	26.41	3.49	9.57	0	0	9.37	5.53

(a) Find the five-number summary for these data.
(b) Make a boxplot.
(c) Make a histogram.

(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.67 A small accounting firm pays each of its five clerks $35,000, two junior accountants $80,000 each, and the firms owner $320,000. What is the mean salary for this firm? How many employees earn less than the mean? What is the median salary?

1.73 Many standard statistical methods that you will study in Part II of this book are intended for use with distributions that are symmetric and have no outliers. These methods start with the mean and standard deviation, $\bar{x}$ and s. Two examples of scientific data for which standard methods should work well are the pH measurements in Exercise 1.36 and Cavendish's measurements of the density of the earth in Exercise 1.40.

a) Summarize each of these data sets by giving $\bar{x}$ and s.

b) Find the median for each data set. Is the median quite close to the mean, as we expect it to be in these examples?

1.123. The variable *Z* has a standard Normal distribution.

(a) Find the number *z* that has cumulative proportion 0.85.

(b) Find the number *z* such that the event $Z > z$ has proportion 0.40.

1.131 Reports on a student's ACT or SAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent; the percent of all scores that were lower than this one. Jacob scores 17 on the ACT. What is his percentile?

1.139 The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

(a) What percent of pregnancies last less than 240 days (that's about 8 months)?

(b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?

(c) How long do the longest 20% of pregnancies last?

1.147 We expect repeated careful measurements of the same quantity be be approximately Normal. Make a Normal quantile plot for Cavendish's measurements in Exercise 1.40 (data reproduced below). Are the data approximately Normal? If not, describe any clear deviations from Normality.

5.5	5.55	5.57	5.34	5.42	5.3
5.61	5.36	5.53	5.79	5.47	5.75

4.88	5.29	5.62	5.1	5.63	5.68
5.07	5.58	5.29	5.27	5.34	5.85
5.26	5.65	5.44	5.39	5.46	

Chapter 2 Exercises

2.7 Here are the data for the second test and the final exam for the same students as in problem 2.6:

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

(a) Explain why you should use the second-test score as the explanatory variable.
(b) Make a scatterplot and describe the relationship.
(c) Why do you think the relationship between the second-test score and the final-exam score is stronger than the relationship between the first-test score and the final-exam score?

2.21 Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

(a) Make a scatterplot of the data, using different symbols or colors for men and women.
(b) Is the association between these variables positive or negative? How strong is the relationship? Does the pattern of the relationship differ for women and men? How do the male subjects as a group differ from the female subjects as a group?

Sex	Mass	Rate	Sex	Mass	Rate
M	62	1792	F	40.3	1189
M	62.9	1666	F	33.1	913
F	36.1	995	M	51.9	1460
F	54.6	1425	F	42.4	1124
F	48.5	1396	F	34.5	1052
F	42	1418	F	51.1	1347
M	47.4	1362	F	41.2	1204
F	50.6	1502	M	51.9	1867
F	42	1256	M	46.9	1439
M	48.7	1614			

2.23 Table 2.3 (reproduced below) shows the progress of world record times (in seconds) for the 10,000 meter run up to mid-2004. concentrate on the women's world record times. Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.

Women's Record Times			
1967	2286.4	1982	1895.3
1970	2130.5	1983	1895.0
1975	2100.4	1983	1887.6
1975	2041.4	1984	1873.8
1977	1995.1	1985	1859.4
1979	1972.5	1986	1813.7
1981	1950.8	1993	1771.8
1981	1937.2		

2.31 Here are the data for the second test and the final exam for the same students as in problem 2.6 (and 2.30):

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

(a) Find the correlation between these two variables.

2.45 Table 1.10 (reproduced below) gives the city and highway gas mileage for 21 two-seater cars, including the Honda Insight gas-electric hybrid car.

(a) Make a scatterplot of highway mileage *y* against city mileage *x* for all 21 cars. There is a strong positive linear association. The Insight lies far from the other points. Does the Insight extend the linear pattern of the other card, ir is it far from the line they form?
(b) Find the correlation between city and highway mileages both without and with the Insight. Based on your answer to (a), explain why r changes in this direction when you add the Insight.

City	Hwy	City	Hwy
17	24	9	13
20	28	15	22
20	28	12	17
17	25	22	28
18	25	16	23
12	20	13	19

11	16	20	26
10	16	20	29
17	23	15	23
60	66	26	32
9	15		

2.59 Here are the data for the second test and the final-exam scores (again).

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

(a) Plot the data with the second-test scores on the *x* axis and the final-exam scores on the *y* axis. .

(b) Find the least-squares regression line for predicting the final-exam score using the second-test score.

(c) Graph the least-squares regression line on your plot.

2.69 Table 2.4 (reproduced below) gives data on the growth of icicles at two rates of water flow. You examined these data in Exercise 2.24. Use least-squares regression to estimate the rate (centimeters per minute) at which icicles grow at these two flow rates. How does flow rate affect growth?

Run 8903				Run 8905			
Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11
30	2.9	150	21	30	1	150	11.9
40	4	160	23.4	40	1.3	160	12.7
50	5	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

2.87 A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life:

Age (months)	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg)	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

(a) Plot weight against time.

(b) A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is:

The regression equation is
Weight = 4.88 + 0.267 age

Plot this line on your graph. Is it an acceptable summary of the overall pattern of growth? Remember that you can calculate the least-squares line for any set of two-variable data. It's up to your to decide if it makes sense to fit a line.

(c) Fortunately, the software also prints out the residuals from the least-squares line. In order of age along the rows, they are

-0.85	-0.31	0.02	0.35	0.58	0.62
0.45	0.18	-0.08	-0.35	-0.32	-0.28

Verify that the residuals have sum zero (except for roundoff error). Plot the residuals against age and add a horizontal line at zero. Describe carefully the pattern that you see.

2.93 Careful statistical studies often include examination of potential lurking variables. This was true of the study of the effect of nonexercise activity (NEA) on fat gain (Example 2.12, page 109), our lead example in Section 2.3. Overeating may lead our bodies to spontaneously increase NEA (fidgeting and the like). Our bodies might also spontaneously increase their basal metabolic rate (BMR), which measures energy use while resting. If both energy uses increased, regressing fat gain on NEA alone would be misleading. Here are data on BMR and fat gain for the same 16 subjects whose NEA we examined earlier:

BMR increase (cal)	117	352	244	-42	-3	134	136	-32
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
BMR increase (cal)	-99	9	-15	-70	165	172	100	35
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

The correlation between NEA and fat gain is r = -0.7786. The slope of the regression line for predicting fat gain from NEA is b_1 = -0.00344 kilogram per calorie. What are the correlation and slope for BMR and fat gain? Explain why these values show that BMR has much less effect on fat gain than does NEA.

2.119 A market research firm conducted a survey of companies in its state. They mailed a questionnaire to 300 small companies, 300 medium-sized companies, and 300 large companies. The rate of nonresponse is important in deciding how reliable survey results are. Here are the data on response to this survey.

Size of company	Response	No Response	Total
Small	175	125	300
Medium	145	155	300
Large	120	180	300

(a) What is the overall percent of nonresponse?

(b) Describe how nonresponse is related to the size of business (Use percents to make your statements precise).

(c) Draw a bar graph to compare the nonresponse percents for the three size categories.

(d) Using the total number of responses as a base, compute the percent of responses that come from each of small, medium, and large businesses.

(e) The sampling plan was designed to obtain equal numbers of responses from small, medium, and large companies. In preparing an analysis \of the survey results, do you think it would be reasonable to preceed as if the responses represented companies of each size equally?

Chapter 3 Exercises

3.27 Doctors identify “chronic tension-type headaches” as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Investigators compared four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. Outline the design of the experiment. The headache sufferers named below have agreed to participate in the study. Use software or Table B at line 151 to randomly assign the subjects to the treatments.

3.43 We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this will speed their recovery. We might measure recovery from intense exercise as follows: Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Because players vary greatly in speed, you plan a matched pairs experiment using 20 football players as subjects. Describe the design of such an experiment to investigate the effect of inhaling oxygen during the rest period. Why should each player’s two trials be on different days? Use Table B at line 140 to decide which players will get oxygen on their first trial.

3.51 The walk to your statistics class takes about 10 minutes, about the amount of time needed t listen to three songs on your iPod. You decide to take a simple random sample of songs from a Billboard list of Rock Songs. Here is the list:

1	Miss Murder	2	Animal I Have Become	3	Steady As She Goes	4	Dani California
5	The Kill (Bury Me)	6	Original Fire	7	When You Were Young	8	MakeD – Sure
9	Vicarious	10	The Diary of Jane				

Select the three songs for your iPod using a simple random sample.

3.57 You are planning a report on apartment living in a college town. You decide to select 5 apartment complexes at random for in-depth interviews with residents. Select a simple random sample of 5 of the following apartment complexes. If you use Table B, start at line 137.

1	Ashley Oaks	2	Country View	3	Mayfair Village
4	Bay Pointe	5	Country Villa	6	Nobb Hill
7	Beau Jardin	8	Crestview	9	Pemberly Courts
10	Bluffs	11	Del-Lynn	12	Peppermill

13	Brandon Place	14	Fairington	15	Pheasant Run
16	Briarwood	17	Fairway Knolls	18	Richfield
19	Brownstone	20	Fowler	21	Sagamore Ridge
22	Burberry	23	Franklin Park	24	Salem Courthouse
25	Cambridge	26	Georgetown	27	Village Manor
28	Chauncey Village	29	Greenacres	30	Waterford Court
31	Country Squire	32	Lahr House	33	Williamsburg

3.67 Stratified samples are widely used to study large areas of forest. Based on satellite images, a forest area in the Amazon basin is divided into 14 types. Foresters studied the four most commercially valuable types: alluvial climax forests of quality levels 1, 2, and 3, and mature secondary forest. They divided the area of each type into large parcels, chose parcels of each type at random, and counted tree species in a 20-by-25 meter rectangle randomly placed within each parcel selected. Here is some detail:

Forest Type	Total Parcels	Sample Size
Climax 1	36	4
Climax 2	72	7
Climax 3	31	3
Secondary	42	4

Choose the stratified sample of 18 parcels. Be sure to explain how you assigned labels to parcels. If you use Table B, start at line 140.

