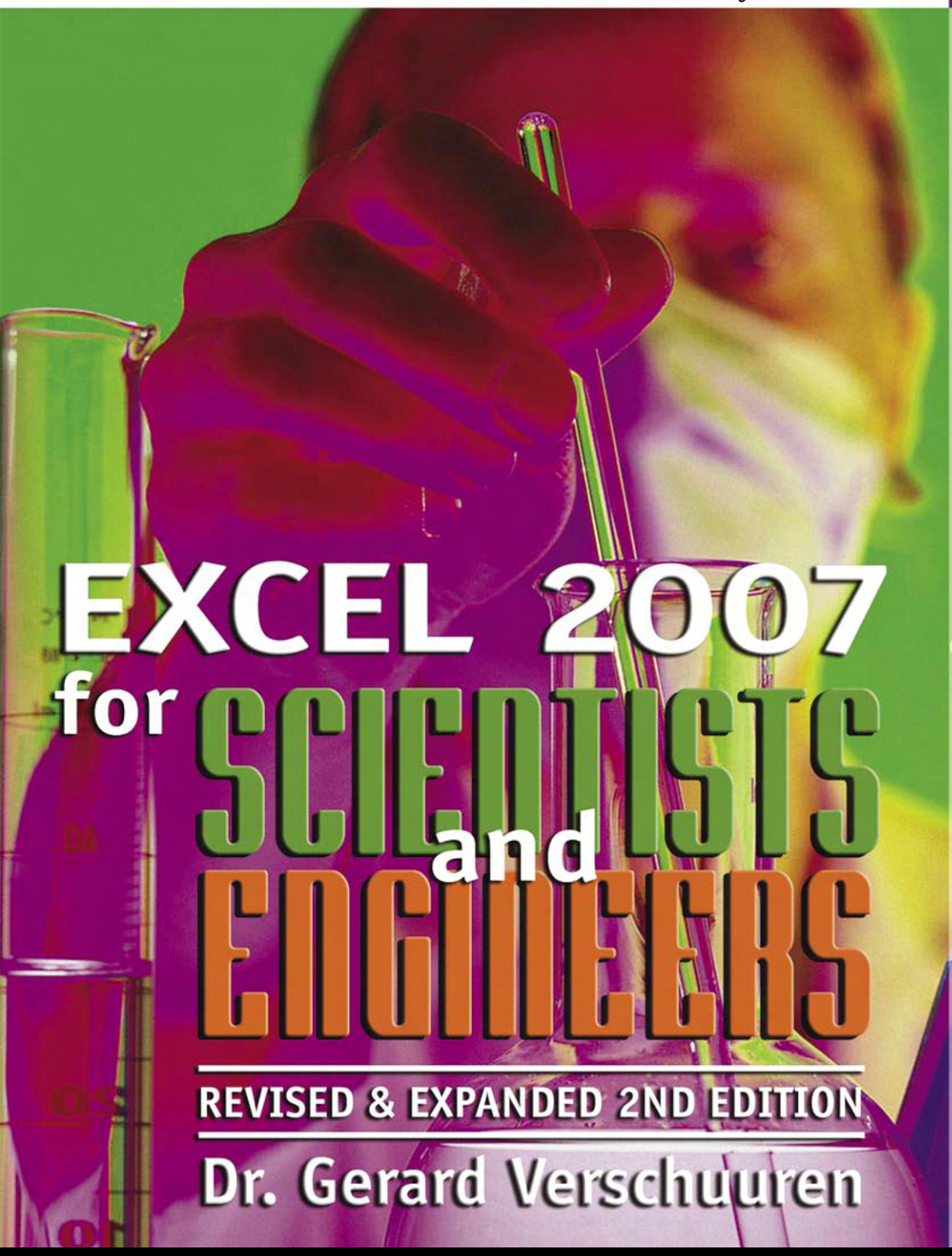
"Excel for Scientists and Engineers gave me the key skills I needed to have for working with Excel in a scientific profession. Dr. Verschuuren has an easy to understand teaching approach, covering both the basics and more in-depth tasks. It's highly informative with great reference material and is great for the beginner and the more advanced user. Step by step instructions will have you up to speed in no time."

-Erin Beal, Wyeth BioPharma

Designed by Scientists for Scientists



EXCEL 2007 for SCIENTISTS and ENGINEERS

REVISED & EXPANDED 2ND EDITION

Dr. Gerard Verschuuren

- Graphs
- Transformations
- Error Bars
- Secondary Axis
- Histograms
- Chart Formulas
- Record Keeping
- If & Nested IF
- VLOOKUP
- Frequency
- DSUM
- Filtering
- Solving
- Array Formulas
- Sampling
- Distributions
- Regression
- Analysis

Excel 2007 for Scientists

by

Dr. Gerard M. Verschuuren

Holy Macro! Books
PO Box 82, Uniontown, OH 44685

Excel 2007 for Scientists

Author: Dr. Gerard M. Verschuuren

Copy Editor: Kitty Jarrett

Technical Editor: Bob Umlas

Cover Design: Shannon Mattiza, 6’4 Productions

Layout: Fine Grains (India) Private Limited, New Delhi, India.

Publisher: Holy Macro! Books. PO Box 82, Uniontown, OH 44685

Distributor: Independent Publishers Group, Chicago, IL

First printing: March 2008

Printed in the United States of America by Malloy

Library of Congress Control Number: 2007943884

ISBN: 978-1-932802-35-1

Contents

Contents

Contents

PART 3: Plotting Data

Contents

Contents

PART 5: Statistical Analysis

Contents

About the Author

Dr. Gerard M. Verschuuren is a Microsoft Certified Professional specialized in VB, VBA, and VB.NET. He has more than 25 years of experience in teaching at colleges and corporations.

Dr. Verschuuren holds master's degrees in biology (human genetics) and philosophy, as well as a doctorate in the philosophy of science from universities in Europe.

He is the author of *Life Scientists*, *Their Convictions*, *Their Activities*, *and Their Values* (1995, Genesis Publishing Company) and the author of *From VBA to VSTO* (2006, Holy Macro! Books).

He is also the author behind the Visual Learning series (MrExcel.com):

- Slide Your Way through Excel VBA (2003)
- Your Access to the World (2004)
- Access VBA Made Accessible (2005)
- Master the Web (2005)
- Excel 2007 Expert (2007)
- See Sharper with C#
- Excel 2007 VBA

Introduction

This book can be used on its own or in conjunction with an interactive CD called Excel 2007 for Scientists, also available from MrExcel.com. This book assumes at least some basic knowledge of Excel. Readers new to Excel may want to familiarize themselves with a basic how-to source such as the interactive CD Excel 2007 Expert, available from MrExcel.com.

Scientists do not want nor do they need verbose explanations. Therefore, I was as concise as possible in the chapters of this book. I also attempted to add some meaningful simple exercises because the proof is still in the pudding. The examples appear at the end of each part, along with their solutions. Because I am a human geneticist myself, most of my simple examples stem from the life sciences.

All files used in this book can be found at `www.genesispc.com/Science2007.htm`. Each file has an original version (to work in) and a finished version (to check your solutions).

Excel was originally created as a financial application, but it has evolved into a rather sophisticated scientific tool. Although other and perhaps more advanced programs exist, many of those have a steep learning curve. Excel may, therefore, still be your best choice. I hope you will soon discover why.

PART 1
General Techniques

Chapter 1

Navigation in Excel 2007

Excel 2007 has plenty of space for your scientific work. Each workbook (or .xlsx file) can hold an unlimited number of worksheets (provided that your computer memory permits), and each worksheet has a capacity of 1,048,576 rows by 16,384 columns. Hopefully, you won't use all this space before retirement.

Scientific spreadsheets can be huge—filled with many numbers. So you need ways to quickly navigate around and to create formulas for giant ranges of cells in a swift and efficient way. That's what this chapter is about.

Most sheets in this book have a modest size, so it is easy to practice with them. But in real life, you probably deal with much larger collections of data. The basic techniques discussed in this chapter will benefit you even more when your tables become larger and larger.

Navigation Shortcuts

The following keystrokes are some important navigation shortcuts:

- Ctrl+Home takes you to the origin of the sheet, which is cell A1.
- Ctrl+arrow key jumps between section borders. (A border is an empty row and/or column.)
- Ctrl+Shift+arrow key jumps and selects what is between the section borders.
- Shift+arrow key expands or reduces whatever has been selected.

Let's use Figure 1.1 to see how these shortcuts work. Based on Figure 1.1, the following would happen:

> **Note:** All files used in this book are available from `www.genesispc.com/Science2007.htm` where you can find each file in its original version (to work on) and in its finished version (to check your solutions).

- **Starting in A1**: Pressing Ctrl+Down Arrow takes you to A24 and then to A1048576. Pressing Ctrl+Up Arrow takes you back, with the same stops on the way.
- **Starting in B1:** Repeatedly pressing Ctrl+Down Arrow jumps to B10, B14, B23, and finally the end.

- **Starting in B1:** Pressing Ctrl+Shift+Down Arrow selects the entire range B1:B10. Pressing Shift+Down Arrow once expands the selection with one more cell. Instead pressing Shift+Up Arrow shortens the selection by one cell. The Shift key keeps all in-between cells included in the selection.
- **Starting in J1:** Typing J24 in the Name box just above column A and then pressing Shift+Enter selects all cells between J1 and J24 (thanks to the Shift key). With the range J1:J24 selected, typing `=ROW()` in the formula bar and then pressing Ctrl+Enter causes all the selected cells to be filled with this formula (thanks to the Ctrl key).

	A	B	C	D	E	F	G	H	I	J
1		1	1			0.69	0.42	1.11		1
2		2	8			0.33	0.83	1.16		2
3		3	27			0.63	0.70	1.33		3
4		4	64			0.50	0.43	0.93		4
5		5	125			0.12	0.23	0.35		5
6		6	216			0.10	0.90	1.00		6
7		7	343			0.86	0.94	1.80		7
8		8	512			0.42	0.98	1.40		8
9		9	729			0.41	0.77	1.18		9
10		10	1000			0.56	0.20	0.76		10
11					SD	0.2402684	0.2951459			11
12										12
13			Column							13
14		0.38	0.38							14
15		0.69	0.69							15
16		0.50	0.50							16
17		0.03	0.03							17
18		0.94	0.94							18
19		0.11	0.11							19
20		0.78	0.78							20
21		0.31	0.31							21
22		0.30	0.30							22
23		0.12	0.12							23
24	COUNTIF >.5	3	3							24

Figure: 1.1

Creating Formulas

Figure 1.2 shows an example of how to create some formulas:

1. Select cell F11 and press the fx button (located just in front of the formula bar).
2. Choose the function `STDEV` and start the first argument by clicking cell F10 and then pressing Ctrl+Shift+Up Arrow; this selects the entire range above cell F11. Often, Excel finds the correct range automatically—but not necessarily so; when it doesn't, you have to be in charge yourself!
3. Press OK in the dialog, and the cell shows the actual standard deviation of these cells.
4. In cell B24, use the function `COUNTIF` and follow these steps:
   i. For the first argument, click cell B23 and then press Ctrl+Shift+Up Arrow.

ii. For the second argument, type >=5 (which changes into a string: ">=5").

iii. Finalize the functions by pressing Ctrl+Enter.

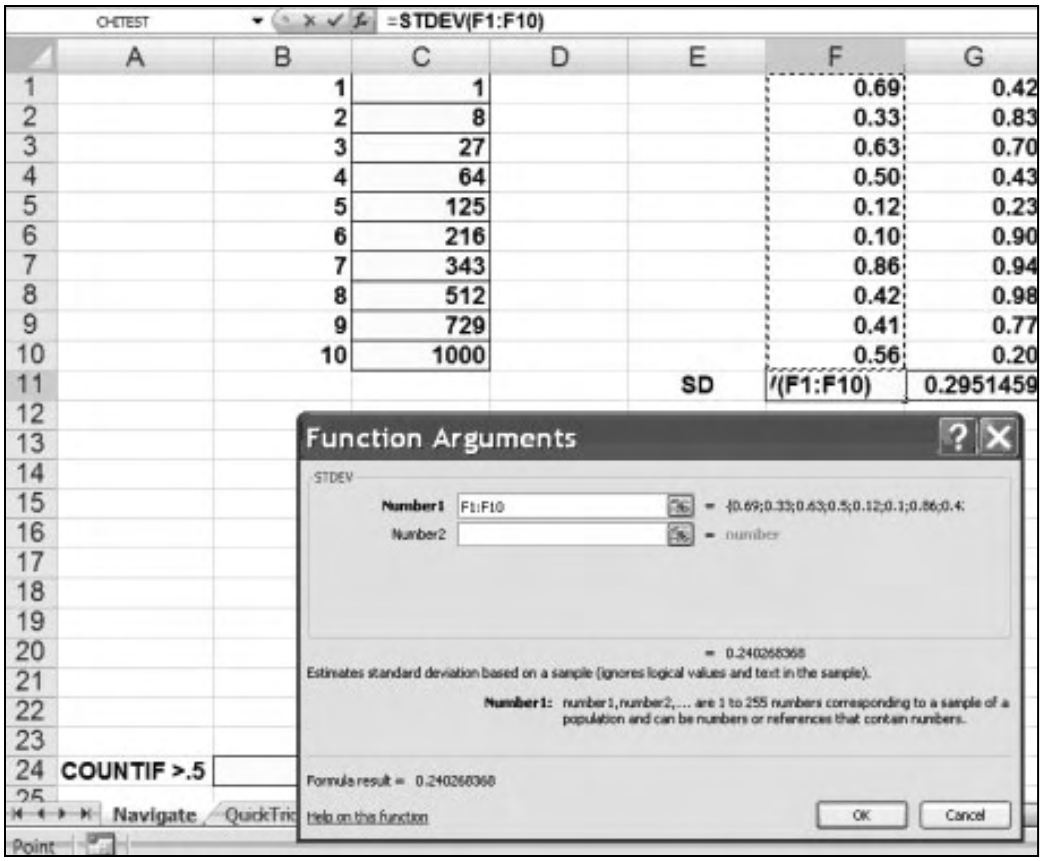
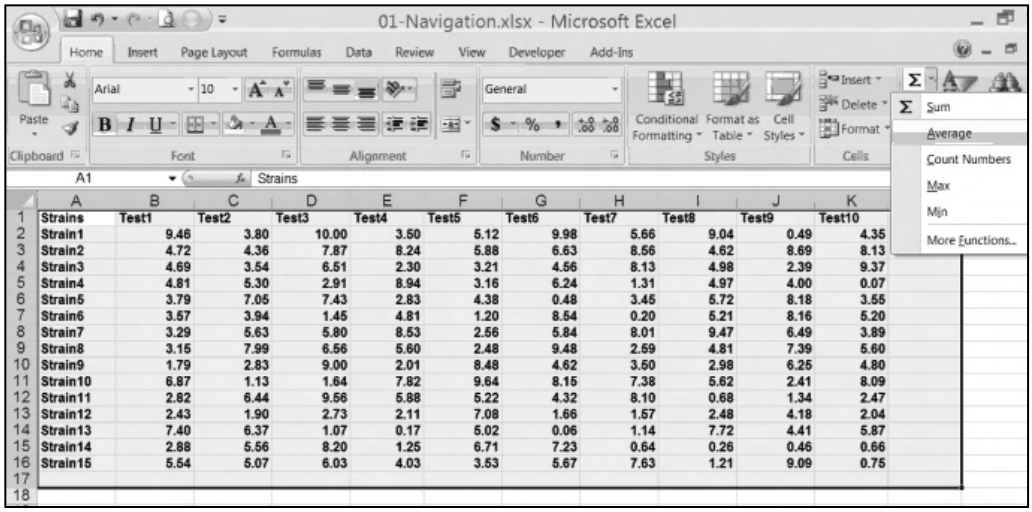


Figure: 1.2

Figure 1.3 shows a quick technique for calculating the mean or standard deviation of certain measurements per strain of bacteria and per test:

1. Select cell A1 (or press Ctrl+Home).
2. Press Ctrl+Shift+Down Arrow and Ctrl+Shift+Right Arrow.
3. You need an extra row and column for the calculations of the means, so to expand the selection with an extra row and column, press Shift+Down Arrow and Shift+Right Arrow.
4. From the drop-down button next to the Sum button, choose the option Average, Min, or Max.

All calculations are done automatically at the end of each numeric row or column. You could do this more tediously; go ahead if you like that route better!

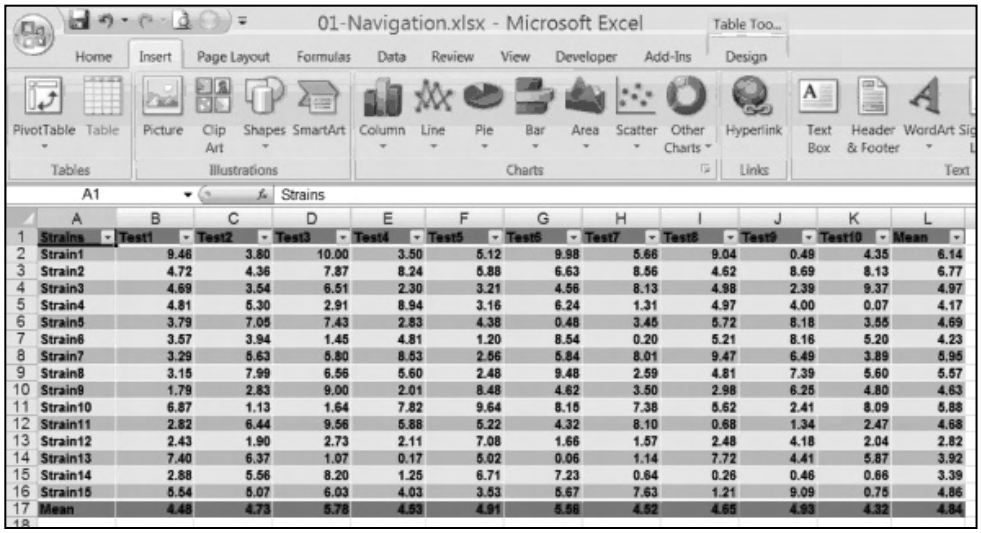


	A	B	C	D	E	F	G	H	I	J	K
1	Strains	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10
2	Strain1	9.46	3.80	10.00	3.50	5.12	9.98	5.66	9.04	0.49	4.35
3	Strain2	4.72	4.36	7.87	8.24	5.88	6.63	8.56	4.62	8.69	8.13
4	Strain3	4.69	3.54	6.51	2.30	3.21	4.56	8.13	4.98	2.39	9.37
5	Strain4	4.81	5.30	2.91	8.94	3.16	6.24	1.31	4.97	4.00	0.07
6	Strain5	3.79	7.05	7.43	2.83	4.38	0.48	3.45	5.72	8.18	3.55
7	Strain6	3.57	3.94	1.45	4.81	1.20	8.54	0.20	5.21	8.16	5.20
8	Strain7	3.29	5.63	5.80	8.53	2.56	5.84	8.01	9.47	6.49	3.89
9	Strain8	3.15	7.99	6.56	5.60	2.48	9.48	2.59	4.81	7.39	5.60
10	Strain9	1.79	2.83	9.00	2.01	8.48	4.62	3.50	2.98	6.25	4.80
11	Strain10	6.87	1.13	1.64	7.82	9.64	8.15	7.38	5.62	2.41	8.09
12	Strain11	2.82	6.44	9.56	5.88	5.22	4.32	8.10	0.68	1.34	2.47
13	Strain12	2.43	1.90	2.73	2.11	7.08	1.66	1.57	2.48	4.18	2.04
14	Strain13	7.40	6.37	1.07	0.17	5.02	0.06	1.14	7.72	4.41	5.87
15	Strain14	2.88	5.56	8.20	1.25	6.71	7.23	0.64	0.26	0.46	0.66
16	Strain15	5.54	5.07	6.03	4.03	3.53	5.67	7.63	1.21	9.09	0.75
17											
18											

Figure: 1.3

Figure 1.4 shows the same table as Figure 1.3, but this time in the dedicated table structure available in Excel 2007. Follow these steps:

1. Click the Table button in the Insert menu to create a table structure with a striping pattern for easy reading.
2. After the table structure is implemented, use the Design menu to change table settings as desired.
3. To create calculations again for mean and standard deviation, use the same technique you used earlier.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Strains	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10	Mean
2	Strain1	9.46	3.80	10.00	3.50	5.12	9.98	5.66	9.04	0.49	4.35	6.14
3	Strain2	4.72	4.36	7.87	8.24	5.88	6.63	8.56	4.62	8.69	8.13	6.77
4	Strain3	4.69	3.54	6.51	2.30	3.21	4.56	8.13	4.98	2.39	9.37	4.97
5	Strain4	4.81	5.30	2.91	8.94	3.16	6.24	1.31	4.97	4.00	0.07	4.17
6	Strain5	3.79	7.05	7.43	2.83	4.38	0.48	3.45	5.72	8.18	3.55	4.69
7	Strain6	3.57	3.94	1.45	4.81	1.20	8.54	0.20	5.21	8.16	5.20	4.23
8	Strain7	3.29	5.63	5.80	8.53	2.56	5.84	8.01	9.47	6.49	3.89	5.95
9	Strain8	3.15	7.99	6.56	5.60	2.48	9.48	2.59	4.81	7.39	5.60	5.57
10	Strain9	1.79	2.83	9.00	2.01	8.48	4.62	3.50	2.98	6.25	4.80	4.63
11	Strain10	6.87	1.13	1.64	7.82	9.64	8.15	7.38	5.62	2.41	8.09	5.88
12	Strain11	2.82	6.44	9.56	5.88	5.22	4.32	8.10	0.68	1.34	2.47	4.68
13	Strain12	2.43	1.90	2.73	2.11	7.08	1.66	1.57	2.48	4.18	2.04	2.82
14	Strain13	7.40	6.37	1.07	0.17	5.02	0.06	1.14	7.72	4.41	5.87	3.92
15	Strain14	2.88	5.56	8.20	1.25	6.71	7.23	0.64	0.26	0.46	0.66	3.39
16	Strain15	5.54	5.07	6.03	4.03	3.53	5.67	7.63	1.21	9.09	0.75	4.86
17	Mean	4.48	4.73	5.78	4.53	4.91	5.56	4.52	4.65	4.93	4.32	4.84

Figure: 1.4

* * *

Chapter 2

The Fill Handle

One of the best-kept secrets in Excel is the fill handle. This tool allows you to copy cells over a contiguous range of cells or to fill such a range with a series of specific values. In addition, it helps you copy formulas over huge ranges.

The fill handle is located in the lower-right corner of your selected cell(s). Whenever you move your mouse to that location, the cursor will change to a small + sign (not to be confused with a crosshairs). That very spot holds the fill handle.

Figure 1.5 shows some examples that help you to explore some of the features of the fill handle:

- Cell A2: Click and drag the fill handle downward to cell A6 in order to stop at Friday. If you keep going, the fill handle goes into Saturday, and so on. If you stop at Friday and then start the fill handle again (with A2:A6 still selected), you can just copy the previous section by holding the Ctrl key down until you are finished.
- Cell B2: To insert the number 8 in column B for every day of the week, double-click the fill handle of cell B2. A double-click on the fill handle copies the content down to the first empty cell in the previous or next column. So the double-click does not work when there is no column with data to the immediate left or right.
- Cell C2: Double-clicking the fill handle of cell C2 gives you a series of 1s. To change this into an incrementing series, click the button that has popped up and select ◉ Fill Series. Now the series increments by 1.
- Cells D2 and D3: For a series that needs to increment by a value different from 1, create a pattern in the first two cells, select both cells, and then double-click.
- Cell E2: If you don't want to create a pattern ahead of time, double-click the fill handle of the first cell only. Now go to the Fill button drop-down (located under the Σ button on the Home tab) and then choose the option Series. Specify any step value (for example, 2).
- Cell F2: To multiply D2 by E2, follow these steps (you will appreciate them someday!):

  1. Select cell F2.
  2. Type the equals sign (=).
  3. Press Left Arrow twice to get to D2 (that is, do not type `D2`).

	A	B	C	D	E	F	G	H
1	Day	Hours	Day#	+10	+3.3	D*E	Anal.	List
2	Monday	8	1	0	7	0	gmv	gmv
3	Tuesday	8	2	10	10.3	103	tjk	tjk
4	Wednesday	8	3	20	13.6	272	bdo	bdo
5	Thursday	8	4	30	16.9	507	gmv	gmv
6	Friday	8	5	40	20.2	808	tjk	tjk
7	Monday	8	6	50	23.5	1175	bdo	bdo
8	Tuesday	8	7	60	26.8	1608	gmv	gmv
9	Wednesday	8	8	70	30.1	2107	tjk	tjk
10	Thursday	8	9	80	33.4	2672	bdo	bdo
11	Friday	8	10	90	36.7	3303	gmv	gmv

Figure: 1.5

4. Type the multiplication sign (*).
5. Press the Left Arrow key once to get to E2.
6. To finish, press Ctrl+Enter (not just Enter).

> Note: What is the advantage of pressing Ctrl+Enter instead of Enter only? You stay in the same cell, so you can just double-click the fill handle to copy the formula all the way down. (Otherwise, you would have to go back to the previous cell first.)

- **Cell G2**: If you always work with the same analysts in the same order, type their names once, select them all, and double-click the fill handle. If you want to use this same list over and over again—especially if it's a long list—use the following technique:
  1. If you have a listing on your sheet already, select that listing first.
  2. Click the Office icon in the left-top corner.
  3. At the bottom of the new box, select Excel Options.
  4. In the Excel Options dialog, choose Popular (in the left panel).
  5. Click Edit Custom Lists.
  6. Accept the highlighted list.
  7. Click the Import button.
  8. Click OK twice.

Now you can use this list anywhere in Excel. Just type the first name of this (potentially long) list and double-click the fill handle—provided that there is a column with contents to the left or right. Excel does the rest!

* * *

Chapter 3

Relative vs. Absolute Cell References

Each cell on a sheet has a certain position. When you copy a cell that contains a formula to another position, the formula's cell references automatically adjust. Those references are called *relative*. Sometimes, you do not want formula references to adapt to their new location; in that case, you make them *absolute*.

To see how relative and absolute cell references work, take a look at Figure 1.6. Cell C1 has a formula in it: `=A1*B1`. You can copy the formula in cell C1 down by double-clicking because, in this case, you *do* want the cell references to change! How can you see all formulas at once? Use the shortcut Ctrl+~ (the tilde is located under the Esc key). Notice that Ctrl+~ causes all cell references to be relative here—which means: "Multiply two-cells-over-to-the-left by one-cell-over-to-the-left."

In cell F2, however, you want to find out what the value in cell E2 is, as a percentage of the mean in cell E11, using the formula `=E2/E11`. You accept the formula by pressing Ctrl+Enter and then double-click the fill handle downward. This time, you get into trouble! Ctrl+~ reveals the problem: The reference to E11 should be absolute; otherwise, the adjusted formula in the downward cells attempts to divide by empty cells, which is an invalid division-by-zero error.

C1 =A1*B1

	A	B	C	D	E	F
1	1	2	2			% of mean
2	2	4	8		1	8%
3	3	6	18		2	15%
4	4	8	32		7	54%
5	5	10	50		8	62%
6	6	12	72		16	123%
7	7	14	98		17	131%
8	8	16	128		21	162%
9	9	18	162		22	169%
10	10	20	200		23	177%
11				Mean	13	100%
12						
13		Dilute	0.02	0.04	0.06	0.08
14		2	0.04	0.08	0.12	0.16
15		4	0.08	0.16	0.24	0.32
16		6	0.12	0.24	0.36	0.48
17		8	0.16	0.32	0.48	0.64
18		10	0.20	0.40	0.60	0.80
19		12	0.24	0.48	0.72	0.96
20		14	0.28	0.56	0.84	1.12
21		16	0.32	0.64	0.96	1.28
22		18	0.36	0.72	1.08	1.44
23		20	0.40	0.80	1.20	1.60
24						

Figure: 1.6

Let's start over in cell F2:

1. Type the equals sign (=) in cell F2.
2. Press the Left Arrow key once to get to E2.
3. Type `/` (in order to divide).
4. Press the Left Arrow key once and then Ctrl+Down Arrow to get to E11.
5. Press F4 to make E11 absolute (that is, $E$11).
6. Press Ctrl+Enter.

As a result, the copy behavior of the cell references is correct now: It is partly relative (E2) and partly absolute ($E$11). $ is a string sign that makes the column number and/or the row number absolute. F4 is a cycle key that works like this:

- Pressing F4 once changes E11 to $E$11.
- Pressing F4 twice changes E11 to E$11.
- Pressing F4 three times changes E11 to $E11.
- Pressing F4 four times takes changes the cell back to E11.

You select the range C14:F23 in order to calculate what the new concentration of a certain solution is if you dilute certain concentrations (in column B) with a particular factor (in row 13). Then you follow these steps:

1. Enter the formula `=$B14*C$13` in cell C14.
2. While building the formula in the formula bar, select a cell and press F4 immediately. If you do this at a later stage, you need to click or double-click the cell address that needs to be changed before you press F4.
3. Accept this formula with Ctrl+Enter so it goes into all the selected cells, where it behaves partially as relative and partially as absolute.

* * *

Chapter 4

Range Names

A cell address is like a street number—and both can be difficult to remember. You might want to replace a cell number with a more meaningful address: a *cell name*. A cell name basically acts like an absolute cell address, but when used in formulas, it may be more informative. You can also name a range of cells. So you can have cell names and range names, but because a cell is basically also a range consisting of only one cell, the term *range name* is more comprehensive.

The top of Figure 1.7 shows a table with a list of readings that several analysts found during several tests. The bottom table marks each combination of a specific analyst and a specific test with a plus sign (+) if that reading was above the grand mean. Instead of comparing each individual reading with the grand mean in $G$12, you could also give cell G12 a more meaningful name—a range name. Here's how you do it:

1. Select cell G12.
2. In the Name box, to be found to the left of the formula bar, type `GrandMean`. Here are a few rules for naming:
   - Don't include spaces in a name, nor dots,@, #, /; underscores are okay.
   - Names are not case-sensitive, so GrandMean is the same as grandmean, GRANDMEAN, and so on.
   - Unique names function in the entire workbook. If you create a duplicate name, the second name will be assigned only to the specific sheet you are in.
3. Press Enter. If you don't, the name does not exist.

Now the name GrandMean has become official, so you can access the cell GrandMean through the drop-down list of the Name box—no matter where you are in this workbook—and Excel will take you there!

From now on, you should be able to call the `IF` function in cell B15. Its first argument is `B2>GrandMean`. You can just type the new range name, or you can click cell G12 to have Excel insert its name automatically.

Unfortunately, the previously installed range name does not kick in when you select multiple cells to fill them with the same formula. You must use an absolute cell address again, or you could use an extra tool: the *Use in Formula* drop-down located on the Formulas tab.

A1 Readings

	A	B	C	D	E	F	G
1	Readings	Test1	Test2	Test3	Test4	Test5	Means
2	Analyst1	3.83	5.06	6.08	4.36	8.37	5.54
3	Analyst2	9.83	8.75	7.21	4.85	1.84	6.50
4	Analyst3	9.20	4.07	3.36	6.52	4.76	5.58
5	Analyst4	9.71	3.51	1.19	4.66	3.48	4.51
6	Analyst5	7.29	9.10	9.46	9.05	3.11	7.60
7	Analyst6	6.61	4.53	0.22	0.41	0.31	2.42
8	Analyst7	9.80	7.55	4.62	8.14	9.19	7.86
9	Analyst8	9.73	5.63	0.49	9.22	1.14	5.24
10	Analyst9	3.19	4.20	8.27	4.06	6.21	5.19
11	Analyst10	3.09	1.57	1.07	2.93	5.16	2.76
12	Means	7.23	5.40	4.20	5.42	4.36	5.32
13							
14	>Mean	Test1	Test2	Test3	Test4	Test5	GrandMean
15	Analyst1			+		+	5.32
16	Analyst2	+	+	+			5.32
17	Analyst3	+			+		5.32
18	Analyst4	+					5.32
19	Analyst5	+	+	+	+		5.32
20	Analyst6	+					5.32
21	Analyst7	+	+		+	+	5.32
22	Analyst8	+	+		+		5.32
23	Analyst9			+		+	5.32
24	Analyst10						5.32

Figure: 1.7

Try getting a copy of the grand mean in the cells G15:G24. When you just type the formula in the formula bar, notice that the name nicely pops up while you type `=Gr`…

To find the range names listed, follow these steps:

1. Select the Formulas tab.
2. Click the Name Manager button.
3. Select the name of your choice.
4. Delete that name (using the button to the right) or expand/change its reference (at the bottom).

Figure 1.8 shows that you can also name ranges of multiple cells. For example, you could name the first range Analysts, the second one Strains, and the third one Readings. Instead of doing all this manually, you can use a handy Excel tool:

1. Select the entire table by selecting A1, pressing Ctrl+Shift+Down Arrow, and then pressing Ctrl+Shift+Right Arrow.
2. From the Defined Names section of the Formulas tab, select Create Names from Selection.

F2 =COUNTIF(Analysts,E2)

	A	B	C	D	E	F	G	H
1	Analysts	Strains	Readings			Count	Sum	Means
2	Analyst1	Strain1	1.22		Analyst1	5	27.39	5.48
3	Analyst1	Strain2	5.19		Analyst2	5	25.65	5.13
4	Analyst1	Strain2	8.81		Analyst3	5	29.24	5.85
5	Analyst1	Strain3	5.71		Analyst4	5	30.99	6.20
6	Analyst1	Strain3	6.46					
7	Analyst2	Strain1	1.23					
8	Analyst2	Strain1	8.22					
9	Analyst2	Strain2	1.70			Count	Sum	Means
10	Analyst2	Strain2	6.96		Strain1	6	38.11	6.36
11	Analyst2	Strain3	7.54		Strain2	7	36.52	5.22
12	Analyst3	Strain1	9.44		Strain3	7	38.64	5.52
13	Analyst3	Strain1	9.86					
14	Analyst3	Strain2	2.41					
15	Analyst3	Strain3	0.67					
16	Analyst3	Strain3	6.86					
17	Analyst4	Strain1	8.14					
18	Analyst4	Strain2	1.80					
19	Analyst4	Strain2	9.65					
20	Analyst4	Strain3	5.48					
21	Analyst4	Strain3	6.92					
22								

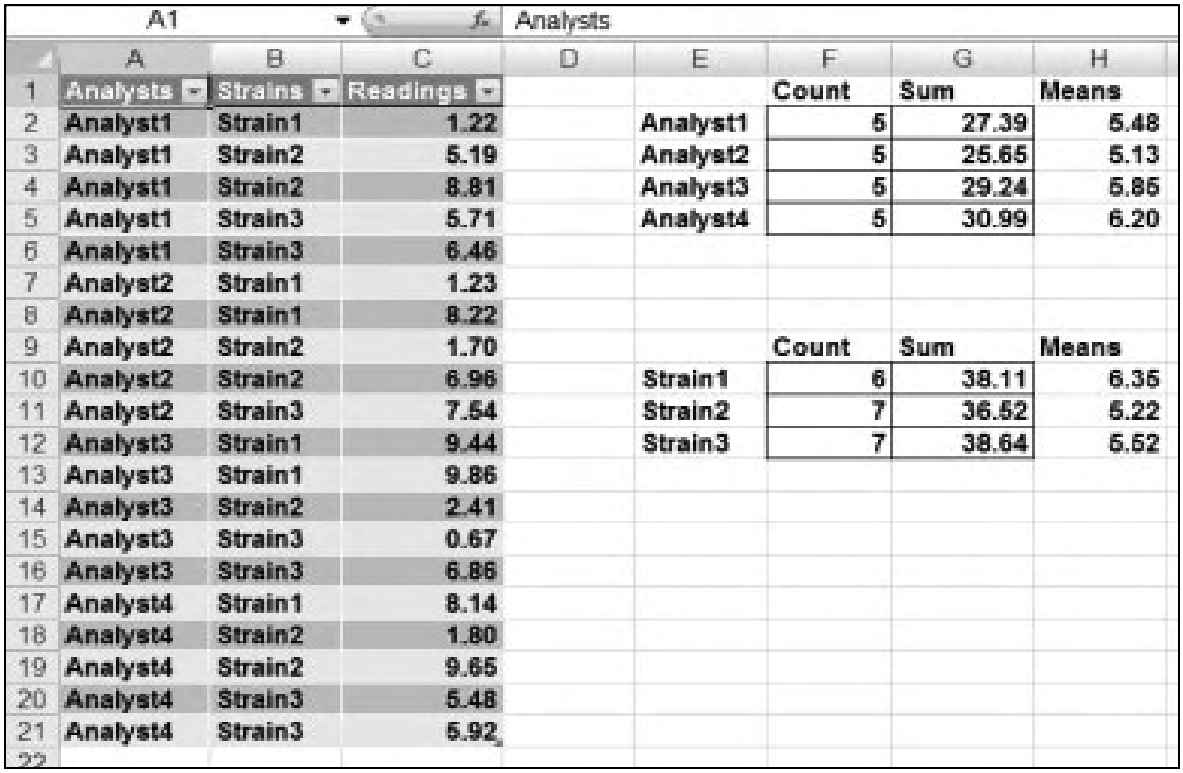
Figure: 1.8

3. Select Create Names from Values in ☑ Top Row
4. Check the Name box to ensure that the three new names appear.

Now try counting in cell F2 how many readings Analyst1 has—by using the `COUNTIF` function: `=COUNTIF(Analysts,E2)`. You can do something similar in cell G2 with the `SUMIF` function: `=SUMIF(Analysts,E2,Readings)`. To find the Mean in cell H2, you need both previous columns—or you could use the Excel function `AVERAGEIF`. Then you can do something similar for the second table: `=COUNTIF(Strains,E10)` and `=SUMIF(Strains,E10,Readings)`.

There is another interesting feature about names that you may benefit from. Instead of using a formula like `=SUM(Readings)`, you can use the word Readings as it is displayed somewhere in a cell. However, you need the function `INDIRECT` to change the word into a name. For example, if cell A1 holds the word Readings, the formula `=SUM(INDIRECT(A1))` would deliver the same results as `=SUM(Readings)`. Why would you want to make such a detour? The answer is simple: Sometimes you want the headers of a summary table somewhere else in your book to be used in your formulas. One of the exercises at the end of this part offers an example of such a scenario.

Range names are great. However, the problem with range names is that a new row added at the end of a table, column, or row is not automatically incorporated into the range name—so formulas based on that range ignore the new entries. You can solve this problem by either inserting cells inside the range or manually fixing the range reference through the Name Manager. Neither solution is ideal.



	A	B	C	D	E	F	G	H
1	Analysts	Strains	Readings			Count	Sum	Means
2	Analyst1	Strain1	1.22		Analyst1	5	27.39	5.48
3	Analyst1	Strain2	5.19		Analyst2	5	25.65	5.13
4	Analyst1	Strain2	8.81		Analyst3	5	29.24	5.85
5	Analyst1	Strain3	5.71		Analyst4	5	30.99	6.20
6	Analyst1	Strain3	6.46					
7	Analyst2	Strain1	1.23					
8	Analyst2	Strain1	8.22					
9	Analyst2	Strain2	1.70			Count	Sum	Means
10	Analyst2	Strain2	6.96		Strain1	6	38.11	6.35
11	Analyst2	Strain3	7.54		Strain2	7	36.52	5.22
12	Analyst3	Strain1	9.44		Strain3	7	38.64	5.52
13	Analyst3	Strain1	9.86					
14	Analyst3	Strain2	2.41					
15	Analyst3	Strain3	0.67					
16	Analyst3	Strain3	6.86					
17	Analyst4	Strain1	8.14					
18	Analyst4	Strain2	1.80					
19	Analyst4	Strain2	9.65					
20	Analyst4	Strain3	5.48					
21	Analyst4	Strain3	5.92					
22								

Figure: 1.9

Figure 1.9 shows how you can handle this problem in a better way, by using Excel's table structure. First of all, the table itself has automatically been given a default name (as you can see when you open the Name box). When you want to "talk" to the table from outside the table, you must use this name. You just type something like the following somewhere in a cell outside the table: `=Table1[`. Yes, the keystroke at the end is a bracket, not a parenthesis! As soon as you type the first bracket, several range names pop up:

- Analysts (A2:A21 in this example)
- Strains (B2:B21 in this example)
- Readings (C2:C21 in this example)
- #All (A1:C21 in this example)
- #Data (A2:C21 in this example)

- #Headers (A1:C1 in this example)
- #Totals (missing in this example)

A great feature of a table name is that it adjusts to a new range whenever you use its drag button in the right-lower corner. Be careful, though! When you have selected the last cell, the drag button is hidden behind the fill handle of the selected cell. So you must select another cell first in order to make the drag button accessible.

Now let's tackle the formulas in the two summary tables to the right. This time, you manually type the `COUNTIF` formula in cell F2. (Notice how much help you get here.) You end up with the following two formulas: `=COUNTIF(Table1[Analysts],E2)` and `=SUMIF(Table1[Analysts],E2,Table1[Readings])`.

But what can you do if you don't want to create a table structure? Is there another way of making range names automatically expand when new rows or columns have been added at the end? Yes, there is; although it is not an easy way, it is sometimes highly advantageous. You use the `OFFSET` function in the Name Manager. The `OFFSET` function has the following syntax: `=OFFSET(start-cell, row-offset, col-offset, number-of-rows, number-of-columns)`. To assign a dynamic name to the table in the columns A:C, you take these steps:

1. Open the Name Manager.
2. Type a new name for the table, such as `MyTable`, `MyColumn`, or `MyRow`.
3. In the Refers To box, type `=OFFSET(A1,0,0,COUNTA(A:A),COUNTA(1:1))`. This is how the formula works:
   - The name's reference starts in A1.
   - You want to keep that cell as a starting point, so you offset by 0 rows and 0 columns.
   - You find the height of the range by counting all non-empty cells in column A. You need to make sure there are no hidden cells below the table.
   - You find the width of the range by counting all non-empty cells in row 1. You need to make sure there are no hidden cells to the right of the table (which is not the case in Figure 1.8!).