3.91 We can construct a sampling distribution by hand in the case of a very small population. The population contains 10 students. Here are their scores on an exam:

Student	0	1	2	3	4	5	6	7	8	9
Score	82	62	80	58	72	73	65	66	74	62

The parameter of interest is the mean score, which is 69.4. The sample is an SRS of $n = 4$ students drawn from this population. The students are labeled 0 to 9 so that a simble random digit from table B chooses one student for the sample.

(a) Use table B to draw an SRS of size 4 from this population. Write the four scores in your sample and calculate the mean $\overline{x}$ of the sample scores. This statistic is an estimate of the population parameter.

(b) Repeat this process 9 more times. Make a histogram of the 10 values of $\overline{x}$. Is the center of your histogram close to 69.4? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students – using different parts of Table B- to obtain several hundred repetitions and make a histogram of the values of $\overline{x}$. This histogram is a better approximation to the sampling distribution.)

Chapter 4 Exercises

4.7 The basketball player Shaquille O'Neal makes about half of his free throws over an entire season. Use Table B or the Probability applet to simulate 100 free throws show independently by a player who as probability 0.5 of making each shot.

(a) What percent of the 100 shots did he hit?

(b) Examine the sequence of hits and misses. How long was the longest run of shots made? Of shots missed? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)

4.51 Spell-checking software catches "nonword errors," which result in a string of letters that is not a word, as when "the" is typed as "the." When undergraduates are asked to write a 250 word essay (without spell checking), the number X of nonword errors has the following distribution:

Value of X	0	1	2	3	4
Probability	0.1	0.3	0.3	0.2	0.1

(a) Sketch the probability distribution for this random variable.

4.65 How many close friends do you have? Suppose that the number of close friends adults claim to have varies from person to person with mean $\mu = 9$ and standard deviation $\sigma = 2.5$. An opinion poll asks this question of an SRS of 1100 adults. We will see in the next chapter that in this situation the sample mean response $\overline{x}$ has approximately the Normal distribution with mean 9 and standard deviation 0.075. What is $P(8 \le \overline{x} \le 10)$, the probability that the statistic $\overline{x}$ estimates the parameter μ to within ±1?

4.73 Example 4.22 gives the distribution of grades (A = 4, B = 3, and so on)in English 210 at North Carolina State University as

Value of X	0	1	2	3	4
Probability	0.05	0.04	0.20	0.40	0.31

Find the average (that is, the mean) grade in this course.

4.89 According to the current Commissioners' Standard Ordinary mortality table, adopted by state insurance regulators in December 2002, a 25-year-old man has these probabilities of dying during the next five years:

Age at death	25	26	27	28	29
Probability	0.00039	0.00044	0.00051	0.00057	0.00060

(a) What is the probability that the man does not die in the next five years?

(b) An online insurance site offers a term insurance policy that will pay $100,000 if a 25-year-old man dies within the next 5 years. The cost is $175 per year. So the insurance company will take in $875 from this policy if the man does not die within five years. If he does die, the company must pay $100,000. Its loss depends on how many premiums were paid, as follows:

Age at death	25	26	27	28	29
Loss	$99,825	$99,650	$99,475	$99,300	$99,125

What is the insurance company's mean cash intake from such policies?

4.137 A grocery store gives its customers cards that may win them a prize when matched with other cards. The back of the card announces the following probabilities of winning various amounts if a customer visits the store 10 times:

Amount	$1000	$250	$100	$10
Probability	1/10,000	1/1000	1/100	1/20

(a) What is the probability of winning nothing?

(b) What is the mean amount won?

(c) What is the standard deviation of the amount won?

Chapter 5 Exercises

5.5 (a) Suppose X has the B(4, 0.3) distribution. Use software or Table C to find P(X = 0) and P($X \geq 3$).

(b) Suppose X has the B(4, 0.7) distribution. Use software or Table C to find P(X = 4) and P($X \leq 1$).

5.7 Suppose we toss a fair coin 100 times. Use the Normal approximation to find the probability that the sample proportion is

(a) between 0.4 and 0.6

(b) between 0.45 and 0.55.

5.13 Typographic errors in a text are either nonword errors (as when "the" is typed as "teh") or word errors that result in a real but incorrect word. Spell-checking software will catch nonword errors but not word errors. Human proofreaders catch 70% of word errors. You ask a fellow student to proofread an essay in which you have deliberately made 10 word errors.

(a) If the student matches the usual 70% rate, what is the distribution of the number of errors caught? What is the distribution of the number of errors missed?

(b) Missing 4 or more out of 10 errors seems a poor performance. What is the probability that a proofreader who catches 70% of word errors misses 4 or more out of 10?

5.17 In the proofreading setting of Exercise 5.13, what is the smallest number of misses m with P($X \geq m$) no larger than 0.05? You might consider m or more misses as evidence that a proofreader actually catches fewer than 70% of word errors.

5.21 Children inherit their blood type from their parents, with probabilities that reflect the parents' genetic makeup. Children of Juan and Maria each have probability ¼ of having blood type A and inherit independently of each other. Juan and Maria plan to have 4 children; let X be the number who have blood type A.

(a) What are n and p in the binomial distribution of X?

(b) Find the probability of each possible value of X, and draw a probability histogram for this distribution.

(c) Find the mean number of children with type A blood, and mark the location of the mean on your probability histogram.

5.25 The Harvard College Alcohol Study finds that 67% of college students support efforts to "crack down on underage drinking." The study took a sample of almost 15,000 students, so the population proportion who support a crackdown is very close to p = 0.67.

The administration of your college surveys an SRS of 200 students and finds that 140 support a crackdown on underage drinking.

(a) What is the sample proportion who support a crackdown on underage drinking?

(b) If in fact the proportion of all students on your campus who support a crackdown is the same as the national 67%, what is the probability that the proportion in an SRS of 200 students is as large or larger than the result of the administration's sample?

5.31 One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known demographic facts about the population. The 2000 census found that 23,772,494 of the 209,128,094 adults (aged 18 and over) in the United States called themselves "Black or African American."

(a) What is the population proportion *p* of blacks among American adults?

(b) An opinion poll chooses 1200 adults at random. What is the mean number of blacks in such samples? (Explain the reasoning behind your calculation.)

(c) Use a Normal approximation to find the probability that such a sample will contain 100 or fewer blacks. Be sure to check that you can safely use the approximation.

5.49 The gypsy moth is a serious threat to oak and aspen trees. A state agriculture department places traps throughout the state to detect the moths. When traps are checked periodically, the mean number of moths trapped is only 0.5, but some traps have several moths. The distribution of moth counts is discrete and strongly skewed, with standard deviation 0.7.

(a) What are the mean and standard deviation of the average number of moths $\overline{x}$ in 50 traps?

(b) Use the central limit theorem to find the probability that the average number of moths in 50 traps is greater than 0.6.

5.53 Sheila's measured glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl. What is the level *L* such that there is probability only 0.05 that the mean glucose level of 3 test results falls above *L* for Sheila's glucose level distribution?

5.55 In response to the increasing weights of airline passenger, the Federal Aviation Administration told airlines to assume that passengers weigh an average of 190 pounds in the summer, including clothing and carry-on baggage. But passengers vary: the FAA gave a mean but not a standard deviation. A reasonable standard deviation is 35 pounds. Weights are not Normally distributed, especially when the population includes both men and women, but they are very non-Normal. A commuter plane carries 25 passengers.

What is the approximate probability that the total weight of the passengers exceeds 5200 pounds?

5.57 The distribution of annual returns on common stocks is roughly symmetric, but extreme observations are more frequent than in a Normal distribution. Because the distribution is not strongly non-Normal, the man return over even a moderate number of years is close to Normal. Annual real returns on the Standard & Poor's 500 stock index over the period 1871 to 2004 have varied with mean 9.2% and standard deviation 20.6%. Andrew plans to retire in 45 years and is considering investing in stocks. What is the probability (assuming that the past pattern of variation continues) that the mean annual return on common stocks over the next 45 years will exceed 15%? What is the probability that the mean return will be less than 5%?

Chapter 6 Exercises

6.5 An SRS of 100 incoming freshmen was taken to look at their college anxiety level. The mean score of the sample was 83.5 (out of 100). Assuming a standard deviation of 4, give a 95% confidence interval for μ, the average anxiety level among all freshmen.

6.7 You are planning a survey of starting salaries for recent marketing majors. In 2005, the average starting salary was reported to be $37,832. Assuming the standard deviation for this study is $10,500, what sample size do you need to have a margin of error equal to $900 with 95% confidence?

6.17 For many important processes that occur in the body, direct measurement of characteristics is not possible. In many cases, however, we can measure a *biomarker*, a biochemical substance that is relatively easy to measure and is associated with the process of interest. Bone turnover is the net effect of two processes: the breaking down on old bone, called resorption, and the building of new bone, called formation. One biochemical measure of bone resorption is tartrate resistant acid phosphatase (TRAP), which can be measured in blood. In a study of bone turnover in young women, serum TRAP was measured in 31 subjects. The units are units per liter (U/l). The mean was 13.2 U/l. Assume that the standard deviation is known to be 6.5 U/l. Give the margin of error and find a 95% confidence interval for the mean for young women represented by this sample.

6.29 A new bone study is being planned that will measure the biomarker TRAP described in Exercise 6.17. Using the value of σ given there, 6.5 U/l, find the sample size required to provide an estimate of the mean TRAP with a margin of error of 2.0 U/l for 95% confidence.

6.43 You will perform a significance test of H_0: $\mu = 25$ based on an SRS of n = 25. Assume $\sigma = 5$.

(a) If $\overline{x}$ = 27, what is the test statistic z?

(b) What is the p-value if H_A: $\mu > 25$?

(c) What is the p-value if H_A: $\mu \neq 25$?

6.57 A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic z = -1.73.

(a) What is the p-value if the alternative is H_A: $\mu > \mu_0$?

(b) What is the p-value if the alternative is H_A: $\mu < \mu_0$?

(c) What is the p-value if the alternative is H_A: $\mu \neq \mu_0$?

6.69 The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\overline{x} = 132.2$.

(a) Assuming that $\sigma = 30$ for the population of older students, carry out a test of H$_0$: $\mu = 115$ and H$_A$: $\mu > 115$. Report the *p*-value of your test, and state your conclusion clearly.

(b) Your test in (a) required two important assumptions in addition to the assumption that the value of σ is known. What are they? Which of these assumptions is most important to the validity of your conclusion in (a)?

6.71 Refer to Exercise 6.26. In addition to the computer computing mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at teach fill-up. The following data are the differences between the computer's and the driver's calculations for that random sample of 20 records. The driver wants to determine if these calculations are different. Assume the standard deviation of a difference to be $\sigma = 3.0$.

5.0	6.5	-0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	-0.6	-4.2

(a) State the appropriate H$_0$ and H$_A$ to test this suspicion.

(b) Carry out the test. Give the *p*-value, and then interpret the result in plain language.

6.95 Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Let us suppose that SAT Mathematics (SATM) scores in the absence of coaching very Normally with mean $\mu = 505$ and $\sigma = 100$. Suppose further that coaching may change μ but does not change σ. An increase in the SATM from 505 to 508 is of no importance in seeking admission to college, but this unimportant change can be statistically very significant. To see this, calculate the *p*-value for the test of H$_0$: $\mu = 505$ against H$_A$: $\mu > 505$ in each of the following situations:

(a) A coaching service coaches 100 students; their SATM scores average $\overline{x} = 508$.

(b) By the next year, this service has coached 1000 students; their SATM scores average $\overline{x} = 508$.

(c) An advertising campaign brings the number of students coached to 10,000; their SATM scores average $\overline{x} = 508$.

6.113 Example 6.16 gives a test of a hypothesis about the SAT scores of California high school students based on an SRS of 500 students. The hypotheses are H_0: $\mu = 450$ and H_A: $\mu > 450$. Assume that the population standard deviation is $\sigma = 100$. The test rejects H_0 at the 1% level of significance when $z \geq 2.326$, where

$$z = \frac{\overline{x} - 450}{100/\sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 10 points in the population mean SAT score? Answer this question by calculating the power of the test against the alternative $\mu = 460$.

Chapter 7 Exercises

7.3 You randomly choose 15 unfurnished one-bedroom apartments from a large number of advertisements in you local newspaper. You calculate that their mean monthly rent of $570 and their standard deviation is $105. Construct a 95% confidence interval for the mean monthly rent of all advertised one-bedroom apartments.

7.5 A test of a null hypothesis versus a two-sided alternative gives $t = 2.35$.

(a) The sample size is 15. Is the test result significant at the 5% level?

(b) The sample size is 6. Is the test result significant at the 5% level?

7.25 A study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, is described in Example 6.1. For each tree in the tract, the researchers measured the diameter at breast height (DBH). This is the diameter of the three at 4.5 feet and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

(a) Use a histogram or stemplot and a boxplot to examine the distribution of DBHs. Include a Normal quantile plots if you have the necessary software. Write a careful description of the distribution.

(b) Is it appropriate to use the methods of this section to find a 95% confidence interval for the mean DBH of all trees in the Wade Tract? Explain why or why not.

(c) Report the mean and margin of error and the confidence interval.

7.29 Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. Then they were asked to rate how luck played a role in determining their scores. This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5
1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

(a) Use graphical methods to display the distribution. Describe any unusual characteristics. Do you think that these would lead you to hesitate before using the Normality-based methods of this section?

(b) Give a 95% confidence interval for the mean luck score.

7.33 Nonexercise activity thermogenesis (NEAT) provides a partial explanation for the results you found in Exercise 7.32. NEAT is energy burned by fidgeting, maintenance of posture, spontaneous muscle contraction, and other activities of daily living. In the study of the previous exercise, the 16 subjects increased their NEAT by 328 calories per day, on average, in response to the additional food intake. The standard deviation was 256.

(a) Test the null hypothesis that there was no change in NEAT versus the two-sided alternative. Summarize the results of the test and five your conclusion.

(b) Find a 95% confidence interval for the change in NEAT. Discuss the additional information provided by the confidence interval that is not evident from the results of the significance test.

7.35 Refer to Exercise 7.24. In addition to the computer calculating mpg, the driver also recorded the mpg by dividing the miles driven by the amount of gallons at fill-up. The driver wants to determine if these calculations are different.

Fill-up	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

(a) State the appropriate H_0 and H_A.

(b) Carry out the test. Give the *p*-value, and then interpret the result.

7.49 Use the sign test to assess whether the computer calculates a higher mpg than the driver in Exercise 7.35. State the hypotheses, give the *p*-value using the binomial table (Table C), and report your conclusion.

7.57 Assume $\bar{x}_1 = 100$, $\bar{x}_2 = 120$, $s_1 = 10$, $s_2 = 12$, $n_1 = 10$, $n_2 = 10$. Find a 95% confidence interval for the difference in the corresponding values of μ Would you reject the null hypothesis that the population means are equal in favor of the two-sided alternate at significance level 0.05? Explain.

7.61 A recent study at Baylor University investigated the lipid levels in a cohort of sedentary university students. A total of 108 students volunteered for the study and met the eligibility criteria. The following table summarizes the blood lipid levels, in milligrams per deciliter (mg/dl), of the participants broken down by gender:

	Females (n = 71)		Males (n = 37)	
	$\overline{x}$	s	$\overline{x}$	s
Total Cholesterol	173.70	34.79	171.81	33.24
LDL	96.38	29.78	109.44	31.05
HDL	61.62	13.75	46.47	7.94

(a) Is it appropriate to use the two-sample t procedures that we studied in this section to analyze these data for gender differences? Give reasons for your answer.
(b) Describe the appropriate null and alternative hypotheses for comparing male and females total cholesterol levels.
(c) Carry out the significance test. Report the test statistic with the degrees of freedom and the p-value. Write a short summary of your conclusion.
(d) Find a 95% confidence interval for the difference between the two means. Compare the information given by the interval with the information given by the test.

7.83 A market research firm supplies manufacturers with estimates of the retail sales of their products form samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 70 stores thismonth shows mean sales of 53 units of a small appliance, with standard deviation 15 units. During the same month last year, an SRS of 55 stores gave mean sales of 50 units, with standard deviation 18 units. An increase from 50 to 53 is 6%. The marketing manager is happy because sales are up 6%.

(a) Use the two-sample t procedure to give a 95% confidence interval for the difference in mean number of units sold at all retail stores.
(b) Explain in language that the marketing manager can understand why he cannot be certain that sales rose by 6%, and that in fact sales may even have dropped.

7.95 The F statistic $F = {s_1^2}/{s_2^2}$ is calculated from samples of size $n_1 = 16$ and $n_2 = 21$.

(a) What is the upper 5% critical value for this F?
(b) In a test of equality of standard deviations against the two-sided alternative, this statistic has the value F = 2.45. Is this value significant at the 10% level? Is it significant at the 5% level?

7.99 Compare the standard deviations of total cholesterol in Exercise 7.61. Give the test statistic, the degrees of freedom, and the *p*-value. Write a short summary of your analysis, including comments on the assumptions of the test.

Chapter 8 Exercises

8.1 In a 2004 survey of 1200 undergraduate students throughout the United States, 89% of the respondents said they owned a cell phone. For 90% confidence, what is the margin of error?

8.3 A 1993 nationwide survey by the National Center for Education Statistics reports that 72% of all undergraduates work while enrolled in school. You decide to test whether this percent is different at your university. In your random sample of 100 students, 77 said they were currently working.

(a) Give the null and alternative hypotheses.

(b) Carry out the significance test. Report the test statistic and p-value.

(c) Does is appear that the percent of students working at your university is different at the $\alpha = 0.05$ level?

8.5 Refer to Example 8.6 (page 499). Suppose the university was interested in a 90% confidence interval with margin of error 0.03. Would the required sample size be smaller or larger than 1068 students? Verify this by performing the calculation.

8.11 Gambling is an issue of great concern to those involved in Intercollegiate athletics. Because of this, the National Collegiate Athletic Association (NCAA) surveyed student-athletes concerning their gambling-related behaviors. There were 5594 Division I male athletes in the survey. Of these, 3547 reported participation in some gambling behavior. This included playing cards, betting on games of skill, buying lottery tickets, and betting on sports.

(a) Find the sample proportion and the large-sample margin of error for 95% confidence. Explain in simple terms the 95%.

8.15 The Pew Poll of n = 1048 U.S. drivers found that 38% of the respondents “shouted, cursed, or made gestures to other drivers” in the last year.

(a) Construct a 95% confidence interval for the true proportion of U.S. drivers who did these actions in the last year.

8.23 For a study of unhealthy eating behaviors, 267 college women aged 18 to 25~years were surveyed. Of these, 69% reported that they were on a diet sometime during the past year. Give a 95% confidence interval for the true proportion of college women aged 18 to 25 years in this population who dieted last year.

8.29 The South African mathematician John Kerrich, while a prisoner of war during world War II, tossed a coin 10,000 times and obtained 5067 heads.

(a) Is this significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5? Use a sketch of the standard Normal Distribution to illustrate the *p*-value.

(b) Use a 95% confidence interval to find the range of probabilities of heads that would not be rejected at the 5% level.

8.31 Suppose after reviewing the results of the previous survey, you proceeded with preliminary development of the product. Now you are at the stage where you need to decide whether or not to make a major investment to produce and market it. You will use another random sample of your customers but not you want the margin of error to be smaller. What sample size wyoud you use if you wanted the 95% margin of error to be 0.075 or less?

8.35 A study was designed to compare two energy drink commercials. Each participant was shown the commercials in random order and asked to select the better one. Commercial A was selected by 45 out of 100 women and 80 out of 140 men. Give an estimate of the difference in gender proportions that favored Commercial A. Also construct a large-sample 95% confidence interval for this difference.