Whenever rows or columns are added to the table or deleted, the two `COUNTA` functions inside `OFFSET` automatically take care of the size of the adjusted range.

Needless to say that you can also name the entire column A in this case. Anything added to the column is automatically included in the name.

* * *

Chapter 5

Nested Functions

Formulas in Excel contain calculations and/or functions. What is the difference between them?

- Calculations work with operators such as `()`, `^`, `*`, `/`, `+`, and `-` (in this order of precedence). So whereas `2+4/2` is 4 in Excel, `(2+4)/2` returns 3. To create the square root of the number 4 by using operators, you need parentheses as well: `=4^(1/2)`.
- Functions are built-in operations, such as `=SUM(A1:A3)`, which is equivalent to the calculation =A1+A2+A3. Most functions accept or require one or more arguments inside their parentheses. For instance, the square root of 4 done with a function would require one argument: `SQRT(4)`.
- Formulas can be a combination of calculations and functions. There can be calculations inside functions (for instance, `=ROUND(A1/B1,1)`) and there can be functions inside calculations (for example, `=SUM(A1:A3)*0.05`).
- Formulas can also nest functions inside functions (for example, `=ROUND(SUM(A1:A3), 2)`). In this case, the `SUM` function is nested inside the `ROUND` function.

Figure 1.10 shows a few of these options:

- **D9 holds a calculation**: `=C9^(1/2)`, which is the square root of 4.
- **D11 holds a calculation**: `=C11*-1`, which returns the absolute value of 4. (Note that this does not work for -4.)
- **F9 holds a function**: `=SQRT(G9)`, which returns the square root of 4.
- **F11 holds a function**: `=ABS(G11)`, which works always, even when G11 is already positive.
- **E14 has a calculation inside a function**: `=ROUND(1/6,4)`.
- **E17 has a function inside a calculation**: `=C17/SQRT(G17)`.
- **E20 has a function nested inside a function**: `=ROUND(RAND()*10,2)`.

You can just type these formulas into a cell, but don't forget to start with an equals sign if you do so. When you type functions, you receive some help as to their syntax: which arguments appear and in which order. Notice also that some arguments are shown in bold, which means

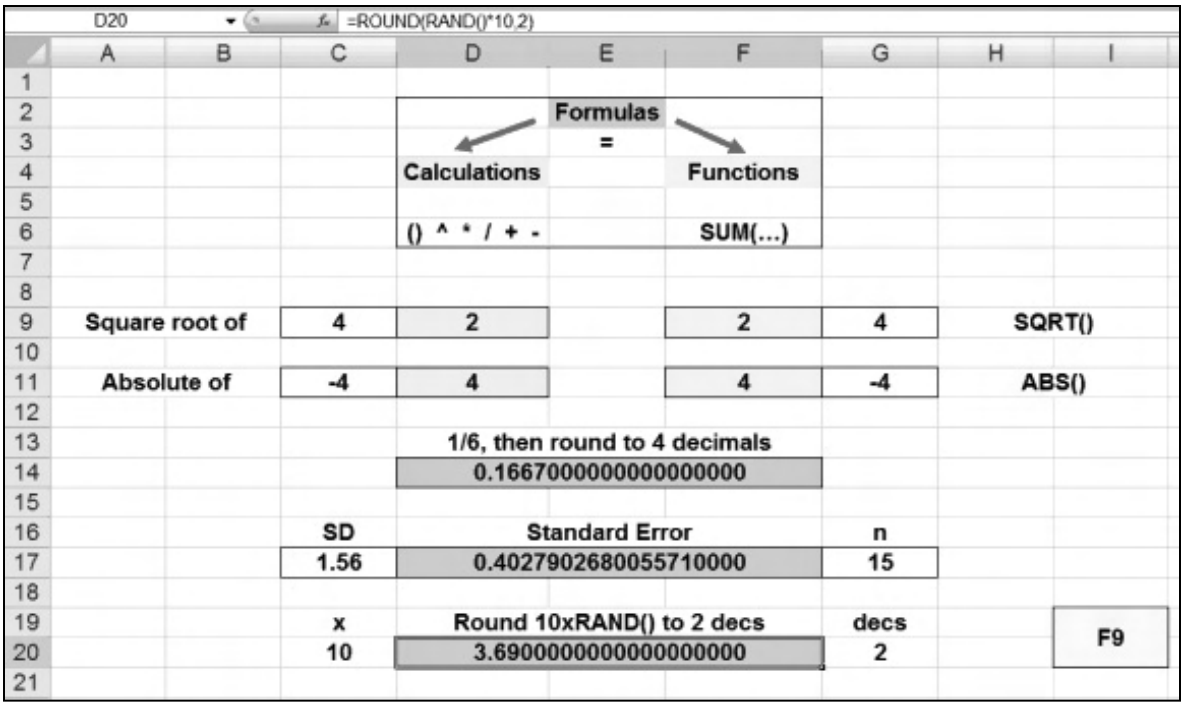


Figure: 1.10

they are required for the function to work; the functions that are not bold are optional and can be skipped most of the time.

To get more extensive help, you should use one of these three options:

- The fx button on the formula bar
- The Σ drop-down button on the Home tab
- One of the buttons on the Formulas tab

Figure 1.11 shows the use of two different functions: in column A, the function `RAND` (a random number between 0 and 1) and in column B, the function `ROUND` (to round `RAND` to two decimals). You could combine these two operations into one column (column C) by using a nested function. Here's how you do it:

1. Start the `ROUND` function, this time from the fx button.
2. Ensure that the first argument of the `ROUND` function is a nested `RAND` function. You get this second function from a new drop-down section located where the Name box used to be. When you do so, the `ROUND` box gets replaced by the `RAND` box.
3. Do not click OK because you are not yet done with `ROUND`. You get back to `ROUND` by clicking its name in the formula bar (not in the Name box). When you are back in the `ROUND` box, you can complete the second argument.

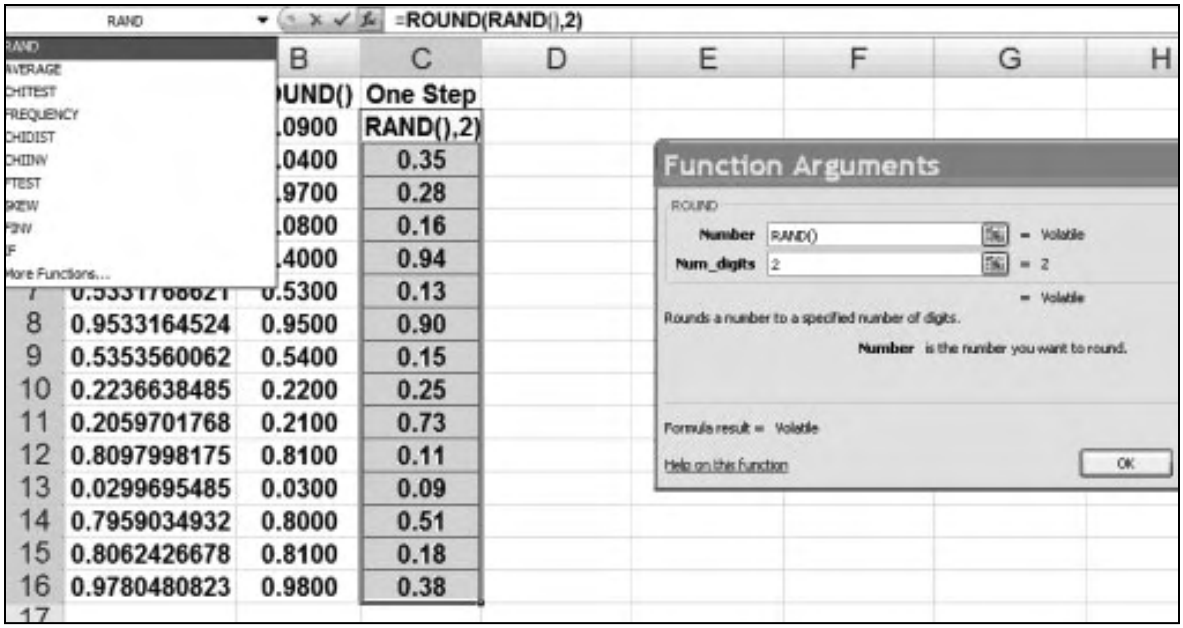


Figure: 1.11

Figure 1.12 shows a spreadsheet in which you mark each analyst with a plus sign if both tests came out above the mean; otherwise, there is no marker. In this case, you need two functions: the function IF to test for values above the mean and the function `AND` to check for both tests at the same time; in other words, `AND` should be nested inside `IF`. Here's how it works:

1. Call the function `IF` from fx.
2. Give the first argument an `AND` function from the former Name box: AND(B2>$B$12,C2>$C$12).

	A	B	C	D
1		Test1	Test2	Both>Avg
2	Analyst1	5.89	4.41	=IF(AND(B2>$B$12,C2>$C$12),"+","")
3	Analyst2	5.87	0.77	=IF(AND(B3>$B$12,C3>$C$12),"+","")
4	Analyst3	7.96	6	=IF(AND(B4>$B$12,C4>$C$12),"+","")
5	Analyst4	8.21	6.62	=IF(AND(B5>$B$12,C5>$C$12),"+","")
6	Analyst5	1.36	9.88	=IF(AND(B6>$B$12,C6>$C$12),"+","")
7	Analyst6	4.77	1.39	=IF(AND(B7>$B$12,C7>$C$12),"+","")
8	Analyst7	6.35	9.34	=IF(AND(B8>$B$12,C8>$C$12),"+","")
9	Analyst8	0.01	2.89	=IF(AND(B9>$B$12,C9>$C$12),"+","")
10	Analyst9	3.83	1.4	=IF(AND(B10>$B$12,C10>$C$12),"+","")
11	Analyst10	5.51	7.03	=IF(AND(B11>$B$12,C11>$C$12),"+","")
12		=AVERAGE(B2:B11)	=AVERAGE(C2:C11)	

Figure: 1.12

3. Instead of clicking OK, click the word `IF` in the formula bar.
4. When you are back in `IF`, finish the formula: `=IF(AND(B2>$B$12,C2>$C$12),"+","")`.

Could you just type this formula from scratch? Sure you could, but when the syntax gets more complicated, those function dialog boxes protect you from making mistakes, and they provide more help when you are dealing with a function you are not familiar with.

Now that you know these basic techniques, you can use them when dealing with the analysis of scientific data—whether you're doing data analysis (discussed in Part 2), regression analysis (Part 4), statistical analysis (Part 5), or data plotting (Part 3). Read on!

* * *

Excercises - Part 1

You can download all the files used in this book from `www.genesispc.com/Science2007.htm`, where you can find each file in its original version (to work on) and in its finished version (to check your solutions).

Exercise 1

1. The Fill Handle

   1.1. Make column D completely identical to column A by using just the fill handle (and, if needed, the Ctrl key).

   2.2. Make column E completely identical to column B by using just the fill handle (and, if needed, the Ctrl key).

D1 — Plate1

	A	B	C	D	E
1	Plate1	Cycle1		Plate1	Cycle1
2	Plate1	Cycle2			
3	Plate1	Cycle3			
4	Plate1	Cycle4			
5	Plate1	Cycle5			
6	Plate1	Cycle6			
7	Plate1	Cycle7			
8	Plate1	Cycle8			
9	Plate1	Cycle9			
10	Plate1	Cycle10			
11	Plate2	Cycle1			
12	Plate2	Cycle2			
13	Plate2	Cycle3			
14	Plate2	Cycle4			
15	Plate2	Cycle5			
16	Plate2	Cycle6			
17	Plate2	Cycle7			
18	Plate2	Cycle8			
19	Plate2	Cycle9			
20	Plate2	Cycle10			
21					

Figure: Ex-1

Exercise 2

2. Relative vs. Absolute Cell References

2.1. Select the range D4:H15 and implement the equation shown below the table in a single step. (Note: The value 1.96 features in cell B2, so use its cell reference.)

2.2. Do the same for range L4:P15 but this time with the value 1.65 (in cell J2).

1.96		margin/mean				
		0.01	0.02	0.05	0.1	0.25
SD/mean	0.1	384	96	15	4	1
	0.2	1537	384	61	15	2
	0.3	3457	864	138	35	6
	0.4	6147	1537	246	61	10
	0.5	9604	2401	384	96	15
	0.6	13830	3457	553	138	22
	0.7	18824	4706	753	188	30
	0.8	24586	6147	983	246	39
	0.9	31117	7779	1245	311	50
	1	38416	9604	1537	384	61
	2	153664	38416	6147	1537	246
	3	345744	86436	13830	3457	553

1.65		margin/mean				
		0.01	0.02	0.05	0.1	0.25
SD/mean	0.1	272	68	11	3	0
	0.2	1089	272	44	11	2
	0.3	2450	613	98	25	4
	0.4	4356	1089	174	44	7
	0.5	6806	1702	272	68	11
	0.6	9801	2450	392	98	16
	0.7	13340	3335	534	133	21
	0.8	17424	4356	697	174	28
	0.9	22052	5513	882	221	35
	1	27225	6806	1089	272	44
	2	108900	27225	4356	1089	174
	3	245025	61256	9801	2450	392

$$n = \left(\frac{1.96}{margin\ /\ \mu}\right)^2 \left(\frac{SD}{\mu}\right)^2$$

Figure: Ex-2

Exercise 3

3. Range Names

3.1. Use Excel's handy Naming tool to name each column in the left table with its label name.

In the table in H2:K5, calculate the correlation between any two variables used in the table to the left. Use the labels of your correlation table in combination with the functions `CORREL` and `INDIRECT`.

Age	Weight	Height	FVC
55	1.84	1.41	2.67
59	1.80	1.41	2.80
44	1.53	1.50	3.21
37	1.50	1.54	3.32
50	1.57	1.54	2.87
48	1.90	1.55	3.01
45	1.89	1.56	2.96
40	1.50	1.57	3.21
26	1.73	1.57	3.56
52	1.92	1.60	2.82
56	1.64	1.41	2.74
44	1.66	1.46	3.14
53	1.88	1.51	2.82
36	1.78	1.54	3.39
46	1.71	1.56	2.98
35	1.71	1.60	3.41
31	1.98	1.65	3.44
23	1.77	1.67	3.73
36	1.71	1.73	3.30
27	1.61	1.74	3.67
37	1.93	1.75	3.28
28	1.99	1.75	3.63
23	1.96	1.77	3.73
21	1.73	1.78	3.77
40	1.79	1.78	3.38
30	1.98	1.82	3.46
24	1.74	1.91	3.77
30	1.72	1.91	3.60
31	1.99	1.91	3.46
22	1.87	1.95	3.71

	Age	Weight	Height	FVC
Age	1	-0.122390	-0.778753	-0.980487
Weight	-0.122390	1	0.328697	0.069172
Height	-0.778753	0.328697	1	0.775915
FVC	-0.980487	0.069172	0.775915	1

Figure: Ex-3

Exercise 4

4. Nested Functions

4.1. In cell B2, use the function `ADDRESS` with the nested functions `ROW` and `COLUMN` to get the address of cell A1.

In cell F2, use the function `CONCATENATE` with the nested functions `LEFT` and `RIGHT`. To get the number 10 in cells F11 and F21 (instead of 0), you may have to also use the function `LEN` nested inside `RIGHT`.

	A	B	C	D	E	F
1		Address				Combine
2	1	$A$2		Plate1	Cycle1	Pla1-Cy1
3	2	$A$3		Plate1	Cycle2	Pla1-Cy2
4	3	$A$4		Plate1	Cycle3	Pla1-Cy3
5	4	$A$5		Plate1	Cycle4	Pla1-Cy4
6	5	$A$6		Plate1	Cycle5	Pla1-Cy5
7	6	$A$7		Plate1	Cycle6	Pla1-Cy6
8	7	$A$8		Plate1	Cycle7	Pla1-Cy7
9	8	$A$9		Plate1	Cycle8	Pla1-Cy8
10	9	$A$10		Plate1	Cycle9	Pla1-Cy9
11	10	$A$11		Plate1	Cycle10	Pla1-Cy10
12	11	$A$12		Plate2	Cycle1	Pla2-Cy1
13	12	$A$13		Plate2	Cycle2	Pla2-Cy2
14	13	$A$14		Plate2	Cycle3	Pla2-Cy3
15	14	$A$15		Plate2	Cycle4	Pla2-Cy4
16	15	$A$16		Plate2	Cycle5	Pla2-Cy5
17	16	$A$17		Plate2	Cycle6	Pla2-Cy6
18	17	$A$18		Plate2	Cycle7	Pla2-Cy7
19	18	$A$19		Plate2	Cycle8	Pla2-Cy8
20	19	$A$20		Plate2	Cycle9	Pla2-Cy9
21	20	$A$21		Plate2	Cycle10	Pla2-Cy10
22						

Figure: Ex-4

Exercise 5

5. Nested Functions

5.1. In column C, round the standard deviation of the two values to the left to two decimals by using the function `STDEV` nested inside the function `ROUND`.

You receive errors in column C, so improve your formula in column G by nesting the previous formula inside an `IF` function that tests for standard deviation errors by using the function `ISERROR`.

	A	B	C	D	E	F	G
1			SD errors				Correct
2	0.45	6.32	4.15		0.45	6.32	4.15
3	10.63	5.86	3.37		10.63	5.86	3.37
4	8.92		#DIV/0!		8.92		
5	8.77	14.22	3.85		8.77	14.22	3.85
6	1.02	6.64	3.97		1.02	6.64	3.97
7		3.90	#DIV/0!			3.90	
8	9.06	9.71	0.46		9.06	9.71	0.46
9	8.96	11.83	2.03		8.96	11.83	2.03
10	1.85	3.85	1.41		1.85	3.85	1.41
11	5.27	10.80	3.91		5.27	10.80	3.91
12			#DIV/0!				
13	3.70	4.70	0.71		3.70	4.70	0.71
14	0.88	2.44	1.10		0.88	2.44	1.10
15	12.36	2.09	7.26		12.36	2.09	7.26
16	9.45	13.37	2.77		9.45	13.37	2.77
17	6.90	3.74	2.23		6.90	3.74	2.23
18	1.20	11.85	7.53		1.20	11.85	7.53
19	10.40	0.12	7.27		10.40	0.12	7.27
20	1.48	13.67	8.62		1.48	13.67	8.62
21	12.57	12.30	0.19		12.57	12.30	0.19
22	Mean SD		#DIV/0!		Mean SD		3.58

Figure: Ex-5

* * *

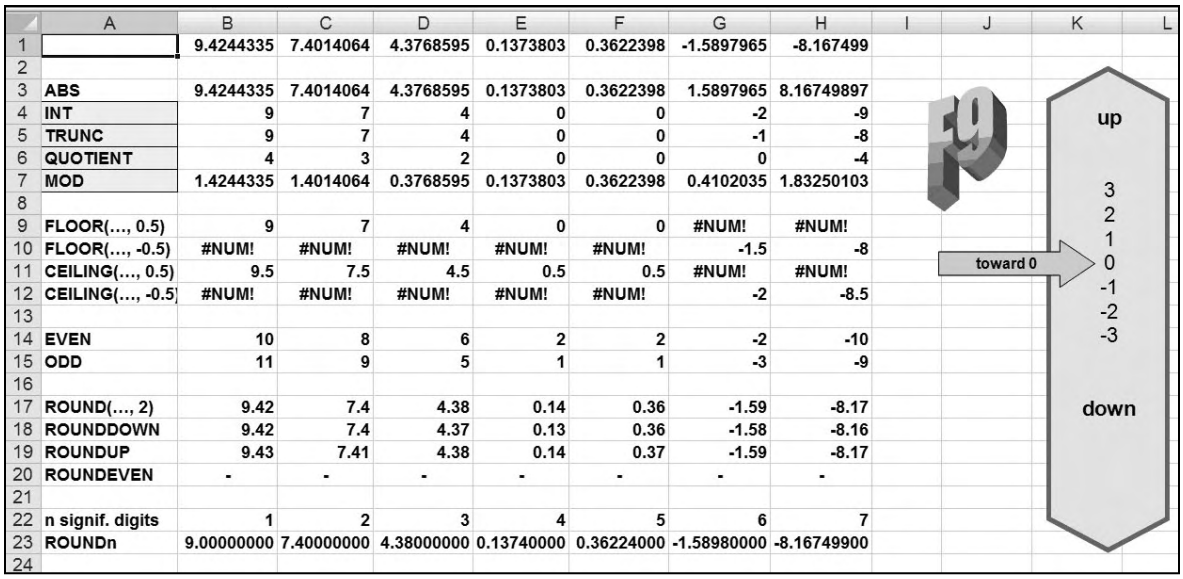
PART 2
Data Analysis

Chapter 6

AUTO-NUMBERING

To make record keeping easier, it may be wise to implement a good numbering system for each row or record. You need to know about some of Excel's good tools and functions before you take on your table's numbering system.

Figure 2.1 provides an overview of Excel's rounding functions. In the value view of the sheet (vs. formula view) , if you press F9, all numbers in the top row change after each click because the function RAND is used in the first row. Also, all the rounding functions change accordingly—but they all do different things: Some always round down; others round toward 0, depending on whether they are in the positive or negative range; and so on.



	A	B	C	D	E	F	G	H
1		9.4244335	7.4014064	4.3768595	0.1373803	0.3622398	-1.5897965	-8.167499
2								
3	ABS	9.4244335	7.4014064	4.3768595	0.1373803	0.3622398	1.5897965	8.16749897
4	INT	9	7	4	0	0	-2	-9
5	TRUNC	9	7	4	0	0	-1	-8
6	QUOTIENT	4	3	2	0	0	0	-4
7	MOD	1.4244335	1.4014064	0.3768595	0.1373803	0.3622398	0.4102035	1.83250103
8								
9	FLOOR(..., 0.5)	9	7	4	0	0	#NUM!	#NUM!
10	FLOOR(..., -0.5)	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-1.5	-8
11	CEILING(..., 0.5)	9.5	7.5	4.5	0.5	0.5	#NUM!	#NUM!
12	CEILING(..., -0.5	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-2	-8.5
13								
14	EVEN	10	8	6	2	2	-2	-10
15	ODD	11	9	5	1	1	-3	-9
16								
17	ROUND(..., 2)	9.42	7.4	4.38	0.14	0.36	-1.59	-8.17
18	ROUNDDOWN	9.42	7.4	4.37	0.13	0.36	-1.58	-8.16
19	ROUNDUP	9.43	7.41	4.38	0.14	0.37	-1.59	-8.17
20	ROUNDEVEN	-	-	-	-	-	-	-
21								
22	n signif. digits	1	2	3	4	5	6	7
23	ROUNDn	9.00000000	7.40000000	4.38000000	0.13740000	0.36224000	-1.58980000	-8.16749900
24								

Figure: 2.1

This chapter focuses on four of the rounding functions:

- `INT`: This function returns the integer part of a number but rounds down; for example, `INT(7/2)` returns 3, and `INT(-7/2)` returns -4.
- `TRUNC`: This function returns the integer part of a number but rounds toward 0; for example, `TRUNC(7/2)` returns 3, and `TRUNC(7/2)` returns -3.

- `QUOTIENT`: This function returns the integer part of a number after division; for example, `QUOTIENT(7,2)` returns 3. However, in Excel 2003 and earlier, `QUOTIENT` is available only through the Analysis Toolpak (see Chapter 35).
- `MOD`: This function returns the remainder of a division; for example, `MOD(7,2)` returns 1

There is one more function you should know about in this context—the ROW function:

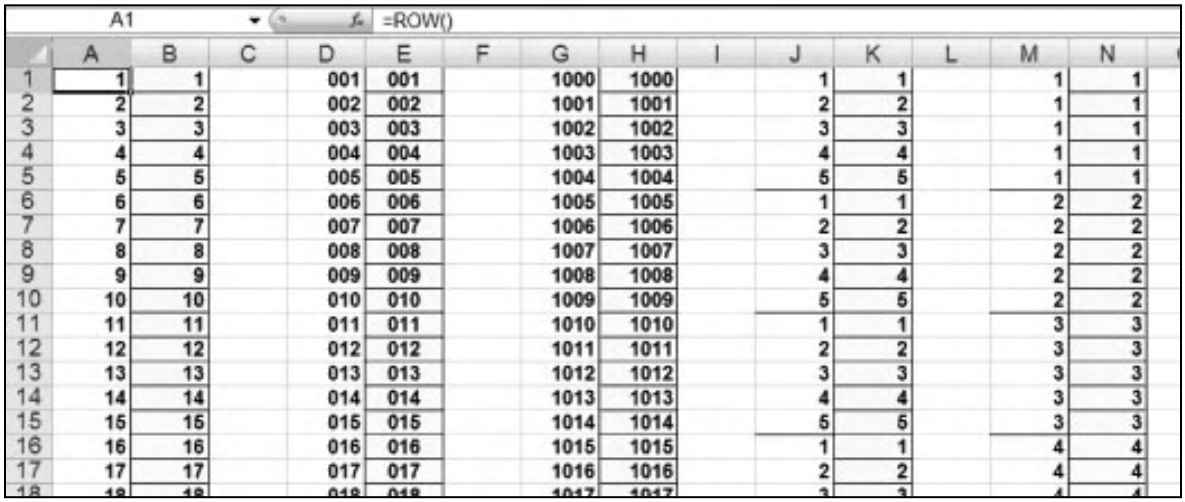
- `=ROW()` returns the number of the row you are in.
- `=ROW(A1)` returns the row number of cell A1, which is 1. When you copy the formula downward, the reference to A1 automatically updates to A2, and so on.

Figure 2.2 shows you some fancy automatic numbering systems:

- **A1**: `=ROW()`
- **D1**: `=RIGHT("000"&ROW(),3)`. The ampersand (`&`) is a string connector to hook things together; `RIGHT` takes the last three digits.
- **G1**: `=ROW(A1000)`. The `ROW` function's argument makes you start at a higher number, and it adjusts to copying.
- **J1**: `=MOD(ROW()-1,5)+1`. After each fifth row, the number starts all over again.
- **M1**: `=QUOTIENT(ROW()-1,5)+1`. The number repeats itself five times.

Here are three important techniques to use in implementing an auto-numbering system:

- You can select the entire numbering range. To do so, you click in the start cell (for example, A1), in the Name box, type the address of the end cell (for example, A1000),



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	1		001	001		1000	1000		1	1		1	1
2	2	2		002	002		1001	1001		2	2		1	1
3	3	3		003	003		1002	1002		3	3		1	1
4	4	4		004	004		1003	1003		4	4		1	1
5	5	5		005	005		1004	1004		5	5		1	1
6	6	6		006	006		1005	1005		1	1		2	2
7	7	7		007	007		1006	1006		2	2		2	2
8	8	8		008	008		1007	1007		3	3		2	2
9	9	9		009	009		1008	1008		4	4		2	2
10	10	10		010	010		1009	1009		5	5		2	2
11	11	11		011	011		1010	1010		1	1		3	3
12	12	12		012	012		1011	1011		2	2		3	3
13	13	13		013	013		1012	1012		3	3		3	3
14	14	14		014	014		1013	1013		4	4		3	3
15	15	15		015	015		1014	1014		5	5		3	3
16	16	16		016	016		1015	1015		1	1		4	4
17	17	17		017	017		1016	1016		2	2		4	4

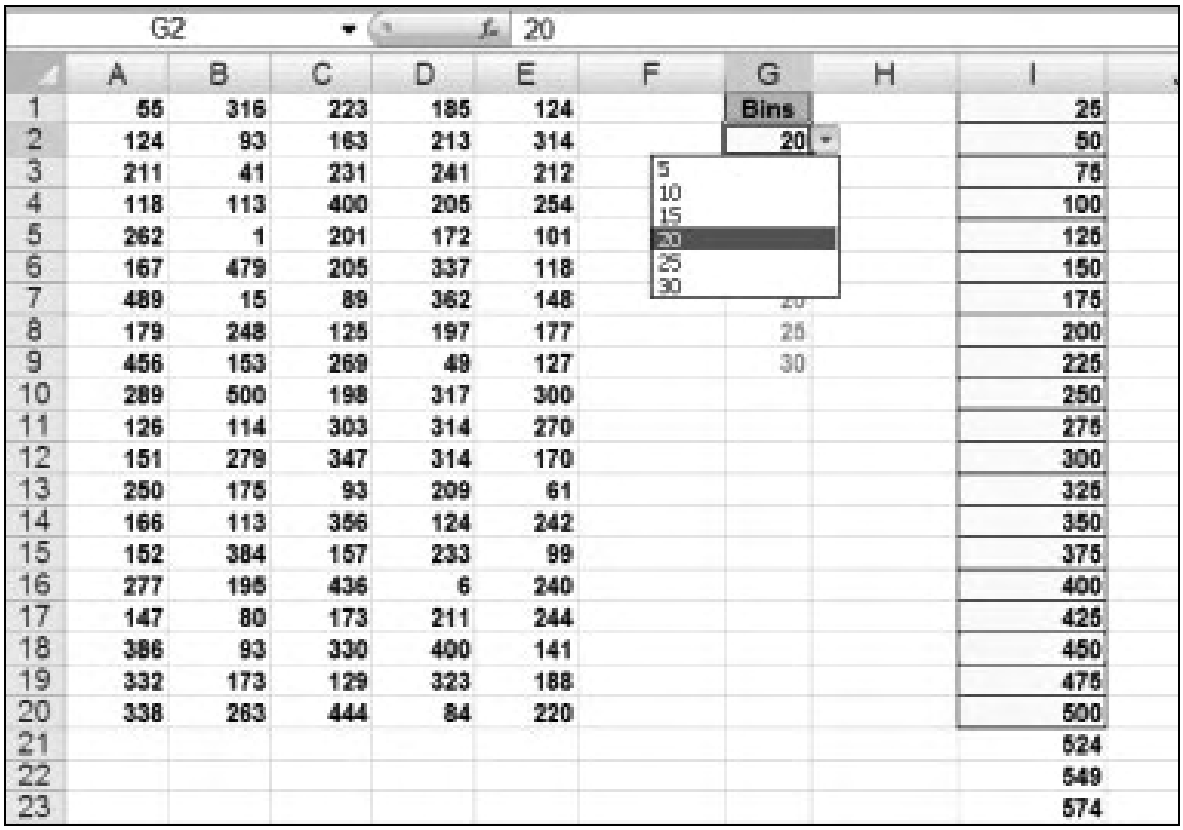
Figure: 2.2

and press Shift+Enter to select the entire range in between.

- You can just start your function or formula, which automatically ends up in the start cell; then you press Ctrl+Enter.
- You can change the formulas into real numbers by copying the entire section, selecting Paste Values from the Paste dropdown.

Figure 2.3 shows a different kind of auto-numbering system: It places a series of values (to the left) into a number of bins (in column I); you can determine how many bins you would like between the minimum and maximum value. Note the following about Figure 2.3:

- The range of the series of values has been named `data`.
- Cell G2 allows you to change the number of bins; Chapter 10 discusses how to create this drop-down button.
- When you change the number of bins, the frequency bins adjust automatically, thanks to the following formula in cell I1: `=INT(MIN(data)+(ROW(A1)*(MAX(data)-MIN(data))/$G$2))`.



	A	B	C	D	E	I
1	55	316	223	185	124	25
2	124	93	163	213	314	50
3	211	41	231	241	212	75
4	118	113	400	205	254	100
5	262	1	201	172	101	125
6	167	479	205	337	118	150
7	489	15	89	362	148	175
8	179	248	125	197	177	200
9	456	153	269	49	127	225
10	289	500	198	317	300	250
11	126	114	303	314	270	275
12	151	279	347	314	170	300
13	250	175	93	209	61	325
14	166	113	356	124	242	350
15	152	384	157	233	99	375
16	277	195	436	6	240	400
17	147	80	173	211	244	425
18	386	93	330	400	141	450
19	332	173	129	323	188	475
20	338	263	444	84	220	500
21						524
22						549
23						574

Figure: 2.3

* * *

Chapter 7

SUBTOTALS

A common step in data analysis is to create subtotals for specific subsets of records. Subtotals can be sums, means, standard deviations, and the like. This step can become a tremendous task in large spreadsheets if you're not familiar with the right tools to do so quickly and efficiently. As you'll learn in this chapter, you have to follow a few basic rules.

Figure 2.4 shows a good way of keeping track of your readings, measurements, observations, and so on. Of course, in your real life, the number of records is much, much larger than is shown here. In addition, you would not want to repeat recurring information in each record

	A	B	C	D	E	F	G	H
1	Plate ID	Analyst	50 ng/mL	%CV	25 ng/mL	%CV	10 ng/mL	%CV
2	8696p08a	gmv	52.3	1.0	26.5	2.0	12.2	2.0
3			49.7	2.0	25.0	1.0	11.2	3.0
4		ksm	56.4	12.0	29.1	11.0	12.8	9.0
5			51.3	2.0	26.9	1.0	11.7	1.0
6	8696p08b	bdo	52.9	3.0	27.5	6.0	13.1	9.0
7			50.1	1.0	26.8	3.0	12.1	2.0
8		ksm	51.0	1.0	26.1	1.0	12.3	2.0
9			48.6	1.0	24.8	0.0	11.4	2.0
10	8697p58b	tjk	47.5	0.0	22.7	2.0	9.6	8.0
11			47.5	1.0	22.9	3.0	9.4	0.0
12		tkm	43.2	4.0	22.6	3.0	9.4	4.0
13			44.3	3.0	22.6	3.0	9.4	6.0
14	8877p58a	gmv	47.7	2.0	22.7	0.0	10.3	0.0
15			48.7	3.0	23.0	3.0	10.0	0.0
16		tjk	49.3	0.0	23.0	1.0	10.4	3.0
17			45.9	2.0	22.9	1.0	9.8	1.0
18	8877p58b	bdo	43.2	4.0	22.6	3.0	9.6	4.0
19			44.3	3.0	20.7	3.0	9.7	6.0
20		tkm	47.5	0.0	22.7	2.0	9.6	8.0
21			45.9	1.0	22.9	3.0	9.4	0.0
22								

Figure: 2.4

as is done here because that would be a time-consuming job—and also a potential source of error. However, the advantage of not replicating repetitive information turns into a grave disadvantage when you need to reshuffle records. If you have incomplete records, sorting records by a specific column would change a neat table into a mess. (So if you have tried to sort, undo the sorting before the power goes off!)

Apparently, you really need to fill all those blank cells. Here is an easy, quick, and efficient way to select all blank cells in this table and fill them with the same data contained in the previous cell:

1. Select columns A and B
2. Click Find & Select.
3. Select Go To Special.
4. Choose ◉ Blanks.
5. Now all blank cells have been selected. Type the = sign, and then press the Up Arrow key (to point to the cell above "me").
6. Press Ctrl+Enter to place this formula in all selected cells.
7. Replace the formulas with their values by selecting Copy, Paste Special, Values Only.

Figure 2.5 shows the completed table. Say that you want to calculate the standard deviations per plate and per analyst who worked on that plate. The first step is to ensure that the table is properly sorted for this purpose. This is an example of multilevel sorting, which you can achieve by following these steps:

1. Click Sort & Filter.
2. Choose Custom Sort.
3. Add Plate ID as the first level.
4. Add Analyst as the second level.

When the sorting order is correct, do the following:

1. Click Subtotal on the Data tab.
2. Enter the plate ID in the Each Change in box.
3. Use the `STDEV` function.
4. Select the three columns for which you want subtotals.

Not only do you get beautiful subtotals, you also get a great outlining tool to the left of the sheet. To add another level of subtotals, you follow these steps:

1. Click Subtotal again.
2. Enter the analyst in the Each Change in box.

	A	B	C	D	E	F	G	H
1	Plate ID	Analyst	50 ng/mL	%CV	25 ng/mL	%CV	10 ng/mL	%CV
2	8696p08a	gmv	52.3	1.0	26.5	2.0	12.2	2.0
3	8696p08a	gmv	49.7	2.0	25.0	1.0	11.2	3.0
4		**gmv StdDev**	1.838478		1.06066		0.707107	
5	8696p08a	ksm	56.4	12.0	29.1	11.0	12.8	9.0
6	8696p08a	ksm	51.3	2.0	26.9	1.0	11.7	1.0
7		**ksm StdDev**	3.606245		1.555635		0.777817	
8	**8696p08a StdDev**		2.858175		1.693861		0.684957	
9	8696p08b	bdo	52.9	3.0	27.5	6.0	13.1	9.0
10	8696p08b	bdo	50.1	1.0	26.8	3.0	12.1	2.0
11		**bdo StdDev**	1.979899		0.494975		0.707107	
12	8696p08b	ksm	51.0	1.0	26.1	1.0	12.3	2.0
13	8696p08b	ksm	48.6	1.0	24.8	0.0	11.4	2.0
14		**ksm StdDev**	1.697056		0.919239		0.636396	
15	**8696p08b StdDev**		1.79722		1.15181		0.699405	
16	8697p58b	tjk	47.5	0.0	22.7	2.0	9.6	8.0
17	8697p58b	tjk	47.5	1.0	22.9	3.0	9.4	0.0
18		**tjk StdDev**	0		0.141421		0.141421	
19	8697p58b	tkm	43.2	4.0	22.6	3.0	9.4	4.0
20	8697p58b	tkm	44.3	3.0	22.6	3.0	9.4	6.0
21		**tkm StdDev**	0.777817		0		0	
22	**8697p58b StdDev**		2.211146		0.141421		0.1	
23	8877p58a	gmv	47.7	2.0	22.7	0.0	10.3	0.0
24	8877p58a	gmv	48.7	3.0	23.0	3.0	10.0	0.0
25		**gmv StdDev**	0.707107		0.212132		0.212132	

Figure: 2.5

3. Select Replace. Do not replace the current ones unless you want to.

> Note: If you ever want to get rid of the subtotals, select Data tab, Subtotal command in the Outline Group, Remove All.

If you prefer Excel's table structure over the table shown in Figure 2.5, be aware that the Subtotal tool doesn't work on a table. But subtotals are really handy, and you can follow these steps to use them in a table:

1. Click inside the table.
2. Click the Design tab.
3. Select Convert to Range. You're now back in a regular set of rows for your subtotals.

Figure 2.6 shows a table that has columns that were created in the "wrong" order. Changing their order— especially if you have a large number of columns—would be another enormous job. But you can use the following trick:

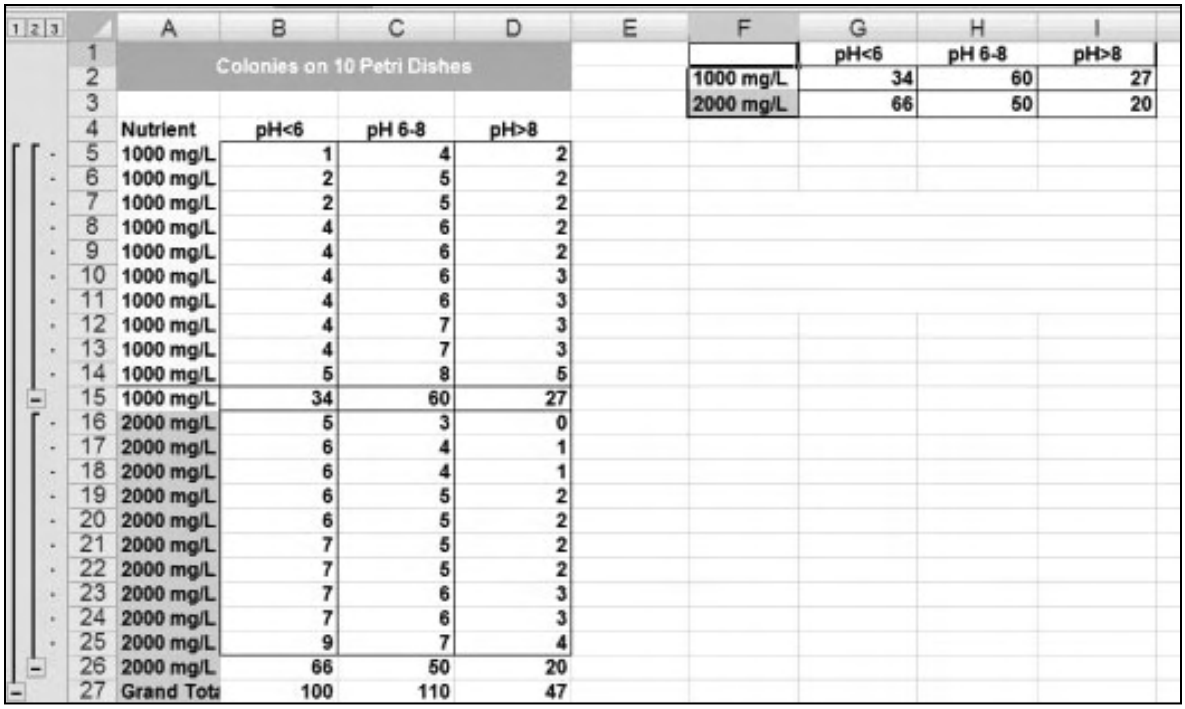
	A	B	C	D	E	F	G	H	I
1	1	2	3	4	5	6	7	8	
2	Plate ID	Analyst	10 ng/mL	%CV	25 ng/mL	%CV	50 ng/mL	%CV	
3	8877p58a	gmv	10.3	0.0	22.7	0.0	47.7	2.0	
4	8877p58a	gmv	10.0	0.0	23.0	3.0	48.7	3.0	
5	8877p58a	tjk	10.4	3.0	23.0	1.0	49.3	0.0	
6	8877p58a	tjk	9.8	1.0	22.9	1.0	45.9	2.0	
7	8877p58b	tkm	9.6	8.0	22.7	2.0	47.5	0.0	
8	8877p58b	tkm	9.4	0.0	22.9	3.0	45.9	1.0	
9	8877p58b	bdo	9.6	4.0	22.6	3.0	43.2	4.0	
10	8877p58b	bdo	9.7	6.0	20.7	3.0	44.3	3.0	
11	8696p08a	ksm	12.8	9.0	29.1	11.0	56.4	12.0	
12	8696p08a	ksm	11.7	1.0	26.9	1.0	51.3	2.0	
13	8696p08a	gmv	12.2	2.0	26.5	2.0	52.3	1.0	

Figure: 2.6

1. Create a dummy row by inserting an empty row before row 1.
2. Give each column a rank number (in the order in which you want the columns to appear).
3. Select Data tab, Sort & Filter, Custom Sort.
4. Click Options.
5. Select Sort Left to Right.
6. Click OK.
7. Sort by Row1.
8. Now you can delete the dummy row.