8.41 In Exercise 8.4, you were asked to compare the 2004 proportion of cell phone owners (89%) with the 2003 estimate (83%). It would be more appropriate to compare these two proportions using the methods of this section. Given that the sample size of each SRS is 1200 students, compare these to years with a significance test, and give an estimate of the difference in proportions of undergraduate cell phone owners with a 95% margin of error.

8.49 A 2005 survey of Internet users reported that 22% downloaded music onto their computers. The filing of lawsuits by the recording industry may be a reason why this percent has decreased from the estimate of 29% from a survey taken two years before. Assume that the sample sizes are both 1421. Using a significance test, evaluate whether or not there has been a change in the percent fo Internet users who download music. Provide all the details for the test and summarize your conclusion. Also report a 95% confidence interval for the difference in proportions and explain what information is provided in the interval that is not in the significance test results.

Chapter 9 Exercises

9.5 M&M Mars Company has varied the mix of colors for M&M's Milk Chocolate Candies over the years. These changes in color blends are the result of consumer preference tests. Most recently, the color distribution is reported to be 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. You open up a 14-ounce bad of M&M's and find 61 frown, 59 yellow, 49 red, 77 orange, 141 blue, and 88 green. Use a goodness of fit test to examine how well this bag fits the percents stated by the M&M Mars company.

9.11 Cocaine addiction is difficult to overcome. Addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to take cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a substance containing no medication, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each treatment. Following are the results:

	Cocaine relapse?	
Treatment	Yes	No
Desipramine	10	14
Lithium	18	6
Placebo	20	4

(a) Compare the effectiveness of the three treatments in preventing relapse using percents and a bar graph. Write a brief summary.

(b) Can we comfortable use the chi-square test to test the null hypothesis that there is no difference between treatments? Explain.

(c) Perform the significance test and summarize the results.

9.17 As part of the 1999 College Alcohol Study, students who drank alcohol in the last year were asked if drinking ever resulted in missing a class. The data are given in the following table:

	Binging Status		
Missed Class	Nonbinger	Occasional	Frequent
No	4617	2047	1176
Yes	446	915	1959

(a) Summarize the results of this table graphically and numerically.

(b) What is the marginal distribution of drinking status? Display the results graphically.

(c) Compute the relative risk of missing a class for occasional bingers versus nonbingers and for frequent bingers versus nonbingers. Summarize these results.

(d) Perform the chi-square test for this two-way table. Give the test statistic, degrees of freedom, the *p*-value, an your conclusion.

9.21 A recent study of undergraduates looked at gender differences in dieting trends. There were 181 women and 105 men who participated in the survey. The table below summarizes whether a student tried a low-fat diet or not by gender.

	Gender	
Tried low-fat diet	Women	Men
Yes	35	8
No		

(a) Fill in the missing cells of the table.

(b) Summarize the data numerically and graphically.

(c) Test that there is no association between gender and the likelihood of trying a low-fat diet. Summarize the results.

9.25 *E. jugularis* is a type of hummingbird that lives in the forest preserves of the Caribbean island of Santa Lucia. The males and the females of this species have bills that are shaped somewhat differently. Researchers who study these birds thought that the bill shape might be related to the shape of the flowers that the visit for food. The researchers observed 49 females and 21 males. Of the females, 20 visited the flowers of *H. bihai*, while none of the males visited these flowers. Display the data in a two-way table and perform the chi-square test. Summarize the results and five a brief statement of your conclusion. Your two-way table has a count of zero in one cell. Does this invalidate your significance test? Explain why or why not.

9.31 The study of shoppers in secondhand stores cited in the previous exercise also compared the income distribution of shoppers in the two stores. Hers is the two-way table of counts:

Income	City 1	City 2
Under $10,000	70	62
$10,000 to $19,999	52	63
$20,000 to $24,999	69	50
$25,000 to $34,999	22	19
$35,000 or more	28	24

Verify that the $\chi2$ statistic for this table is $\chi2 = 3.955$. Give the degrees of freedom and the *p*-value. Is there good evidence that the customers at the two stores have different income distributions?

9.35 In one part of the study described in Exercise 9.34, students were asked to respond to some questions regarding their interests and attitudes. Some of these questions form a scale called PEOPLE that measures altruism, or an interest in the welfare of others. Each student was classified as low, medium, or high on this scale. Is there an association between PEOPLE score and field of study? Here are the data:

	PEOPLE score		
Field of Study	Low	Medium	High
Agriculture	5	27	35
Child Dev. and Fam. Studies	1	32	54
Engineering	12	129	94
Liberal arts and education	7	77	129
Management	3	44	28
Science	7	29	24
Technology	2	62	64

Analyze the data and summarize your results. Are there some fields of study that have very large or very small proportions of students in the high-PEOPLE category?

9.41 The 2005 National Survey of Student Engagement reported on the use of campus services during the first year of college. In terms of academic assistance (for example tutoring, writing lab), 43% never used the services, 35% sometimes used the services,, 15% often used the services, and 7% very often used the services. You decide to see if your large university has this same distribution. You survey first-year students and obtain the counts 79m 83, 36, and 12 respectively. Use a goodness of fit test to examine how well your university reflects the national average.

Chapter 10 Exercises

10.5 The National Science Foundation collects data on the research and development spending by universities and colleges in the United States. Here are the data for the years 1999 to 2001 (using 1998 dollars):

Year	1999	2000	2001
Spending (billions of dollars)	26.4	28.0	29.7

Do the following by hand or with a calculator and verify your results with a software package.

(a) Make a scatterplot that shows the increase in research hand development spending over time. Does the pattern suggest that the spending in increasing linearly over time?

(b) Find the equation of the least-squares regression line for prediction spending from year. Add this line to your scatterplot.

(c) For each of the three years, find the residual. Use these residuals to calculate the standard error s.

(d) Write the regression model for this setting What are your estimates of the unknown parameters in this model?

(e) Compute a 95% confidence interval for the slope and summarize what this interval tells you about the increase in spending over time.

10.9 For each of the settings in the previous exercise, test the null hypothesis that the slope is zero versus the two-sided alternate.

(a) $n = 25,\ \hat{y} = 1.3 + 12.10x,$ and $SE_{b1} = 6.31$

(b) $n = 25,\ \hat{y} = 13.0 + 6.10x,$ and $SE_{b1} = 6.31$

10.11 Refer to Exercise 10.10.

(a) Construct a 95% confidence interval for the slope. What does this interval tell you about the percent increase in tuition between 2000 and 2005?

(b) The tuition at Stat U was $5000 in 2000. What is the predicted tuition in 2005?

(c) Find a 95% prediction interval for the 2005 tuition at Stat U and summarize the results.

Table 10.1 In-state tuition and fees (in dollars) for 32 public universities

University	Year 2000	Year 2005	University	Year 2000	Year 2005
Penn State	7018	11508	Purdue	3872	6458
Pittsburgh	7002	11436	Cal-San Diego	3848	6685
Michigan	6926	9798	Cal-Santa Barbara	3832	6997
Rutgers	6333	9221	Oregon	3819	5613
Michigan State	5432	8108	Wisconsin	3791	6284
Maryland	5136	7821	Washington	3761	5610
Illinois	4994	8634	UCLA	3698	6504
Minnesota	4877	8622	Texas	3575	6972
Missouri	4726	7415	Nebraska	3450	5540
Buffalo	4715	6068	Iowa	3204	5612
Indiana	4405	7112	Colorado	3188	5372
Ohio State	4383	8082	Iowa State	3132	5634
Virginia	4335	7370	North Carolina	2768	4613
Cal-Davis	4072	7457	Kansas	2725	5413
Cal-Berkeley	4047	6512	Arizona	2348	4498
Cal-Irvine	3970	6770	Florida	2256	3094

10.17 Consider the data in Table 10.3 and the relationship between IBI and the percent of watershed area that was forest. The relationship between these two variables is almost significant at the .05 level. In this exercise you will demonstrate the potential effect of an outlier on statistical significance. Investigate what happens when you decrease the IBI to 0.0 for (1) an observation with 0% forest and (2) an observation with 100% forest.

Table 10.3 Percent forest and index of biotic integrity

Forest	IBI	Forest	IBI	Forest	IBI	Forest	IBI	Forest	IBI
0	47	9	33	25	62	47	33	79	83
0	61	10	46	31	55	49	59	80	82
0	39	10	32	32	29	49	81	86	82
0	59	11	80	33	29	52	71	89	86
0	72	14	80	33	54	52	75	90	79
0	76	17	78	33	78	59	64	95	67
3	85	17	53	39	71	63	41	95	56
3	89	18	43	41	55	68	82	100	85
7	74	21	88	43	58	75	60	100	91
8	89	22	84	43	71	79	84		

10.23 *Storm Data* is a publication of the National Climatic Data Center that contains a listing of tornadoes, thunderstorms, floods, lightning, temperature extremes, and other weather phenomena. Table 10.4 summarizes the annual number of tornadoes in the United States between 1953 and 2005.

(a) Make a plot of the total number of tornadoes by year. Does a linear trend over the years appear reasonable?
(b) Are there any outliers or unusual patterns? Explain your answer.
(c) Run the simple linear regression and summarize the results, making sure to construct a 95% confidence interval for the average annual increase in the number of tornadoes.
(d) Obtain the residuals and plot them versus year. Is there anything unusual in the plot?
(e) Are the residuals Normal? Justify your answer.

Table 10.4 Annual number of tornadoes in the United States between 1953 and 2005

Year	Count	Year	Count	Year	Count	Year	Count
1953	421	1967	926	1981	783	1995	1235
1954	550	1968	660	1982	1046	1996	1173
1955	593	1969	608	1983	931	1997	1148
1956	504	1970	653	1984	907	1998	1449
1957	856	1971	888	1985	684	1999	1340
1958	564	1972	741	1986	764	2000	1076
1959	604	1973	1102	1987	656	2001	1213
1960	616	1974	947	1988	702	2002	934
1961	697	1975	920	1989	856	2003	1372
1962	657	1976	835	1990	1133	2004	1819
1963	464	1977	852	1991	1132	2005	1194
1964	704	1978	788	1992	1298		
1965	906	1979	852	1993	1176		
1966	585	1980	866	1994	1082		

10.25 In Exercise 7.26 we examined the distribution of C-reactive protein (CRP) in a sample of 40 children from Papua New Guinea. Serum retinol values for the same children were studied in Exercise 7.28. One important question that can be addressed with these data is whether or not infections, as indicated by CRP, cause a decrease in the measured values of retinol, low values of which indicate a vitamin A deficiency. The data are given in Table 10.5.

(a) Examine the distribution of CRP and serum retinol. Use graphical and numerical methods.
(b) Forty percent of the CRP values are zero. Does this violate any assumptions that we need to do a regression analysis using CRP to predict serum retinol? Explain your answer.
(c) Run the regression, summarize the results, and write a short patagraph explaining your conclusions.

(d) Explain the assumptions needed for your results to be valid. Examine the data with respect to these assumptions and report your results.

Table 10.5 C-reactive protein and serum retinol

CRP	RETINOL	CRP	RETINOL	CRP	RETINOL	CRP	RETINOL	CRP	RETINOL
0	1.15	30.61	0.97	22.82	0.24	5.36	1.19	0	0.83
3.9	1.36	0	0.67	0	1	0	0.94	0	1.11
5.64	0.38	73.2	0.31	0	1.13	5.66	0.34	0	1.02
8.22	0.34	0	0.99	3.49	0.31	0	0.35	9.37	0.56
0	0.35	46.7	0.52	0	1.44	59.76	0.33	20.78	0.82
5.62	0.37	0	0.7	0	0.35	12.38	0.69	7.1	1.2
3.92	1.17	0	0.88	4.81	0.34	15.74	0.69	7.89	0.87
6.81	0.97	26.41	0.36	9.57	1.9	0	1.04	5.53	0.41

10.37 We assume that our wages will increase as we gain experience and become more valuable to our employers. Wages also increase because of inflation. By examining a sample of employees at a given point in time, we can look at part of the picture. How does length of service (LOS) relate to wages? Table 10.8 gives data on the LOS in months and wages for 60 women who work in Indiana banks. Wages are yearly total income divided by the number of weeks worked. We have multiplied wages by a constant for reasons of confidentiality.

(a) Plot wages versus LOS. Describe the relationship. There is one woman with relatively high wages for her length of service. Circle this point and do not use it in the rest of this exercise.

(b) Find the least-squares line. Summarize the significance test for the slope. What do you conclude?

(c) State carefully what the slope tells you about the relationship between wages and length of service.

(d) Give a 95% confidence interval for the slope.

Table 10.8 Bank wages and length of service (LOS)

Wages	LOS	Wages	LOS	Wages	LOS
48.3355	94	64.1026	24	41.2088	97
49.0279	48	54.9451	222	67.9096	228
40.8817	102	43.8095	58	43.0942	27
36.5854	20	43.3455	41	40.7	48
46.7596	60	61.9893	153	40.5748	7
59.5238	78	40.0183	16	39.6825	74
39.1304	45	50.7143	43	50.1742	204
39.2465	39	48.84	96	54.9451	24
40.2037	20	34.3407	98	32.3822	13
38.1563	65	80.5861	150	51.713	30

50.0905	76	33.7163	124	55.8379	95
46.9043	48	60.3792	60	54.9451	104
43.1894	61	48.84	7	70.2786	34
60.5637	30	38.5579	22	57.2344	184
97.6801	70	39.276	57	54.1126	156
48.5795	108	47.6564	78	39.8687	25
67.1551	61	44.6864	36	27.4725	43
38.7847	10	45.7875	83	67.9584	36
51.8926	68	65.6288	66	44.9317	60
51.8326	54	33.5775	47	51.5612	102

10.39 The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the differences between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer.

Year	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

(a) Plot the data. Does the trend in lean over time appear to be linear?
(b) What is the equation of the least-squares line? What percent of the variation in lean is explained by this line?
(c) Give a 99% confidence interval for th average rate of change (tenths of a millimeter per year) of the lean.

10.51 A study reported a correlation r = 0.5 based on a sample of size n = 20; another reported the same correlation based on a sample size of n = 10. For each, perform the test of the null hypothesis that $\rho = 0$. Describe the results and explain why the conclusions are different.

Chapter 11 Exercises

11.3 Recall Exercise 11.1. Due to missing values for some students, only 86 students were used in the multiple regression analysis. The following table contains the estimated coefficients and standard errors:

Variable	Estimate	SE
Intercept	–0.764	0.651
SAT Math	0.00156	0.00074
SAT Verbal	0.00164	0.00076
High school rank	1.4700	0.430
Bryant College placement	0.889	0.402

(a) All the estimated coefficients for the explanatory variables are positive. Is this what you would expect? Explain.
(b) What are the degrees of freedom for the model and error?
(c) Test the significance of each coefficient and state your conclusions.

11.35 Let's use regression methods to predict VO+, the measure of bone formation.
(a) Since OC is a biomarker of bone formation, we start with a simple linear regression using OC as the explanatory variable. Run the regression and summarize the results. Be sure to include an analysis of the residuals.
(b) because the processes of bone formation and bone resorption are highly related, it is possible that there is some information in the bone resorption variables that can tell us something about bone formation. Use a model with both OC and TRAP, the biomarker of bone resorption, to predict VO+. Summarize the results. IN the context of this model, it appears that TRAP is a better predictor of bone formation, VO+, than the biomarker of bone formation, OC. Is this view consistent with the pattern of relationships that you described in the previous exercise? One possible explanation is that, while all of these variables are highly related, TRAP is measured with more precision than OC.

11.51 For each of the four variables in the CHEESE data set. Find the mean, median, standard deviation, and interquartile range. Display each distribution by means of a stemplot and use a Normal quantile plot to assess Normality of the data. Summarize your findings. Note that when doing regressions with these data, we do not assume that these distributions are Normal. Only the residuals from our model need to be (approximately) Normal. The careful study of each variable to be analyzed in nonetheless an important first step in any statistical analysis.

11.53 Perform a simple linear regression analysis using Taste as the response variable and Acetic as the explanatory variable. Be sure to examine the residuals carefully.

Summarize your results. Include a plot of the data with the least-squares regression line. Plot the residuals versus each of the other two chemicals. Are any patterns evident? (The concentrations of the other chemicals are lurking variables for the simple linear regression.)

11.55 Repeat the analysis of Exercise 11.53 using Taste as the response variable and Lactic as the explanatory variable.

11.57 Carry out a multiple regression using Acetic and H2S to predict Taste. Summarize the results of your analysis. Compare the statistical significance of Acetic in this model with its significance in the model with Acetic alone as a predictor (Exercise 11.53). Which model do you prefer? Give a simple explanation for the fact that Acetic alone appears to be a good predictor of Taste, but with H2S in the model, it is not.

11.59 Use the three explanatory variables Acetic, H2S, and Lactic in a multiple regression to predict Taste. Write a short summary of your results, including an examination of the residuals. Based on all of the regression analyses you have carried out on these data, which model do you prefer and why?

Chapter 12 Exercises

12.15 A study compared 4 groups with 8 observations per group. An *F* statistic of 3.33 was reported.

(a) Give the degrees of freedom for this statistic and the entries in Table E that correspond to this distribution.
(b) Sketch a picture of this *F* distribution with the information from the table included.
(c) Based on the table information, how would you report the *p*-value?
(d) Can you conclude that all pairs of means are different? Explain your answer.

12.17 For each of the following situations, find the *F* statistic and the degrees of freedom. Then draw a sketch of the distribution under the null hypothesis and shade in the portion corresponding to the *p*-value. State how you would report the *p*-value.

(a) Compare 5 groups with 9 observations per group, MSE = 50, and MSG = 127.
(b) Compare 4 groups with 7 observations per group, SSG = 40, and SSE = 153.

12.23 The National Intramural-Recreational Sports Association (NIRSA) performed a survey to look at the value of recreational sports to college satisfaction and success. Responses were on a 10-point scale with 1 indicating total lack of importance and 10 indicating very high importance. The following table summarizes these results:

Class	N	Mean Score
Freshman	724	7.6
Sophomore	536	7.6
Junior	593	7.5
Senior	437	7.3

(a) To compare the mean scores across classes, what are the degrees of freedom for the ANOVA *F* statistic?
(b) The MSG =11.806. If s_p = 2.16, what is the *F* statistic?
(c) Give an approximate or exact *p*-value. What do you conclude?

12.25 An experimenter was interested in investigating the effects of two stimulant drugs (labeled A and B). She divided 20 rats equally into 5 groups (placebo, Drug A low, Drug A high, Drug B low, Drug B high) and 20 minutes after injection of the drug, recorded each rat's activity level (higher score is more active). The following table summarizes the results:

Treatment	$\overline{x}$	s
Placebo	14.00	8.00
Low A	15.25	12.25

High A	15.25	12.25
Low B	16.75	6.25
High B	22.50	11.00

(a) Plot the means versus the type of treatment. Does there appear to be a difference in the activity level? Explain.

(b) Is it reasonable to assume that the variances are equal? Explain your answer, and if reasonable, compute s_p.

(c) Give the degrees of freedom for the *F* statistic.

(d) The *F* statistic is 4.35. Find the associated *p*-value and state your conclusions.

12.29 Does bread lose its vitamins when stored? Small loaves of bread were prepared with flour that was fortified with a fixed amount of vitamins. After baking, the vitamin C content of two loaves was measured. Another two loaves were baked at the same time, stored for one day,, and then the vitamin C content was measured. In a similar manner, two loaves were stored for three, five, and seven days before measurements were taken. The units are milligrams of vitamin C per humdred grams of flour (mg/100 g). Here are the data:

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

(a) Give a table with sample size, mean, standard deviation, and standard error for each condition.

(b) Perform a one-way ANOVA for these data. Be sure to state your hypotheses, the test statistic with degrees of freedom, and the *p*-value.

(c) Summarize the data and the means with a plot. Use the plot and the ANOVA results to write a short summary of your conclusions.