Figure 2.7 shows the number of colonies growing on 10 Petri dishes with two different nutrient levels. The table to the left is good for record keeping but not for data analysis. That's what the table to the right is for. So you need to transfer the data from the table on the left into the table on the right. How can you transport the subtotals from the left into the table on the right? You use the following technique:

1. Collapse the left table to its subtotals only.
2. Select B15:D26 (subtotals only).
3. Select Home, Find & Select.
4. Select Go To Special.
5. Select Visible Cells Only (otherwise, you would copy all that's in between as well).
6. Press Ctrl+C (to copy what is visible).
7. Select G2 and press Ctrl+V (to paste the subtotals only).



	A	B	C	D	E	F	G	H	I
1	Colonies on 10 Petri Dishes						pH<6	pH 6-8	pH>8
2						1000 mg/L	34	60	27
3						2000 mg/L	66	50	20
4	Nutrient	pH<6	pH 6-8	pH>8					
5	1000 mg/L	1	4	2					
6	1000 mg/L	2	5	2					
7	1000 mg/L	2	5	2					
8	1000 mg/L	4	6	2					
9	1000 mg/L	4	6	2					
10	1000 mg/L	4	6	3					
11	1000 mg/L	4	6	3					
12	1000 mg/L	4	7	3					
13	1000 mg/L	4	7	3					
14	1000 mg/L	5	8	5					
15	1000 mg/L	34	60	27					
16	2000 mg/L	5	3	0					
17	2000 mg/L	6	4	1					
18	2000 mg/L	6	4	1					
19	2000 mg/L	6	5	2					
20	2000 mg/L	6	5	2					
21	2000 mg/L	7	5	2					
22	2000 mg/L	7	5	2					
23	2000 mg/L	7	6	3					
24	2000 mg/L	7	6	3					
25	2000 mg/L	9	7	4					
26	2000 mg/L	66	50	20					
27	Grand Tota	100	110	47					

Figure: 2.6

Now you can remove the subtotals from the table on the left. The only problem is that the values on the right are hard coded, so they don't update when the values to the left change. Chapter 16 discusses a way to copy the formulas.

* * *

Chapter 8

Summary Functions

Data analysis often requires summary functions that provide summary information for a particular subset of records only. Excel calls these functions *Database functions*. They supply regular summary operations, according to conditions you have specified somewhere. Database functions require filters that identify the criteria for filtering the "database," using the table's labels as identifiers.

The rules for filters are pretty simple:

- A table must have labels or headers on top (usually in row 1).
- A table cannot contain completely empty rows or columns.
- The filter uses labels, which are usually identical to the table's labels.
- Criteria in the filter that appear on the same row act as an `AND` condition.
- Criteria in the filter that appear in the same column act as an `OR` condition.
- A filter cannot contain completely empty rows or columns.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Patient	Gender	DOB	Age	Weight	Systolic		Patient	Gender	DOB	Age	Weight	Systolic	
2	Bush	M	1/3/1975	32	160	178					>50		>140	
3	Carter	F	10/22/1937	69	192	151								
4	Clinton	M	7/15/1971	35	171	175		Mean			66	169	161	
5	Eisenhower	F	11/24/1934	72	154	128		SD			7	22	23	
6	Ford	M	2/9/1950	57	164	141		Count			5	5	5	
7	Johnson	F	6/13/1965	41	152	125								
8	Kennedy	M	11/27/1973	33	165	193		Patient	Gender	DOB	Age	Weight	Systolic	
9	Lincoln	M	8/1/1972	34	188	166					>60			
10	Nixon	F	9/15/1930	76	189	146					<40			
11	Reagan	M	3/28/1939	68	140	170								
12	Roosevelt	M	9/29/1945	61	159	196		Mean			53	169	167	
13	Truman	F	10/20/1962	44	151	139		SD			19	18	22	
14	Washington	F	10/15/1963	43	209	180		Count			9	9	9	
15														
16								Patient	Gender	DOB	Age	Weight	Systolic	Systolic
17													>140	<180
18														
19								Mean			53	172	161	
20								SD			19	19		
21								Count			7	7		
22														
23														

Figure: 2.8

Figure 2.8 shows a table named Data and three filters named Filter1, Filter2, and Filter3. In cell K4, the function DAVERAGE would calculate the mean of the age (K1), weight (L1), and systolic blood pressure (M1) for records in Data according to Filter1. This is the formula: =DAVERAGE(Data,K$1,Filter1). You just copy the formula downward and to the right. All you have to change is DAVERAGE into DSTDEV and DCOUNT. According to the table in Figure 2.8, the mean age of people older than 50 with a systolic blood pressure over 140 is 66.

> **Note**: If you receive a #VALUE message for the third argument in the function box, just ignore it.

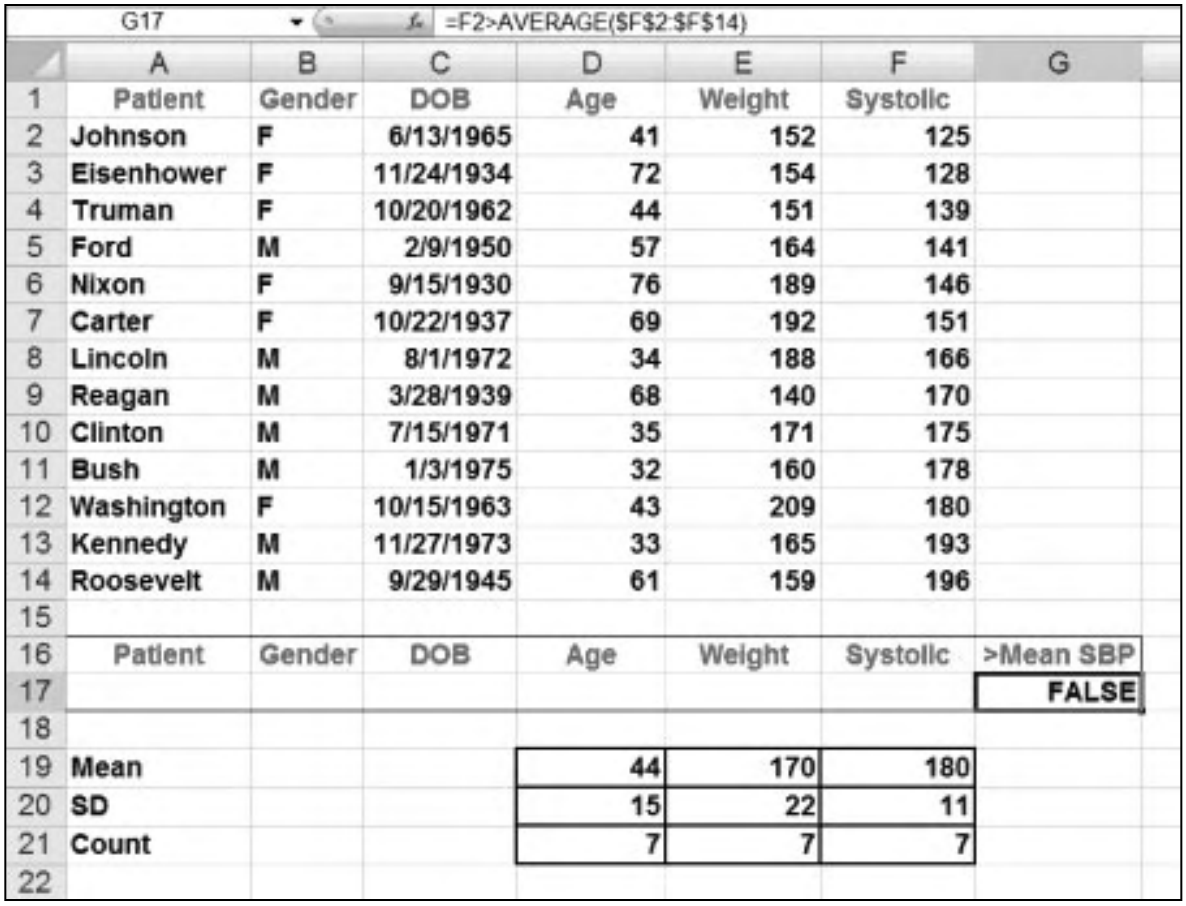
Any changes in the settings of Filter1 are automatically reflected in the results of the summary calculations. The same thing happens when you implement the two other filters. Note that Filter2 is set up as an OR filter, and Filter3 acts as an AND filter.

The real power of these filters is that they can include fields that contain filter formulas. You can just expand a filter with a newly invented field name that holds a formula for an even more customized filter. The idea behind the formula is that if the formula evaluates to TRUE, the record qualifies for the filtered subset. You should know, though, that the formula you create in a filter always works on the first record, but internally it runs through all records in the table.

Figure 2.9 has a formula in its filter to select all records that have a systolic blood pressure above the mean systolic blood pressure: =F2>AVERAGE($F$2:$F$14). You need to make sure this formula checks the first record (F2) and compares it with the mean of the entire, absolute range ($F$2:$F$14). Calculated or computed filters always evaluate to TRUE or FALSE. They actually show the evaluation for the first record—which happens to be TRUE in this case because 178 is in fact greater than the mean of 161. Notice that the database functions in the summary table reflect the subset they are based on, and they follow this rule: If the formula evaluates to TRUE, the record qualifies.

The following are examples of some other filters:

- **A filter for records between the 25th and 75th percentiles**: =AND(E2>=PERCENTILE($E$2:$E$14,0.75),F2>=PERCENTILE($F$2:$F$14,0.75))
- **A filter that excludes records with missing systolic blood pressure readings**: =F2<>""
- **An alternative filter for the same purpose**: =ISBLANK(F2)=FALSE
- **A filter that skips non-numeric entries**: =ISERROR(F2)=FALSE



G17 =F2>AVERAGE($F$2:$F$14)

	A	B	C	D	E	F	G
1	Patient	Gender	DOB	Age	Weight	Systolic	
2	Johnson	F	6/13/1965	41	152	125	
3	Eisenhower	F	11/24/1934	72	154	128	
4	Truman	F	10/20/1962	44	151	139	
5	Ford	M	2/9/1950	57	164	141	
6	Nixon	F	9/15/1930	76	189	146	
7	Carter	F	10/22/1937	69	192	151	
8	Lincoln	M	8/1/1972	34	188	166	
9	Reagan	M	3/28/1939	68	140	170	
10	Clinton	M	7/15/1971	35	171	175	
11	Bush	M	1/3/1975	32	160	178	
12	Washington	F	10/15/1963	43	209	180	
13	Kennedy	M	11/27/1973	33	165	193	
14	Roosevelt	M	9/29/1945	61	159	196	
15							
16	Patient	Gender	DOB	Age	Weight	Systolic	>Mean SBP
17							FALSE
18							
19	Mean			44	170	180	
20	SD			15	22	11	
21	Count			7	7	7	
22							

Figure: 2.9

> **Note**: Non-numeric entries such as `NA` or `#N/A` would interfere with most calculations.

> **Note**: Filter formulas cannot use references outside the data range of the table. So you cannot compare the cell A2 of the first record with cell A1 (which is its label) or with cell A15 (which is outside the table you are filtering). In cases like these, you may have to expand your formula with functions such as `IF`.

* * *

Chapter 9

Unique Lists

Summary overviews often depend on listings of unique entries. Duplicate readings may have to be skipped. When you add or import records coming from an external data source, you may end up with duplicates that need to be eliminated in order to avoid miscalculations. In all these cases, you need lists of unique values or records. You can often create unique lists by removing duplicates.

Excel has a simple tool for removing duplicates. Say that you want to remove all records that have duplicate readings for the same plate. This is what you do:

1. On the Data tab, click Remove Duplicates.
2. Select the correct columns—in this case Plate ID and C-Value.

> **Note**: To remove only completely identical records, you must select all columns.

	A	B	C	D	E	F	G	H	I
1	Date	Plate ID	Analyst	C Value			etv	kpm	luv
2	09/14/01	8696p08b	etv	62.5		8696p08b	1	1	1
3	09/14/01	8696p08b	kpm	73.6		8877p63b	1	1	1
4	09/14/01	8696p08b	luv	63.3		8877p70d	2	1	1
5	09/26/01	8877p63b	etv	47.5		8877p78b	1	1	1
6	09/14/01	8877p63b	kpm	50.2		8877p84b	1	2	1
7	09/26/01	8877p63b	luv	45.4					
8	09/19/01	8877p70d	etv	79.4					
9	09/20/01	8877p70d	kpm	58.4					
10	09/21/01	8877p70d	luv	65.8					
11	09/18/01	8877p70d	etv	39.8					
12	09/21/01	8877p78b	kpm	60.8					
13	09/25/01	8877p78b	luv	78.4					
14	09/25/01	8877p78b	etv	64.9					
15	10/03/01	8877p84b	kpm	62.5					
16	09/12/01	8877p84b	luv	58.3					
17	09/11/01	8877p84b	etv	64.8					
18	09/11/01	8877p84b	kpm	70.2					
19									

Figure: 2.10

Figure 2.10 shows a case in which you want to create a summary table (to the right) based on unique entries in the records table (to the left). These are the steps:

1. Copy and paste the Plates column and the Analysts column. The original columns remain untouched but we will eliminate duplicates in the copied columns .
2. Make sure the Plates column and the Analysts column are surrounded by an empty column on either side, so Excel treats them as isolated tables.
3. On the Data tab, click Remove Duplicates.
4. To turn the Analysts column by 90 degrees, we could use the Transpose feature. Select the unique Analysts entries, use Copy, click where you want the transposed version, then use Paste Special and choose Transpose. Now you have a two-dimensional summary table with row entries and column entries.
5. Use the function `COUNTIFS` in G2: `=COUNTIFS($B$2:$B$18,$F2,$C$2:$C$18,G$1)`.

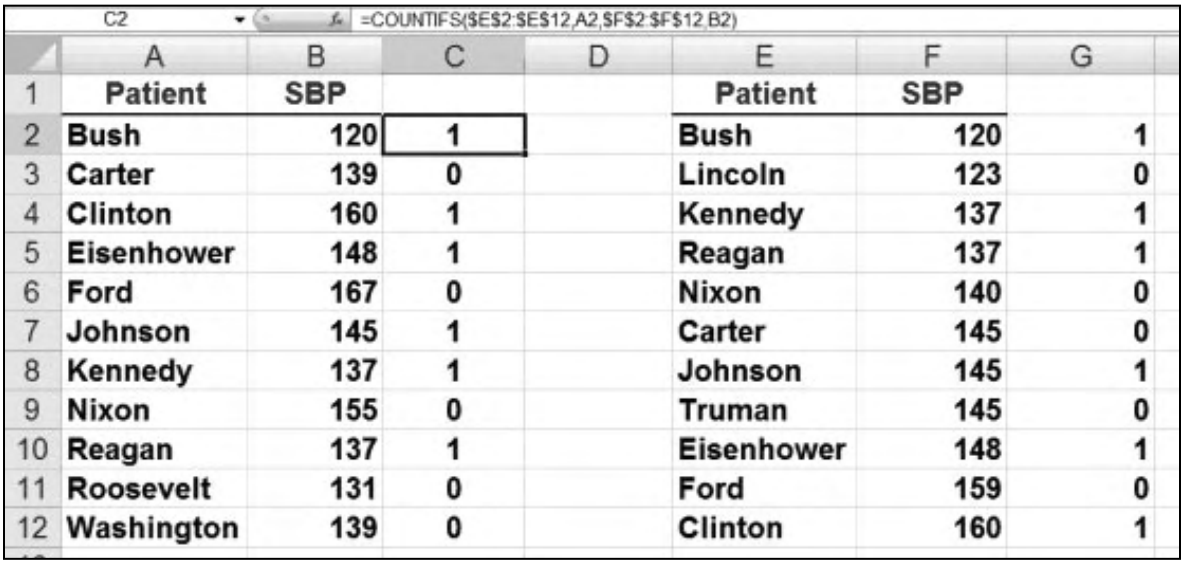
Figure 2.11 shows another situation. Sometimes you receive lists with updated records, and you need to compare what is missing in the old list (to the left) and what is gone in the new list (to the right). Here's what you do:

1. Add the following formula in C2: `=COUNTIF($E$2:$E$12,A2)`. The zeros represent missing records.
2. Add the following formula in G2: `=COUNTIF($A$2:$A$12,E2)`.
3. If needed, sort by 0 or 1 and delete what should be deleted.

	A	B	C	D	E	F	G
1	Patient	SBP			Patient	SBP	
2	Bush	120	1		Bush	120	1
3	Carter	139	1		Carter	139	1
4	Clinton	160	1		Clinton	160	1
5	Eisenhower	148	1		Eisenhower	148	1
6	Ford	167	1		Ford	167	1
7	Johnson	145	1		Johnson	145	1
8	Kennedy	137	1		Kennedy	137	1
9	Lincoln	123	0		Nixon	155	1
10	Nixon	155	1		Reagan	137	1
11	Reagan	137	1		Roosevelt	131	0
12	Truman	145	0		Washington	139	0
13							

Figure: 2.11

Figure 2.12 shows a case that's a bit more complicated. To determine whether any part of the record has changed, you can use function `COUNTIFS`: `=COUNTIFS($E$2:$E$12,A2,$F$2:$F$12,B2)`. The function returns `1` for no change and `0` if there are changes. As you can imagine, this function is especially useful for huge listings.



	A	B	C	D	E	F	G
1	Patient	SBP			Patient	SBP	
2	Bush	120	1		Bush	120	1
3	Carter	139	0		Lincoln	123	0
4	Clinton	160	1		Kennedy	137	1
5	Eisenhower	148	1		Reagan	137	1
6	Ford	167	0		Nixon	140	0
7	Johnson	145	1		Carter	145	0
8	Kennedy	137	1		Johnson	145	1
9	Nixon	155	0		Truman	145	0
10	Reagan	137	1		Eisenhower	148	1
11	Roosevelt	131	0		Ford	159	0
12	Washington	139	0		Clinton	160	1

Figure: 2.12

* * *

Chapter 10

Data Validation

Good data analysis depends on reliable records. Of course, no one can prevent inaccurate data entry, but you do have the power to subject data entry to some kind of validation check. A table without data validation may contain very unreliable information. With Excel, you can set up your own rules for checking data entry. For instance, if dates are important to you, you want to make sure no one can enter a date in the past or somewhere in the future. This is where data validation comes in. You can apply validation by using a button on the Data tab. This activates a dialog box with three tabs:

- The first tab is for what you want to allow. The default option is Any Value. The most flexible option is Custom because you can use it to implement your own formulas.
- The second tab creates a message that kicks in whenever the user enters a validated cell; this setting is usually annoying.
- The third tab implements a specific error alert

Caution: Validation settings can be deleted by the Copy and Paste operations; a regular paste also replaces the cell's validation settings with the validation settings of the copied cell. Therefore, invalid entries that were pasted may elude detection. Data validation applied later on to a column that already has invalid entries does not detect those violations until you click Circle Invalid Data (under the Validation dropdown button).

You can apply data validation in two different ways:

- **Selecting the entire column**: This is a great option when you keep adding records at the bottom of the list, but it takes a bit more memory in your file.
- **Selecting a specific range of data**: If you ever need to find out later where your validation range ended, you can use either of the following routes:
  - From the Home tab, select Find & Select, Data Validation
  - From the Home tab, select Find & Select, GoTo Special, Data Validation. You can then choose between highlighting all ranges and highlighting just the one around your selected cell(s).

Again, you can use your own formulas; for example, you can use the formula `=LEN(B2)=3` when you want a text length of exactly three characters. Formulas always deal with the first selected cell (B2, in this case) but adjust to all selected cells (so B2 should be relative). Unfortunately, you can only type the formula; other help, such as fx, is not available. The idea behind the validation formula is that if the formula evaluates to `TRUE`, the entry is valid.

Figure 2.13 shows many candidates for data validation:

- In column A, you accept only dates between a certain date in the past and the current date. The settings are Allow Date and Between any start date and `=TODAY()`.
- In column B, you want to make sure that each analyst has a three-character designation. You could use the option Text or choose Custom and enter the formula `=LEN(B2)=3`.
- In column C, you want to prevent duplicate plate numbers, so each plate number can only be used once. The custom setting would be `=COUNTIF($C$2:$C$21,C2)=1`.
- In column D, say that you've never found readings outside the range 40–60. You can use the following formula to trap any values outside the range 40–60: `=AND(D2>40,D2<60)`.
- In column E, you want to validate for five characters and no duplicates. This should work: `=AND(LEN(E2)=5,COUNTIF($E$2:$E$21,E2)=1)`.

Some of these settings could be achieved in a simpler way, but the Custom option is more flexible—and often your only choice.

	A	B	C	D	E
1	Date	Analyst	Plate ID	50 ng/mL	Plate ID
2	1/3/2006	gmv	2321a	47.7	2321a
3	1/4/2006	gmv	2321b	48.7	2321b
4	1/5/2006	tjk	2322a	49.3	2322a
5	1/6/2006	tjk	2322b	45.9	2322b
6	1/9/2006	tkm	2323a	47.5	2323a
7	1/10/2006	tkm	2323b	45.9	2323b
8	1/11/2006	bdo	2324a	43.2	2324a
9	1/12/2006	bdo	2324b	44.3	2324b
10	1/13/2006	ksm	2325a	56.4	2325a
11	1/16/2006	ksm	2325b	51.3	2325b

Figure: 2.13

	A	B	C
1	Date	Patient	New SBP
2	4/16/2007	Bush	127
3	4/13/	Bush	152
4	4/12/	Carter	165
5	4/11/	Clinton	180
6	4/10/	Eisenhower	158
7	4/9/	Ford	160
8	4/6/2007	Johnson	142
9	4/5/2007	Kennedy	190
10	4/4/2007	Nixon	125
11	4/3/2007	Reagan	141
12	4/2/2007	Bush	191
13	3/30/2007	Carter	155
14	3/29/2007	Clinton	160
15	3/28/2007	Eisenhower	141
16	3/27/2007	Ford	132
17	3/26/2007	Johnson	139
18	3/23/2007	Kennedy	147
19	3/22/2007	Nixon	151
		Reagan	

Figure: 2.14

Figure 2.14 may not look like a validation issue, but it actually is. In column B, the user can only choose from a list of participants that is located somewhere else in this workbook; its range has already been named Patients. Using this list as a validation tool prevents typos during data entry and thus guarantees more reliable data analysis later on. To get to this point, you do the following in column B:

1. Select either B2:B19 or the entire column.
2. Start data validation.
3. Set Allow to `List`.
4. Set Source to `=Patients`. and click OK.

Now when you enter each cell, you see a button, and the column rejects any entries that are invalid according to this list.

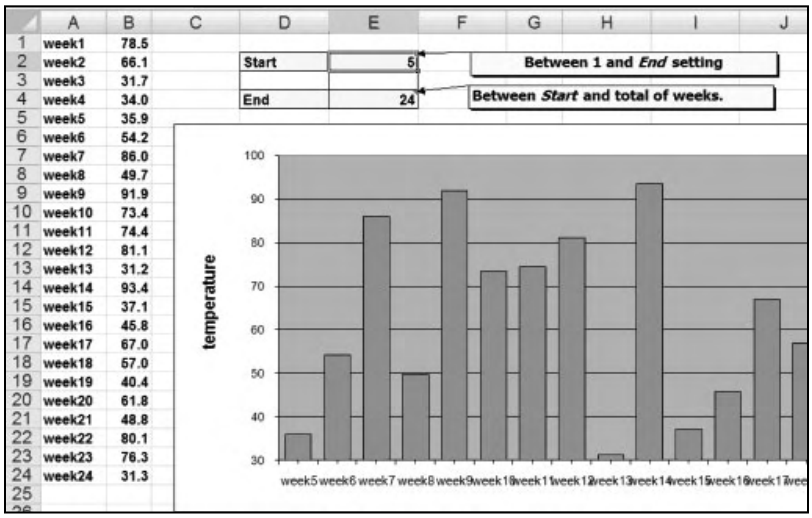


Figure: 2.15

Figure 2.15 shows another validation example applied to temperature readings per week. Later, we'll discuss other features of this sheet—for instance, how the graph adjusts automatically to new settings. For now, take a look at cells E2 and E4, which regulate where to start the graph and where to end the plot:

- The validation setting of cell E2 is rather obvious: It is a whole number between `1` and `=E4`.
- The setting for cell E4 could be a whole number between `=E2` and `=COUNTA(A:A)`.

* * *

Chapter 11

Conditional Formatting

Among the numerous details in a spreadsheet, you may want to flag, mark, or highlight specific values that should stick out for some reason. Conditional formatting allows you to do this according to particular rules you set up yourself. You are the master of your sheets.

Figure 2.16 shows some examples of conditional formatting that you can practically do automatically, thanks to Excel's new tools:

- **Column D**: You can "grade" values by simply selecting Home, Conditional Formatting, Data Bars or Color Scales.
- **Column E**: You can flag all values that are above (or below) the mean by selecting Home, Conditional Formatting, Top/Bottom Rules, Above Average.
- **Column F**: You can flag even maximum and minimum values by using the top or bottom one item.
- **Column B**: You flag duplicate (or unique) values by selecting Highlight Cells Rules, Duplicate Values.

A1 | Plate ID

	A	B	C	D	E	F
1	Plate ID	Date	Analyst	Max OD	C-Value	SD
2	8696p08d	09/14/01	cod	2.3	59.9	7.9
3	8877p63a	09/14/01	cod	2.8	45.0	2.2
4	8877p120	10/03/01	ejs	1.5	62.5	7.4
5	8877p66b	09/26/01	ejs	2.6	47.5	7.1
6	8877p78b	09/21/01	ejs	2.8	60.8	7.0
7	8696p08b	09/14/01	etv	2.4	62.5	7.6
8	8877p60a	09/12/01	etv	3.2	58.3	0.9
9	8877p71b	09/19/01	etv	2.5	79.4	4.4
10	8877p84b	09/25/01	etv	1.9	78.4	0.3
11	8877p70c	09/18/01	gmv	1.4	57.1	1.3
12	8877p83a	09/25/01	gmv	2.0	83.7	2.0
13	8696p08e	09/14/01	kpm	2.5	73.6	7.3
14	8877p58a	09/11/01	kpm	2.6	64.8	4.0
15	8877p63b	09/14/01	kpm	2.7	50.2	7.5
16	8877p75b	09/20/01	kpm	1.7	58.4	8.3
17	8696p08f	09/14/01	luv	2.4	63.3	4.7
18	8877p66a	09/26/01	luv	2.8	45.4	0.2
19	8877p78a	09/21/01	luv	2.6	65.8	8.2
20	8877p58b	09/11/01	tjk	2.2	70.2	8.6
21	8877p70d	09/18/01	tjk	1.1	39.8	7.0
22	8877p83b	09/25/01	tjk	1.7	64.9	1.3

Figure: 2.16

Excel's options are extremely rich, but you may sometimes need your own formulas, especially for scientific work. With formulas, the sky is the limit. The idea behind formulas should be familiar by now: If a formula evaluates to `TRUE`, the conditional format is applied. The steps are simple:

1. From the Home tab, select Conditional Formatting.
2. Select New Rule (which is almost at the bottom).
3. Choose the last option: Use a formula.
4. Type the formula, making sure to correct the default absolute cell reference if necessary.
5. Determine the format you would like.

Note: If you discover a mistake later on, do not go to New Rule (that would add another rule); instead, go to Manage Rules (where you can add, edit, and delete rules).

Figure 2.17 offers some interesting examples:

- To mark the C-values between the 25th and 75th percentiles, the formula is quite long: `=AND(D2>PERCENTILE($D$2:$D$22,0.25),D2<PERCENTILE($D$2:$D$22,0.75))`.
- To flag the analysts in column C who have more than three readings, you use

	A	B	C	D	E	F
1	Plate ID	Date	Analyst	C-Value	Max OD	SD
2	8877p70d	09/18/01	tjk	39.8	1.1	7.0
3	8877p63a	09/14/01	cod	45.0	2.8	2.2
4	8877p66a	09/26/01	luv	45.4	2.8	0.2
5	8877p66b	09/26/01	ejs	47.5	2.6	7.1
6	8877p63b	09/14/01	kpm	50.2	2.7	7.5
7	8877p70c	09/18/01	gmv	57.1	1.4	1.3
8	8877p60a	09/12/01	etv	58.3	3.2	0.9
9	8877p75b	09/20/01	kpm	58.4	1.7	8.3
10	8696p08d	09/14/01	cod	59.9	2.3	7.9
11	8877p78b	09/21/01	ejs	60.8	2.8	7.0
12	8877p120(	10/03/01	ejs	62.5	1.5	7.4
13	8696p08b	09/14/01	etv	62.5	2.4	7.6
14	8696p08f	09/14/01	luv	63.3	2.4	4.7
15	8877p58a	09/11/01	kpm	64.8	2.6	4.0
16	8877p83b	09/25/01	tjk	64.9	1.7	1.3
17	8877p78b	09/21/01	luv	65.8	2.6	8.2
18	8877p58b	09/11/01	tjk	70.2	2.2	8.6
19	8696p08e	09/14/01	kpm	73.6	2.5	7.3
20	8877p84b	09/25/01	etv	78.4	1.9	0.3
21	8877p71b	09/19/01	etv	79.4	2.5	4.4
22	8877p63a	09/25/01	gmv	83.7	2.0	2.0

Figure: 2.17

`=COUNTIF($C$2:$C$22,C2)>3`. Don't forget to set `COUNTIF` to `>3`.

- When you try to mark values in columns D and E that are above the 50th percentile in the range, you may encounter a surprise: `=AND(D2>PERCENTILE($D$2:$D$22,0.5), E2>PERCENTILE($E$2:$E$22,0.5))` formats only the left column. To correct, this you use `=AND($D2>PERCENTILE($D$2:$D$22,0.5), $E2>PERCENTILE($E$2:$E$22,0.5))`. See the difference? Cells D2 and E2 need an absolute column reference!

Figure 2.18 shows a similar problem: The left table (A:E) simply marks column C where a new analyst appears in column C in the right table. But the right table (G:K) needs an adjusted formula to create borders for the entire record. This is how you do it:

- For C2:C20 (do not start in C1!), use `=C2<>C1`.
- For G2:K20 (again, not G1!), use `=$C2<>$C1`.

	A	B	C	D	E	F	G	H	I	J	K
1	Plate ID	Date	Analyst	C Value	Max OD		Plate ID	Date	Analyst	C Value	Max OD
2	8696p08d	09/14/01	cod	59.9	2.3		8696p08d	09/14/01	cod	59.9	2.3
3	8877p63a	09/14/01	cod	45.0	2.8		8877p63a	09/14/01	cod	45.0	2.8
4	8877p120e	10/03/01	ejs	62.5	1.5		8877p120e	10/03/01	ejs	62.5	1.5
5	8877p66b	09/26/01	ejs	47.5	2.6		8877p66b	09/26/01	ejs	47.5	2.6
6	8696p08b	09/14/01	etv	62.5	2.4		8696p08b	09/14/01	etv	62.5	2.4
7	8877p60a	09/12/01	etv	58.3	3.2		8877p60a	09/12/01	etv	58.3	3.2
8	8877p84b	09/25/01	etv	78.4	1.9		8877p84b	09/25/01	etv	78.4	1.9
9	8877p70c	09/18/01	gmv	57.1	1.4		8877p70c	09/18/01	gmv	57.1	1.4
10	8877p83a	09/25/01	gmv	83.7	2.0		8877p83a	09/25/01	gmv	83.7	2.0
11	8696p08e	09/14/01	kpm	73.6	2.5		8696p08e	09/14/01	kpm	73.6	2.5
12	8877p58a	09/11/01	kpm	64.8	2.6		8877p58a	09/11/01	kpm	64.8	2.6
13	8877p63b	09/14/01	kpm	50.2	2.7		8877p63b	09/14/01	kpm	50.2	2.7
14	8877p75b	09/20/01	kpm	58.4	1.7		8877p75b	09/20/01	kpm	58.4	1.7
15	8696p08f	09/14/01	luv	63.3	2.4		8696p08f	09/14/01	luv	63.3	2.4
16	8877p66a	09/26/01	luv	45.4	2.8		8877p66a	09/26/01	luv	45.4	2.8
17	8877p78a	09/21/01	luv	65.8	2.6		8877p78a	09/21/01	luv	65.8	2.6
18	8877p58b	09/11/01	tjk	70.2	2.2		8877p58b	09/11/01	tjk	70.2	2.2
19	8877p70d	09/18/01	tjk	39.8	1.1		8877p70d	09/18/01	tjk	39.8	1.1
20	8877p83b	09/25/01	tjk	64.9	1.7		8877p83b	09/25/01	tjk	64.9	1.7

Figure: 2.18

Figure 2.19 may not look like it has a conditional formatting issue, but it does. You could reach this striping effect by using Excel's table structure, but you might not always want to do so. Instead, you can use conditional formatting based on a formula with the `MOD` function, as discussed in Chapter 6: `=MOD(ROW(A1),2)=1`. An advantage of using this formula is that you can make the striping alternate at any step, such as at every fifth row. This is something a table structure cannot achieve.

	A	B	C	D	E
1	Plate ID	Date	Analyst	C Value	Max OD
2	8696p08d	09/14/01	cod	59.9	2.3
3	8877p63a	09/14/01	cod	45.0	2.8
4	8877p120e	10/03/01	ejs	62.5	1.5
5	8877p66b	09/26/01	ejs	47.5	2.6
6	8696p08b	09/14/01	etv	62.5	2.4
7	8877p60a	09/12/01	etv	58.3	3.2
8	8877p84b	09/25/01	etv	78.4	1.9
9	8877p70c	09/18/01	gmv	57.1	1.4
10	8877p83a	09/25/01	gmv	83.7	2.0
11	8696p08e	09/14/01	kpm	73.6	2.5
12	8877p58a	09/11/01	kpm	64.8	2.6
13	8877p63b	09/14/01	kpm	50.2	2.7
14	8877p75b	09/20/01	kpm	58.4	1.7
15	8696p08f	09/14/01	luv	63.3	2.4
16	8877p66a	09/26/01	luv	45.4	2.8
17	8877p78a	09/21/01	luv	65.8	2.6
18	8877p58b	09/11/01	tjk	70.2	2.2
19	8877p70d	09/18/01	tjk	39.8	1.1
20	8877p83b	09/25/01	tjk	64.9	1.7
21					

Figure: 2.19

G2 fx 20

	A	B	C	D	E	F	G	H	I	J
1	55	316	223	185	124		Bins		25	3
2	124	93	163	213	314		20		50	2
3	211	41	231	241	212				75	2
4	118	113	400	205	254		5		100	7
5	262	1	201	172	101		10		125	10
6	167	479	205	337	118		15		150	6
7	489	15	89	362	148		20		175	12
8	179	248	125	197	177		25		200	7
9	456	153	269	49	127		30		225	10
10	289	500	198	317	300				250	8
11	126	114	303	314	270				275	5
12	151	279	347	314	170				300	4
13	250	175	93	209	61				325	7
14	166	113	356	124	242				350	5
15	152	384	157	233	99				375	2
16	277	195	436	6	240				400	4
17	147	80	173	211	244				425	0
18	386	93	330	400	141				450	2
19	332	173	129	323	188				475	1
20	338	263	444	84	220				500	3
21									524	0
22									549	0

Figure: 2.20

In Figure 2.20, it would be nice if the number of bins in columns I and J would be highlighted according to the number of bins chosen in cell G2. Here is the simple formula for making this happen: `=ROW()<=$G$2`.

* * *

Chapter 12

Filtering Tools

In the chapters to this point, you have marked a subset of records inside the total collection, so the entire set of records remains visible and may thus obscure the specific records you want to focus on. Filtering tools allow you to display just the filtered subset on its own, so you can study and analyze the subset without being distracted by the immense surroundings. These tools help you combat the famous forest and trees problem.

Say that in Figure 2.21, you forgot to validate the systolic blood pressure values as being between 100 and 200 mmHg. You can correct this by applying an advanced filter on the Data tab, similar to the one you used earlier for summary or database functions. This filter creates a subset of invalid records so they can be corrected or deleted. These are the steps:

	A	B	C	D	E	F
1	Patient	Gender	DOB	Age	Weight	Systolic
2	Bush	M	1/3/1975	32	160	205
3	Carter	F	10/22/1937	69	192	151
4	Clinton	M	7/15/1971	35	171	175
5	Eisenhower	F	11/24/1934	72	154	128
6	Ford	M	2/9/1950	57	164	141
7	Johnson	F	6/13/1965	41	152	94
8	Kennedy	M	11/27/1973	33	165	193
9	Lincoln	M	8/1/1972	34	188	166
10	Nixon	F	9/15/1930	76	189	146
11	Reagan	M	3/28/1939	67	140	170
12	Roosevelt	M	9/29/1945	61	159	211
13	Truman	F	10/20/1962	44	151	139
14	Washington	F	10/15/1963	43	209	180
15						
16						
17	Patient	Gender	DOB	Age	Weight	Systolic
18						<100
19						>200
20						

Figure: 2.21

1. Create the correct settings in the filter for violations; `OR` conditions should be in separate rows (F18 and F19).
2. Click inside the table first (so Excel can automatically detect the table range).
3. On the Data tab, click the Advanced button. The List Range is already displayed, thanks to step 2.
4. Select the criteria range A17:F19.
5. Click OK and watch the subset of "violators":
   - If you want to correct the filtered violators, do so.
   - If you want to delete the filtered violators, select all their rows and then select the Home tab, Find& Select, Go To Special, Visible Cells Only; then right-click rows and select Delete Rows.
   - If you want all records back, click the Clear button in the Sort & Filter group of the Data tab.

As with summary filters, you can use computed criteria here as well. If the formula evaluates to `TRUE`, the record qualifies for the subset. Here are some examples of possible filters:

- **For blood pressures above the mean in column F, starting in F2**:
  `=F2>AVERAGE($F$2:$F$14)`
- **For records that have no weight or systolic blood pressure values**:
  `=OR(E2<>"",F2<>"")`

You can also place filters on a sheet separate from the database itself in order to gather several filtered subsets on separate sheets. If you regularly use certain filters, you can give each one its own sheet. Be aware, though, that unlike DB functions, filters do not automatically update. You must apply a filter again.

When using filters on a separate sheet, you must perform the following actions:

1. Click inside the filter first (not in the database).
2. Start the advanced filter.
3. Highlight the database or type its name.
4. Select Copy to Another Location.
5. Indicate the location of the copied subset.

Figure 2.22 shows a database on its own sheet; it has been named Data. (Do not name a database `Database` because Excel uses that name internally for its filters.) On separate sheets, you could use the following filters, among many others:

- `=Separate!D2>AVERAGE(Separate!$D$2:$D$22)`

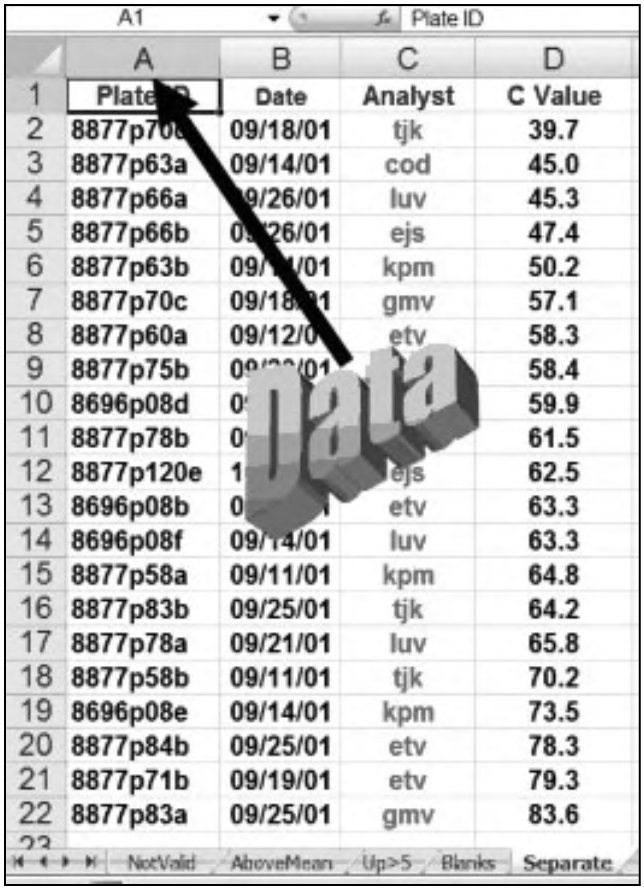


Figure: 2.22

- `=AND(Separate!D2>PERCENTILE(Separate!$D$2:$D$22,0.25), Separate!D2<PERCENTILE(Separate!$D$2:$D$22,0.75))`
- `=Separate!D2=MEDIAN(Separate!$D$2:$D$22)`
- `=Separate!D2=MODE(Separate!$D$2:$D$22)`
- `=Separate!D2=AVERAGE(Separate!$D$2:$D$22)` (but this record will probably never be found unless there is a reading that happens to be equal to the mean)

After working with all these examples, you should have a better understanding of how to use formulas for filtering, validation, and conditional formatting.

* * *

Chapter 13

LOOKUPS

Because data analysis depends on using reliable data, it also depends on the use of lookup functions, which allow you to quickly and efficiently locate specific data in another list and to automatically ensure that you are using the correct and latest information.