12.35 Different varieties of the tropical flower Heliconia are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:

41.9	42.01	41.93	43.09	41.47	41.69	39.78	40.57
H. bihai							
39.63	42.18	40.66	37.87	39.16	37.4	38.2	38.07
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
38.1	37.97	38.79	38.23	38.87	37.78	38.01	
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
H. caribaea yellow							
H. caribaea red							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1

35.17	36.82	36.66	35.68	36.03	34.57	34.63

Do a complete analysis that includes description of the data and a significance test to compare the mean lengths of the flowers for the three species.

12.39 Kudzu is a plant that was imported to the United States from Japan and now covers over seven million acres in the South. The plant contains chemicals called isoflavones that have been shown to have beneficial effects on bones. One study used three groups of rats to compare a control group with rats that were fed wither a low dose or a high dose of isoflavones from kudzu. One of the outcomes examined was the bone mineral density in the femur (in grams per square centimeter). Here are the data:

Treatment	Bone mineral density (g/cm^2)					
Control	0.228	0.221	0.234	0.220	0.217	0.228
	0.209	0.221	0.204	0.220	0.203	0.219
	0.218	0.245	0.210			
Low dose	0.211	0.220	0.211	0.233	0.219	0.233
	0.226	0.228	0.216	0.225	0.200	0.208
	0.198	0.208	0.203			
High dose	0.250	0.237	0.217	0.206	0.247	0.228
	0.245	0.232	0.267	0.261	0.221	0.219
	0.232	0.209	0.203			

(a) Use graphical and numerical methods to describe the data.
(b) Examine the assumptions necessary for ANOVA. Summarize your findings.
(c) Use a multiple-comparisons method to compare the three groups.

12.45 Recommendations regarding how long infants in developing countries should be breast-fed are controversial. If the nutritional quality of the breast milk is inadequate because the mothers are malnourished, then there is risk in inadequate nutrition for the infant. On the other hand, the introduction of other foods carries the risk of infection from contamination. Further complicating the situation is the fact that companies that produce infant formulas and other foods benefit when these foods are consumed by large numbers of customers. One question related to this controversy concerns the amount of energy intake for infants who have other foods introduced into the diet at different ages. Part of one study compared the energy intakes, measured in kilocalories per day (kcal/d), for infants who were breast-fed exclusively for 4, 5, or 6 months. Here are the data:

Breast-fed for	Energy intake (kcal/d)						
4 months	499	620	469	485	660	588	675
	517	649	209	404	738	628	609
	617	704	558	653	548		
5 months	490	395	402	177	475	617	616

	587 368	528 538	518 519	370 506	431	518	639
6 months	585 465	647	477	445	485	703	538

(a) Make a table giving the sample size, mean, and standard deviation for each group of infants. Is it reasonable to pool the variance?

(b) Run the analysis of variance. Report the F statistic with its degrees of freedom and *p*-value. What do you conclude?

12.47 Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the cones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats. There were three treatments: a control with no jumping, a low-jump condition (the jump was 30 centimeters), and a high jump condition (the ump was 60 centimeters). After 8 weeks of 10 jumps per day, 5 days per week, the bone density of the rats (expressed in mg/cm^3) was measured. Here are the data:

Group	Bone density (mg/cm^3)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

(a) Make a table giving the sample size, mean, and standard deviation for each group of rats. Is it reasonable to pool the variances?

(b) Run the analysis of variance. Report the F statistic with its degrees of freedom and *p*-value. What do you conclude?

12.53 Refer to Exercise 12.25. There are two comparisons of interest to the experimenter: They are (1) Placebo versus the average of the 2 low-dose treatments; and (2) the difference between High A and Low A versus the difference between High B and low B.

(a) Express each contrast in terms of the means (μ's) of the treatments.

(b) Give estimates with standard errors for each of the contrasts.

(c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

12.63 Refer to the price promotion study that we examined in Exercise 12.40. The explanatory variable in this study is the number of price promotions in a 10 week period, with possible values of 1, 3, 5, and 7. When using analysis of variance, we treat the explanatory variable as categorical. An alternative analysis is to use simple linear regression. Perform this analysis and summarize the results. Plot the residuals from the regression model versus the number of promotions. What do you conclude?

Chapter 13 Exercises

13.7 A recent study investigated the influence that proximity and visibility of food have on food intake. A total of 40 secretaries from the University of Illinois participated in the study. A candy dish full of individually wrapped chocolates was placed either at the desk of the participant or at a location 2 meters from the participant. The candy dish was either a clear (candy visible) or opaque (candy not visible) covered bowl. After a week, the researchers noted not only the number of candies consumed per day but also the self-reported number of candies consumed by each participant. The table summarizes the mean differences between these two values (reported minus actual).

Proximity	Clear	Opaque
Proximate	-1.2	-0.8
Less proximate	0.5	0.4

Make a plot of the means and describe the patterns that you see. Does the plot suggest an interaction between visibility and proximity?

13.9 The National Crime Victimization Survey estimates that there were over 400,000 violent crimes committed against women by their intimate partner that resulted in physical injury. An intervention study designed to increase safety behaviors of abused women compared the effectiveness of six telephone intervention sessions with a control group of abused women who received standard care. Fifteen different safety behaviors were examined. One of the variables analyzed was that total number of behaviors (out of 15) that each woman performed. Here is a summary of the means of this variable at baseline (just before the first telephone call) an at follow-up 3 and 6 months later:

Group	Baseline	3 months	6 months
Intervention	10.4	12.5	11.9
Control	9.6	9.9	10.4

(a) Find the marginal means. Are they useful for understanding the results of this study?

(b) Plot the means. Do you think there is an interaction? Describe the meaning of an interaction for this study.

13.11 Lying is a common component of all human relationships. To investigate the acceptability of lying under various scenarios, researchers questioned 229 high school students fro a West Coast public high school and 261 college students for a state university in the Midwest. As part of the questioning, participants were asked to read a vignette in which the protagonist lies to his or her parents and to evaluate the acceptability of lying on a 4-point scale (1= totally unacceptable, 4 = totally acceptable). Each participant was randomly assigned to read the vignette with either a male or female protagonist. The following chart summarizes the mean response across age and protagonist.

	Age	
Protagonist	H.S.	College
Male	2.25	2.18
Female	2.35	1.82

(a) Plot the means and describe the pattern that you see.

(b) Suppose the F statistic for the interaction was 3.26. What are the degrees of freedom for this statistic and the approximate P-value? Is there a significant interaction?

(c)This study involved participants from one high school and one college. To what extent do you think this limits the generalizability of the conclusions? Explain.

13.13 Analysis of data for a 3 × 2 ANOVA with f observations per cell gave the F statistics in the following table:

Effect	F
A	1.53
B	3.87
AB	2.94

What can you conclude from the information given?

13.25 One way to repair serious wounds is to insert some material as a scaffold for the body's repair cells to use as a template for new tissue. Scaffolds made form extracellular material (ECM) are particularly promising for this purpose. Because they are made form biological material, they serve as an effective scaffold and are then resorbed. Unlike biological material that includes cells, however, they do not trigger tissue rejection reactions in the body. One study compared 6 types of scaffold material. Three of these were ECMs and the other three were made of inert materials. There were three mice used per scaffold type. The response measure was the percent of glucose phosphated isomerase (Gpi) cells in the region of the wound. A large value is good, indicating that there are many bone marrow cells sent by the body to repair the tissue. In Exercise 12.51 we analyzed the data for rats whose tissues were measured 4 weeks after the repair. The experiment included additional groups of rats who received the same types of scaffold but were measured at different times. The data in the table below are for 4 weeks and 8 weeks after the repair:

(a) Make a table giving the sample size, mean, and standard deviation for each of the material-by-time combinations. Is it reasonable to pool the variances?

Because the sample sizes in this experiment are very small, we expect a large amount of variability in the sample standard deviations. Although they vary more than we would prefer, we will proceed with the ANOVA.

(b) Make a plot of the means. Describe the main features of the plot.

(c) Run the analysis of variance. Report the F statistics with degrees of freedom and *p*-values for each of the main effects and the interaction. What do you conclude?

Material	4 weeks			6 weeks		
ECM1	55	70	70	60	65	65
ECM2	60	65	65	60	70	60
ECM3	75	70	75	70	80	70
MAT1	20	25	25	15	25	25
MAT2	5	10	5	10	5	5
MAT3	10	15	10	5	10	10

13.27 Refer to the previous exercise. Analyze the data for each time period separately using a one-way ANOVA. Use a multiple comparisons procedure where needed. Summarize the results. (The data are reproduced below for convenience).

Material	2 weeks			4 weeks			6 weeks		
ECM1	70	75	65	55	70	70	60	65	65
ECM2	60	65	70	60	65	65	60	70	60
ECM3	80	60	75	75	70	75	70	80	70
MAT1	50	45	50	20	25	25	15	25	25
MAT2	5	10	15	5	10	5	10	5	5
MAT3	30	25	25	10	15	10	5	10	10

13.31 One step in the manufacture of large engines requires that holes of very precise dimensions be drilled. The tools that do the drilling are regularly examined and are adjusted to ensure that the holes meet the required specifications. Part of the examination involves measurement of the diameter of the drilling tool. A team studying the variation in the sizes of the drilled holes selected this measurement procedure as a possible cause of variation in the drilled holes. They decided to use a designed experiment as one part of this examination. Some of the data are given in Table 13.2. The diameters in millimeters (mm) of five tools were measured by the same operator at three times (8:00 am, 11:00 am, and 3:00 pm). The person taking the measurements could not tell which tool was being measured, and the measurements were taken in random order.

(a) Make a table of means and standard deviations for each of the 5×3 combinations of the two factors.

(b) Plot the means and describe how the means vary with tool and time. Note that we expect the tools to have slightly different diameters. These will be adjusted as needed. It is the process of measuring the diameters that is important.

(c) Use a two-way ANOVA to analyze these data. Report the test statistics, degrees of freedom, and *p*-values for the significance tests.

Tool	Time	Diameter (mm)		
1	1	25.030	25.030	25.032
1	2	25.028	25.028	25.028
1	3	25.026	25.026	25.026
2	1	25.016	25.018	25.016
2	2	25.022	25.020	25.018
2	3	25.016	25.016	25.016
3	1	25.005	25.008	25.006
3	2	25.012	25.012	25.014
3	3	25.010	25.010	25.008
4	1	25.012	25.012	25.012
4	2	25.018	25.020	25.020
4	3	25.010	25.014	25.018
5	1	24.996	24.998	24.998
5	2	25.006	25.006	25.006
5	3	25.000	25.002	24.999

13.35 A study of the question “Do left-handed people live shorter lives than right-handed people?” examined a sample of 949 death records and contacted next of kin to determine handedness. Note that there are many possible definitions of ”left-handed.” The researchers examined the effects of different definitions on the results of their analysis and found that their conclusions were not sensitive to the exact definition used. For the results presented here, people were defined to be right-handed if they wrote, drew, and threw a ball with the right hand. All others were defined to be left-handed. People were classified by gender (female or male), and a 2 × 2 ANOVA was run with the age at death as the response variable. The *F* statistics were 22.36 (handedness), 37.44 (gender), and 2.10 (interaction). The following marginal mean ages at death (in years) were reported: 77.39 (females), 71.32 (males), 75.00 (right-handed), and 66.03 (left-handed).

(a) For each of the *F* statistics given above find the degrees of freedom and an approximate *p*-value. Summarize the results of these tests.

Chapter 14 Exercises

14.1 If you deal one card from a standard deck, the probability that the card is a heart is 0.25. Find the odds of drawing a heart.

14. 3 A study was designed to compare two energy drink commercials. Each participant was shown to commercials, A and B, in random order and asked to select the better one. There were 100 women and 140 men who participated in the study. Commercial A was selected by 45 women and by 80 men. Find the odds of selecting Commercial A for the men. Do the same for the women.

14.5 Refer to Exercise 14.3. Find the log odds for the mean and the log odds for the women.

14.7 Refer to Exercises 14.3 and 14.5. Find the logistic regression equation and the odds ratio.

14.11 Following complaints about the working conditions in some apparel factories both in the United States and abroad, a joint government and industry commission recommended in 1998 that companies that monitor and enforce proper standards be allowed to display a "No Sweat" label on their products. Does the presence of these labels influence consumer behavior?

A survey of U.S. residents asked 18 or older asked a series of questions about how likely they would be to purchase a garment under various conditions. For some conditions, it was stated that the garment had a "No Sweat" label; for other, there was no mention of such a label. On the basis of the responses, each person was classified as a "label user" or a "label nonuser." Suppose we want to examine the data for a possible gender effect. Here are the data for comparing men and women:

Gender	*n*	Number of Label users
Women	296	63
Men	251	27

(a) For each gender, find the proportion of label users.
(b) Convert each of the proportions that you found in part (a) to odds.
(c) Find the log of each of the odds that you found in part (b).

14.13 Refer to Exercise 14.11. Use $x = 1$ for women and $x = 0$ for men.

(a) Wind the estimates b_0 and b_1.

(b) Give the fitted logistic regression model.

(c) What is the odds ratio for men versus women?

14.21 Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the Directory of Public High Technology Corporations, and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

(a) What proportion of the high-tech companies offer stock options to their key employees? What are the odds?

(b) What proportion of the non-high-tech companies offer stock options to their key employees? What are the odds?

(c) Find the odds ratio using the odds for the high-tech companies in the numerator. Describe the result in a few sentences.

14.25 There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men from the low-blood-pressure group and 55 in the high-blood-pressure group died from cardiovascular disease.

(a) Find the proportion of men who died from cardiovascular disease in the high-bllod-pressure group. Then calculate the odds.

(b) Do the same for the low-blood-pressure group.

(c) Now calculate the odds ratio with the odds for the high-blood-pressure group in the denominator. Describe the result in words.

14.27 Refer to the study of cardiovascular disease and blood pressure in Exercise 14.25. Computer output for a logistic regression analysis of these data gives an estimated slope $b_1 = 0.7505$ with standard error $SE_{b1} = 0.2578$.

(a) Five a 95% confidence interval for the slope.

(b) Calculate the X^2 statistic for testing the null hypothesis that the slope is zero and use Table F to find an approximate p-value.

14.35 A study of alcohol use and deaths due to bicycle accidents collected data on a large number of fatal accidents. For each of these, the individual who died was classified according to whether or not there was a positive test for alcohol and by gender. Here are the data:

Gender	n	X (tested positive)
Female	191	27
Male	1520	515

Use logistic regression to study the question of whether or not gender is related to alcohol use in people who are fatally injured in bicycle accidents.

Chapter 15 Exercises

15.3 Refer to Exercise 15.1. State appropriate null and alternative hypotheses for this setting and calculate the value of *W*, the test statistic.

Group A	552	448	68	243	30
Group B	329	780	560	540	240

15.5 Refer to Exercises 15.1 and 15.3. Find μ_W, σ_W, and the standardized rank sum statistic. Then give the approximate p-value using the Normal approximation. What do you conclude?

15.11 How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:

2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

(a) Make a back-to-back stemplot. Does it appear reasonable to assume that the two distributions have the same shape?

(b) Is there evidence that the breaking strengths are lower for the strips buried longer?

15.13 "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo. Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:

Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged	17	4	18	14	18	15	15	10	12			

(a) Make a back-to-back stemplot of the data. Does there appear to be a difference in species counts for logged and unlogged plots?

(b) Does logging significantly reduce the number of species in a plot after 8 years? State hypotheses, do a Wilcoxon test, and state your conclusion.

15.19 Refer to the previous exercise. Here are the scores for a random sample of 7 spas that ranked between 19 and 36:

Spa	1	2	3	4	5	6	7
Diet/Cuisine	77.3	85.7	84.2	85.3	83.7	84.6	78.5
Program/Facilities	95.7	78.0	87.2	85.3	93.6	76.0	86.3

Is food, expressed by the Diet/Cuisine score, more important than activities, expressed as the Program/Facilities score, for a top ranking? Formulate this question in terms of null and alternative hypotheses. Then compute the differences and find the value of the Wilcoxon signed rank statistic, W^+.

15.21 Refer to exercise 15.19. Find μ_{W+}, σ_{W+}, and the Normal approximation for the p-value for the Wilcoxon signed rank test.

15.25 Can the full moon influence behavior? A study observed at nursing home patients with dementia. The number of incidents of aggressive behavior was recorded each dat for 12 weeks. Call a day a "moon day" if it is the day of a full moon or the day before or after a full moon. Here are the average numbers of aggressive incidents for moon days and other days for each subject:

Patient	Moon days	Other days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30
8	2.67	0.40
9	6.00	1.59
10	4.33	0.60
11	3.33	0.65
12	0.67	0.69
13	1.33	1.26
14	0.33	0.23
15	2.00	0.38

The matched pairs t test (Example 7.7) gives $p < 0.000015$ and a permutation test (Example 16.14) gives $p = 0.0001$. Does the Wilcoxon signed rank test, based on ranks rather than means, agree that there is strong evidence that there are more aggressive behaviors on moon days?

15.31 Exercise 7.32 presents the data below on the weight gains (in kilograms) of adults who were fed an extra 1000 calories per day for 8 weeks.

(a) Use a rank test to test the null hypothesis that the median weight gain is 16 pounds, as theory suggests. What do you conclude?

Subject	Before	After
1	55.7	61.7
2	54.9	58.8
3	59.6	66
4	62.3	66.2
5	74.2	79
6	75.6	82.3
7	70.7	74.3
8	53.3	59.3
9	73.3	79.1
10	63.4	66
11	68.1	73.4
12	73.7	76.9
13	91.7	93.1
14	55.9	63
15	61.7	68.2
16	57.8	60.3

15.33 Many studies suggest that exercise causes bones to get stronger. One study examined the effect of jumping on the bone density of growing rats. Ten rats were assigned to each of three treatments: a 60-centimeter “high jump,” a 30-cedntimeter “low jump,” and a control group with no jumping. Here are the bone densities (in milligrams per cubic centimeter) after 8 weeks of 10 jumps per day:

Group	Bone density (mg/cm^3)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

(c) Do the Kruskal-Wallis test. Explain the distinction between the hypotheses tested by Kruskal-Wallis and ANOVA.

Chapter 16 Exercises

16.1 To illustrate the bootstrap procedure, let's bootstrap a small random subset of the Verizon data:

26.47	0.00	5.32	17.30	29.78	3.67

a) Sample *with replacement* from this initial SRS by rolling a die. Rolling a 1 means select the first member of SRS (26.47), a 2 means select the second member (0.00), and so on. (You can also use Table B of random digits, responding only to digits 1 to 6.) Create 20 resamples of size $n = 6$.

b) Calculate the sample mean of each of the resamples.

c) Make a stemplot of the means of the 20 resamples. This is the bootstrap distribution.

d) Calculate the standard deviation of the bootstrap distribution.

16.5 The distribution of carbon diozinde (CO2) emissions in Table 1.6 is strongly skewed to the right. The United States and several other countries appear to he high outliers. Generate a bootstrap distribution for the mean of C-reactive protein; construct a histogram and Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of $\bar{x}$ to be close to Normal?

16.7 The measurements of C-reactive protein in 40 children (Exercise 7.26) are very strongly skewed. We were hesitant to use *t* procedures for these data. Generate a bootstrap distribution for the mean of C-reactive protein; construct a histogram and Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of $\bar{x}$ to be close to Normal?

16.9 We have two ways to estimate the standard deviation of a sample mean $\bar{x}$: use the formula $s/\sqrt{n}$ for the standard error, or use the bootstrap standard error.

(b) Find the sample standard deviation s for the CO2 emissions in Exercise 16.5 and use it to find the standard error $s/\sqrt{n}$ of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.5?

16.13 Return to or create the bootstrap distribution resamples on the sample mean for the audio file lengths in Exercise 16.8. In Example 7.11, the *t* confidence interval for the average length was constructed.

(a) Inspect the bootstrap distribution. Is a bootstrap *t* confidence interval appropriate? Explain why or why not.
(b) Construct the 95% bootstrap t confidence interval.
(c) Compare the bootstrap results with the *t* confidence interval reported in Example 7.11.