Pretend your analysts are assigned to groups and you want to analyze how each group is performing. Your best solution is to look up the information in a lookup table. First, the use of a lookup function guarantees correct information. Second, changes in the lookup table will immediately cascade through your records. When you look up information in a vertical way, you can use the `VLOOKUP` function; if the table is horizontally structured, you use `HLOOKUP` instead. `VLOOKUP` has the following syntax: `=VLOOKUP(lookup-value, table, column#, T/F)`.

`VLOOKUP`'s last argument (`T/F`) determines whether your `lookup-value` has an exact match in the lookup table (`0` or `FALSE`), or not (`1` or `TRUE`). If you go for an exact match, the first column of the lookup table doesn't have to be sorted; `VLOOKUP` can find an answer anyway. If you cannot guarantee an exact match, the first column must be sorted in ascending order because `VLOOKUP` may have to go for the previous closest match in ascending order.

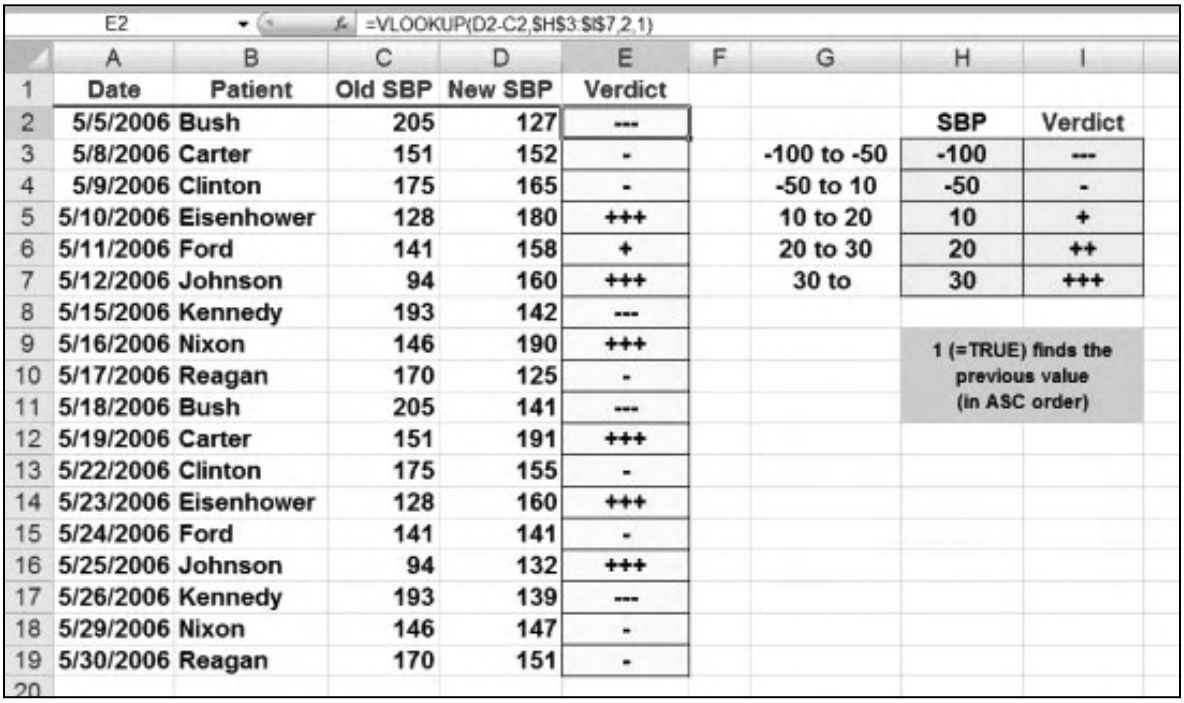
A1 Plate ID

	A	B	C	D	E	F	G	H	I
1	Plate ID	Date	Analyst	Group	C Value		Analyst	Group	
2	8877p58a	09/11/01	kpm	102	64.8		cod	102	
3	8696p08d	09/14/01	cod	102	59.9		kpm	102	
4	8696p08e	09/14/01	kpm	102	73.6		luv	102	
5	8696p08f	09/14/01	luv	102	63.3		tmv	107	
6	8877p63a	09/14/01	cod	102	45.0		etv	107	
7	8877p63b	09/14/01	kpm	102	50.2		wow	107	
8	8877p75b	09/20/01	kpm	102	58.4		ejs	115	
9	8877p78a	09/21/01	luv	102	65.8		gmv	115	
10	8877p58c	09/11/01	tmv	107	60.7		tjk	115	
11	8877p60a	09/12/01	etv	107	58.3				
12	8877p60b	09/12/01	wow	107	73.6				
13	8696p08a	09/14/01	tmv	107	72.4				
14	8696p08b	09/14/01	etv	107	62.5				
15	8696p08c	09/14/01	wow	107	56.7				
16	8877p71a	09/19/01	tmv	107	66.9				
17	8877p71b	09/19/01	etv	107	79.4				
18	8877p75a	09/20/01	wow	107	61.8				
19	8877p84a	09/25/01	tmv	107	63.0				
20	8877p58b	09/11/01	tjk	115	70.2				
21	8877p70c	09/18/01	gmv	115	57.1				
22	8877p70d	09/18/01	tjk	115	39.8				
24	8877p78b	09/21/01	ejs	115	60.8				
25	8877p83a	09/25/01	gmv	115	83.7				
26	8877p83b	09/25/01	tjk	115	64.9				
27									

Figure: 2.23

Figure 2.23 shows an example of a simple lookup situation: Column D finds its answer in the lookup table to the right (or wherever). In cell D2, you look up information for analyst kpm (C2) in a vertical lookup table (G2:H10), and you find an answer in column 2 of the table: `=VLOOKUP(C2,$G$2:$H$10,2,0)`. You can now perform analysis operations such as creating subtotals per group. If someone were assigned to the wrong group, updates in the lookup table would automatically permeate the records as well.

Figure 2.24 presents a different situation. In trying to find a verdict for column E in the lookup table, you may not always be able to find an exact match in the lookup table. Fortunately, `VLOOKUP` can also search for the closest previous value in an ascending list. So you have to structure the lookup table in such a way that the first bin can handle the lowest value; the last bin with the highest value needs to handle everything exceeding that value. Now you should be able to determine whether the blood pressure–lowering pill made the systolic blood pressure go up or down and by how much: `=VLOOKUP(D2-C2,$H$3:$I$7,2,1)`. The last argument is set to `1` or `TRUE`.



E2 =VLOOKUP(D2-C2,$H$3:$I$7,2,1)

	A	B	C	D	E	F	G	H	I
1	Date	Patient	Old SBP	New SBP	Verdict				
2	5/5/2006	Bush	205	127	---			SBP	Verdict
3	5/8/2006	Carter	151	152	-		-100 to -50	-100	---
4	5/9/2006	Clinton	175	165	-		-50 to 10	-50	-
5	5/10/2006	Eisenhower	128	180	+++		10 to 20	10	+
6	5/11/2006	Ford	141	158	+		20 to 30	20	++
7	5/12/2006	Johnson	94	160	+++		30 to	30	+++
8	5/15/2006	Kennedy	193	142	---				
9	5/16/2006	Nixon	146	190	+++			1 (=TRUE) finds the previous value (in ASC order)	
10	5/17/2006	Reagan	170	125	-				
11	5/18/2006	Bush	205	141	---				
12	5/19/2006	Carter	151	191	+++				
13	5/22/2006	Clinton	175	155	-				
14	5/23/2006	Eisenhower	128	160	+++				
15	5/24/2006	Ford	141	141	-				
16	5/25/2006	Johnson	94	132	+++				
17	5/26/2006	Kennedy	193	139	---				
18	5/29/2006	Nixon	146	147	-				
19	5/30/2006	Reagan	170	151	-				
20									

Figure: 2.24

> **Note**: If you ever decide to make your categories stricter or more lenient, you just adjust the lookup table.

Figure 2.25 shows a slightly different case. You have measured the forced vital capacity (FVC; a pulmonary function) in relationship to gender and body length, and you would like to compare each individual with known reference values, as found in a reference table to the right. This time, you need to distinguish readings by gender, so you must look in the proper column by using an IF function nested inside a `VLOOKUP` function: `=VLOOKUP(A3,$H$2:$J$18,IF(B3="F",2,3),1)`.

So far, you have specified a hard-coded column number for `VLOOKUP` to find the proper answer. However, there is also a function that can find column and row numbers on its own: `MATCH`. The `MATCH` function finds the relative position of a row or column in a range. Thanks to `MATCH`,

D3 =VLOOKUP(A3,$H$2:$J$18,IF(B3="F",2,3),1)

	A	B	C	D	E	F	G	H	I	J
1	Pulmonary Function Forced Vital Capacity (FVC)							Length/m	F	M
2	Length/m	Gender	FVC/L	Reference	Ratio			1.50	2.3	
3	1.77	F	3.9	4.9	80%			1.52	2.5	
4	1.80	M	3.7	5.4	69%			1.54	2.7	
5	1.80	M	5.0	5.4	93%			1.56	2.9	3.0
6	1.71	F	2.8	4.3	65%			1.58	3.1	3.2
7	1.56	F	3.5	2.9	121%			1.60	3.3	3.4
8	1.58	F	3.7	3.1	119%			1.62	3.5	3.6
9	1.71	F	2.3	4.3	53%			1.64	3.7	3.8
10	1.80	M	3.4	5.4	63%			1.66	3.9	4.0
11	1.75	F	3.1	4.7	66%			1.68	4.1	4.2
12	1.75	F	2.5	4.7	53%			1.70	4.3	4.4
13	1.55	F	3.3	2.7	122%			1.72	4.5	4.6
14	1.73	M	2.8	4.6	61%			1.74	4.7	4.8
15	1.58	M	2.0	3.2	63%			1.76	4.9	5.0
16	1.78	M	2.4	5.2	46%			1.78		5.2
17	1.65	M	2.5	3.8	66%			1.80		5.4
18	1.62	M	2.9	3.6	81%			1.82		5.6
19	1.78	M	1.6	5.2	31%					
20	1.76	M	3.9	5.0	78%					
21	1.71	M	3.6	4.4	82%					
22	1.54	F	2.2	2.7	81%					
23										

Figure: 2.25

columns can still be located, even after they are moved around. Its syntax is `=MATCH(lookup-value, table, -1/0/1)`. Figure 2.26 shows how the last argument in this syntax works:

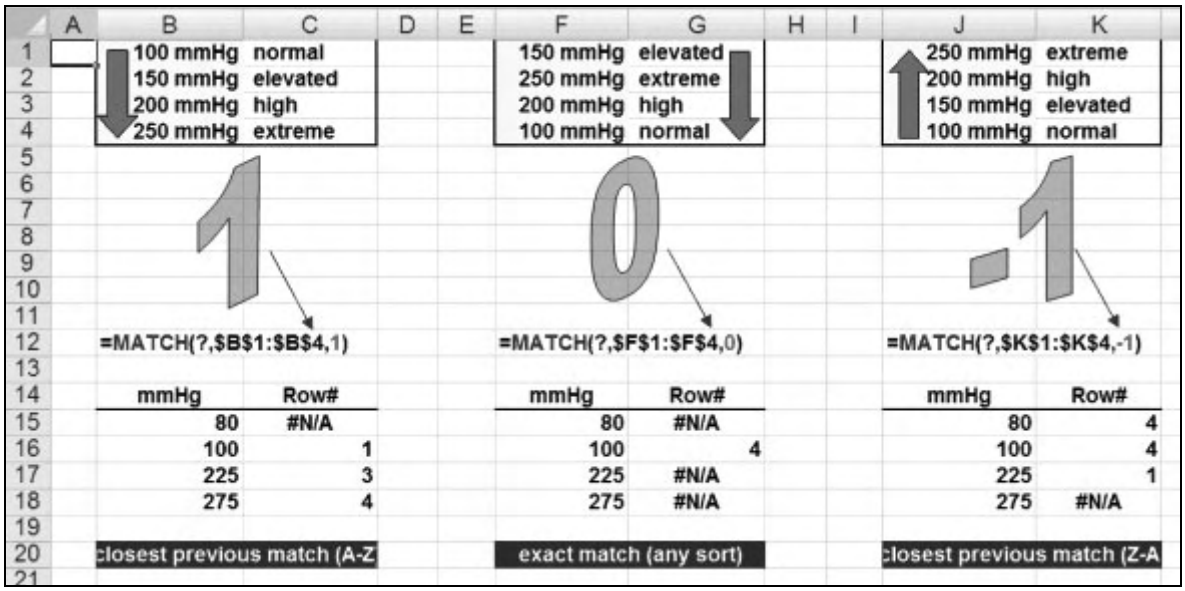


Figure: 2.26

- **1**: Finds the closest previous match in ascending, A–Z, order. (If there is none, the result is `#N/A`.) In Figure 2.26, `MATCH` has found every row number except for 80 because there is no previous match before 100.
- **0**: Finds an exact match. In Figure 2.26, this is possible only for 100 mmHg. All other cases are not found: #N/A.
- `-1`: Finds the closest previous match in a descending, Z–A, order. In Figure 2.26, `MATCH` has found every row number except for 275 mmHg because there is no previous match before 250 (therefore, `#N/A`).

In Figure 2.27, you use a drop-down box in cell I1 to select a patient's name (see Chapter 10). Based on that selected name, you can locate all other information in the main table by using `MATCH` nested inside `VLOOKUP`:

- The function `MATCH` would find the column number of the Gender column in the main table's headers (A1:F1): `MATCH(H2,$A$1:$F$1,0)`.
- The formula in I2 would be `=VLOOKUP($I$1,$A$2:$F$14,MATCH(H2,$A$1:$F$1,0),0)`.

I1 Reagan

	A	B	C	D	E	F	G	H	I
1	Patient	Gender	DOB	Age	Weight	Systolic		Patient	Reagan
2	Bush	M	1/3/1975	32	160	205		Gender	M
3	Carter	F	10/22/1937	69	192	151		DOB	3/28/1939
4	Clinton	M	7/15/1971	35	171	175		Age	67
5	Eisenhower	F	11/24/1934	72	154	128		Weight	140
6	Ford	M	2/9/1950	57	164	141		Systolic	170
7	Johnson	F	6/13/1965	41	152	94			
8	Kennedy	M	11/27/1973	33	165	193			
9	Lincoln	M	8/1/1972	34	188	166			
10	Nixon	F	9/15/1930	76	189	146			
11	Reagan	M	3/28/1939	67	140	170			
12	Roosevelt	M	9/29/1945	61	159	211			
13	Truman	F	10/20/1962	44	151	139			
14	Washington	F	10/15/1963	43	209	180			

Figure: 2.27

> **Note:** An advantage of not using a hard-coded column in VLOOKUP's third argument is that you could change the order of columns in the table without affecting the outcome of the `VLOOKUP` function. But you cannot change the column's label!

`VLOOKUP` works great, but it has two big limitations:

- It can only search in the first column of a table and find corresponding information in the next columns. So `VLOOKUP` cannot find corresponding information in previous columns. Of course, you could move columns around.

- It accepts column numbers (which can be done with `MATCH`), but it does not acknowledge row numbers. In other words, you could never offset a row number by a certain amount.

Fortunately, Excel has a function without these limitations: `INDEX`. Its syntax is much more flexible than that of VLOOKUP: `=INDEX(table, row#, col#)`. `INDEX` needs to know the row position and the column position, which is a nice challenge for the `MATCH` function. Then it can search in any row and in any column to find a value at their intersection. When calling the function `INDEX`, you find out that it has two versions: One version returns the value at a certain intersection (that's the one used here); the other one returns the reference of the cell at that intersection (the address of the intersection, not its value). You normally use the first version.

Figure 2.28 shows an example of `INDEX`. `VLOOKUP` could never find the ID in cell J1 because it is located in a column before the lookup value.

1. Use the INDEX function in cell J1 to find the ID information for a specific Patient: =INDEX($A$2:$G$14,MATCH($J$2,$B$2:$B$14,0),MATCH(I1,$A$1:$G$1,0)).
2. The first MATCH finds row 11 for Roosevelt.
3. The second MATCH finds column 1 for the ID field.

	A	B	C	D	E	F	G	H	I	J
1	ID	Patient	Gender	DOB	Age	Weight	Systolic		ID	011
2	001	Bush	M	1/3/1975	32	160	205		Patient	Roosevelt
3	002	Carter	F	10/22/1937	69	192	151		Gender	M
4	003	Clinton	M	7/15/1971	35	171	175		DOB	29/9/45
5	004	Eisenhower	F	11/24/1934	72	154	128		Age	61
6	005	Ford	M	2/9/1950	57	164	141		Weight	159
7	006	Johnson	F	6/13/1965	41	152	94		Systolic	211
8	007	Kennedy	M	11/27/1973	33	165	193			
9	008	Lincoln	M	8/1/1972	34	188	166			
10	009	Nixon	F	9/15/1930	76	189	146			
11	010	Reagan	M	3/28/1939	67	140	170			
12	011	Roosevelt	M	9/29/1945	61	159	211			
13	012	Truman	F	10/20/1962	44	151	139			
14	013	Washington	F	10/15/1963	43	209	180			

Figure: 2.28

> **Note**: All `MATCH` positions are relative, so row 12 on the sheet is row 11 within range A2:G14.

Another function, `INDIRECT`, can make your life easier here. As discussed briefly in Chapter 4, INDIRECT returns the value of a cell referenced by its name. Here are some examples:

- `INDIRECT("A1")` returns Plate ID (or whatever is in the cell named A1).
- `INDIRECT("E4")` returns 60.7 (or whatever is in the cell named E4).
- `INDIRECT(Total)` returns whatever is in the cell named Total.

	A	B	C	D	E	F	G	H
1	Plate ID	Date	Analyst	Group	C Value		Analyst	Group
2	8877p58a	09/11/01	kpm	102	64.8		cod	102
3	8877p58b	09/11/01	tjk	115	70.2		kpm	102
4	8877p58c	09/11/01	tmv	107	60.7		luv	102
5	8877p60a	09/12/01	etv	107	58.3		tmv	107
6	8877p60b	09/12/01	wow	107	73.6		etv	107
7	8696p08a	09/14/01	tmv	107	72.4		wow	107
8	8696p08b	09/14/01	etv	107	62.5		ejs	115
9	8696p08c	09/14/01	wow	107	56.7		gmv	115
10	8696p08d	09/14/01	cod	102	59.9		tjk	115
11	8696p08e	09/14/01	kpm	102	73.6			
12	8696p08f	09/14/01	luv	102	63.3			
13	8877p63a	09/14/01	cod	102	45.0			
14	8877p63b	09/14/01	kpm	102	50.2			
15	8877p70c	09/18/01	gmv	115	57.1			
16	8877p70d	09/18/01	tjk	115	39.8			
17	8877p71a	09/19/01	tmv	107	66.9			
18	8877p71b	09/19/01	etv	107	79.4			
19	8877p75a	09/20/01	wow	107	61.8			
20	8877p75b	09/20/01	kpm	102	58.4			
21	8877p78a	09/21/01	luv	102	65.8			
22	8877p78b	09/21/01	ejs	115	60.8			
24	8877p83a	09/25/01	gmv	115	83.7			
25	8877p83b	09/25/01	tjk	115	64.9			
26	8877p84a	09/25/01	tmv	107	63.0			

Figure: 2.29

Figure2.29showsINDIRECTinaction. If you give cell H2 the name `cod`, the formula `INDIRECT(cod)` returns the value of cell H2, which happens to be 102. So let's give the cells H2:H10 the names from the cells to their left (G2: G10). Here is an easy way to do so:

1. Select G2:H10.
2. Click Create from Selection on the Formulas tab.
3. Check Left Column.
4. Click OK. Notice all the names that were given to H2:H10 in the Name box
5. Call each group number through its name (for example, `=INDIRECT(C2)`).

Figure 2.30 shows a similar situation. In cell I3, you can use `INDIRECT` to find the reading for a specific strain (selected in cell I1) and a specific test (selected in cell I2). As you can see in this example, `INDIRECT` finds this reading at the intersection of the ranges named `Strain5` and `Test3`:

1. Select A1:F16.
2. Click Create from Selection for both the top row and the left column.
3. Use in cell I3 both row and column labels as names by using `INDIRECT` and separating them with a space: `=INDIRECT(I1) INDIRECT(I2)`.

D1 Test3

	A	B	C	D	E	F	G	H	I
1		Test1	Test2	Test3	Test4	Test5		Strain	Strain5
2	Strain1	1.49	1.23	1.01	1.60	1.88		Test	Test3
3	Strain2	1.11	1.01	1.94	1.97	1.74		Reading	1.45
4	Strain3	1.79	1.21	1.04	1.58	1.68			
5	Strain4	1.04	1.29	1.41	1.19	1.68			
6	Strain5	1.70	1.65	1.45	1.59	1.86			1.45
7	Strain6	1.01	1.74	1.34	1.99	1.15			1.45
8	Strain7	1.22	1.13	1.23	1.75	1.13			#VALUE!
9	Strain8	1.97	1.73	1.09	1.75	1.85			
10	Strain9	1.21	1.30	1.47	1.89	1.74			
11	Strain10	1.49	1.40	1.58	1.99	1.54			
12	Strain11	1.39	1.40	1.89	1.31	1.13			
13	Strain12	1.65	1.52	1.29	1.81	1.21			
14	Strain13	1.46	1.27	1.77	1.93	1.39			
15	Strain14	1.09	1.14	1.42	1.39	1.72			
16	Strain15	1.86	1.25	1.77	1.45	1.34			

Figure: 2.30

Note: The space keystroke (spacebar) acts as an intersection operator.

* * *

Chapter 14

Working with Trends

Based on values that you have observed, you might want to predict values you have not observed. Predictions are based on trend or regression analysis, as discussed in Part 4. Here we focus only on how we may have to look up existing, observed values first.

Let's use Figure 2.31 as an example. (This sheet has many features we don't discuss yet: Controls are discussed in Chapter 42; and graphs and charts are discussed in Part 3.) The scrollbar control in A19:B19 regulates C19, which determines D19 through a formula, so in turn the graph gets updated. The values you have measured and observed show a rather clear trend: When the percentage of C-G bonds in DNA goes up, the temperature of denaturation goes up as well—according to the formula y = 42.97x + 69.50 (more on this later). You have a relatively easy situation here because you know the formula, so you can predict for any (unobserved) x value what the corresponding y value would be. The graph shows where that point is.

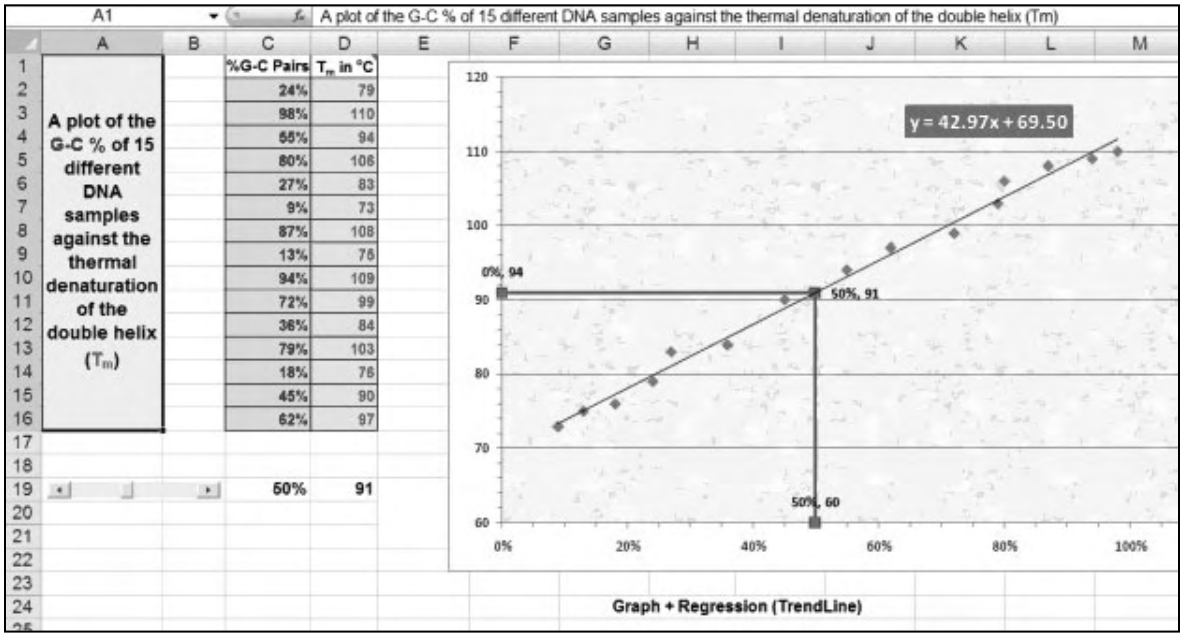


Figure: 2.31

Figure 2.32 shows an example of EC50 determination, where you do not yet have a formula. To predict an unobserved x value, you need a linear trend between the two nearest observations. In other words, you need to locate the two nearest observations—for -9.21, these would be -9.00 and -9.52—plus their corresponding mean values. `VLOOKUP` cannot do the job because it does not work with a row number (only a column number). Instead, you need a combination of `INDEX` and `MATCH`. These are the formulas you need, based on the Names Logs for B2:B10, Means for C2:C10, and Determ for the entire table:

- **Cell B15**: Search in column 1 of the table `Determ`:
  `=INDEX(Determ,MATCH($B$14,Logs,-1),1)`.
- **Cell B16**: Do the same but go one row farther:
  `=INDEX(Determ,MATCH($B$14,Logs,-1)+1,1)`.
- **Cell C15**: Search in column 2 of the table:
  `=INDEX(Determ,MATCH($B$14,Logs,-1),2)`.
- **Cell C16**: Do the same again but down one row:
  `=INDEX(Determ,MATCH($B$14,Logs,-1)+1,2)`.

Now that you know the nearest observations in the table, you can predict the y value that corresponds with the new x value of -9.21—by using the `TREND` function: `=TREND(C15:C16,B15:B16,B14)`. Thanks to `TREND`, `INDEX`, and `MATCH`, you can predict new values by just manipulating the Scrollbar control.

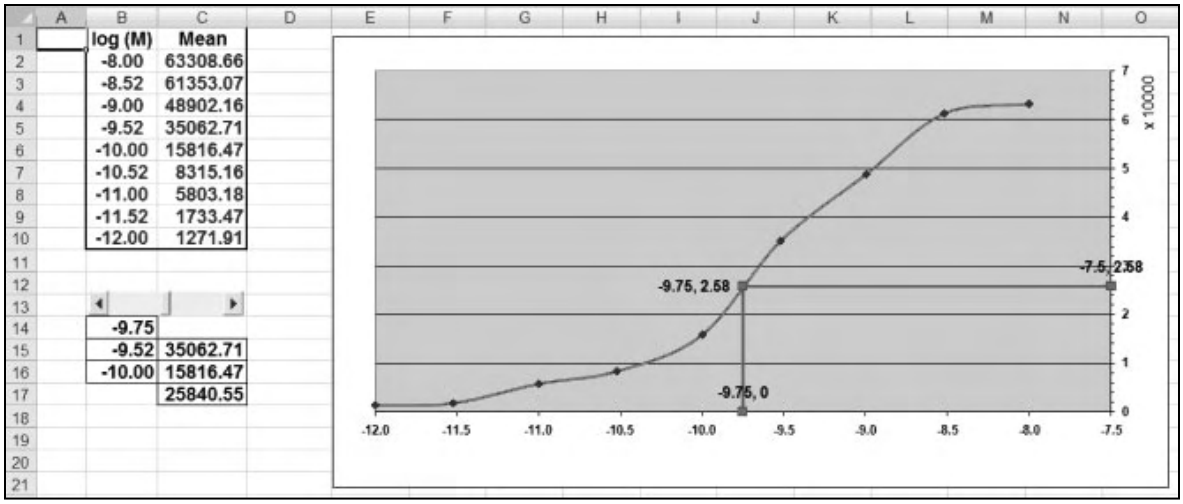


Figure: 2.32

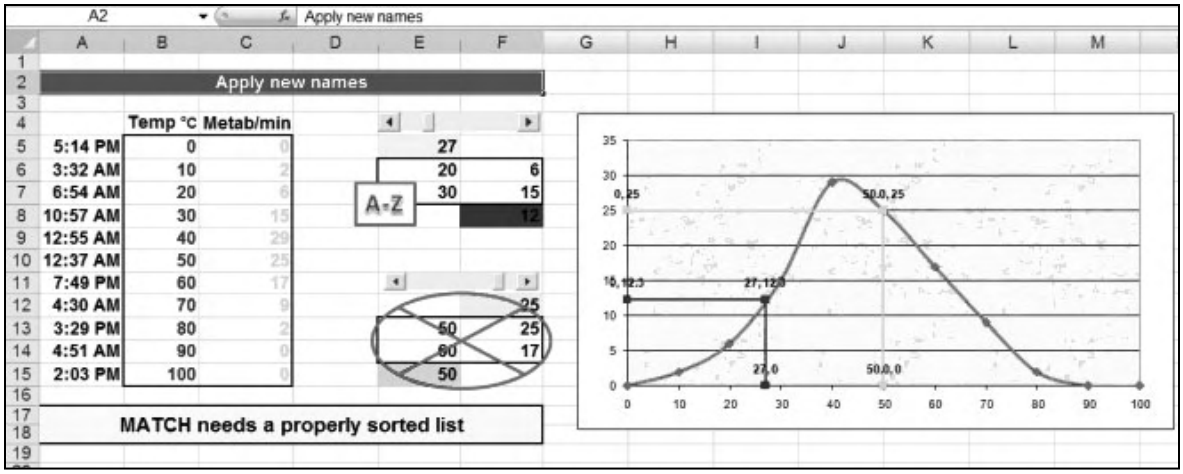


Figure: 2.33

Caution: The first control on Figure 2.33 doesn't work until column B is properly sorted. Remember that `MATCH` often needs a sorted range. But the second control does not work when column B is sorted because column C then has an ascending section followed by a descending section. If you sort column C in an ascending order, the predicted value follows an "imaginary" path that connects the dots in a zigzag way. (In addition, the first control would be in trouble.) The bottom line is that you need to be careful when you use `INDEX`, `MATCH`, and `TREND` together. `MATCH` cannot handle certain situations; when that's the case, you must find the formula behind the trend, which you will learn about in Part 4.

* * *

Chapter 15

Fixing Numbers

Sometimes you want your figures to look a bit different, and yet you need them to be numeric for calculation purposes. On the other hand, your figures might look great, but you cannot use them for calculations. How do you handle situations like these in your data analysis?

Figure 2.34 shows two tables. The table on the left is perfect, but for some reason, you would rather have it look like the one on the right. Here's what you can do:

- Say that in cell I2, you want bdo in cell B2 and combined with 105 in cell C2 into one entry: "bdo, 105". You can do this with the `CONCATENATE` function: `=CONCATENATE(B2,", ",C2)`. It can have many arguments, but in this case, it has only three: B2 and C2, separated by a literal comma and space (`", "`).
- Say that in column J, you want to show 43.2 as 43.2 ng/mL. Instead of using `CONCATENATE`, you could apply the ampersand as a concatenation operator: `=D2 & " ng/mL"`. Generally, the ampersand operator (`&`) is easier to use than `CONCATENATE`.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Plate ID	Analyst	Group	50 ng/mL	%CV	25 ng/mL	%CV		Both	50 ng/mL	CV	25 ng/mL	CV
2	8877p58b	bdo	105	43.2	4	22.6	3		bdo, 105	47.7 ng/mL	2%	22.7 ng/mL	0%
3	8877p58b	bdo	105	44.3	3	20.7	3		bdo, 105	48.7 ng/mL	3%	23 ng/mL	3%
4	8696p08b	bdo	105	52.9	3	27.5	6		bdo, 105	49.3 ng/mL	0%	23 ng/mL	1%
5	8696p08b	bdo	105	50.1	1	26.8	3		bdo, 105	45.9 ng/mL	2%	22.9 ng/mL	1%
6	8877p58a	gmv	107	47.7	2	22.7	0		gmv, 107	47.5 ng/mL	0%	22.7 ng/mL	2%
7	8877p58a	gmv	107	48.7	3	23.0	3		gmv, 107	45.9 ng/mL	1%	22.9 ng/mL	3%
8	8696p08a	gmv	107	52.3	1	26.5	2		gmv, 107	43.2 ng/mL	4%	22.6 ng/mL	3%
9	8696p08a	gmv	107	49.7	2	25.0	1		gmv, 107	44.3 ng/mL	3%	20.7 ng/mL	3%
10	8696p08a	ksm	121	56.4	12	29.1	11		ksm, 121	56.4 ng/mL	12%	29.1 ng/mL	11%
11	8696p08a	ksm	121	51.3	2	26.9	1		ksm, 121	51.3 ng/mL	2%	26.9 ng/mL	1%
12	8696p08b	ksm	121	51.0	1	26.1	1		ksm, 121	52.3 ng/mL	1%	26.5 ng/mL	2%
13	8696p08b	ksm	121	48.6	1	24.8	0		ksm, 121	49.7 ng/mL	2%	25 ng/mL	1%
14	8877p58a	tjk	113	49.3	0	23.0	1		tjk, 113	52.9 ng/mL	3%	27.5 ng/mL	6%
15	8877p58a	tjk	113	45.9	2	22.9	1		tjk, 113	50.1 ng/mL	1%	26.8 ng/mL	3%
16	8697p58b	tjk	113	47.5	0	22.7	2		tjk, 113	51.0 ng/mL	1%	26.1 ng/mL	1%
17	8697p58b	tjk	113	47.5	1	22.9	3		tjk, 113	48.6 ng/mL	1%	24.8 ng/mL	0%
18	8877p58b	tkm	209	47.5	0	22.7	2		tkm, 209	47.5 ng/mL	0%	22.7 ng/mL	2%
19	8877p58b	tkm	209	45.9	1	22.9	3		tkm, 209	47.5 ng/mL	1%	22.9 ng/mL	3%
20	8697p58b	tkm	209	43.2	4	22.6	3		tkm, 209	43.2 ng/mL	4%	22.6 ng/mL	3%
21	8697p58b	tkm	209	44.3	3	22.6	3		tkm, 209	44.3 ng/mL	3%	22.6 ng/mL	3%
22	Mean			48.4	2	24.2	3						

Figure: 2.34

- You can do something similar for column K. Notice that both new columns contain text now, not numbers. Consequently, the mean calculations at the bottom do not work. You can solve the calculation problem by leaving the value as is but applying a different format. You can therefore change the format of the cells in column L to `Custom: 0.0 "ng/mL"`. (The number of zeros determines the number of decimals.) Notice that calculations at the bottom work now because formatting is a matter of appearance and does not affect the number itself.
- You can change the format for column M to custom as well: `0"%"`. After you do this, watch the calculation.

Unfortunately, oftentimes, some columns contain non-numeric entries that are supposed to be numeric. You may have received such "values" through e-mail or through another program that does not work with numbers. How can you fix them so you can perform calculations? There are basically two solutions: You can use Excel functions or you can an Excel tool. Let's look at both options.

Figure 2.35 shows how to fix some troubled "values" by using functions:

- In cell E2, you can use the function LEFT: `=LEFT(D2,4)`. Although this cuts off the text part, the remaining part is still a text string—until you add the `VALUE` function: `=VALUE(LEFT(D2,4))`. In addition to using VALUE, one can multiply the result by 1:

E2 =VALUE(LEFT(D2,4))

	A	B	C	D	E	F	G	H	I	J	K
1	Plate ID	Analyst	Group	50 ng/mL	50 ng/mL	CV	CV	25 ng/mL	25 ng/mL	CV	CV
2	8877p58a	gmv	105	47.7 ng/mL	47.7	2%	2	22.7 ng/mL	22.7	0%	0
3	8877p58a	gmv	105	48.7 ng/mL	48.7	3%	3	23 ng/mL	23.0	3%	0.03
4	8877p58a	tjk	105	49.3 ng/mL	49.3	0%	0	23 ng/mL	23.0	1%	0.01
5	8877p58a	tjk	105	45.9 ng/mL	45.9	2%	2	22.9 ng/mL	22.9	1%	0.01
6	8877p58b	tkm	107	47.5 ng/mL	47.5	0%	0	22.7 ng/mL	22.7	2%	0.02
7	8877p58b	tkm	107	45.9 ng/mL	45.9	1%	1	22.9 ng/mL	22.9	3%	0.03
8	8877p58b	bdo	107	43.2 ng/mL	43.2	4%	4	22.6 ng/mL	22.6	3%	0.03
9	8877p58b	bdo	107	44.3 ng/mL	44.3	3%	3	20.7 ng/mL	20.7	3%	0.03
10	8696p08a	ksm	121	56.4 ng/mL	56.4	12%	12	29.1 ng/mL	29.1	11%	0.11
11	8696p08a	ksm	121	51.3 ng/mL	51.3	2%	2	26.9 ng/mL	26.9	1%	0.01
12	8696p08a	gmv	121	52.3 ng/mL	52.3	1%	1	26.5 ng/mL	26.5	2%	0.02
13	8696p08a	gmv	121	49.7 ng/mL	49.7	2%	2	25 ng/mL	25.0	1%	0.01
14	8696p08b	bdo	113	52.9 ng/mL	52.9	3%	3	27.5 ng/mL	27.5	6%	0.06
15	8696p08b	bdo	113	50.1 ng/mL	50.1	1%	1	26.8 ng/mL	26.8	3%	0.03
16	8696p08b	ksm	113	51.0 ng/mL	51.0	1%	1	26.1 ng/mL	26.1	1%	0.01
17	8696p08b	ksm	113	48.6 ng/mL	48.6	1%	1	24.8 ng/mL	24.8	0%	0
18	8697p58b	tjk	209	47.5 ng/mL	47.5	0%	0	22.7 ng/mL	22.7	2%	0.02
19	8697p58b	tjk	209	47.5 ng/mL	47.5	1%	1	22.9 ng/mL	22.9	3%	0.03
20	8697p58b	tkm	209	43.2 ng/mL	43.2	4%	4	22.6 ng/mL	22.6	3%	0.03
21	8697p58b	tkm	209	44.3 ng/mL	44.3	3%	3	22.6 ng/mL	22.6	3%	0.03
22				#DIV/0!	48.365	#DIV/0!	2.3	#DIV/0!	24.2	#DIV/0!	0.026

Figure: 2.35

1*LEFT(D2,4).

- Be aware that the columns F and G hold text, not numbers, but because they were formatted right-aligned, they look like numbers.
- In column G, the `LEFT` function cannot just take the first character because there may be more than one digit. To get around this problem, you also need the `LEN` function, which determines the length of the string: `=LEFT(F2,LEN(F2)-1)`.
- Cell I2 needs the formula `=VALUE(LEFT(H2,LEN(H2)-6))`.
- In column K, you use the bare VALUE function—and nothing else. It can handle simple situations such as `VALUE("5%")` and `VALUE("$5")`.

Figure 2.36 shows how to fix troubled "values" by using an Excel tool, Text to Columns, instead of functions. You find the Text to Columns tool on the Data tab. Say that someone had combined the series number and its sub IDs into a single Plate ID, so you cannot sort properly. You take these steps to split the ID:

	A	B	C	D	E	F
1	Plate ID	Sub ID	Date	Analyst	Group	C Value
2	8877p58a		09/11/01	kpm, 102		64.8
3	8877p58b		09/11/01	tjk, 115		70.2
4	8877p58c		09/11/01	tmv, 107		60.7
5	8877p60a		09/12/01	etv, 107		58.3
6	8877p60b		09/12/01	wow, 107		73.6
7	8696p08a		09/14/01	tmv, 107		72.4
8	8696p08b		09/14/01	etv, 107		62.5
9	8696p08c		09/14/01	wow, 107		56.7
10	8696p08d		09/14/01	cod, 102		59.9
11	8696p08e		09/14/01	kpm, 102		73.6
12	8696p08f		09/14/01	luv, 102		63.3
13	8877p63a		09/14/01	cod, 102		45.0
14	8877p63b		09/14/01	kpm, 102		50.2
15	8877p70c		09/18/01	gmv, 115		57.1
16	8877p70d		09/18/01	tjk, 115		39.8
17	8877p71a		09/19/01	tmv, 107		66.9
18	8877p71b		09/19/01	etv, 107		79.4
19	8877p75a		09/20/01	wow, 107		61.8
20	8877p75b		09/20/01	kpm, 102		58.4
21	8877p78a		09/21/01	luv, 102		65.8
22	8877p78b		09/21/01	ejs, 115		60.8
24	8877p83a		09/25/01	gmv, 115		83.7
25	8877p83b		09/25/01	tjk, 115		64.9
26	8877p84a		09/25/01	tmv, 107		63.0

1. Make sure you have a receptacle ready for each section you cut off. In this case, you need an empty column to the right of column A.
2. Choose Text to Columns from the Data tab.
3. Because in this case, you are dealing with a fixed width situation—all IDs have the same length—split column A after five characters.
4. Repeat steps 1–3 for column D, and don't forget to create an empty column for the second part of the split. This time, however, choose Delimited Text with both comma and space as delimiters.

Figure: 2.36

Figure 2.37 shows another way to fix numbers. Let's pretend that the instruments of some analysts needed calibration, and for two analysts, a correction factor is required. Instead of using formulas in extra columns, you can multiply directly by using the Paste Special tool. These are the steps:

1. Copy I1 (the correction factor for gmv).
2. Select the readings done by gmv.
3. Use Paste Special, selecting the option Multiply.
4. Repeat steps 1–3 for analyst tkm.

This trick works quickly and efficiently.

	A	B	C	D	E	F	G	H	I
1	Plate ID	Analyst	50 ng/mL	%CV	25 ng/mL	%CV		gmv	1.1
2	8877p58b	bdo	43.2	4	22.6	3		tkm	0.9
3	8877p58b	bdo	44.3	3	20.7	3			
4	8696p08b	bdo	52.9	3	27.5	6			
5	8696p08b	bdo	50.1	1	26.8	3			
6	8877p58a	gmv	52.5	2	25.0	0			
7	8877p58a	gmv	53.6	3	25.3	3			
8	8696p08a	gmv	57.5	1	29.2	2			
9	8696p08a	gmv	54.7	2	27.5	1			
10	8696p08a	ksm	56.4	12	29.1	11			
11	8696p08a	ksm	51.3	2	26.9	1			
12	8696p08b	ksm	51.0	1	26.1	1			
13	8696p08b	ksm	48.6	1	24.8	0			
14	8877p58a	tjk	49.3	0	23.0	1			
15	8877p58a	tjk	45.9	2	22.9	1			
16	8697p58b	tjk	47.5	0	22.7	2			
17	8697p58b	tjk	47.5	1	22.9	3			
18	8877p58b	tkm	42.8	0	20.4	2			
19	8877p58b	tkm	41.3	1	20.6	3			
20	8697p58b	tkm	38.9	4	20.3	3			
21	8697p58b	tkm	39.9	3	20.3	3			
22									

Figure: 2.37

Figure 2.38 tackles the rounding issue. The table on the right has the correction factors for each analyst. Here's what you do:

1. In cell D2, use the formula `=ROUND(C2*VLOOKUP(B2,$J$1:$K$5,2,0),1)`.
2. In cell G2, use the formula `=ROUND(F2*VLOOKUP(B2,$J$1:$K$5,2,0),1)`.
3. Use Paste Value before you get rid of the previous, old columns.

	A	B	C	D	E	F	G	H	I	J	K
1	Plate ID	Analyst	50 ng/mL	50 Corr.	%CV	25 ng/mL	25 Corr.	%CV		bdo	0.9
2	8877p58b	bdo	43.2	38.90000	4	22.6	20.30000	3		gmv	1.1
3	8877p58b	bdo	44.3	39.90000	3	20.7	18.60000	3		ksm	1.0
4	8696p08b	bdo	52.9	47.60000	3	27.5	24.80000	6		tjk	0.8
5	8696p08b	bdo	50.1	45.10000	1	26.8	24.10000	3		tkm	0.9
6	8877p58a	gmv	47.7	52.50000	2	22.7	25.00000	0			
7	8877p58a	gmv	48.7	53.60000	3	23.0	25.30000	3			
8	8696p08a	gmv	52.3	57.50000	1	26.5	29.20000	2			
9	8696p08a	gmv	49.7	54.70000	2	25.0	27.50000	1			
10	8696p08a	ksm	56.4	56.40000	12	29.1	29.10000	11			
11	8696p08a	ksm	51.3	51.30000	2	26.9	26.90000	1			
12	8696p08b	ksm	51.0	51.00000	1	26.1	26.10000	1			
13	8696p08b	ksm	48.6	48.60000	1	24.8	24.80000	0			
14	8877p58a	tjk	49.3	39.40000	0	23.0	18.40000	1			
15	8877p58a	tjk	45.9	36.70000	2	22.9	18.30000	1			
16	8697p58b	tjk	47.5	38.00000	0	22.7	18.20000	2			
17	8697p58b	tjk	47.5	38.00000	1	22.9	18.30000	3			
18	8877p58b	tkm	47.5	42.80000	0	22.7	20.40000	2			
19	8877p58b	tkm	45.9	41.30000	1	22.9	20.60000	3			
20	8697p58b	tkm	43.2	38.90000	4	22.6	20.30000	3			
21	8697p58b	tkm	44.3	39.90000	3	22.6	20.30000	3			

Figure: 2.38

C2 — 47.7

	A	B	C
1	Plate ID	Analyst	C-Value
2	8877p58a	gmv	47.7
3	8877p58a	gmv	48.7
4	8877p58a	tjk	49.3
5	8877p58a	tjk	45.9
6	8877p58b	tkm	47.5
7	8877p58b	tkm	45.9
8	8877p58b	bdo	43.2
9	8877p58b	bdo	44.3
10	8696p08a	ksm	56.4
11	8696p08a	ksm	51.3
12	8696p08a	gmv	52.3
13	8696p08a	gmv	49.7
14	8696p08b	bdo	52.9
15	8696p08b	bdo	50.1
16	8696p08b	ksm	51
17	8696p08b	ksm	48.6
18	8697p58b	tjk	47.5
19	8697p58b	tjk	47.5
20	8697p58b	tkm	43.2
21	8697p58b	tkm	44.3
22		Mean	#DIV/0!
23		SD	#DIV/0!

Figure: 2.39

Figure 2.39 shows a sheet with figures that were copied from an e-mail message or that were manipulated with text functions such as LEFT. No matter how you obtained them, they are text, and, therefore, cannot be used in calculations. Of course, you could use a new column with the function VALUE, as you did before. But there is actually a quicker solution: You can use Paste Special and then multiply or divide by 1. This forces the text to be a number.

> **Note:** Make sure to type the multiplier 1 somewhere on the sheet, copy it, and then use Paste Special based on a multiplication or division operation.

* * *

Chapter 16

Copying Formulas

Copying formulas can be a tedious process. When you copy a formula, its (relative) cell references adjust to their new location. Sometimes, that's great. If formulas (with relative references) did not adjust to their new locations, you would never be able to copy your formulas downward in the columns of your tables—which you have probably been doing for years. You can even use this formula behavior to your advantage when creating a transposed version of a table; it will be nicely rotated by 90 degrees, and all formulas will beautifully adjust. So you should be grateful for such performance.

However, adjusting references are not so great when you want a formula pasted into another location to come up with the same formula result as the original formula. Is such a thing possible? Yes, it is. Although it is not easy to reach this result, you may need to do it at times.

Figure 2.40 shows different ways of copying formulas that should keep the formulas' result the same (that is, if you want copies of formulas that do not change their references):

- Row 3 shows that Copy and Paste are not going to work in cell G3. Are there alternatives? Yes, but don't expect miracles.

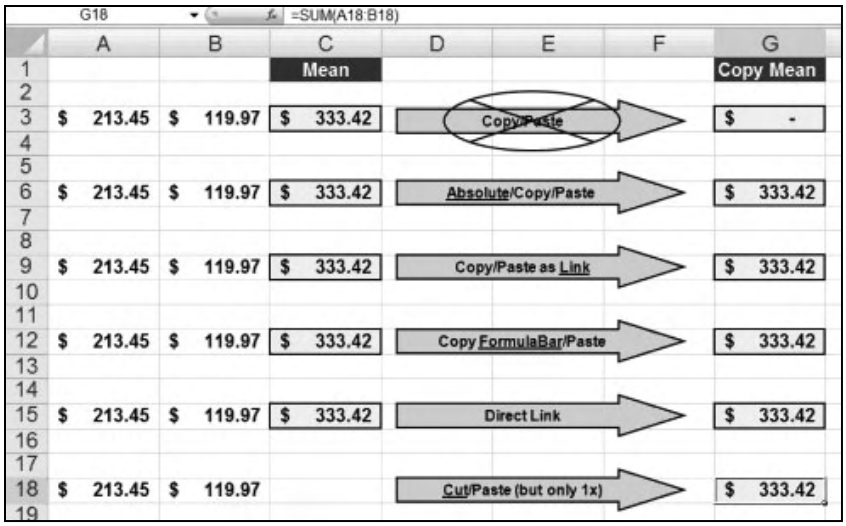


Figure: 2.40

- Row 6 uses absolute cell references. So cell C6 has the formula `=SUM($A$6:$B$6)`, which remains unchanged in the Copy and Paste process.
- Row 9 uses the Paste Link option from the Paste command in the Home tab. This automatically creates the following reference in cell G9: `=$C$9`. This is great,

but you can never delete the original cell(s).

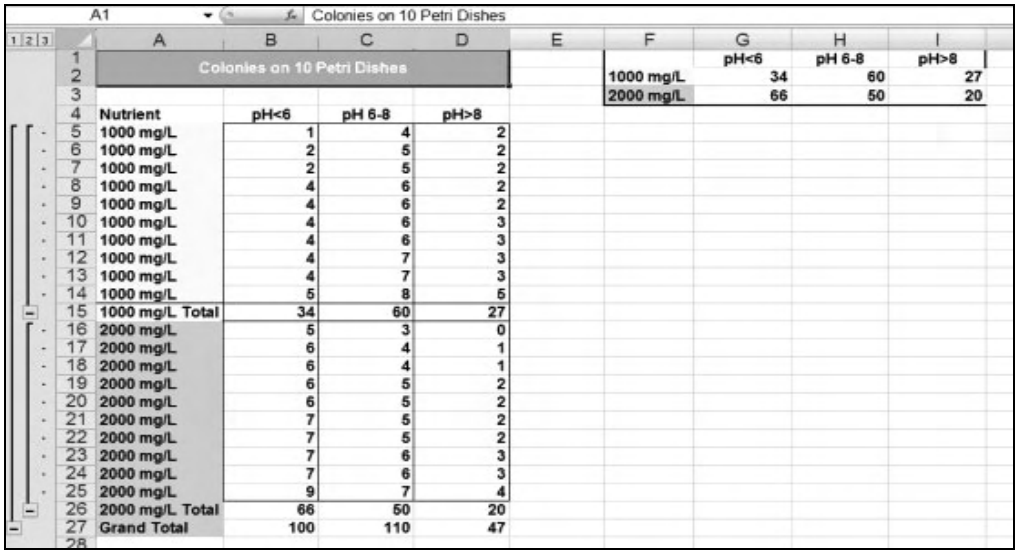
- In row 12, you apply a copying technique to the formula bar: Select C12's formula in the formula bar, copy it, press Esc, and paste it into cell G12. The formula does not update in its new location.
- In row 15, you create own link. Just type the following in G15: `=C15` or `=$C$15`.
- If you use the Cut option, the formula of cells that you cut and paste does not adjust, but the Paste operation works only once, and you lose the original. Sometimes, that's a high price to pay.

What is the best method? It depends on your preferences, but the formula bar method is often the best.

Figure 2.41 revisits a situation we discussed earlier. The summary table on the right needs to be created from records on the left. The easiest way to get started is to use the Subtotal tool. You collapse the table to its subtotals and copy only the visible cells into the summary table on the right (refer to Chapter 7). But no matter how you paste them, there isn't an ideal solution here: The link does not work when you decide to delete the subtotals, and the values do not update when records on the left are altered. The best solution is not so ideal in terms of time and work:

1. Copy the first formula from the formula bar.
2. Press Esc.
3. Paste the copy into G2.
4. Use the fill handle to copy G2 to the right.
5. Repeat steps 1–4 for B26 and other cells that need to be changed.

Now you can delete the subtotals to the left, and cell references adjust automatically. In addition, changes made to the table on the left replicate to the table on the right.



	A	B	C	D	E	F	G	H	I
1							pH<6	pH 6-8	pH>8
2	Colonies on 10 Petri Dishes					1000 mg/L	34	60	27
3						2000 mg/L	66	50	20
4	Nutrient	pH<6	pH 6-8	pH>8					
5	1000 mg/L	1	4	2					
6	1000 mg/L	2	5	2					
7	1000 mg/L	2	5	2					
8	1000 mg/L	4	6	2					
9	1000 mg/L	4	6	2					
10	1000 mg/L	4	6	3					
11	1000 mg/L	4	6	3					
12	1000 mg/L	4	7	3					
13	1000 mg/L	4	7	3					
14	1000 mg/L	5	8	5					
15	1000 mg/L Total	34	60	27					
16	2000 mg/L	5	3	0					
17	2000 mg/L	6	4	1					
18	2000 mg/L	6	4	1					
19	2000 mg/L	6	5	2					
20	2000 mg/L	6	5	2					
21	2000 mg/L	7	5	2					
22	2000 mg/L	7	5	2					
23	2000 mg/L	7	6	3					
24	2000 mg/L	7	6	3					
25	2000 mg/L	9	7	4					
26	2000 mg/L Total	66	50	20					
27	Grand Total	100	110	47					

Figure: 2.41

* * *

Chapter 17

Multi-cell Arrays

Most of the time, a formula returns a single value or result. Some formulas, though, return multiple values at the same time. Consequently, these array formulas, called *multi-cell array formulas*, need multiple cells to display their formula results. With these formulas, you need to select multiple cells ahead of time and accept each formula by pressing Ctrl+Shift+Enter (not just Enter or even Ctrl+Enter). If you forget that final step, you receive a `#VALUE` error. To correct that error, you click in the formula bar and then press Ctrl+Shift+Enter.

Figure 2.42 shows an example of a multi-cell array formula. As you learned in Chapter 8, filters work through labels that need to be exact replicas of the table labels. Each one of these filters is a perfect candidate for a multi-cell array formula. You use one as follows:

1. Select multiple cells, such as G1:K1.
2. Delete the contents of these cells.
3. Type `=A1:E1`. (A multi-cell array formula does not need absolute cell references.)
4. Press Ctrl+Shift+Enter

When you press Ctrl+Shift+Enter, Excel adds braces (`{}`). You should never actually type braces; they appear when you press Ctrl+Shift+Enter, and they change the formula into an array formula. Using an array formula has two advantages: First, no one can delete part of the array; second, when the table labels change, the filter labels change accordingly. Any changes will then propagate, and filter labels cannot be deleted.

G1 {=A1:E1}

	A	B	C	D	E	F	G	H	I	J	K
1	Patient	DOB	Age	Weight	Systolic		Patient	DOB	Age	Weight	Systolic
2	Bush	6/21/1970	36	152	126				>50		
3	Carter	1/26/1940	67	179	151						
4	Clinton	10/22/1976	30	185	160						
5	Eisenhower	8/21/1971	35	163	138		Patient	DOB	Age	Weight	Systolic
6	Ford	7/22/1949	57	172	144						>150
7	Johnson	6/21/1959	47	156	128						
8	Kennedy	2/26/1941	66	145	120						
9	Nixon	12/24/1948	58	165	140		Patient	DOB	Age	Weight	Systolic
10	Reagan	9/22/1955	51	159	132				>50		>150
11											

Figure: 2.42

Because parts of an array cannot be deleted, arrays are a nice tool for protecting certain cells from being changed inadvertently. To take advantage of this tool, you select the cell with a formula that you want to protect plus an empty cell next to it, click in the formula bar, and then press Ctrl+Shift+Enter. You probably want to hide the contents of the neighboring cell by making its font color white (to match its background). Now, unaware users (including yourself) cannot inadvertently delete the visible formula because it is attached to the neighboring cell.

Not only can you make your own array formulas, but there are also some preexisting array functions. The most important ones are `TRANSPOSE`, `FREQUENCY`, and `TREND`. There are a few more; we discuss some of them in Chapters 37 and 40.

Figure 2.43 shows how multi-cell array functions work. Let's start with `TRANSPOSE`:

1. Select I2:I6.
2. Type or call the function `TRANSPOSE`, whose range is A1:E1 (that is, no absolute references needed): `=TRANSPOSE(A1:E1)`.
3. Press Ctrl+Shift+Enter to get a transposed copy.

You follow these steps to find the frequencies per age category:

1. Select J9:J14.
2. Type or call `FREQUENCY`, whose data range is C2:C10 and bins range is I9:I13.
3. Press Ctrl+Shift+Enter to get the frequencies per category: `=FREQUENCY(C2:C10,I9:I13)`.

Finally, say that you want to predict what the systolic blood pressure would be if there is a linear relationship between systolic blood pressure and weight. You can easily achieve this by using the multi-cell array function `TREND`:

1. Select F2:F10.
2. Type or call `TREND`, using the formula `=TREND(E2:E10,D2:D10)`.
3. Press Ctrl+Shift+Enter to get predictions based on a linear regression line.

	A	B	C	D	E	F	G	H	I	J
1	Patient	DOB	Age	Weight	SBP_{obs}	SBP/Wgt_{pred}			Field	Value
2	Bush	6/21/1970	36	152	126	126			Patient	Bush
3	Carter	1/26/1940	67	179	151	152			DOB	6/21/70
4	Clinton	10/22/1976	30	185	160	158			Age	36
5	Eisenhower	8/21/1971	35	163	138	137			Weight	152
6	Ford	7/22/1949	57	172	144	146			SBPobs	126
7	Johnson	6/21/1959	47	156	128	130				
8	Kennedy	2/26/1941	66	145	120	119			Age	Freq
9	Nixon	12/24/1948	58	165	140	139		Age to 30	30	1
10	Reagan	9/22/1955	51	159	132	133		30 to 40	40	2
11								40 to 50	50	1
12								50 to 60	60	3
13								60 to 70	70	2
14										0
15										

Figure: 2.43

* * *

Chapter 18

Single-cell Arrays

Whereas a multi-cell array formula returns multiple values, a single-cell array formula returns a single value. But in order to calculate this single value, a single-cell array formula needs to work with arrays in the background; it can therefore perform calculations that are otherwise impossible.

Figure 2.44 shows a situation in which you might want to use a single-cell array because, following the rules of significant digits, the sum of the diluted volumes is not correct anymore: It should be 23.1 (E7) and not 23.2 (C7). In order to correct for this, you need a new column E that rounds all volumes to one digit first before you calculate the total in column F. A single-cell array formula could do all this work in the background by rounding the values in a "hidden" array and then performing a sum operation on the values from this rounded array. It's all done in one formula of the single-cell array type.

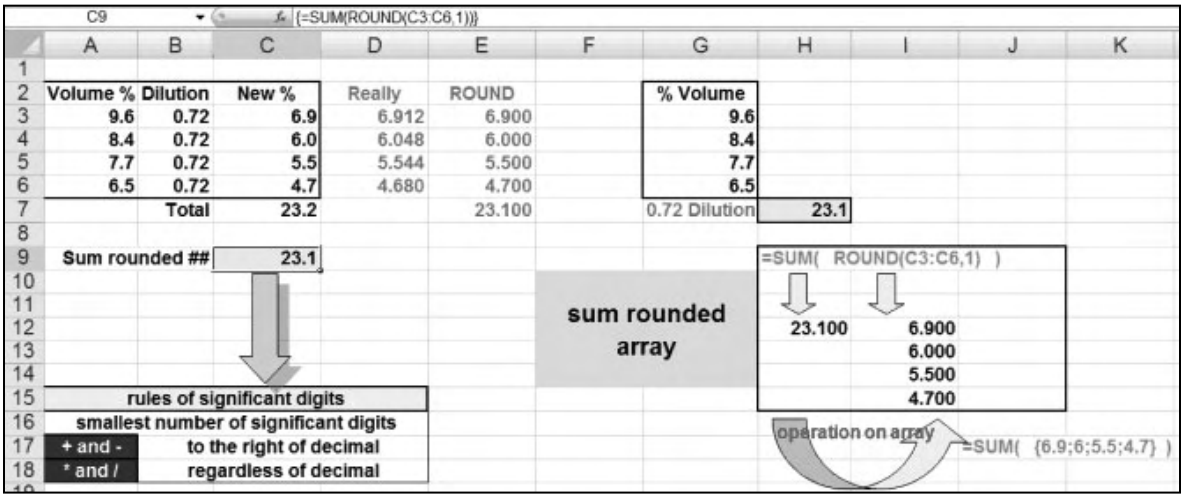


Figure: 2.44

This is the array formula for cell C9: `=SUM(ROUND(C3:C6,1))`. Don't forget to accept the formula by pressing Ctrl+Shift+Enter. If you want to check how the formula actually performs, you can do the following:

1. Highlight the `ROUND(…)` part of the formula in the formula bar. Make sure you don't highlight more or fewer characters.

2. Press F9 to run the formula, and you get the rounded array.
3. Press Esc to stop the run.
4. Highlight the `SUM(…)` part of the formula and press F9 to get the sum of this array.
5. Press Esc. (If you don't, you replace the ranges with values. If you press Enter instead of Esc, use Undo.)

> **Note**: In this case, the single-cell array formula combines (and, therefore, eliminates) intermediate columns by using internal arrays. You could skip even more columns! In cell H7, you could use one single-cell array formula based on only one column of figures: `=SUM(ROUND((G3:G6)*0.72,1))`.

Figure 2.45 presents a similar situation. In this case, the effect of a diet pill was measured for 16 participants in the project. You could use extra columns for in-between calculations, but all you need is some summary information. So you can use single-cell array formulas instead:

- In cell C19, set the mean change with `=AVERAGE((C2:C17)-(B2:B17))`.
- In cell C20, set the largest positive change with `=MAX((C2:C17)-(B2:B17))`.
- In cell C21, set the largest negative change with `=MIN((C2:C17)-(B2:B17))`.
- In cell C22, set the mean negative change with `=AVERAGE(IF((C2:C17)-(B2:B17)<0,(C2:C17)-(B2:B17)))`.

C19 {=AVERAGE((C2:C17)-(B2:B17))}

	A	B	C	D	E
1		Before	After		
2	Patient1	213.4	200.1		
3	Patient2	225.0	216.4		
4	Patient3	217.0	195.6		
5	Patient4	183.7	175.0		
6	Patient5	197.2	202.3		
7	Patient6	223.6	214.8		
8	Patient7	224.2	215.7		
9	Patient8	215.2	200.7		
10	Patient9	202.4	211.7		
11	Patient10	217.7	216.1		
12	Patient11	221.0	208.5		
13	Patient12	219.9	188.4		
14	Patient13	205.4	211.4		
15	Patient14	195.1	180.9		
16	Patient15	218.0	184.1		
17	Patient16	207.6	202.3		
18					
19	Mean Change		-10.15		
20	Largest Pos. Change		9.3		
21	Largest Neg. Change		-33.9		
22	Mean Negative Change		-14.06		
23					

Figure: 2.45

Figure 2.46 shows another example. Calculating the standard deviation at the bottom of column C will never work here because some observations were marked as `#NA`. You could solve this problem by creating extra columns: first a column D that checks for errors by using the function `ISERROR` or `IFERROR`, and then a column E that uses `IF` to replace error values with no value (`""`). However, a single-cell array formula can do all this in one step: `=STDEV(IFERROR(C2:C11,""))` – or the longer one: `=STDEV(IF(ISERROR(C2:C11),"",C2:C11))`.

C12 {=STDEV(IF(ISERROR(C2:C11),"",C2:C11))}

	A	B	C	D	E
1	Date	Analyst	C Value	IsError	Final
2	09/18/01	tjk	39.8	FALSE	39.80
3	09/14/01	cod	45.0	FALSE	45.00
4	09/26/01	luv	#N/A	TRUE	
5	09/26/01	ejs	47.5	FALSE	47.50
6	09/14/01	wow	50.2	FALSE	50.20
7	09/14/01	kpm	50.2	FALSE	50.20
8	09/18/01	gmv	57.1	FALSE	57.10
9	09/12/01	etv	#N/A	TRUE	
10	09/20/01	kpm	58.4	FALSE	58.40
11	09/14/01	cod	59.9	FALSE	59.90
12	Standard Deviation		7.03		7.03

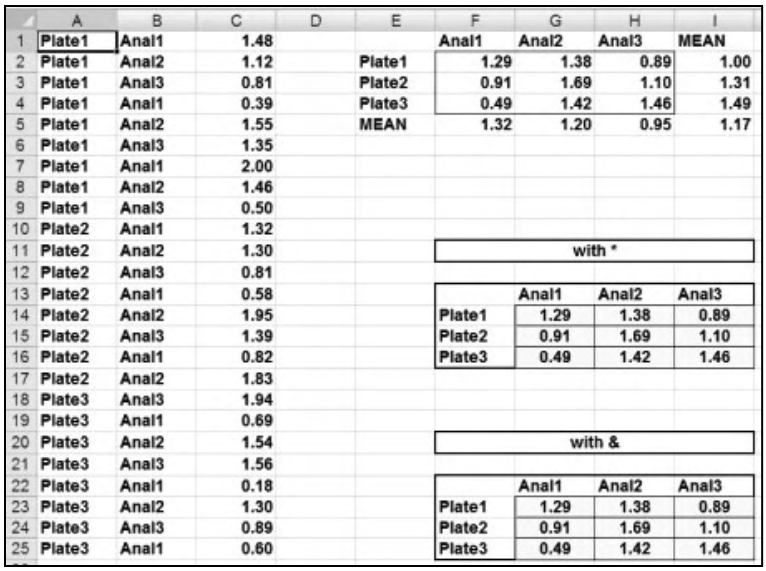
Figure: 2.46

> **Note**: Don't forget `""` in the nested `IF` function. Otherwise, `STDEV` will use zeros in its performance.

Single-cell array formulas can do even more. They use three rather unusual operators:

- For `AND`, you use *. You cannot use `AND()`.
- For `OR`, you use +. You cannot use `OR()`.
- For `CONCATENATE`, you use `&` (but without space on either side).

With these operators, you can perform miracles. Summary tables often require Excel functions such as `SUMIFS`, `COUNTIFS`, and `AVERAGEIFS`, but there is no such thing as `STDEVIFS`. However, you can create your own single-cell array formula. Thanks to the above-mentioned operators, single-cell array formulas can do anything you want.



	A	B	C	D	E	F	G	H	I
1	Plate1	Anal1	1.48			Anal1	Anal2	Anal3	MEAN
2	Plate1	Anal2	1.12		Plate1	1.29	1.38	0.89	1.00
3	Plate1	Anal3	0.81		Plate2	0.91	1.69	1.10	1.31
4	Plate1	Anal1	0.39		Plate3	0.49	1.42	1.46	1.49
5	Plate1	Anal2	1.55		MEAN	1.32	1.20	0.95	1.17
6	Plate1	Anal3	1.35						
7	Plate1	Anal1	2.00						
8	Plate1	Anal2	1.46						
9	Plate1	Anal3	0.50						
10	Plate2	Anal1	1.32						
11	Plate2	Anal2	1.30			with *			
12	Plate2	Anal3	0.81						
13	Plate2	Anal1	0.58			Anal1	Anal2	Anal3	
14	Plate2	Anal2	1.95		Plate1	1.29	1.38	0.89	
15	Plate2	Anal3	1.39		Plate2	0.91	1.69	1.10	
16	Plate2	Anal1	0.82		Plate3	0.49	1.42	1.46	
17	Plate2	Anal2	1.83						
18	Plate3	Anal3	1.94						
19	Plate3	Anal1	0.69						
20	Plate3	Anal2	1.54			with &			
21	Plate3	Anal3	1.56						
22	Plate3	Anal1	0.18			Anal1	Anal2	Anal3	
23	Plate3	Anal2	1.30		Plate1	1.29	1.38	0.89	
24	Plate3	Anal3	0.89		Plate2	0.91	1.69	1.10	
25	Plate3	Anal1	0.60		Plate3	0.49	1.42	1.46	

Figure 2.47 calls for a summary table. Say that the main table has three named ranges: `Plates` for A1:A25, `Alysts` for B1:B25, and `Readings` for C1:C25. You have two options:

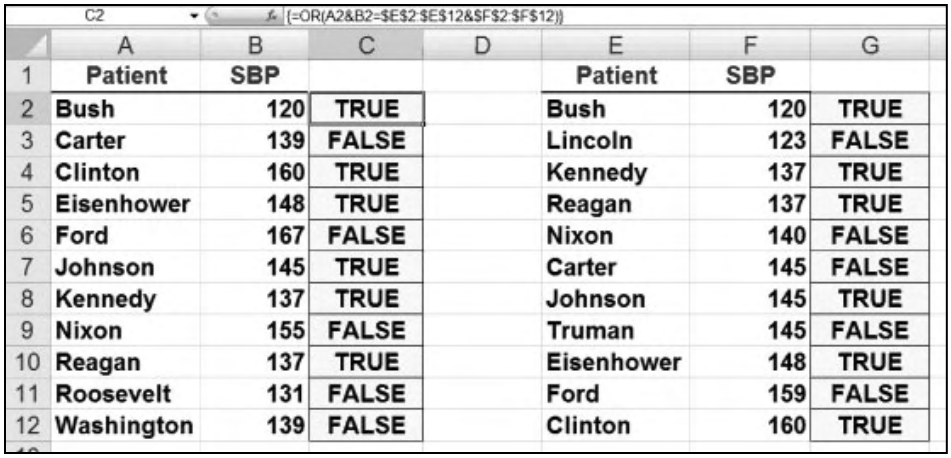
- You can use the operator `*` in cell G14: `=STDEV(IF((Plates=$F14)*(Alysts=G$13),Readings))`.
- You can use the operator `&` in cell G23: `=STDEV(IF(($F23&G$22)=(Plates)&(Alysts),Readings))`.

Figure: 2.47

Figure 2.48 offers another example in which you might want to consider using single-cell array formulas. Suppose you need to compare two lists—a new one versus an old one—such as to find out whether the exact combination of a specific name and specific systolic blood pressure reading can be found anywhere in the other list. The ideal array formula would check whether the combination of two cells exists in another range of multiple, paired cells by using the function `OR` (which returns `TRUE` if at least one of its components is true):

- **In C2**: `=OR(A2&B2=$E$2:$E$12&$F$2:$F$12)`
- **In G2**: `=OR(E2&F2=$A$2:$A$12&$B$2:$B$12)`

You can use F9 to test how the formula builds up.



	A	B	C	D	E	F	G
1	Patient	SBP			Patient	SBP	
2	Bush	120	TRUE		Bush	120	TRUE
3	Carter	139	FALSE		Lincoln	123	FALSE
4	Clinton	160	TRUE		Kennedy	137	TRUE
5	Eisenhower	148	TRUE		Reagan	137	TRUE
6	Ford	167	FALSE		Nixon	140	FALSE
7	Johnson	145	TRUE		Carter	145	FALSE
8	Kennedy	137	TRUE		Johnson	145	TRUE
9	Nixon	155	FALSE		Truman	145	FALSE
10	Reagan	137	TRUE		Eisenhower	148	TRUE
11	Roosevelt	131	FALSE		Ford	159	FALSE
12	Washington	139	FALSE		Clinton	160	TRUE

Figure: 2.48

* * *

Chapter 19

Date Manipulation

If your records require dates, you probably shouldn't skip this chapter. Excel handles dates in a special way—so you can easily calculate with them if you know what is going on in the background.

There are two important Excel date/time functions:

- `TODAY()` returns the current date.
- `NOW()` returns current date and time.

Each time you press F9 or change something on the sheet, these two functions update to the latest date/time information from your computer system. Obviously, `TODAY` and `NOW` are not good for record keeping because they change constantly (and are therefore called `volatile`). For record keeping, you instead need dates that are fixed, so you use the following shortcuts:

- `Ctrl+`; returns the current date (fixed).
- `Ctrl+Shift+`; returns the current time (fixed).
- `Ctrl+`; `[space] Ctrl+Shift+`; returns the current date and time combined.

You can use Ctrl+~ to check what is "behind" these fixed dates and times. You find that Excel assigns a serial number, with or without decimals, to each day. For example, 1/3/07 was day 39085 since Jan 1, 1900. Excel uses these weird numbers because it is easier for Excel to calculate with dates this way. You don't want to see the serial numbers, so Excel can easily convert them into the more human-readable date formats.

Dates can contain time information. For example, day 39085 is 1/3/07, and 39085.5 is 1/3/07 12:00 p.m. So 6:00 p.m. would be 39085.75. The number 39085 is therefore 12:00 a.m. on 1/3/07 in terms of time.

> **Note**: If you want to quickly glimpse at the serial numbers behind dates, you use Ctrl+~.

Figure 2.49 offers an overview of date manipulation:

- Cells B1 and C1 contain formulas that update, whereas cells B4:B6 have fixed dates created with shortcuts.
- You can fill column B with weekdays in either of two ways:
  - Select B8:B12 and copy that section down with the fill handle (refer to Chapter 2).
  - Select B12, copy downward with a right-mouse drag of the fill handle, and choose Fill Weekdays from the drop-down.
- You can place in front of today's weekday (column A) a fixed date for the current date.
- You can copy that date upward and downward—but without weekend dates. To do this, you click the drop-down at the bottom.
- You can place a fixed current date in cell E8 by double-clicking its fill handle.
- In this case, we need to do more work in order to go back by 7 days:
  1. Click the Fill button (located under the Σ button) on the Home tab.
  2. Choose the Series option.
  3. Specify a step value of -7.
- If it is important for you to find out how long ago a certain date is from now (in days, months, and years etc.), you can use the function `DATEDIF` in H8:J13. Although `DATEDIF` is nowhere to be found in Excel, you can type it according to the following syntax: `=DATEDIF(earlier,later,"y")`. The formula in H8 could be `=DATEDIF($G8,TODAY(),H$7)`.

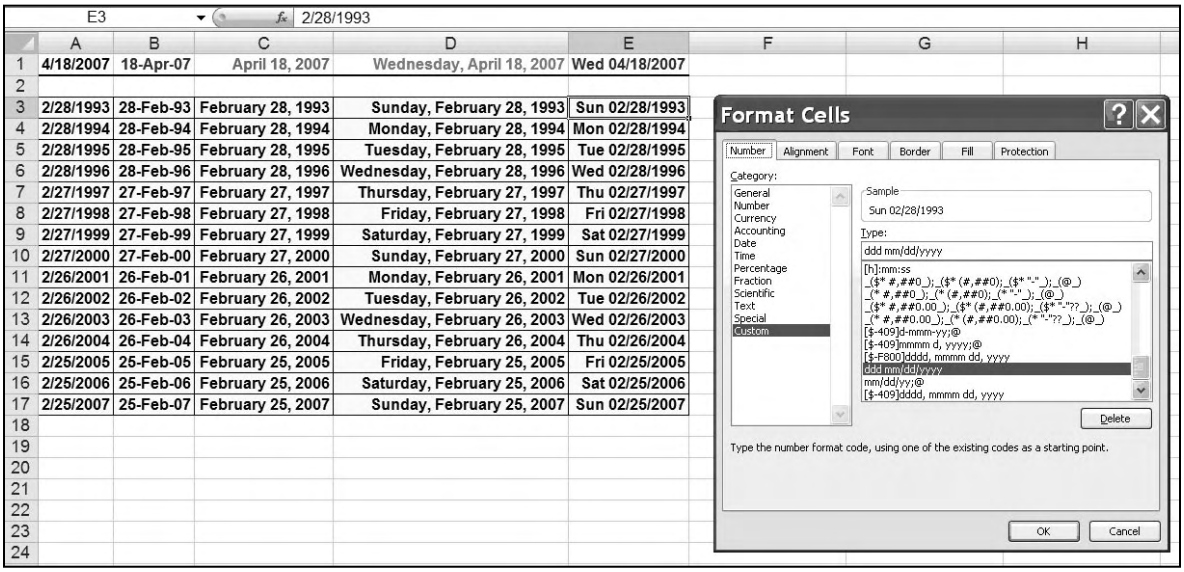
B1 =TODAY()

	A	B	C	D	E	F	G	H	I	J
1	TODAY()	4/18/2007								
2	NOW()	4/18/2007 10:43								
3										
4	Ctr ;	4/18/2007								
5	Ctr Sh ;	10:43 AM						DATEDIF(earlier,later,"y")		
6	combination	4/18/2007 10:43								
7								d	m	y
8		Monday		Today's date			4/18/2007			
9		Tuesday		1 weeks ago			4/10/2007			
10		Wednesday		2 weeks ago			3/11/2007			
11		Thursday		3 weeks ago			3/10/2006			
12		Friday		4 weeks ago			3/9/2005			
13				5 weeks ago			4/18/2004			
14				6 weeks ago						
15				7 weeks ago						
16				8 weeks ago						
17				9 weeks ago						
18										
19										
20										
21										
22										
23										

Figure: 2.49

Figure 2.50 has in column A dates that increase by 365 days. In the columns B:E, you can find an exact replica of the dates in column A—but their looks are rather different. No matter how fancy the looks are, however, the serial number behind them didn't change. You can change the appearance of dates through the Format dialog box. Most of the ones used in these columns are preexisting Excel options, except for the last column. Column E has a Custom format that is based on the following rules:

- dd is 01; ddd is Sun; ddd is Sunday
- mm is 02; mmm is Mar; mmmm is March



	A	B	C	D	E
1	4/18/2007	18-Apr-07	April 18, 2007	Wednesday, April 18, 2007	Wed 04/18/2007
2					
3	2/28/1993	28-Feb-93	February 28, 1993	Sunday, February 28, 1993	Sun 02/28/1993
4	2/28/1994	28-Feb-94	February 28, 1994	Monday, February 28, 1994	Mon 02/28/1994
5	2/28/1995	28-Feb-95	February 28, 1995	Tuesday, February 28, 1995	Tue 02/28/1995
6	2/28/1996	28-Feb-96	February 28, 1996	Wednesday, February 28, 1996	Wed 02/28/1996
7	2/27/1997	27-Feb-97	February 27, 1997	Thursday, February 27, 1997	Thu 02/27/1997
8	2/27/1998	27-Feb-98	February 27, 1998	Friday, February 27, 1998	Fri 02/27/1998
9	2/27/1999	27-Feb-99	February 27, 1999	Saturday, February 27, 1999	Sat 02/27/1999
10	2/27/2000	27-Feb-00	February 27, 2000	Sunday, February 27, 2000	Sun 02/27/2000
11	2/26/2001	26-Feb-01	February 26, 2001	Monday, February 26, 2001	Mon 02/26/2001
12	2/26/2002	26-Feb-02	February 26, 2002	Tuesday, February 26, 2002	Tue 02/26/2002
13	2/26/2003	26-Feb-03	February 26, 2003	Wednesday, February 26, 2003	Wed 02/26/2003
14	2/26/2004	26-Feb-04	February 26, 2004	Thursday, February 26, 2004	Thu 02/26/2004
15	2/25/2005	25-Feb-05	February 25, 2005	Friday, February 25, 2005	Fri 02/25/2005
16	2/25/2006	25-Feb-06	February 25, 2006	Saturday, February 25, 2006	Sat 02/25/2006
17	2/25/2007	25-Feb-07	February 25, 2007	Sunday, February 25, 2007	Sun 02/25/2007
18					
19					
20					
21					
22					
23					
24					

Figure: 2.50

Figure 2.51 starts with "problem" dates in column A. Once in a while, you may still be given records with dates that Excel doesn't recognize—and therefore cannot use in calculations. There are at least 3 different ways:

1. The easiest way to convert these text items into Excel dates is to use the button Text to Columns on the Data tab. Just skip step 1 and 2, but don't forget to mark in step 3 which type of date format you are dealing with – e.g. YMD. This tool will replace the "old" dates with Excel dates.
2. If you do use the steps 1 and 2, you have to make sure you have two blank columns waiting so you have space for the three split components. When something like 990615 has been properly split into 99 + 06 + 15, you still do not have a date that Excel understands until you use the `DATE` function: `=DATE(cell1,cell2,cell3)`.
3. Instead of using Excel's splitting tool, you could do all this in a single step with a heavily nested `DATE` function, such as `=DATE(LEFT(H1,2),MID(H1,3,2),RIGHT(H1,2))`.

	A	B	C	D	E	F	G	H	I
1	990615		99	6	15	06/15/99		990615	06/15/99
2	990616		99	6	16	06/16/99		990616	06/16/99
3	990617		99	6	17	06/17/99		990617	06/17/99
4	990618		99	6	18	06/18/99		990618	06/18/99
5	990619		99	6	19	06/19/99		990619	06/19/99
6	990615		99	6	15	06/15/99		990615	06/15/99
7	990616		99	6	16	06/16/99		990616	06/16/99
8	990617		99	6	17	06/17/99		990617	06/17/99
9	990618		99	6	18	06/18/99		990618	06/18/99
10	990619		99	6	19	06/19/99		990619	06/19/99
11	990620		99	6	20	06/20/99		990620	06/20/99
12	981111		98	11	11	11/11/98		981111	11/11/98
13	981112		98	11	12	11/12/98		981112	11/12/98
14	981113		98	11	13	11/13/98		981113	11/13/98
15	981114		98	11	14	11/14/98		981114	11/14/98
16	981115		98	11	15	11/15/98		981115	11/15/98
17	970101		97	1	1	01/01/97		970101	01/01/97
18	970102		97	1	2	01/02/97		970102	01/02/97
19	970103		97	1	3	01/03/97		970103	01/03/97
20	970104		97	1	4	01/04/97		970104	01/04/97

Figure: 2.51

Figure 2.52 uses conditional formatting to highlight the current date (refer to Chapter 11). When you know that there are also functions such as `DAY`, `MONTH`, and `YEAR`, you can highlight the current date in a calendar and use conditional formatting. The formula would be `=AND(ROW()=DAY(TODAY())+1,COLUMN()=MONTH(TODAY())+1)`. The `+1` addition is needed because rows and columns have headers.

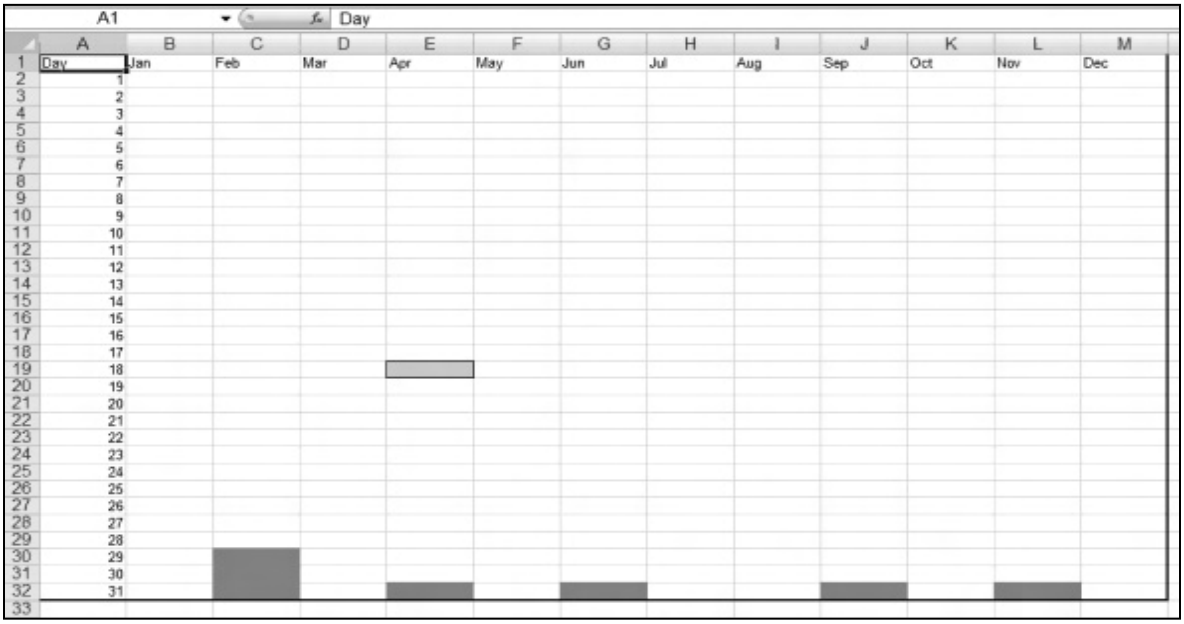


Figure: 2.52

* * *

Chapter 20

Time Manipulation

If the timing of experiments is important to you, you need to know some basics about the way Excel handles time information. You may have to do some manual work before you can perform calculations with your timed data. Time is a value that ranges from 0 to 0.999988425925926, representing the times from 0:00:00 (12:00:00 a.m.) to 23:59:59 (11:59:59 p.m.). You can see the value of a particular time under General Format or by using Ctrl+~. The advantage of using decimal values for time is that you can then easily add and subtract time values. You can even use functions such as `SUM`, `AVERAGE`, and so on.