16.25 Each year, the business magazine *Forbes* publishes a list of the world's billionaires. In 2006, the magazine found 793 billionaires. Here is the wealth, as estimated by *Forbes* and rounded to the nearest 100 million, of an SRS of 20 of these billionaires:

2.9	15.9	4.1	1.7	3.3	1.1	2.7	13.6	2.2	2.5
3.4	4.3	2.7	1.2	2.8	1.1	4.4	2.1	1.4	2.6

Suppose you are interested in "the wealth of typical billionaires." Bootstrap an appropriate statistic, inspect the bootstrap distribution, and draw conclusions based on this sample.

16.31 Consider the small random subset of the Verizon data in Exercise 16.1. Bootstrap the sample mean using 1000 resamples. The data are reproduced below:

26.47	0.00	5.32	17.30	29.78	3.67

(a) Make a histogram and Normal quantile plot. Does the bootstrap distribution appear close to Normal? Is the bias small relative to the observed sample mean?
(b) Find the 95% bootstrap t confidence interval.
(c) Five the 95% bootstrap percentile confidence interval and compare it with the interval in (b).

16.45 Figure 2.7 (page 96) shows a very weak relationship between returns on Treasury bills and returns on common stocks. The correlation is $r = –0.113$. We wonder if this is significantly different from 0. To find out, bootstrap the correlation. (The data are in the file ex16-045.)

(a) Describe the shape and bias of the bootstrap distribution. It appears that even simple bootstrap inference (t and percentile confidence intervals) is justified. Explain why.

16.59 Exercise 7.41 gives data on a study of the effect of a summer language institute on the ability of high school language teachers to understand spoken French. This is a matched pairs study, with scores for 20 teachers at the beginning (pretest) and end (posttest) of the institute. We conjecture that the posttest scores are higher on the average.

(a) Carry out the matched pairs t test. That is, state the hypotheses, calculate the test statistic, and give its p-value.

(b) Make a Normal quantile plot of the gains: posttest score – pretest score. The data have a number of ties and a low outlier. A permutation test can help check the *t* test result.

(c) Carry out the permutation test for the difference in means in matched pairs, using 9999 resamples. The Normal quantile plot shows that the permutation distribution is reasonably Normal, but the histogram looks a bit odd. What explains the appearance of the histogram? What is the p-value for the permutation test? Do your tests in here and in part (a) lead to the same practical conclusion?

Chapter 17 Exercises

17.5 A sandwich shop owner takes a daily sample of 6 consecutive sandwich orders at random times during the lunch rush and records the time it takes to complete each order. Past experience indicates that the process mean should be $\mu = 168$ seconds and the process standard deviation should be $\sigma = 30$ seconds. Calculate the center line and control limits for an $\bar{x}$ control chart.

17.13 A meat-packaging company produces 1-pound packages of ground beef by having a machine slice a long circular cylinder of ground beef as it passes through the machine. The timing between consecutive cuts will alter the weight of each section. Table 17.3 gives the weight of 3 consecutive sections of ground beef taken each hour over two 10-hour days. Past experience indicates that the process mean is 1.03 and the weight varies with $\sigma = 0.02$ lb.

(a) Calculate the center line and control limits for an $\bar{x}$ chart.

(b) What are the center line and control limits for an s chart for this process?

(c) Create the $\bar{x}$ and s chards for these 20 consecutive samples.

(d) Does the process appear to be in control? Explain.

Sample	Weight (pounds)			$\bar{x}$	s
1	0.999	1.071	1.019	1.030	0.0373
2	1.030	1.057	1.040	1.043	0.0137
3	1.024	1.020	1.041	1.028	0.0108
4	1.005	1.026	1.039	1.023	0.0172
5	1.031	0.995	1.005	1.010	0.0185
6	1.020	1.009	1.059	1.029	0.0263
7	1.019	1.048	1.050	1.039	0.0176
8	1.005	1.003	1.047	1.018	0.0247
9	1.019	1.034	1.051	1.035	0.0159
10	1.045	1.060	1.041	1.049	0.0098
11	1.007	1.046	1.014	1.022	0.0207
12	1.058	1.038	1.057	1.051	0.0112
13	1.006	1.056	1.056	1.039	0.0289
14	1.036	1.026	1.028	1.030	0.0056
15	1.044	0.986	1.058	1.029	0.0382
16	1.019	1.003	1.057	1.026	0.0279
17	1.023	0.998	1.054	1.025	0.0281
18	0.992	1.000	1.067	1.020	0.0414
19	1.029	1.064	0.995	1.029	0.0344
20	1.008	1.040	1.021	1.023	0.0159

17.15 A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each lot of tables is measured in order to control the compression process. The process has been operating in control with mean at the target value $\mu = 11.5$ and estimated standard deviation $\sigma = 0.2$. Table 17.4 gives three sets of data, each representing $\bar{x}$ for 20 successive samples of n = 4 tablets. One set of data remains in control at the target value. In a second set, the process mean μ shifts suddenly to a new value. In a third, the process mean drifts gradually.

(a) What are the center line and control limits for an $\bar{x}$ chart for this process?

(b) Draw a separate $\bar{x}$ chart for each of the three data sets. Mark any points that are beyond the control limits.

(c) Based on your work in (b) and the appearance of the control charts, which set of data comes from a process that is in control? In which case does the process mean shift suddenly, and at about which sample do you think that the mean changed? Finally, in which case does the mean drift gradually?

Sample	Data A	Data B	Data C
1	11.602	11.627	11.495
2	11.547	11.613	11.475
3	11.312	11.493	11.465
4	11.449	11.602	11.497
5	11.401	11.360	11.573
6	11.608	11.374	11.563
7	11.471	11.592	11.321
8	11.453	11.458	11.533
9	11.446	11.552	11.486
10	11.522	11.463	11.502
11	11.664	11.383	11.534
12	11.823	11.715	11.624
13	11.629	11.485	11.629
14	11.602	11.509	11.575
15	11.756	11.429	11.730
16	11.707	11.477	11.680
17	11.612	11.570	11.729
18	11.628	11.623	11.704
19	11.603	11.472	12.052
20	11.816	11.531	11.905

17.19 Figure 17.10 reproduces a data sheet from the floor of a factory that makes electrical meters. The sheet shows measurements of the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for 0.6044 inch. All the measurements end in 4. Although we don't know why this is true, it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch.

Calculate $\bar{x}$ and s for the first two samples. The data file *ex*17_19 contains $\bar{x}$ and s for all 18 samples. Based on long experience with this process, you are keeping control charts based on $\mu = 43$ and $\sigma = 12.74$. Make s and $\bar{x}$ charts for the data in Figure 17.10 and describe the state of the process.

17.31 The $\bar{x}$ and s control charts for the mesh-tensioning example (Figures 17.4 and 17.7) were based on $\mu = 275\,\text{mV}$ and $\sigma = 43\,\text{mV}$. Table 17.1 gives the 20 most recent samples from this process.

(a) Estimate the process μ and σ based on these 20 samples.

(b) Your calculations suggest that the process σ may now be less than 34 mV. Explain why the s chart in Figure 17.7 (page 17-15) suggests the same conclusion. (If this pattern continues, we would eventually update the value of σ used for control limits.)

17.35 Do the losses on the 120 individual patients in Table 17.7 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss what it shows. Are the natural tolerances you found in the previous exercise trustworthy?

17.37 The center of the specification for mesh tension is 250 mV, but the center of our process is 275 mV. We can improve capability by adjusting the process to have center 250 mV. This is an easy adjustment that does not change the process variation. What percent of monitors now meet the new specifications? (From the preceding exercise, the specifications are 150 to 350 mV; the standard deviation is 38.4 mV.)

17.39 Figure 17.10 (page 17-21) displays a record sheet of 18 samples of distances between mounting holes in an electrical meter. The data file *ex17_19* adds and for each sample in Exercise 17.19, you found that sample 5 was out of control on the process-monitoring s chart. The special cause responsible was found and removed. Based on the 17 samples that were in control, what are the natural tolerances for the distance between the holes?

17.41 The record sheet in Figure 17.10 gives specifications as 0.6054 ± 0.0010 inch. That's 54 ± 10 as the data are coded on the record. Assuming that the distance varies Normally from meter to meter, about what percent of meters meet specifications?

17.43 Make a Normal quantile plot of the 85 distances in data file *ex*17_19 that remain after removing sample 5. How does the plot reflect the limited precision of the measurements (all of which end in 4)? Is there any departure from Normality that would lead you to discard your conclusions from Exercise 17.39?

17.53 Table 17.1 (page 17-10) gives 20 process control samples of the mesh tension of computer monitors. En Examples 17.13, we estimated from these samples that $\hat{\mu} = \overline{\overline{x}} = 275.065$ MV and $\hat{\sigma} = s = 38.38$ mV.

(a) The original specifications for mesh tension were LSL = 100 mV and USL = 400 mV. Estimate C_p and C_{pk} for this process.

(b) A major customer tightened the specifications to LSL = 150 mV and USL = 350 mV. Now what are C_p and C_{pk}?

17.69 The controller's office of a corporation is concerned that invoices that remain unpaid after 30 days are damaging relations with vendors. To assess the magnitude of the problem, a manager searches payment records for invoices that arrived in the last 10 months. The average number of invoices is 2875 per month, with relatively little month-to-monthly variation. Of all these invoices, 960 remained unpaid after 30 days.

(a) What is the total number of opportunities for unpaid invoices? What is $\overline{p}$?

(b) Give the center line and control limits of a *p* chart on which to plot the future monthly proportions of unpaid invoices.

17.83 You have just installed a new system that uses an interferometer to measure the thickness of polystyrene film. To control the thickness, you plan to measure 3 film specimens every 10 minutes and keep $\overline{x}$ and *s* charts. To establish control, you measure 22 samples of 3 films each at 10-minute intervals. Table 17.12 gives $\overline{x}$ and *s* for these samples. The units are millimeters × 10^{-4}. Calculate control limits for s, make an s chart, and comment on control of short-term process variation.