When the difference in time values is more than 24 hours, the decimal time values go beyond 0.9999999. This causes trouble when you try to force the time decimal into the format hh:mm:ss. Time values beyond 0.9999999 get truncated when forced into the hh:mm:ss format. If the sum is 1.5, Excel shows only its decimal part, 0.5, which is 12:00:00. To solve this problem, you must change the format of this number from h:mm:ss to [h]:mm.ss. Then a number such as 1.5 will show up as 1.5 (in the proper time format, of course: 36:00:00). Thanks to [h]:mm:ss, you can calculate with time values beyond the duration of 1 day, which is usually necessary for sum operations.

Some people prefer to use hours with decimals—where, for example, 13.50 is 13 hours and 30 minutes, as opposed to 13:50, which is 13 hours and 50 minutes. To convert these decimals to Excel's time decimals, you need to divide by 24 because Excel works with day units of 24 hours, 60 minutes, and 60 seconds.

A1 start time

	A	B	C	D
1	start time	end time	duration	
2	9:09:15	9:10:05	0:00:50	
3	10:10:11	11:10:57	1:00:46	
4	13:13:30	16:14:32	3:01:02	
5		average duration	1:20:53	
6				
7				
8				
9	start date+time	end date+time	duration as day decimal	duration in hrs/mins/secs
10	10/23/2004 10:10:30	10/26/2004 11:30:15	3.055381944	73:19:45
11	10/26/2004 10:10:30	10/27/2004 11:30:15	1.055381944	25:19:45
12			total duration	98:39:30
13				
14				
15				
16		hours with decimals	as day decimal	in hrs/mins/secs
17		13.50	0.5625	13:30:00
18		8.25	0.34375	8:15:00
19		7.75	0.322916667	7:45:00
20	total	29.50	1.229166667	29:30:00
21				

Figure: 2.53

Figure 2.53 shows you how to perform calculations with time values:

- **Cell C2**: Uses a regular subtraction, `=B2-A2`, and a regular format, `(<1): h:mm:ss`.
- **Cell C5**: Uses a regular function, `=AVERAGE(C2:C4)`, and a regular format, `(<1): h:mm:ss`.
- **Cell C10**: Has a regular subtraction as well, but if you want to show decimals greater than 1 in a time notation, you

must use the format `[h]:m:ss`.

- **Cell B17**: Uses a decimal time notation, and you need to convert these decimals to Excel's time system. To do so, you divide by 24. Hence, cell C17 uses the formula `=B17/24`.
- **Cell D20**: Contains the function `=SUM(D17:D19)` and the format `[h]:mm:ss` because Excel's time decimal is greater than 1.

Figure 2.54 deals with time information that Excel cannot handle. Excel has no problem with the times in column B because they are done with Excel's time decimals. But column D is another story because those "values" came in as text. You transform them as follows:

- In column F, you extract the hours part: `=LEFT(D4,2)/24`.
- In column G, you extract the minutes part: `=MID(D4,4,2)/60/24`.
- In column H, you extract the seconds part: `=RIGHT(D4,2)/60/60/24`.
- When you have Excel-time decimals, you can sum them in F12:H12 and find a grand total in G14.
- You make cell G16 is the same as cell G14, but formatted as `[h]:mm:ss`. In Chapter 19, you applied the `DATE` function to split text dates. But you cannot apply the `TIME` function to split text times because `TIME` does not go beyond 1 day. The end result in G16 is the same as B12, which you get from "real" Excel time values. But you have to make quite a detour.

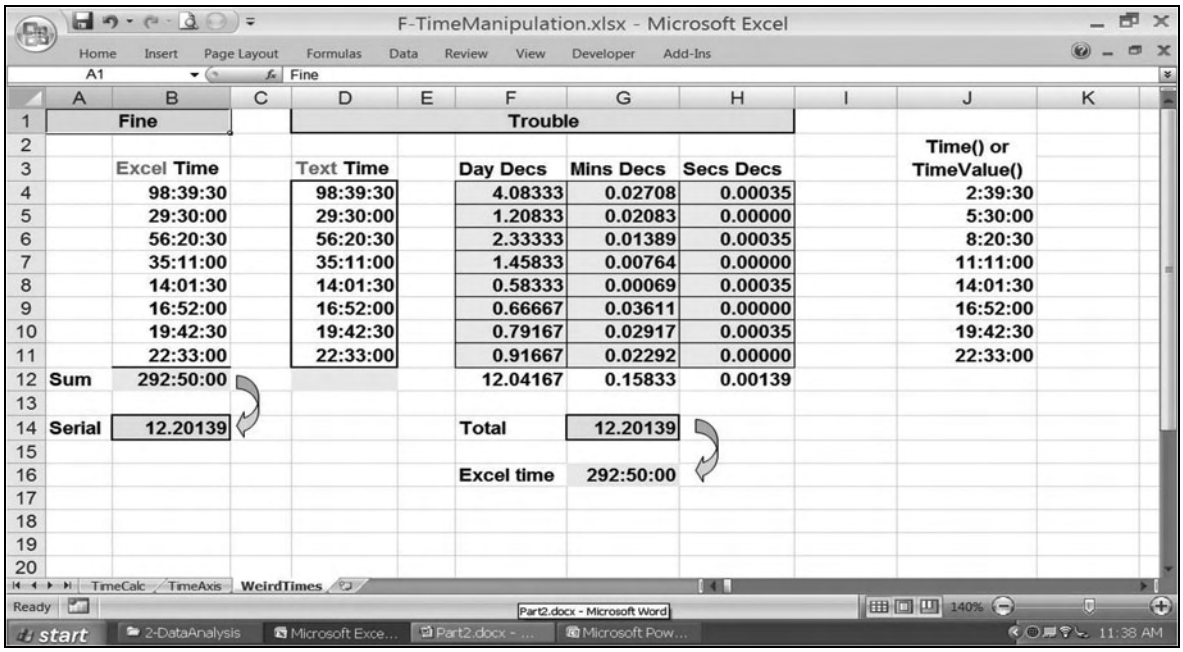


Figure: 2.54

By now you should have some insight into what to do with your own time values, especially if they are part of your data analysis.

* * *

Excercises - Part 2

You can download all the files used in this book from www.genesispc.com/Science2007.htm, where you can find each file in its original version (to work on) and in its finished version (to check your solutions).

Exercise 1

1. Subtotals

   1.1. Sort at two levels: per analyst and per plate (in this order).

   1.2. Create a summary table with subtotals—split per analyst.

   1.3. Add subtotals per plate—without losing the previous subtotals.

	A	B	C	D	E	F	G	H
1	Plate ID	Analyst	50 ng/mL	%CV	25 ng/mL	%CV	10 ng/mL	%CV
2	8696p08b	bdo	52.9	3.0	27.5	8.0	13.1	9.0
3	8696p08b	bdo	50.1	1.0	26.8	3.0	12.1	2.0
4	**8696p08b Average**		51.5		27.2		12.6	
5	8877p58b	bdo	43.2	4.0	22.6	3.0	9.6	4.0
6	8877p58b	bdo	44.3	3.0	20.7	3.0	9.7	6.0
7	**8877p58b Average**		43.8		21.7		9.7	
8		**bdo Average**	47.6		24.4		11.1	
9	8696p08a	gmv	52.3	1.0	26.5	2.0	12.2	2.0
10	8696p08a	gmv	49.7	2.0	25.0	1.0	11.2	3.0
11	**8696p08a Average**		51.0		25.8		11.7	
12	8877p58a	gmv	47.7	2.0	22.7	0.0	10.3	0.0
13	8877p58a	gmv	48.7	3.0	23.0	3.0	10.0	0.0
14	**8877p58a Average**		48.2		22.9		10.2	
15		**gmv Average**	49.6		24.3		10.9	
16	8696p08a	ksm	56.4	12.0	29.1	11.0	12.8	9.0
17	8696p08a	ksm	51.3	2.0	26.9	1.0	11.7	1.0
18	**8696p08a Average**		53.9		28.0		12.3	
19	8696p08b	ksm	51.0	1.0	26.1	1.0	12.3	2.0
20	8696p08b	ksm	48.6	1.0	24.8	0.0	11.4	2.0
21	**8696p08b Average**		49.8		25.5		11.9	
22		**ksm Average**	51.8		26.7		12.1	
23	8697p58b	tjk	47.5	0.0	22.7	2.0	9.6	8.0
24	8697p58b	tjk	47.5	1.0	22.9	3.0	9.4	0.0
25	**8697p58b Average**		47.5		22.8		9.5	

Figure: Ex-1

Exercise 2

2. Summary Functions

2.1. Create the formulas in D19:F21.

2.2. Implement a calculated filter in cell G17 for patients who are above the 75th percentile for both weight and systolic blood pressure.

2.3. Create a subset of records in the table by applying the filter.

	A	B	C	D	E	F	G
1	Patient	Gender	DOB	Age	Weight	Systolic	
2	Bush	M	1/3/1975	32	160	178	
3	Carter	F	10/22/1937	69	192	191	
4	Clinton	M	7/15/1971	35	171	175	
5	Eisenhower	F	11/24/1934	72	154	128	
6	Ford	M	2/9/1950	57	164	141	
7	Johnson	F	6/13/1965	41	152	125	
8	Kennedy	M	11/27/1973	33	185	193	
9	Lincoln	M	8/1/1972	34	188	166	
10	Nixon	F	9/15/1930	76	189	146	
11	Reagan	M	3/28/1939	68	140	170	
12	Roosevelt	M	9/29/1945	61	159	196	
13	Truman	F	10/20/1962	44	151	139	
14	Washington	F	10/15/1963	43	209	189	
15							
16	Patient	Gender	DOB	Age	Weight	Systolic	W+SBP 75th
17							FALSE
18							
19	Mean			56	201	190	
20	SD			75	242	233	
21	Count			2	2	2	
22							

Figure: Ex-2

Exercise 3

3. Summary Functions

3.1. Create the formulas in D19:F21.

3.2. Implement a calculated filter in cell G17 for patients who are above the 75th percentile for both weight and systolic blood pressure.

3.3. Create a subset of records in the table by applying the filter.

	A	B	C	D	E	F	G	H	I	J
1	Date	Patient	Old SBP	New SBP		Date	Patient	Old SBP	New SBP	Up>10
2	5/24/2006	Ford	167	141						FALSE
3	5/25/2006	Johnson	145	132						
4	5/17/2006	Reagan	137	125						
5	5/11/2006	Ford	167	158			Mean	141	166	
6	5/29/2006	Nixon	155	147			SD	11	19	
7	5/22/2006	Clinton	160	155			Count	8	8	
8	5/26/2006	Kennedy	137	139						
9	5/9/2006	Clinton	160	165						
10	5/15/2006	Kennedy	137	142						
11	5/5/2006	Bush	120	127						
12	5/23/2006	Eisenhow	148	160						
13	5/8/2006	Carter	139	152						
14	5/30/2006	Reagan	137	151						
15	5/12/2006	Johnson	145	160						
16	5/18/2006	Bush	120	141						
17	5/10/2006	Eisenhow	148	180						
18	5/16/2006	Nixon	155	190						
19	5/19/2006	Carter	139	191						

Figure: Ex-3

Exercise 4

4. Conditional Formatting

   4.1. Mark the cells in columns C:D for cases in which the systolic blood pressure went up by more than 10 mmHg—by using conditional formatting formulas.

   4.2. Make sure both cells are marked for each record.

   4.3. Locate the section of conditional formatting by using GoTo.

   4.4. Hide values in J:K where systolic blood pressure went down.

B1 Patient

	B	C	D	E	F	G	H	I	J	K
1	Patient	Old SBP	New SBP		Up by more than		Date	Patient	Old SBP	New SBP
2	Bush	120	127		10		5/5/2006	Bush	120	127
3	Bush	120	141				5/18/2006	Bush	120	141
4	Carter	139	152				5/8/2006	Carter	139	152
5	Carter	139	191				5/19/2006	Carter	139	191
6	Clinton	160	165				5/9/2006	Clinton	160	165
7	Clinton	160	155				5/22/2006	Clinton		
8	Eisenhower	148	180				5/10/2006	Eisenhower	148	180
9	Eisenhower	148	160				5/23/2006	Eisenhower	148	160
10	Ford	167	158				5/11/2006	Ford		
11	Ford	167	141				5/24/2006	Ford		
12	Johnson	145	160				5/12/2006	Johnson	145	160
13	Johnson	145	132				5/25/2006	Johnson		
14	Kennedy	137	142				5/15/2006	Kennedy	137	142
15	Kennedy	137	139				5/26/2006	Kennedy	137	139
16	Nixon	155	190				5/16/2006	Nixon	155	190
17	Nixon	155	147				5/29/2006	Nixon		
18	Reagan	137	125				5/17/2006	Reagan		
19	Reagan	137	151				5/30/2006	Reagan	137	151

Figure: Ex-4

Exercise 5

5. Filtering Tools

   5.1. Filter only for records with readings in both column E and column F (so no rows 4, 7, 9, and 12).

   5.2. Do the opposite: Filter for records that have readings missing in column E and/or column F.

	A	B	C	D	E	F	G
1	Patient	Gender	DOB	Age	Weight	Systolic	
2	Bush	M	1/3/1975	32	160	178	
3	Carter	F	10/22/1937	69	192	151	
4	Clinton	M	7/15/1971	35	171		
5	Eisenhower	F	11/24/1934	72	154	128	
6	Ford	M	2/9/1950	57	164	141	
7	Johnson	F	6/13/1965	41			
8	Kennedy	M	11/27/1973	33	165	193	
9	Lincoln	M	8/1/1972	34	188		
10	Nixon	F	9/15/1930	76	189	146	
11	Reagan	M	3/28/1939	67	140	170	
12	Roosevelt	M	9/29/1945	61	159		
13	Truman	F	10/20/1962	44	151	139	
14	Washington	F	10/15/1963	43	209	180	
15							
16							
17	Patient	Gender	DOB	Age	Weight	Systolic	W+SBP
18							TRUE

Figure: Ex-5

Exercise 6

6. Filtering Tools
   6.1. Calculate the mean per plate in column E by using `SUMIF/COUNTIF`.
   6.2. Filter in such a way that you show each plate only once.
   6.3. Filter in such a way that you show each plate only once, but this time with the filter and subset on a separate sheet.

	A	B	C	D	E	F	G
1	Plate ID	Date	Analyst	C Value	Mean/plate	Mean per plate: SUMIF/COUNTIF.	
2	8877p70d	09/18/01	tjk	39.7			
3	8877p70d	09/14/01	cod	45.0			
4	8877p70d	09/26/01	luv	45.3			
5	8877p66b	09/26/01	ejs	47.4			
6	8877p66b	09/14/01	kpm	50.2			
7	8877p66b	09/18/01	gmv	57.1			
8	8877p66b	09/12/01	etv	58.3			
9	8877p75b	09/20/01	kpm	58.4			
10	8877p75b	09/14/01	cod	59.9			
11	8877p78b	09/21/01	ejs	61.5			
12	8877p78b	10/03/01	ejs	62.5			
13	8877p78b	09/14/01	etv	63.3			
14	8877p78b	09/14/01	luv	63.3			
15	8877p78b	09/11/01	kpm	64.8			
16	8877p83b	09/25/01	tjk	64.2			
17	8877p83b	09/21/01	luv	65.8			
18	8877p83b	09/11/01	tjk	70.2			
19	8696p08e	09/14/01	kpm	73.5			
20	8696p08e	09/25/01	etv	78.3			
21	8696p08e	09/19/01	etv	79.3			
22	8696p08e	09/25/01	gmv	83.6			
23						Only when $A2 differs from $A1.	
24							
25	Plate ID	Date	Analyst	C Value	Mean/plate	Once	
26							

Figure: Ex-6

Exercise 7

7. Lookups
   7.1. Find in column D whether the analyst is pre- or post- according to table H1:J10.
   7.2. In column E, fill in the group to which the analyst belongs.

	A	B	C	D	E	F	G	H	I	J
1	Plate ID	Date	Analyst	Pre/Post	Group	C Value		Group	Analyst	Pre/Post
2	8877p58a	09/11/01	kpm	pre	102	64.8		102	cod	pre
3	8877p58b	09/11/01	tjk	pre	115	70.2		102	kpm	pre
4	8877p58c	09/11/01	tmv	post	107	60.7		102	luv	post
5	8877p60a	09/12/01	etv	pre	107	58.3		107	tmv	post
6	8877p60b	09/12/01	wow	pre	107	73.6		107	etv	pre
7	8696p08a	09/14/01	tmv	post	107	72.4		107	wow	pre
8	8696p08b	09/14/01	etv	pre	107	62.5		115	ejs	post
9	8696p08c	09/14/01	wow	pre	107	56.7		115	gmv	post
10	8696p08d	09/14/01	cod	pre	102	59.9		115	tjk	pre
11	8696p08e	09/14/01	kpm	pre	102	73.6				
12	8696p08f	09/14/01	luv	post	102	63.3				
13	8877p63a	09/14/01	cod	pre	102	45.0				
14	8877p63b	09/14/01	kpm	pre	102	50.2				
15	8877p70c	09/18/01	gmv	post	115	57.1				
16	8877p70d	09/18/01	tjk	pre	115	39.8				
17	8877p71a	09/19/01	tmv	post	107	66.9				
18	8877p71b	09/19/01	etv	pre	107	79.4				
19	8877p75a	09/20/01	wow	pre	107	61.8				
20	8877p75b	09/20/01	kpm	pre	102	58.4				
21	8877p78a	09/21/01	luv	post	102	65.8				
22	8877p78b	09/21/01	ejs	post	115	60.8				
23	8877p83a	09/25/01	gmv	post	115	83.7				
24	8877p83b	09/25/01	tjk	pre	115	64.9				
25	8877p84a	09/25/01	tmv	post	107	63.0				
26										

Figure: Ex-7

Exercise 8

8. Trends

8.1. Make the first control functional by creating the formulas in D4:E6.

8.2. Make the second control functional by creating the formulas in D11:E13.

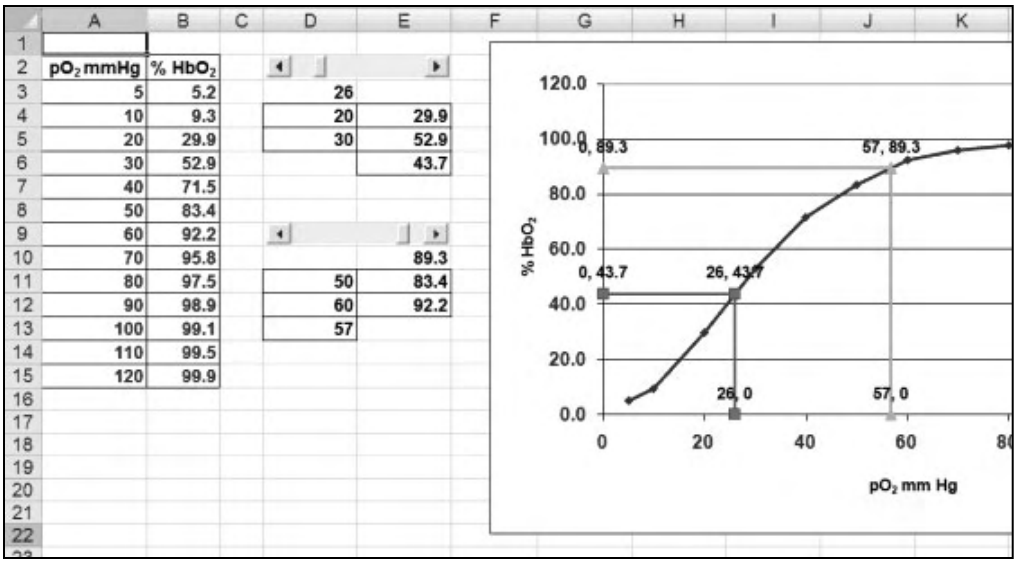


Figure: Ex-8

Exercise 9

9. Multi-cell Array Formulas

9.1. Create the first column of the second table by using multi-cell array formulas, starting in A8:A11, based on cell A2 in the first table.

9.2. Create the second column of the second table, starting in B8:B11 and using the TRANSPOSE function. When the first array is done, you can copy the entire array down.

9.3. Use TRANSPOSE again for the third column. This time copying is not an option.

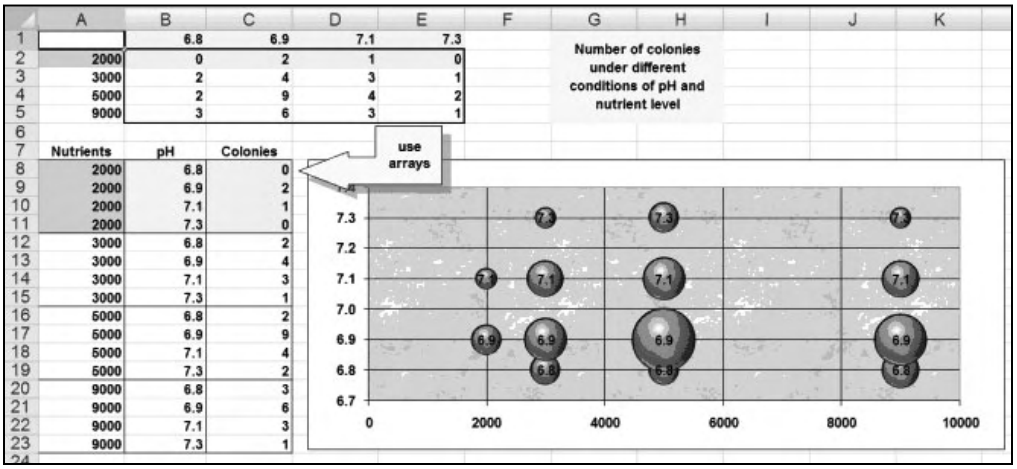


Figure: Ex-9

	A	B	C
1	Plate1	Anal1	1.48
2	Plate1	Anal2	0.00
3	Plate1	Anal3	0.81
4	Plate1	Anal1	0.39
5	Plate1	Anal2	1.55
6	Plate1	Anal3	1.35
7	Plate1	Anal1	2.00
8	Plate1	Anal2	0.00
9	Plate1	Anal3	0.50
10	Plate2	Anal1	1.32
11	Plate2	Anal2	1.30
12	Plate2	Anal3	0.81
13	Plate2	Anal1	0.00
14	Plate2	Anal2	1.95
15	Plate2	Anal3	1.39
16	Plate2	Anal1	0.82
17	Plate2	Anal2	0.00
18	Plate3	Anal3	1.94
19	Plate3	Anal1	0.69
20	Plate3	Anal2	1.54
21	Plate3	Anal3	0.00
22	Plate3	Anal1	0.18
23	Plate3	Anal2	1.30
24	Plate3	Anal3	0.89
25	Plate3	Anal1	0.60
26	Mean (no 0)		1.14

Figure: Ex-10

Exercise 10

10. Single-cell Array Formulas

Calculate the mean in cell C26 in one step. Ignore zero readings.

	A	B	C	D	E	F	G	H
1	55	316	223	185	124		Bins	Percent
2	124	93	163	213	314		50	5.0%
3	211	41	231	241	212		100	9.0%
4	118	113	400	205	254		150	16.0%
5	262	1	201	172	101		200	19.0%
6	167	479	205	337	118		250	18.0%
7	489	15	89	362	148		300	9.0%
8	179	248	125	197	177		350	12.0%
9	456	153	269	49	127		400	6.0%
10	289	500	198	317	300		450	2.0%
11	126	114	303	314	270		500	4.0%
12	151	279	347	314	170			
13	250	175	93	209	61			
14	166	113	356	124	242			
15	152	384	157	233	99			
16	277	195	436	6	240			
17	147	80	173	211	244			
18	386	93	330	400	141			
19	332	173	129	323	188			
20	338	263	444	84	220			

Figure: Ex-11

Exercise 11

11. Single-cell Array Formulas

Calculate the percentage of occurrences per value bin by using `FREQUENCY` and `COUNT` in one formula.

Exercise 12

12. Single-cell Array Formulas

Use `INDEX`, `MATCH`, and the `&` operator in cell J8 to find the value for a specific plate (J6) and a specific cycle (J7).

	A	B	C	D	E	F	G	H	I	J
1	Plate	Cycle	Anal1	Anal2	Anal3	Anal4	Mean			
2	ab10	cycl1	0.649	0.053	0.119	0.398	0.305			
3	ab10	cycl2	0.268	0.607	0.666	0.736	0.569			
4	ab10	cycl3	0.852	0.553	0.366	0.261	0.508		Array INDEX	
5	ab10	cycl4	0.002	0.229	0.019	0.970	0.305			
6	bc11	cycl1	0.522	0.106	0.865	0.228	0.430		Plate	ab10
7	bc11	cycl2	0.348	0.368	0.097	0.261	0.269		Cycle	cycl2
8	bc11	cycl3	0.662	0.623	0.341	0.685	0.578		Total	0.569
9	bc11	cycl4	0.291	0.755	0.361	0.725	0.533			
10	de12	cycl1	0.818	0.469	0.399	0.334	0.505			
11	de12	cycl2	0.298	0.552	0.638	0.529	0.504			
12	de12	cycl3	0.529	0.628	0.175	0.634	0.492			
13	de12	cycl4	0.196	0.528	0.129	0.750	0.401			
14	fg13	cycl1	0.400	0.471	0.945	0.324	0.535			
15	fg13	cycl2	0.020	0.783	0.478	0.138	0.355			
16	fg13	cycl3	0.815	0.588	0.118	0.327	0.462			
17	fg13	cycl4	0.812	0.538	0.112	0.981	0.611			
18										

Figure: Ex-12

Exercise 13

13. Time Manipulation

13.1. Calculate the metabolic rate per hour in column E.

13.2. Calculate the metabolic rate per minute in column F.

13.3. Create the correct units for the x axis scale in the graph, if you know how to deal with graphs (see Part 3).

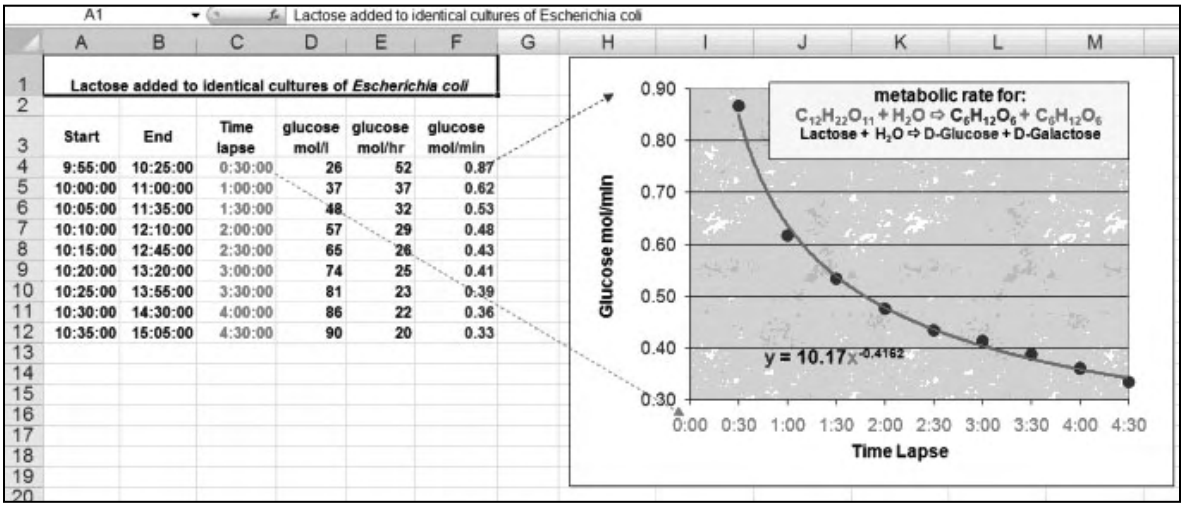


Figure: Ex-13

* * *

PART 3
Plotting Data

Chapter 21

Types of Graphs

Excel offers a good array of graph types, all of which fall into four main types: Pie, Column, Line, and XY. All the other types are essentially subtypes of these four. What are the differences between the various graph types, and when should you use which type?

Figure 3.1 explains a bit of the terminology related to tables and their graphs:

- This table has four *categories*; they end up on the horizontal axis.
- This table has two *series* of values; they determine the number of columns for each category.
- The values of the data series are on the vertical axis, if the graph does have axes.
- The labels of the data series end up in the legend.

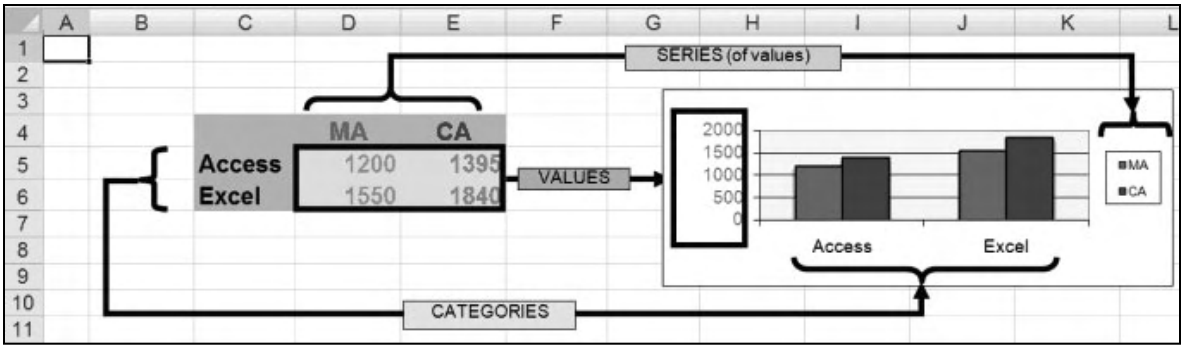


Figure: 3.1

Figure 3.2 shows the four main graph types Excel has to offer:

- Pie (and Doughnut)
- Column (and Bar)
- Line (plus Area and Surface)
- XY (or Scatter)

Here is an overview of the main characteristics of each major graph type:

- **Pie and Doughnut graphs**: These graphs have no axes. A Pie graph is based on a category

and is limited to only one data series; a Doughnut graph may display more than one data series.

- **Column and Line graphs**: These graphs have two axes: one for category and one for values. They can display multiple data series.
- **XY and Scatter graphs**: These graphs have two axes, both of which are for values. An XY or Scatter graph can display more than one data series, but each series represents pairs of x values and y values.

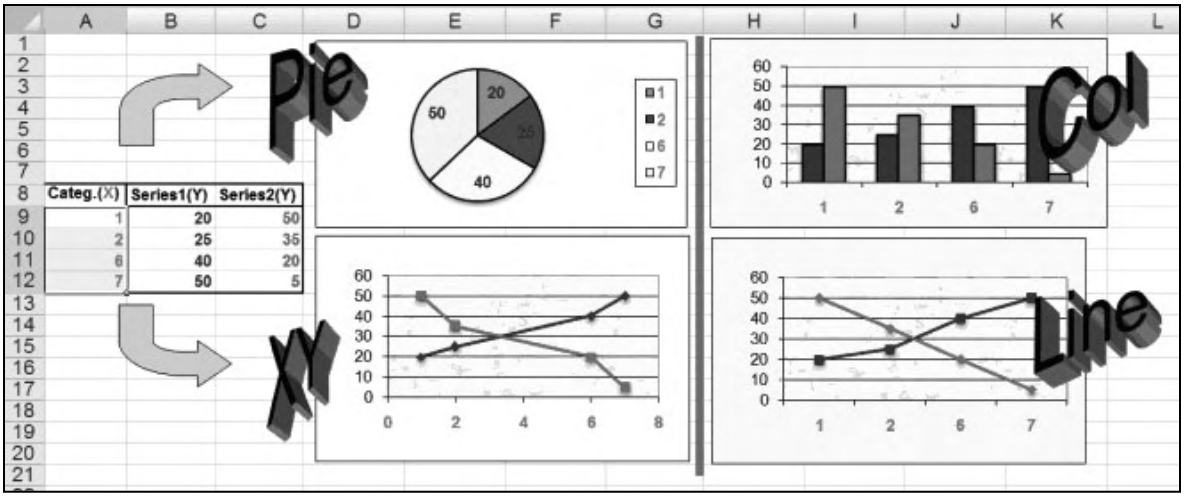


Figure: 3.2

Figure 3.3 deals with the first type, the Pie type, and its close "relative" the Doughnut. Pie and/or Doughnut graphs are ideal for showing the contribution of each value to a total (for

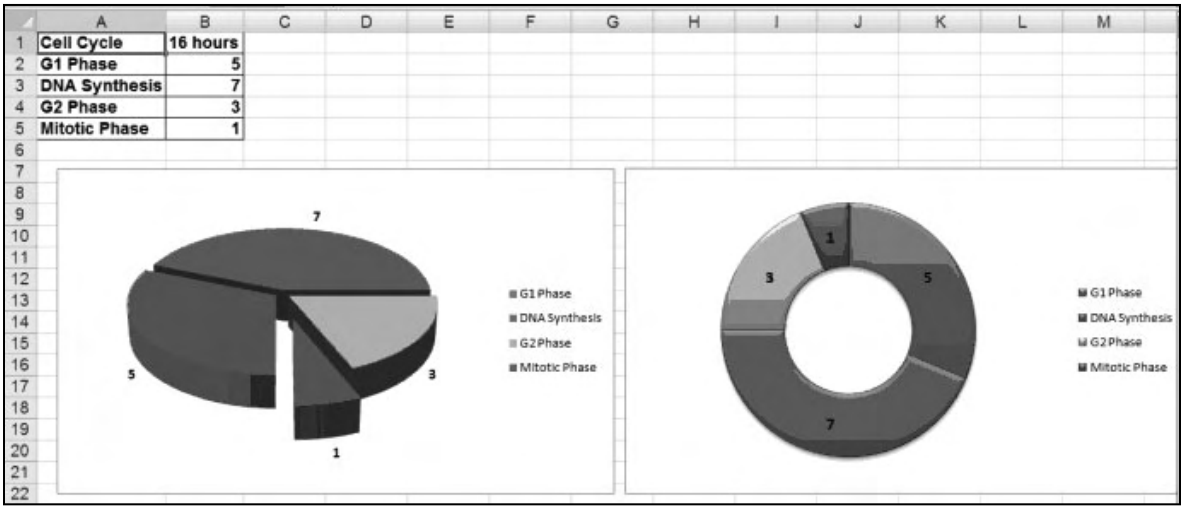


Figure: 3.3

example, the phases of a cell cycle). To create a Pie graph, you follow these steps:

1. Select any cell inside the table. (Excel knows the table's boundaries: the first empty row and column.)
2. Select Insert, Pie, Exploded 3D.
3. If you don't need or like the title, select it and press Delete.
4. To manipulate the Pie, select Layout, 3D-Rotation, and use the buttons to rotate the pie.

> **Note:** You must be inside the graph to get the Layout menu.

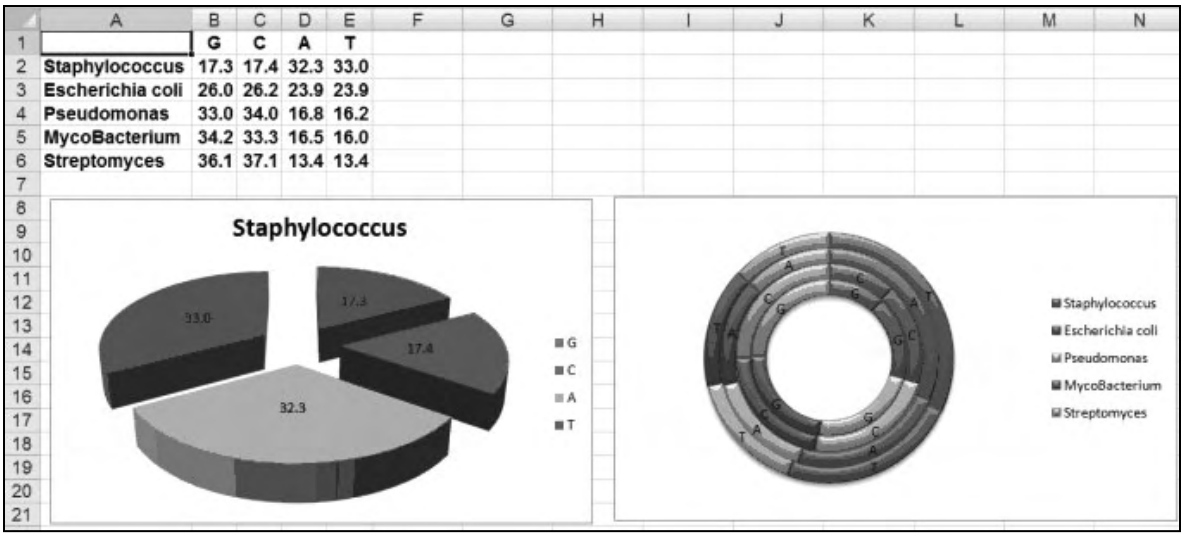
You can open and close gaps by dragging pieces in or out. You need to follow these rules when doing so:

- Click and drag any piece to move all the pieces in or out together.
- Click and drag again to move a specific piece only in or out. (The second click selects one piece only.)
- Click outside any piece to deselect the pieces.
- Click twice on any piece to select that piece specifically.

Here's how you add a Doughnut graph:

1. Click in the table. (Otherwise, you replace the Pie graph.)
2. Select Doughnut under Other Charts.
3. Change the look of the graph by selecting another chart style under Design. (This is available only when you are inside the graph.)
4. If you want data labels,
   - Select the graph.
   - Select Layout, Data Labels.
   - Use the last choice, More, for fine-tuning.
5. To fix the labels, click any label once to select all labels and click any label twice to select one only.

Figure 3.4 lists the DNA components G (guanine), C (cytosine), A (adenine), T (thymine) for some bacteria. The big limitation of Pie graphs is the fact that they cannot display more than one data series. In this case, the data series comes from row 2 rather than from column B because it is more informative. But Excel always takes the series from the longest set of values, which in this case is in the column (with five values), not in the rows (with only four values).



	G	C	A	T
Staphylococcus	17.3	17.4	32.3	33.0
Escherichia coli	26.0	26.2	23.9	23.9
Pseudomonas	33.0	34.0	16.8	16.2
MycoBacterium	34.2	33.3	16.5	16.0
Streptomyces	36.1	37.1	13.4	13.4

Figure: 3.4

When you let Excel decide where to get the data series, it takes its first and only series from column B. That's great if you want to show how the guanine component varies per species—but then a Pie graph would not be very appropriate because this type of graph is best for showing contributions to the total. So you should really use the button Switch Row/Column to make the graph more instructive; that button is located on the Design tab when you are inside the graph.

But what do you do with the other species of bacteria that are not shown in the Pie graph? You have two options if you want to stay in this "food" category of graphs:

- Create five separate Pie graphs for each species of bacteria.
- Create one concentric Doughnut graph for all species together.

Doughnut graphs are difficult to read and interpret, and they are also difficult to fix. (For instance, if you want to display the series values, you must first activate the labels and then select them for each category before you can change them.) So let's go for a better type!

Figure 3.5 shows an example of Column and Bar graphs. Column and Bar graphs are best for comparing values across categories. At any time, you can change the type of the left or right graph by clicking the button Change Chart Type. In addition, you can exchange the axes of the left or right graph by using the Row/Column Switch button.

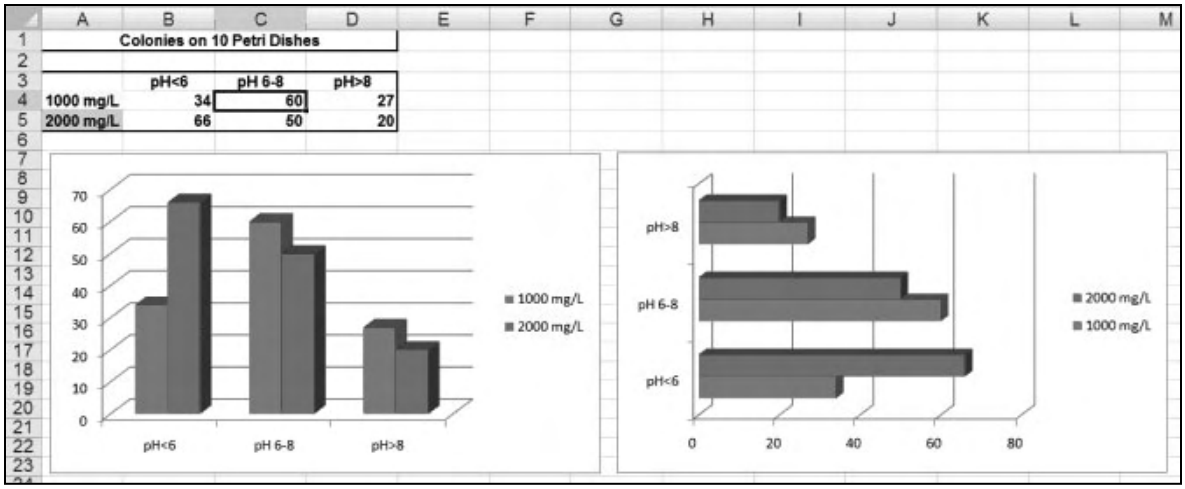


Figure: 3.5

Figure 3.6 shows data plotted in Stacked graphs, which are subtypes of the Column type. *Stacked* means that they resemble a Pie graph in showing contributions to a total: either in values (see the right graph) or in percentages (see the left graph). Which one is best? It depends on your needs or on your hypothesis! The left graph in Figure 3.6, for instance, shows that almost all Ca-ions are in the blood, not in the cells. The right graph, on the other hand, shows that there is a relatively low concentration of Ca-ions anyway.

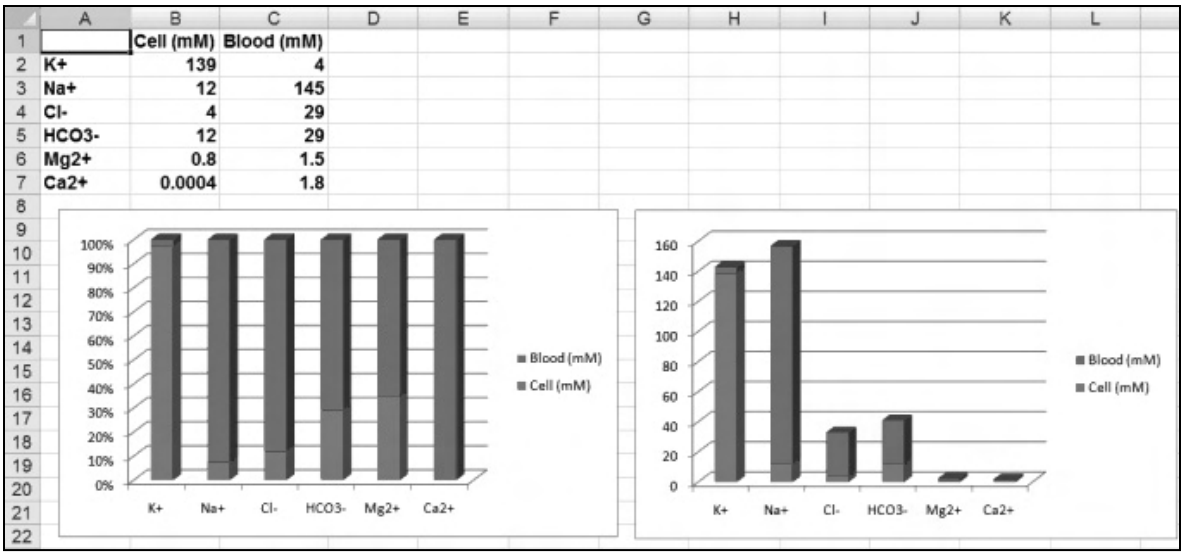
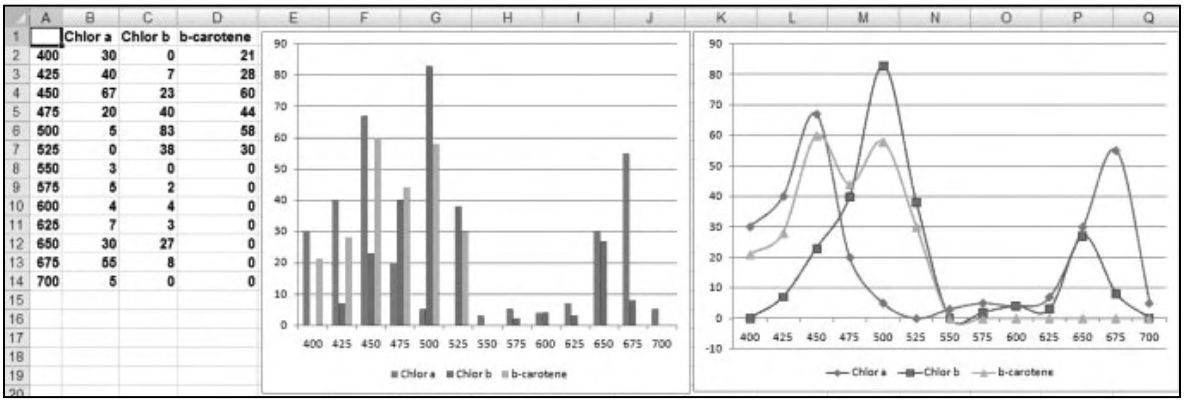


Figure: 3.6

Figure 3.7 compares the photosynthetic activity of three different pigments. Sometimes, as you can see in this case, Line graphs do a better job than Column graphs. Let us experiment how the data can be represented:

1. Create the Column graph first, but move its legend to the bottom. You can drag it around in the plot area, but to move it outside the plot area, you must use the Legend button on the tab.
2. Create a Line graph with markers. Use the first type shown under its dropdown button (because the other two are subtypes of stacked graphs).
3. For each series, create a smoother line. You do this by right-clicking and then selecting Format Data Series and then Line Style. (You can keep the Format Data Series box open while selecting the next Series curve.)



	Chlor a	Chlor b	b-carotene
400	30	0	21
425	40	7	28
450	67	23	60
475	20	40	44
500	5	83	58
525	0	38	30
550	3	0	0
575	5	2	0
600	4	4	0
625	7	3	0
650	30	27	0
675	55	8	0
700	5	0	0

Figure: 3.7

> **Note:** In this case, a Line graph is a better bet than a Column graph. There are no clear rules that dictate when to use one type over the other; you simply have to develop a feel for this issue.

Figure 3.8 plots the hemoglobin percentage versus the erythrocyte count of human blood, using the XY or Scatter type—once with and once without connecting lines. Each dot represents one pair of values: an x value and a y value. So you are dealing here with one series, but the series contains pairs of values this time. The right graph (with connecting lines) may look a little better after column A has been sorted. But still, you probably prefer the left graph; it would be even better if it were outfitted with a regression line (as discussed in Part 4). That's why the XY type is also called the Scatter type.

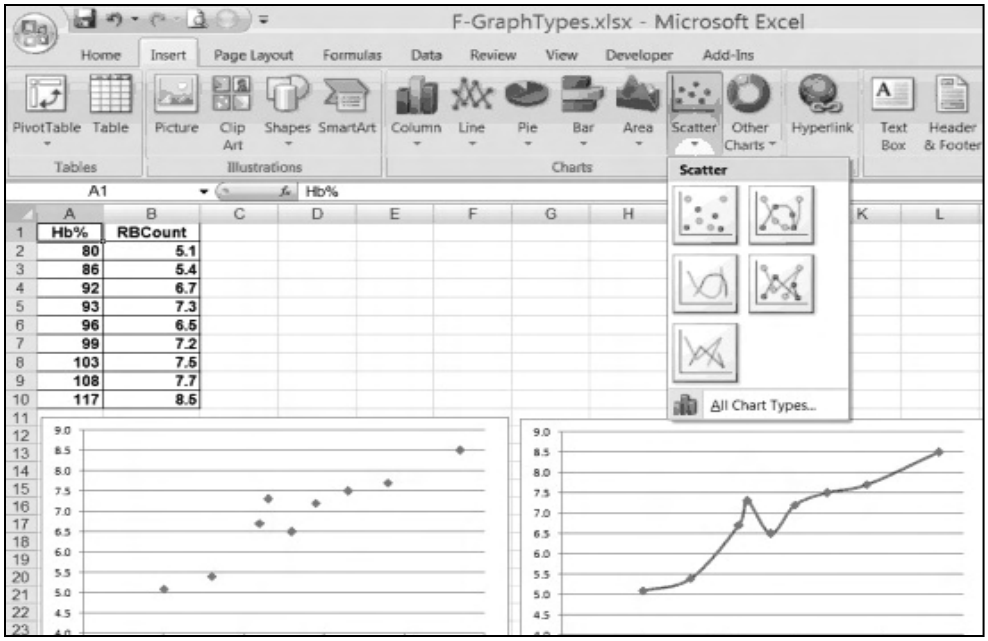


Figure: 3.8

Figure 3.9 makes a comparison between Line graphs (on the left) and XY graphs (on the right). They are almost identical when the intervals in column A are equal (the two top graphs). But when the intervals are unequal, these two types paint very different pictures (the two bottom graphs). The reason for this difference is simple: Line graphs use categories on the horizontal axis (they act like labels, even if they are numeric), whereas XY graphs use real values on the horizontal axis (so they acknowledge the distances in between).

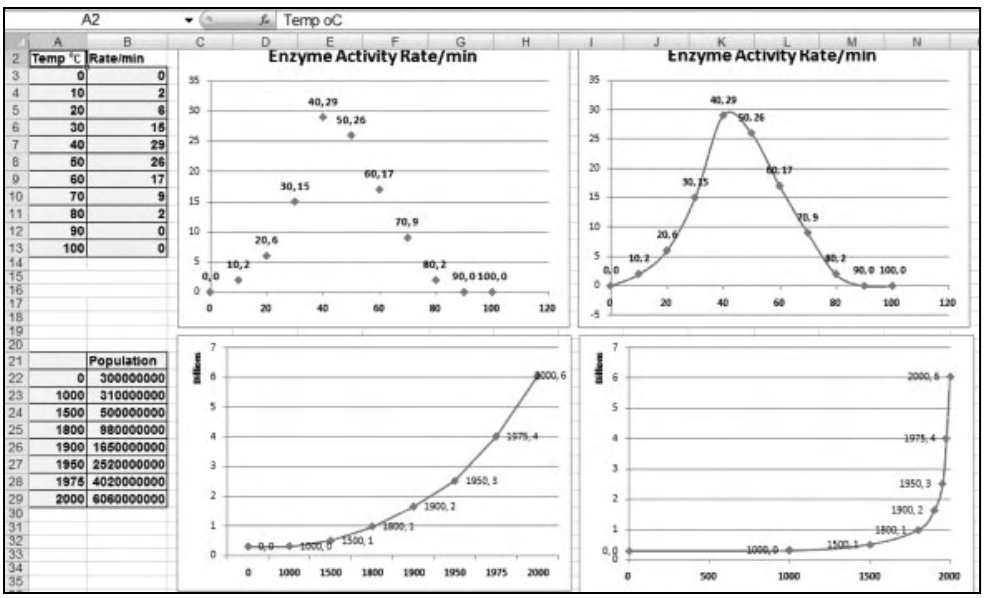


Figure: 3.9

Figure 3.10 measures how a person's eyesight changes with age. The left graph shows how the near point (NP) changes with age. The right graph shows how the focal power (=1/NP) changes with age. You can see that both graphs require the XY type because the values in column A have unequal intervals. Here's how you create the left graph:

1. Select A1:A11 and B1:B11.
2. Select Insert, Scatter.

Here's how you create the right graph:

1. Select A1:A11 and C1:C11.
2. Select Insert, Scatter.

Here's how you add data labels:

1. Select the Layout tab (while in the graph).
2. Select Data Labels, More Data Label Options.

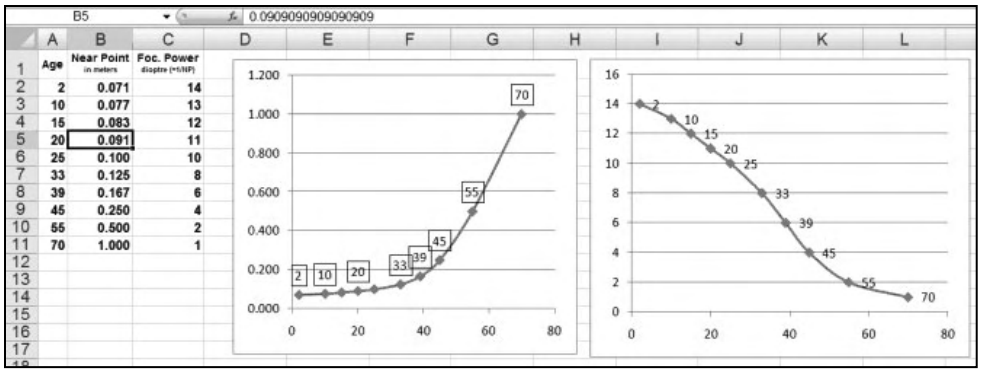


Figure: 3.10

Figure 3.11 presents a unique type of graph: the Radar type. It is basically a subtype of the Line graph, but it works well for very specific occasions—such as the one shown in this figure. A Radar graph shows clustering well (left graph), and it can also be good for cyclic events that have a particular pattern. Remember, though, that Radar graphs are basically Line graphs, so they work with categories. If you use numbers in the first column, you need to make sure to create equal intervals.

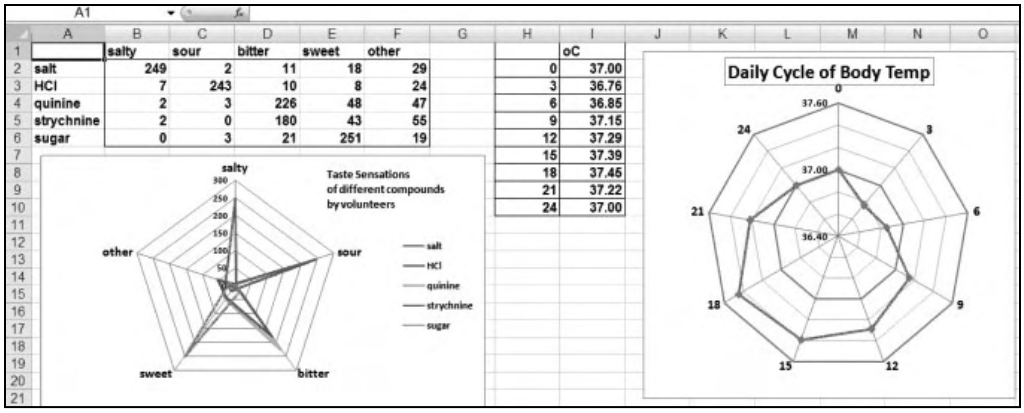


Figure: 3.11

* * *

Chapter 22

A Graph's Data Source

Excel calls the data that a graph is based on the *data source*. Very often, you need to manipulate a graph's data source sometime after you create the graph—in order to correct, add, or remove series or categories. You can do this easily.

When your data source is huge—which is often the case—your graph may get overloaded by the details of the data. One of the simplest ways to eliminate details is to hide certain rows or columns in the table because anything that is hidden in the table is also hidden in the graph. The disadvantage of hiding rows or columns is that the table is affected as well. There must be a way to get the same results just for the graph—and that's where the data source comes in. That is where you can remove any series from the graph alone.

Figure 3.12 has a large table that was originally all part of the graph—until most of the series were manually removed in the graph itself. You remove series from a graph by using these steps:

1. Click inside the table and select Insert, Area with 3D-Area. (Area with 3D-Area is basically a subtype of a Line graph.)
2. Select Switch Row/Column.
3. From the Design tab, select Data.

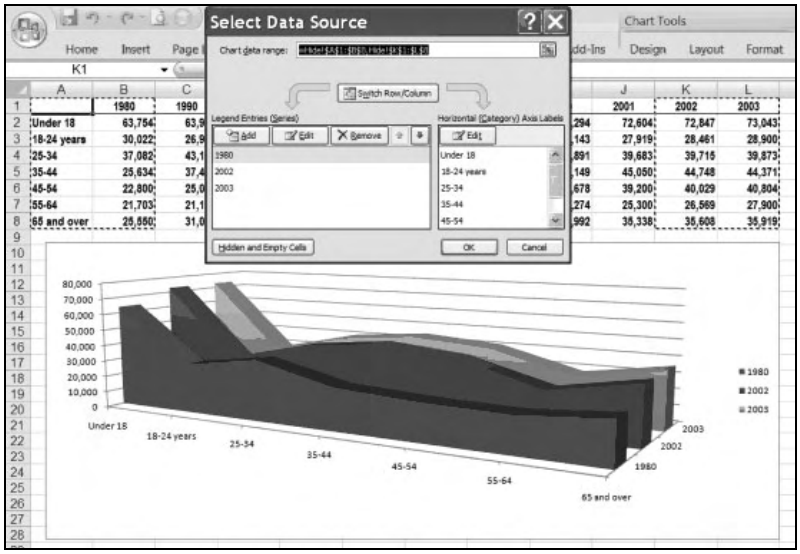
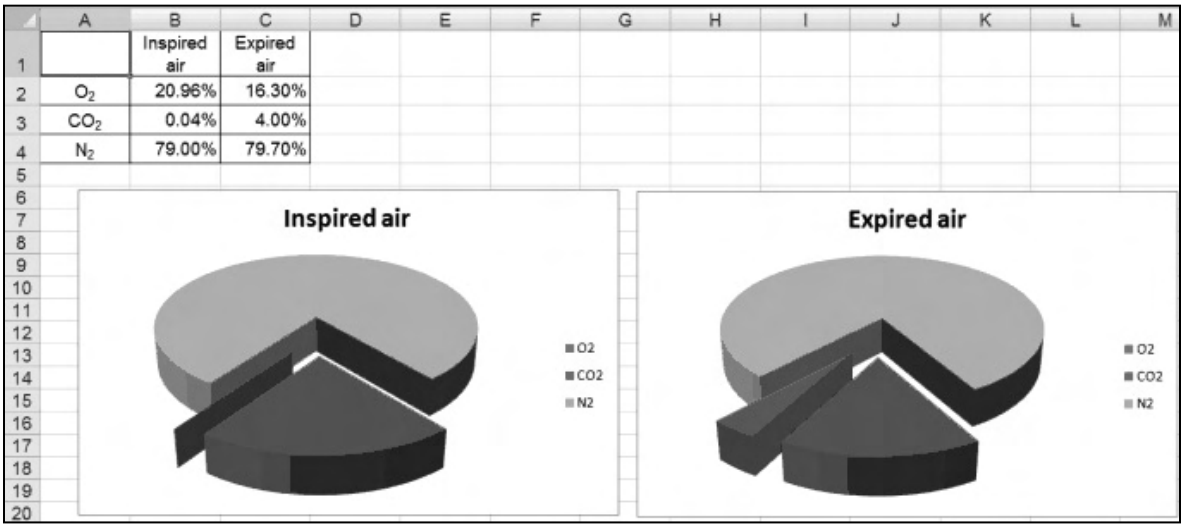


Figure: 3.12

4. In the left section of the box, start removing all unwanted series. It is wise to start with the last one. To remove series, click Remove repeatedly until you have a few series left; alternatively, you can select each series in the graph and then press the Delete key. When you are done, the graph has its own number of series, distinct from the number in the table.

Note: Also notice that when you select a series in the graph, its corresponding table section becomes highlighted automatically.

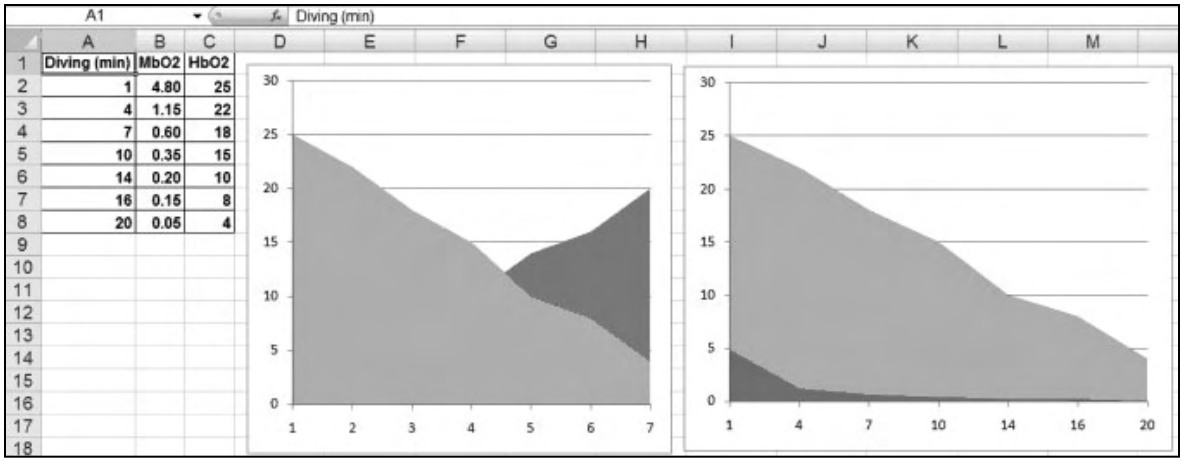
Figure 3.13 uses two Pie graphs to display the composition of inspired and expired air in human respiration. Both graphs are based on the same table, so their data source includes both series, but each graph can display only one of the series, while the other one is hidden. All you have to do is remove one series from each graph to make it distinct and appropriate.



	Inspired air	Expired air
O_2	20.96%	16.30%
CO_2	0.04%	4.00%
N_2	79.00%	79.70%

Figure: 3.13

Figure 3.14 shows a decline in HbO_2 (in the blood) and MbO_2 (in the muscles) with prolonged diving time. Both graphs are of the regular Area type, but the left graph was created automatically, whereas the right graph is a manually adjusted version. In creating the graph on the left, the wizard assumed that the first column is a data series and not a set of



Diving (min)	MbO2	HbO2
1	4.80	25
4	1.15	22
7	0.60	18
10	0.35	15
14	0.20	10
16	0.15	8
20	0.05	4

Figure: 3.14

categories, so it made a graph with three data series and then created its own set of categories (labeled 1 to 7). To create the right graph, you make the following changes:

1. Open the Select Data dialog box.
2. In the left panel, delete the first data series (because those values are really categories).
3. In the right panel, replace the categories the wizard created with the categories from the first column. Click the Edit button and then drag across the cells in Column A.
4. To ensure that the last series does not partially hide the series behind it, change their order by using the appropriate button.

> **Note:** The Area type may not be the best choice here because the intervals between the categories are not completely equal.

Figure 3.15 shows a complicated situation. Suppose you want only the highlighted table sections to be plotted. The wizard cannot figure this out by itself, so you need to start from scratch:

1. Click in neutral territory—nowhere near the table—and choose a regular Line graph. Excel adds a blank graph frame to your sheet.
2. Use the Select Data dialog to add three series manually.
3. Add the categories from the first column.
4. Notice the dips where the series has values that are missing. Replace those `NA` entries in the table with a formula based on the function `NA`: `=NA()`.

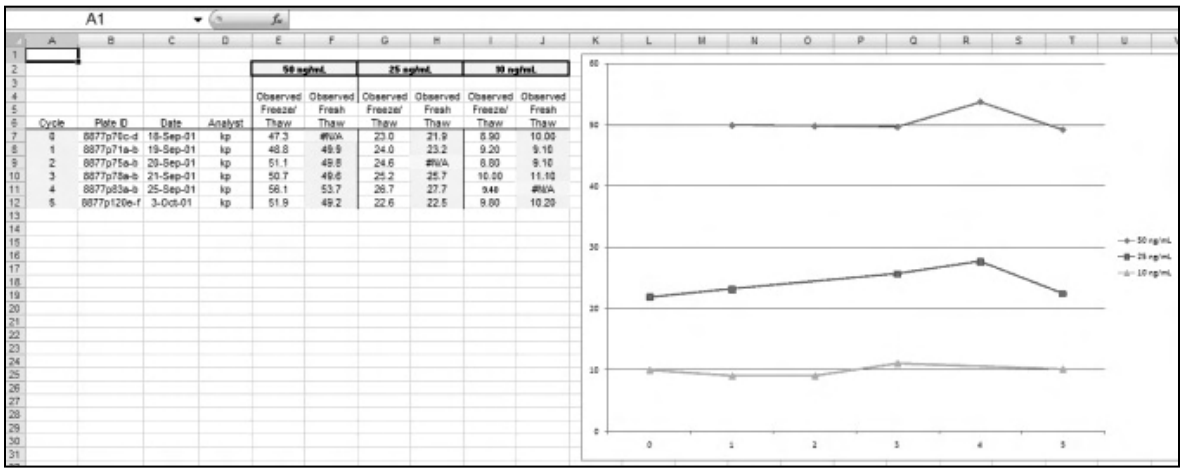


Figure: 3.15

> **Note**: The advantage of using the `NA` function is that the graph does not plot these values as zeros. Changing the graph type to no markers would hide those `#NA` values completely, but this is not good policy in science because each marker is supposed to represent a real observation.

By now, you may have discovered why some tables cause so much trouble: The culprit is often the header, caption, or label that appears on top of the first column! The problem is that the Chart Wizard uses the following built-in rules:

- The longest set of values is considered a series (either by row or by column).
- Each longest set of values with headers becomes a series (even if they are meant to be categories).

In other words, if you want a graph with categories (that is, Line, Column, Bar, or Area graphs), you should not give the category column (or row) a header. If you prefer to include a header, you have to correct the problems manually through the Select Data dialog box. Or you type the first column header after (!) creating the graph.

* * *

Chapter 23

Combining Graph Types

It is possible—and often desirable—to combine different types of graphs into one single graph. It is very common policy, for instance, to add calculations to graphs of observations. However, you don't want calculated values to look the same as observed values. To visually separate the two types of values, you may have to combine two different graph types.

You cannot just combine any graph types. There are a few important rules. The most important rule is that the axes of the graph types you want to combine should not conflict with each other. In other words, you cannot combine a graph based on two value axes (such as XY) with a graph (such as Line, Bar, Column, or Area) that has two very different axes—such as a category axis and a value axis. Another axis-related rule is that you need to be careful in mixing regular Stacked subtypes with 100% Stacked subtypes because they may fight each other.

Figure 3.16 shows a situation in which calculated values are mixed with observed values. To create a graph that looks like the one shown here, you follow these steps:

1. Calculate the mean per day in column F.
2. Add the column of means to the graph either by using the Select Data dialog and adding a new series or by clicking in the plot area and expanding the corresponding hairlines in the table to the next column.
3. Right-click one of the columns of the new series of means and select Change Series Chart

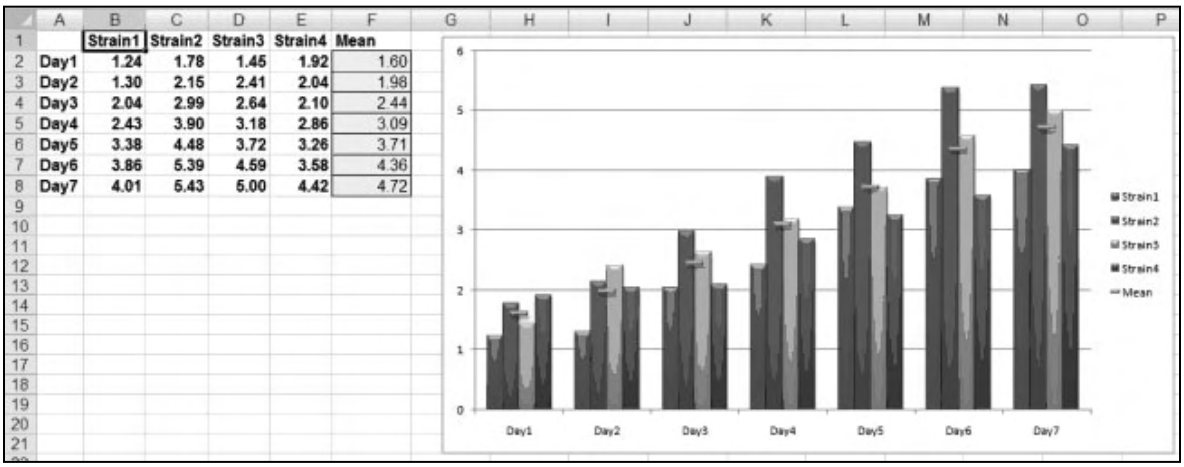


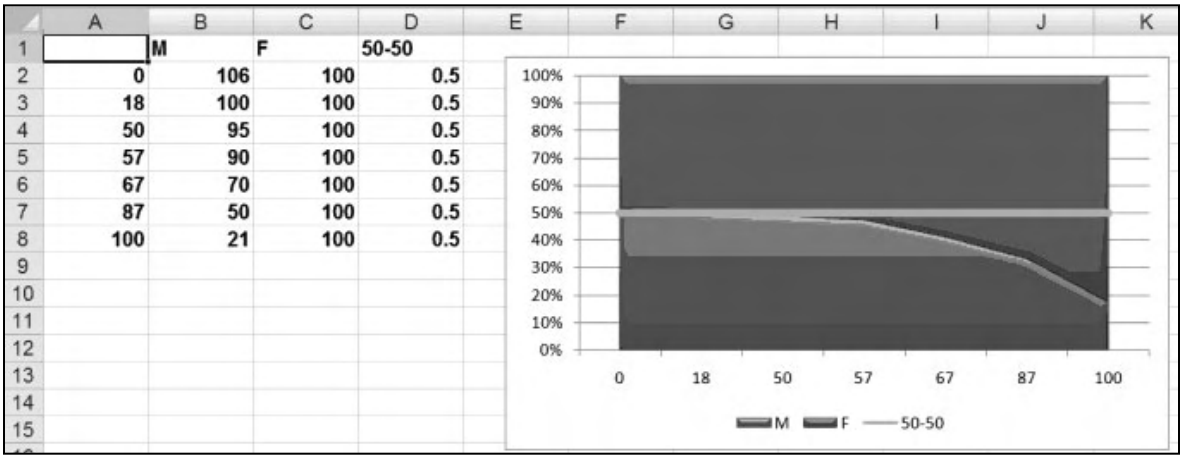
Figure: 3.16

Type. (One of the choices is a Line type, but don't ever try an XY type.)

4. Lines usually suggest a development in time, so if you want to, take a line itself out and change its markers. You do this by right-clicking the line, selecting Format Series, and then selecting No Line under Line Color and Built-in under Marker Options.

Figure 3.17 shows how the sex ratio changes in a human population with increasing age. Instead of an XY graph, this example uses an Area graph. To demarcate the 50% division line, the table includes column D. These are the steps to take to create this graph:

1. Click inside the table.
2. Insert a 100% Stacked Area type graph. The series of column D appears as a line stacked way on top.
3. Right-click the series on top and change its type to Line. Another way of selecting a specific series is selecting it from the Layout tab: the dropdown in its left top corner allows you to select any series.
4. Use Chart Styles from the Design tab to change the appearance, as desired.



	A	B	C	D
1		M	F	50-50
2	0	106	100	0.5
3	18	100	100	0.5
4	50	95	100	0.5
5	57	90	100	0.5
6	67	70	100	0.5
7	87	50	100	0.5
8	100	21	100	0.5

Figure: 3.17

Figure 3.18 shows a graph that is often called a *histogram*. It is a combination of two types: Column and Line. To create a histogram, you could take the following steps:

1. Insert a Column graph.
2. Remove bins as a series, by using the Remove button in the Select Data Source dialog box.
3. Edit the categories so they show the bin values by using the Edit button.
4. Add the frequency values as a new series by using the Add button.
5. Change the second series into a Line or Area graph.

6. Make the graph's line style smooth by selecting Smoothed line in the Format Data Series dialog.
7. Give the first series a smaller gap width through Format Data Series dialog (accessible with a right-click).

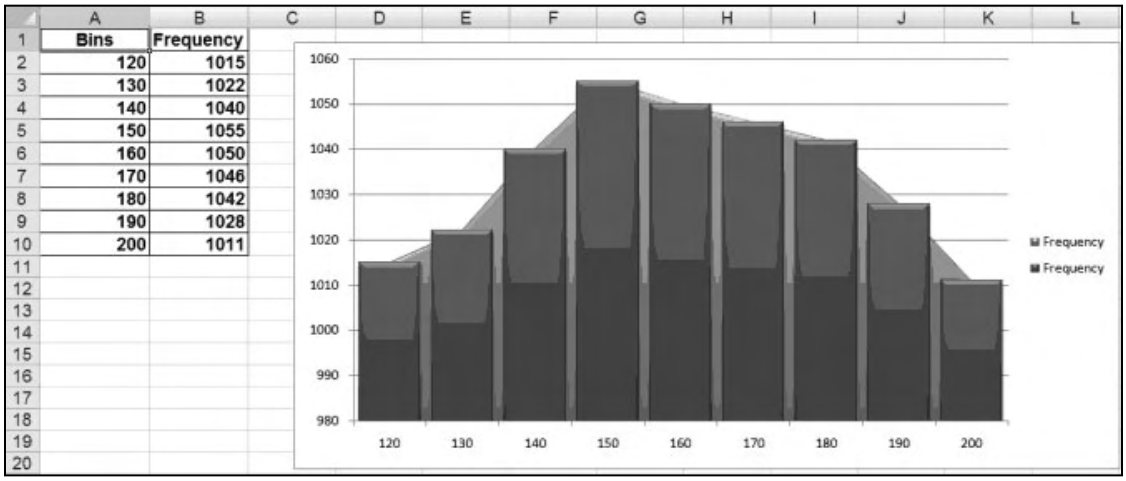


Figure: 3.18

Figure 3.19 shows a *comparative histogram*. Its secret is negative values in the first series. Here's how you make it:

1. Create a Stacked Bar graph.
2. Move the legend to the bottom.
3. Close the gap between the bars.
4. Right-click the vertical axis, and select Format Axis, Axis Options, and set Axis Labels to Low.
5. If you don't like the negative percentages on the (horizontal) value axis, right-click the horizontal axis and select Format Axis. Then, for Number, select Custom, and for Type, select 0%;0%;0%. Finally, click Add.

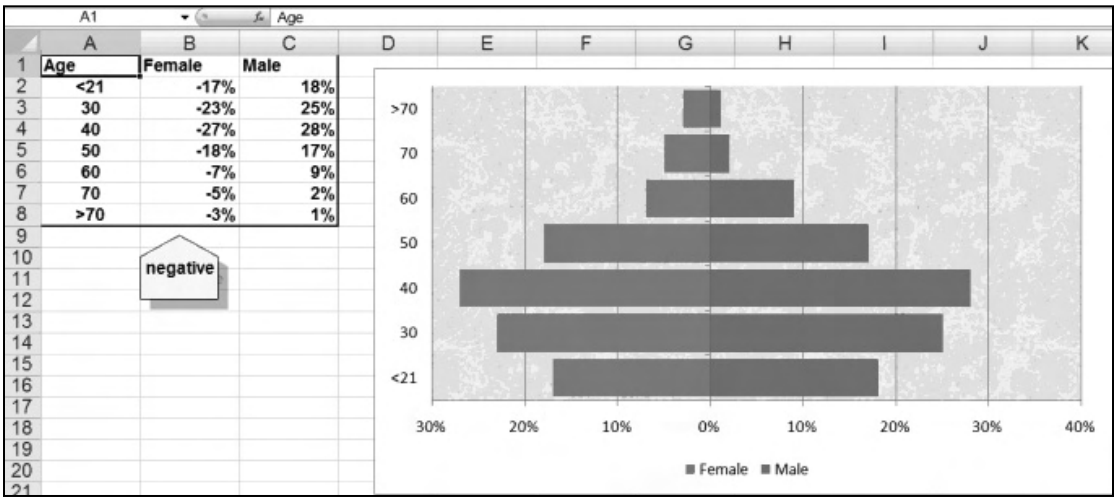


Figure: 3.19

* * *

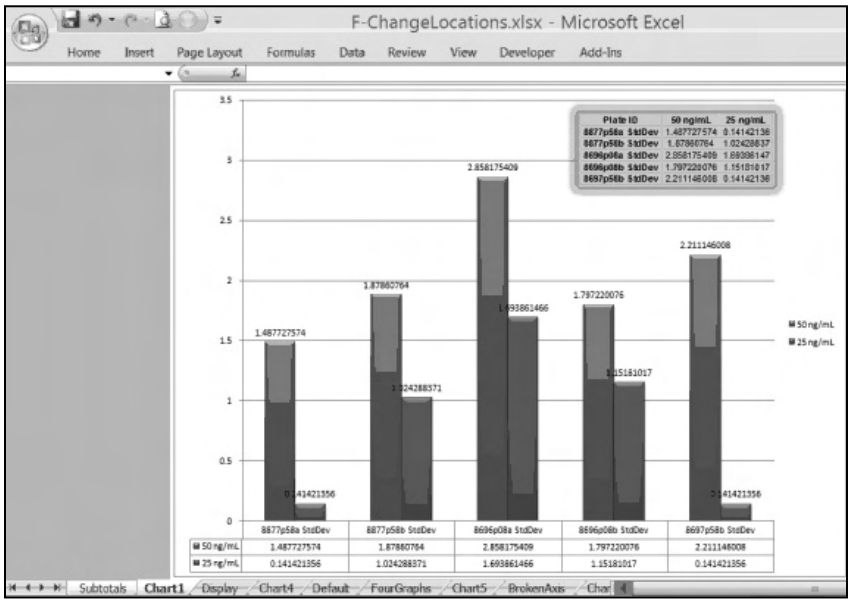
Chapter 24

Changing Graph Locations

So far in this part, in most cases, a graph has been located on the same sheet as its table. When graphs get bigger or when you want to combine several graphs, it may be better to place them on a separate sheet—called a *chart sheet*. A graph located next to a very detailed table—even if it is a collapsed table—is generally difficult to read and work on, and it takes space away from the table. So you probably want to move it to its own sheet.

> **Note**: When moving graphs, it helps to know two handy shortcuts:
>
> - If you decide to keep a graph on a table sheet, but it's temporarily in the way, you can press Ctrl+6 (yes 6, not F6) to toggle from visible to hidden graphs and reversed. This shortcut does not work on a separate chart sheet.
> - You can press F11 to automatically create a default graph on a separate sheet. Chapter 25 discusses default graphs in more detail.

Figure 3.20 is an example in which you would probably want a separate chart sheet. It is based on a very detailed table with a structure of collapsed subtotals (refer to Chapter 7). Here's how you get this result:



1. Insert a Column graph on the same sheet as the table.
2. Click Move Chart on the Design tab and choose a new (separate) sheet.

> **Note**: You can press F11 while in the table to accomplish steps 1 and 2 with one keystroke.

Figure: 3.20

3. Get rid of the grand total by hiding the corresponding row in the table.
4. There is no table in sight, so if you want to add labels in the graph, do so on the Layout tab.
5. To add the original figures to the bottom of the graph in table format, click the Data Table button on the Layout tab.
6. If desired, add a picture of the table. You do so as follows:
    - In the table, you select Copy As Picture (for visible cells only).
    - In the graph, you select Paste As Picture. (Note that pictures don't update.)

Figure 3.21 shows a situation in which you would want to combine several graphs—perhaps to show the data in four different views. You can do this easily, as follows:

1. Click outside the table and press F11. This creates an empty chart sheet as a receptacle.

2. Move the first graph into the empty receptacle by using Move Chart from the Design tab.

3. For each of the other three graphs, click inside the table and click Insert, and choose the type of graph you want.

4. Rearrange the four graphs on the chart sheet, if necessary.

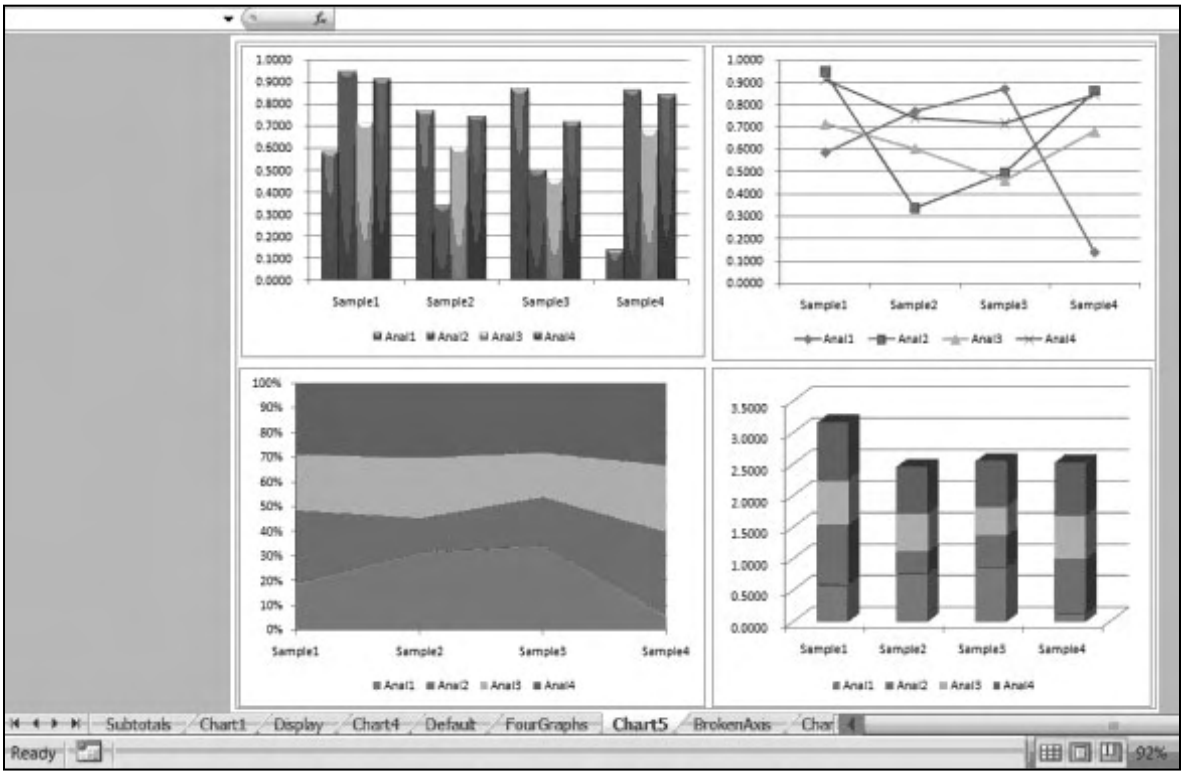


Figure: 3.21

Note: You may want to try to copy and paste the first graph onto this sheet, but don't do it. Or, if you want to see what happens, do try it!

Figure 3.22 shows a situation similar to what Figure 3.21 shows. In this case, the readings before and after treatment are so far apart that it might be better to use a broken axis. Because Excel does not provide such an option, you must instead work with two graphs: one for the top section and another one for the bottom section. Here's how you do it:

1. Click outside the table and press F11.
2. Move two identical graphs into the empty sheet.
3. Set the maximum value for the value axis of the second graph to, say, 5. You do this by right-clicking the value axis, Format Axis, Axis Options, and set the Maximum to 5.
4. Manually fine-tune the graph by moving borders and so on.

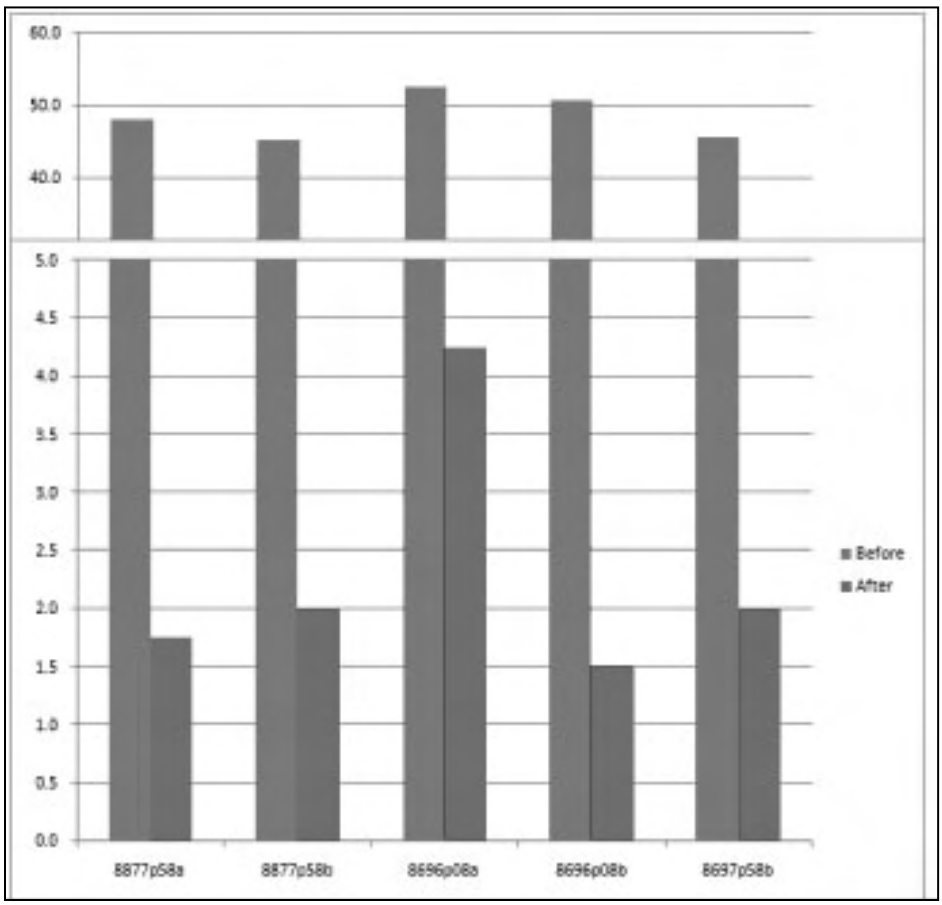


Figure: 3.22

* * *

Chapter 25

Templates and Defaults

If you often use the same type of graph, you can make it your default. If you have a favorite graph type that you have heavily customized to meet your needs, you can save that graph as a chart template (*.crtx) in the charts template folder and then reuse it whenever you need it.

Earlier, we discussed the fact that when you press F11 while in a table, a graph of that table is placed on a separate chart sheet. The graph that appears is of the *default graph type*, which in Excel is initially a Column graph that has a particular Microsoft layout. Go to the Insert tab, use the dropdown of any type of graph, and select the bottom option: All Chart Types. That is where you can see your current default graph highlighted. At the bottom of the box is the Set As Default Chart button, which offers you a chance to make another type the default type. The type you set as your default chart is what F11 will give you next time around.

If you want more than just the "plain" default graph, you can completely customize your own graph templates, like so:

1. Create your favorite graph layout—a specific type with customized axes, colors, and so on.
2. While in the graph, click Save As Template on the Design tab.
3. Store your template in the Chart Templates folder.

You apply the new template as follows:

1. From the Insert Chart Dialog box, choose Templates (in the left panel).
2. If your template is not in the proper folder, you can browse for it by clicking the button at the bottom.
3. If desired, click the button to make your new chart template your default graph.

Figure 3.23 offers an added attraction: pictures as a part of templates. Here's how you use them:

1. Create a Column graph
2. Right-click the series, select Format Data Series, and choose Fill in the Left Panel. From the dialog that appears, make the following selections:
   - Select Picture.

- For the file, select Micro.wmf.
- Choose Stack and Scale.
- In the Apply To box, choose Front.

3. If needed, rotate the graph with the 3D Format option (in the left panel).
4. While in the chart, click Save As Template on the Design tab.
5. Make sure to store your template in the Chart Templates folder.
6. If you want to, make this new template a default graph (so that F11 uses it).

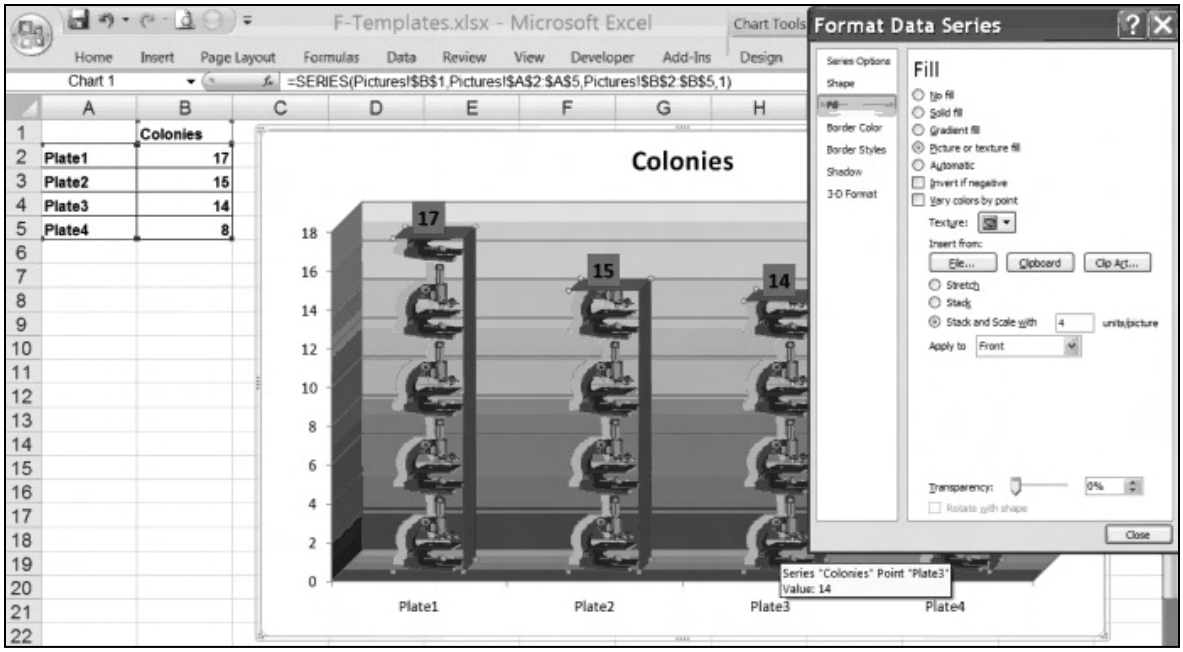


Figure: 3.23

* * *

Chapter 26

Axis Scales

The value axis usually needs more attention than the category axis. You need to decide on the maximum and minimum values, the value steps, the format, and the gridlines. It is detailed work that will pay off in the end when the graph conveys better information. And, of course, you can save your graph settings in a template or make them part of your default graph (refer to Chapter 25).

To work on any axis, you do one of the following:

- Right-click it and choose Format Axis.
- Click the Axes button on the Layout tab.

You manage gridlines in one of two ways:

- By right-clicking the axis and then choosing Add Major and/or Minor Gridlines.
- By clicking the Gridlines button on the Layout tab.

The category axis kicks in automatically, even when there are categories at different levels. Sometimes there are so many details, though, that the category axis does not do a good job. In such cases, you may have to manually decrease the font size and/or decrease the gap between columns.

The value axis, on the other hand, usually requires more attention. The most common procedure on value axes is setting ranges and steps in the Format Axis dialog box, which involves the following:

- To change axis options, you click Fixed and then set Min, Max, Major Unit, and Minor Unit.
- The Minor Unit setting does not kick in until you set its tick mark type.
- Sometimes you can solve a cramped space problem by displaying units in thousands or other large units.

Figure 3.24 shows a relatively complicated problem. To get from the left graph to the right graph takes a few extra steps:

1. Set the horizontal axis and set Axis Labels to High.
2. Select Values in Reverse Order.

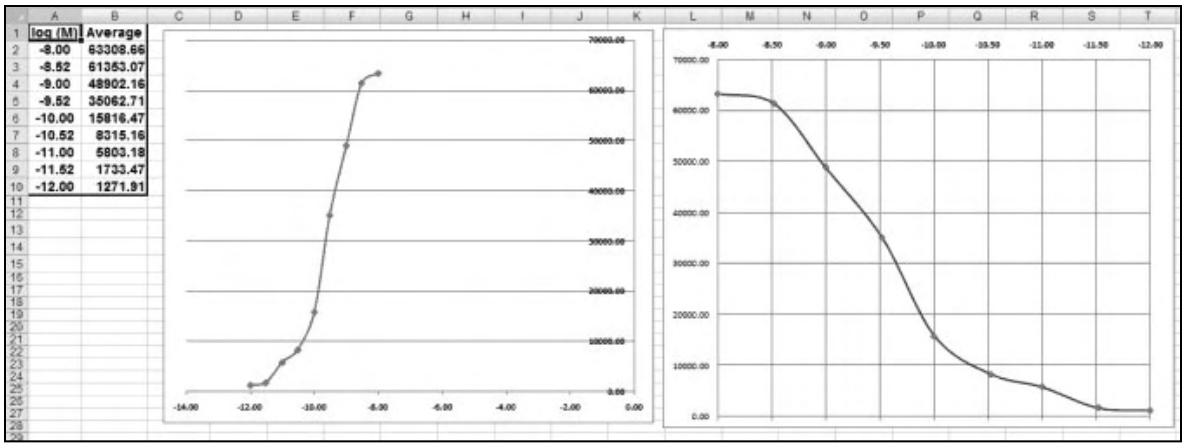


Figure: 3.24

3. Select the vertical axis and set Axis Labels to High.
4. For the top axis, select Add Major Gridlines and set Max to -8.
5. Manually fine-tune the graph as needed.

Figure 3.25 shows two different XY graphs based on the same table. What is the difference between them? You could have used two broken axes because one value is way out in the top-right corner. But in this case, it is probably much better to make both axes logarithmic—by right-clicking the axis, choosing Format Axis, and then selecting Logarithmic Axis. (Part 4 discusses this issue in more detail.) Excel has no simple way of adding the names from column A as labels to the graph. But you can download Rob Bovey's free utility for easily labeling XY points: `www.appspro.com/Utilities/ChartLabeler.htm`. After you install this utility, you can find it in the AddIns menu.

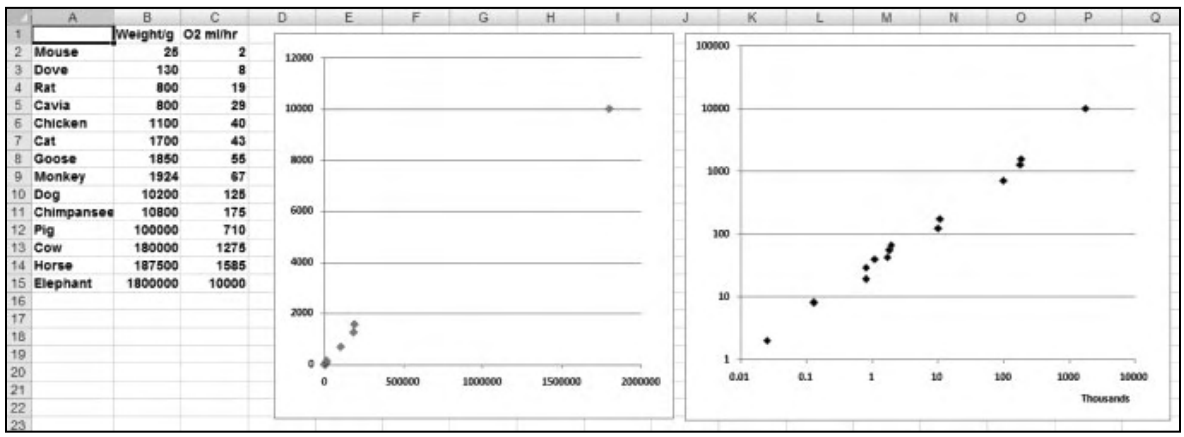
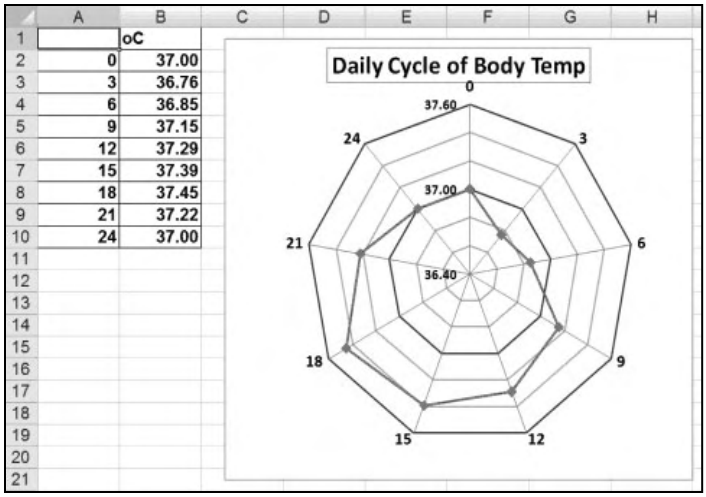


Figure: 3.25



	oC
0	37.00
3	36.76
6	36.85
9	37.15
12	37.29
15	37.39
18	37.45
21	37.22
24	37.00

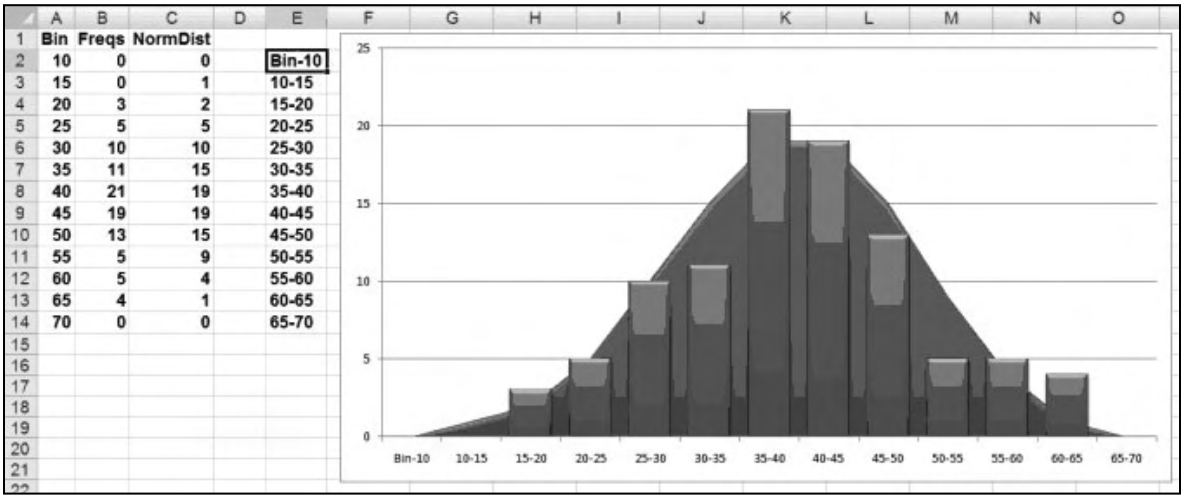
Figure: 3.26

Figure 3.26 shows a Radar graph. You give the 37 degrees marker in this graph extra emphasis as follows:

1. Make it a major gridline (different from the rest) by adjusting four Format Axis options:
   - Set Min Fixed to 36.4.
   - Set Max Fixed to 37.6.
   - Set Major Fixed to 0.6.
   - Set Minor Fixed to 0.2.
2. Add major and minor gridlines by using the Layout menu and clicking the Gridlines option.
3. Give the major gridlines a more pronounced line color and style.

Figure 3.27 tackles another axis issue. In this example, column B is based on the `FREQUENCY` function. Because this function requires numeric values in column A as a top boundary value, you cannot gather from the horizontal axis what the boundaries are for each bin if the categories are based on column A. Because the category axis does not properly describe the bin, you would like 20 to 25 instead of 20. Here's what you do:

1. Create an extra column (in E) and type the following formula in E2: `=A1 & "-" & A2`. (Or you could use the `CONCATENATE` function instead, as described in Chapter 15.)
2. Use the Select Data dialog box to change the category labels (right panel) from column A to column E.



Bin	Freqs	NormDist	Bin-10
10	0	0	
15	0	1	10-15
20	3	2	15-20
25	5	5	20-25
30	10	10	25-30
35	11	15	30-35
40	21	19	35-40
45	19	19	40-45
50	13	15	45-50
55	5	9	50-55
60	5	4	55-60
65	4	1	60-65
70	0	0	65-70

Figure: 3.27

* * *

Chapter 27

More Axes

You are not limited to having only two axes in a graph. First of all, each axis can have a *secondary axis*, in cases when you are dealing with two sets of values of different magnitude; in addition, you can create a third axis on its own.

Let's examine the problem of having values at the extreme ends of a scale. Because Excel does not have a broken axis option, you found a way in Chapter 24 to solve this problem by using two graphs with adjusted axis scales. Another good—and perhaps better—solution might be to give the high (or the low) values their own axis.

Figure 3.28 shows the case you worked on in Chapter 24. But in this case, instead of creating two graphs, you work with one graph and add an extra axis. Day0 has extravagant values that deserve their own secondary axis. This is what you do:

1. Right-click the high series of Day0.
2. Select. Format Data Series.
3. In the Format Data Series dialog box, select Secondary Axis.
4. Change the graph type of this series to Line or Area.

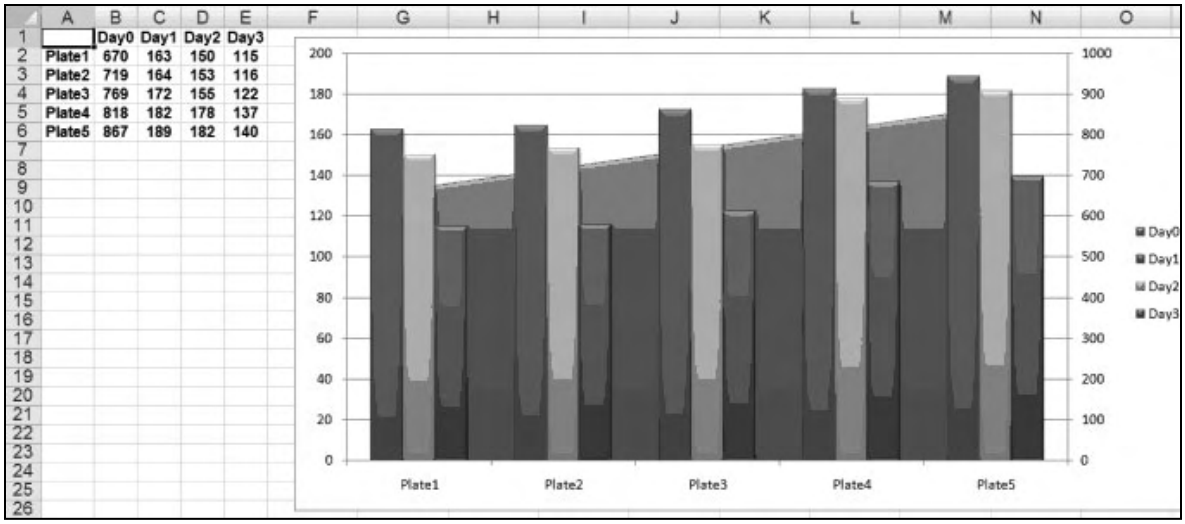


Figure: 3.28

5. Add axis titles if you like by selecting Axis Titles on the Layout tab.
6. Connect each series with its axis by using one of the following methods:
   - Give each series and its axis the same the same font color.
   - Use axis labels.
   - Use arrow shapes to select Insert, Shapes

Figure 3.29 shows a case that deals with extremely high and low values in both dimensions. This example calls for two secondary axes. Although the procedure to do so is a little different this time, using secondary axes may be a good solution for displaying extremely disparate readings in the same graph (instead of using multiple graphs). Here's how you do it:

1. Create two series for your XY graph: A2:B4 for the low end and A5:B7 for the high end.
2. For the second series, select Format Data and then select Secondary Axis. Now you have a secondary vertical axis.
3. Use the Axes button on the Layout tab (which appears only after you have created a secondary axis for the second series) to make a secondary horizontal axis.

In addition to creating secondary axes, you can also create a third axis on its own. A graph with a third axis is also called a 3D graph. But not every 3D graph has a third axis. Most 3D graphs have only a 3D appearance. Remember those beautiful 3D Pie graphs? They look 3D, but they don't have any axis at all. And then there are those great-looking 3D Column, Area, and Surface graphs. But even if a graph seems to have a third axis, that axis is often merely a glorified legend. So does Excel let you add a third value axis? Not really. But it lets you get close by using a Bubble graph. The Bubble type can handle X, Y, and Z values at the same time.

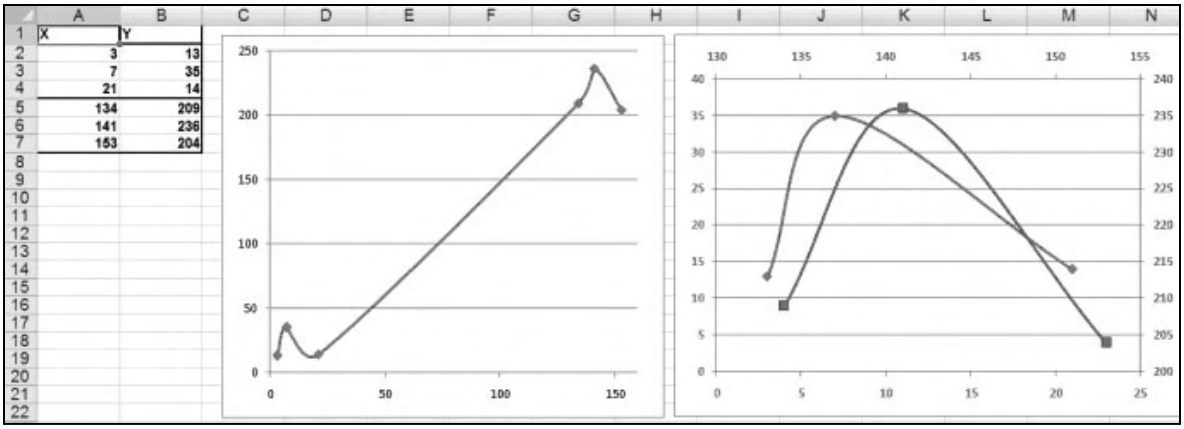


Figure: 3.29

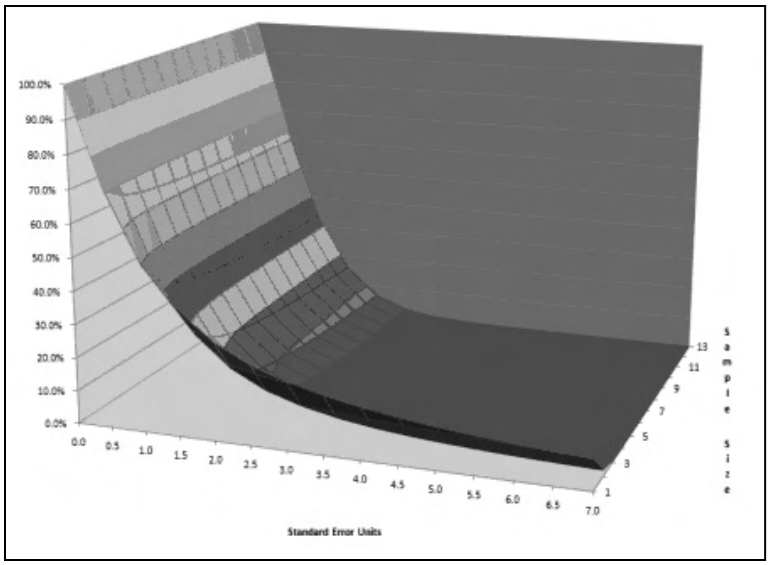


Figure: 3.30

Figure 3.30 plots the probability (according to the Student's t-distribution; see Chapter 46) against the t-value and the sample size. It is a Surface type graph—which is basically a subset of the Line type. Doesn't it look like you have three value axes here? Yes, it does, but two of them are category axes. There is only one value axis here—the vertical one; the other two constitute category axes but happen to have equal intervals. Consequently, the third axis is just a glorified legend.

Figure 3.31, on the other hand, does have something like a third value axis. It is a graph of the Bubble type. Notice that its data source has x values, y values, and z values (called Bubble Size)—in other words, the size of the bubbles represents the z values. It's not really an elegant solution, but it is the closest Excel can come to a "real" 3D graph with a numeric third dimension.

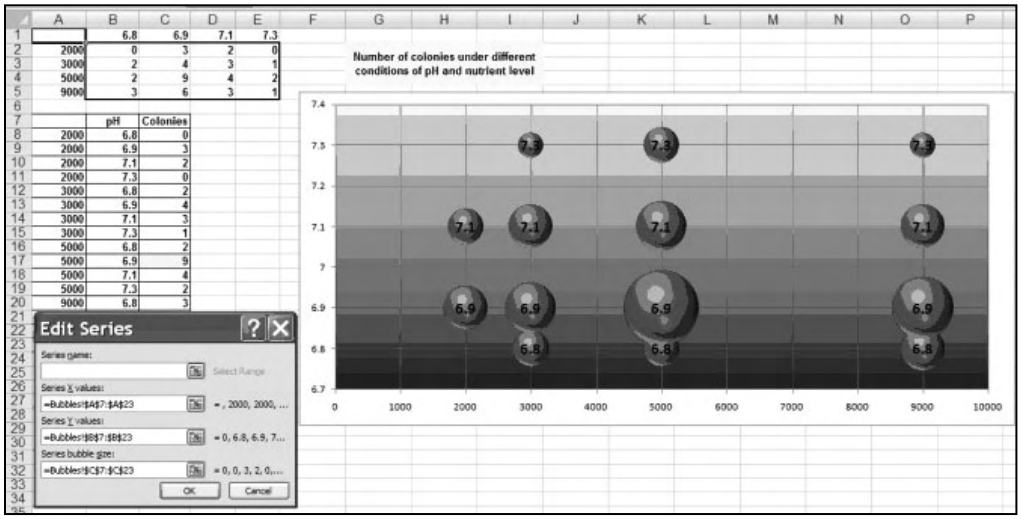


Figure: 3.31

Note: When you create the graph shown in Figure 3.31 on your own, you may have to stretch the vertical axis so the bubbles get more spread out without covering each other.

* * *

Chapter 28

Error Bars

One of the main reasons for adding bars to a scientific graph is to display the standard deviation or standard error. That's why they are often called error bars. Creating error bars is easy, but coming up with the right structure for the graph itself can be difficult. Let's look at the reasons.

Figure 3.32 displays two different ways of using error bars. Both graphs in the figure show the mean monthly temperature for two different locations, plus the standard deviation around the mean through error bars. Yet they are different:

- Using the graph on the left makes more sense if you want to show whether each actual temperature is within one standard deviation. Creating this graph is simple:
  1. Select each series.
  2. Select Layout, Error Bars with SD.
  3. If you want something other than one standard deviation (Excel's default), change it.
  4. Do the formatting through the Format section or by right-clicking the error bar.
- The graph on the right is really rather confusing because it suggests that the standard

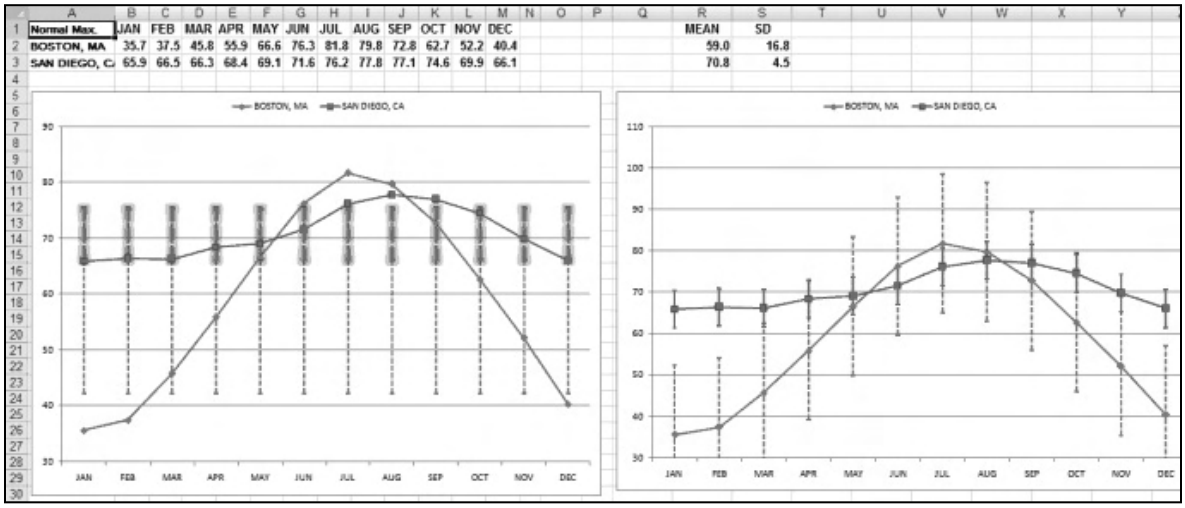


Figure: 3.32

deviation range is specific for each temperature. You would create the graph on the right manually, based on results from calculations in column S:

1. Select Layout, Error Bars, More Error Bar Options.
2. In the Format Error Bars dialog that appears, select Both for the direction, then select the Specify Value button next to Custom, and set both positive and negative values to either cell S2 or S3 (they both have a calculated standard deviation).

It is obvious that the standard deviation in the graph on the right should not move up or down with the temperature moving up or down. However, the technique applied to this graph might be better or even necessary for other occasions.

Figure 3.33 shows two graphs: The graph on the right shows the standard deviation range for each individual strain. The graph on the left was created automatically and doesn't really make sense in this situation. So here's how you create the graph on the right:

1. Start a Column graph based on the range A1:E4 only.
2. To display the standard deviation range for each individual strain, select Switch Row/ Column.
3. Change the mean into a Line graph.
4. Select No Line as the format and format the marker to your liking.
5. Add custom bars to the mean like we discussed earlier.
6. Specify the custom error bar values by setting the Positive and Negative Error value to F2:F4.

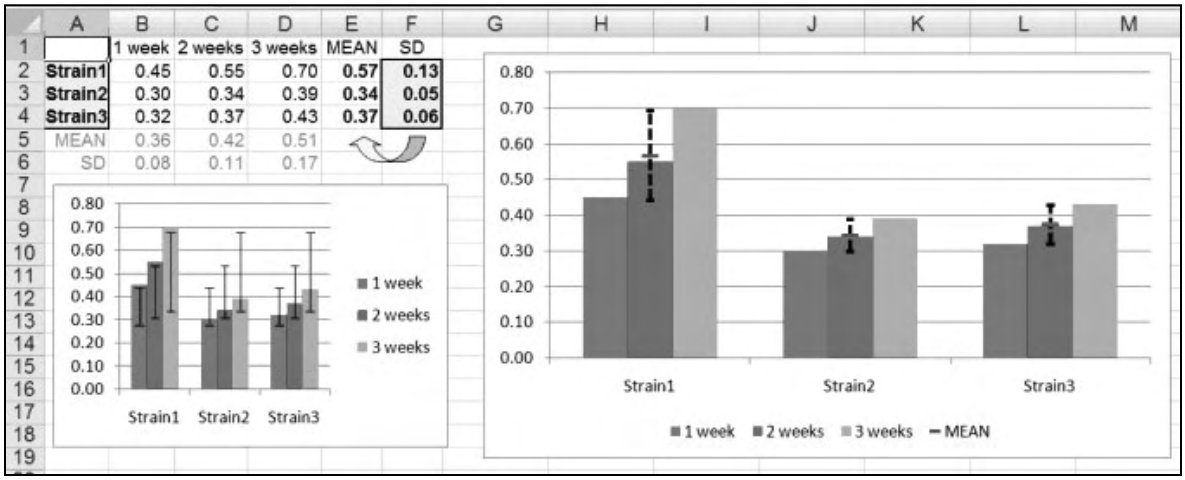
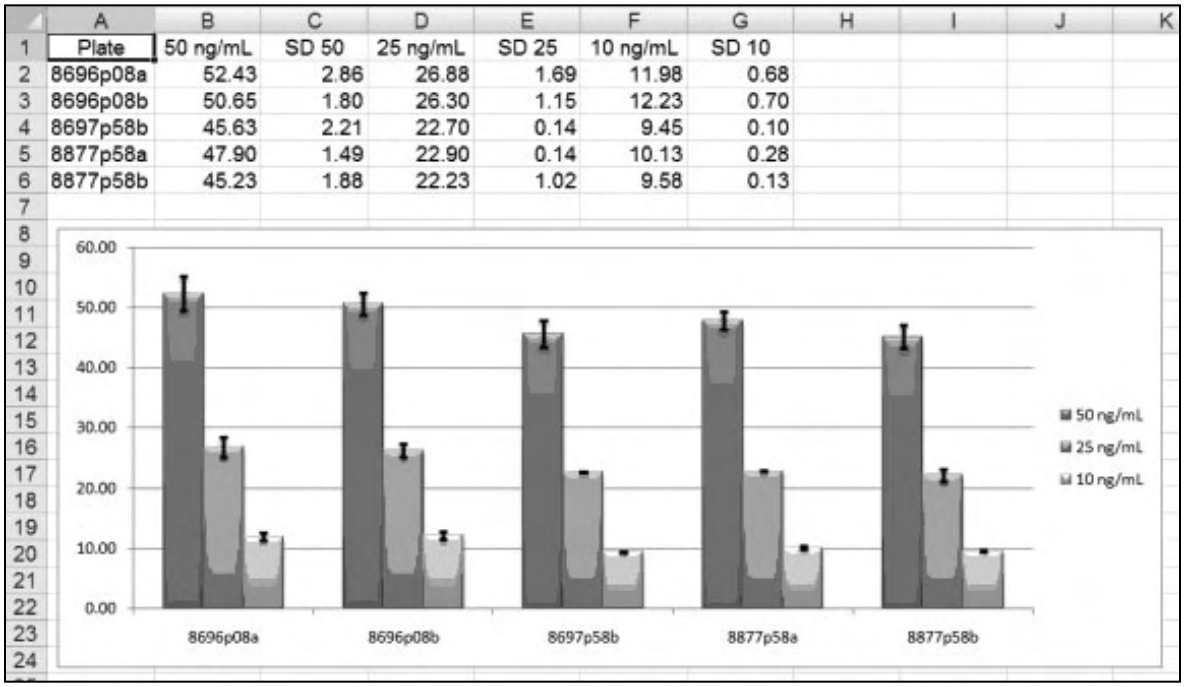


Figure: 3.33

Note: If you hadn't created your own mean and standard deviation, a click on error bars would have given you the graph to the left, which is actually the same result as if you had based it on range A5:D6.

Figure 3.34 shows a summary of repeated measurements on certain plates. You can only make these error bars manually:

1. Insert a Column graph.
2. Select Switch Row/Column and then remove the standard deviations from the data source.
3. Start adding standard deviations: Click on the first series/Layout tab/Error Bars/Error Bars/ More Error bar Options…/Custom/Click Specify Value, then Drag across C2:C6 for both + and -.



	A	B	C	D	E	F	G
1	Plate	50 ng/mL	SD 50	25 ng/mL	SD 25	10 ng/mL	SD 10
2	8696p08a	52.43	2.86	26.88	1.69	11.98	0.68
3	8696p08b	50.65	1.80	26.30	1.15	12.23	0.70
4	8697p58b	45.63	2.21	22.70	0.14	9.45	0.10
5	8877p58a	47.90	1.49	22.90	0.14	10.13	0.28
6	8877p58b	45.23	1.88	22.23	1.02	9.58	0.13

Figure: 3.34

* * *

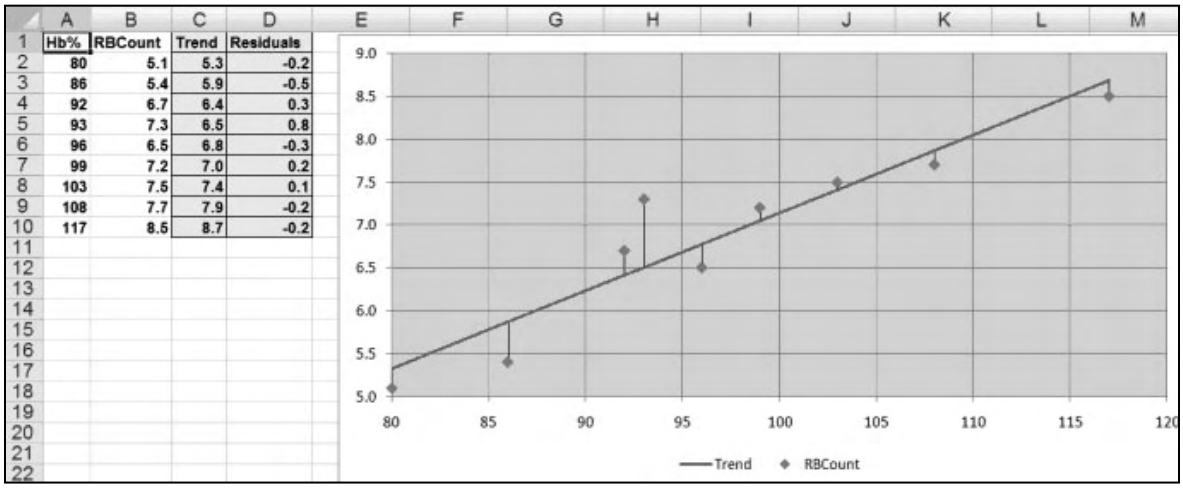
Chapter 29

More Bars

You can use the bars that Excel calls error bars for many more purposes than just indicating statistical error ranges. You can also apply them to anything that needs to be offset against certain values, such as residuals, percentiles, and drop lines.

Figure 3.35 plots the relationship between the hemoglobin percentage and the erythrocytes count in a Scatter graph. Column C predicts or estimates the red blood cell count, as if there were a linear relationship between the count and the hemoglobin percentage. (Part 4 talks about this further.) This calculation is done with the multiple-cell array function `TREND` (which you use by selecting multiple cells and then pressing Ctrl+Shift+Enter; see Chapters 14 and 17). Column D calculates how far off the predictions are—that is, the *residuals*—using the formula `=B2-C2` (that is, observed-minus-predicted). Now you can add the residuals to the graph by using error bars:

1. Select the trend series in the graph, Choose Error Bars from the Layout menu, and then MoreError Bar Options.
2. Choose ◉ Plus (only), then ◉ Custom, click on Specify Value.
3. D2:D10 is the range for both positive and negative values, and click OK.
4. Click the horizontal bars (which you do not need) and then press Delete.



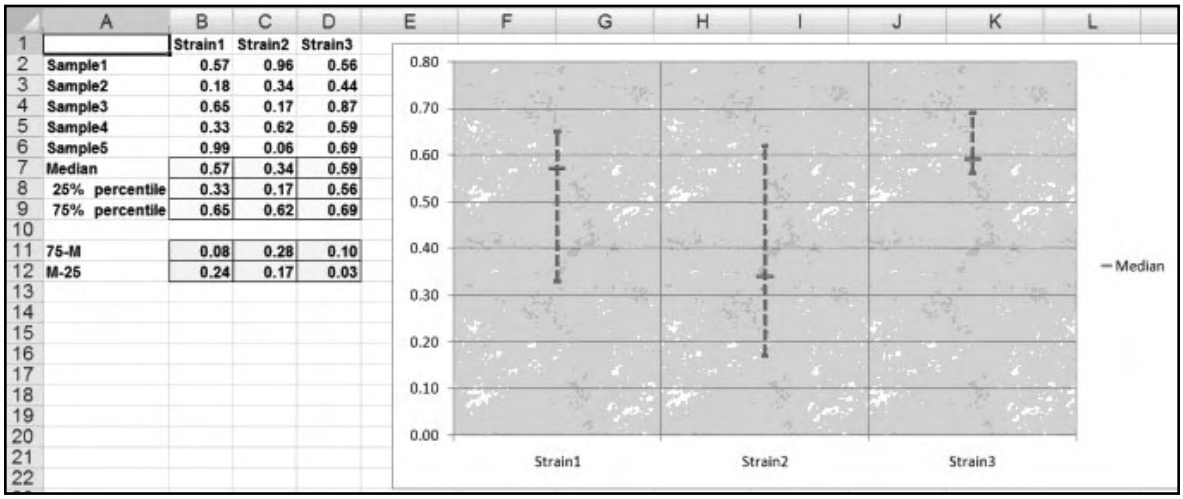
	A	B	C	D
1	Hb%	RBCount	Trend	Residuals
2	80	5.1	5.3	-0.2
3	86	5.4	5.9	-0.5
4	92	6.7	6.4	0.3
5	93	7.3	6.5	0.8
6	96	6.5	6.8	-0.3
7	99	7.2	7.0	0.2
8	103	7.5	7.4	0.1
9	108	7.7	7.9	-0.2
10	117	8.5	8.7	-0.2

Figure: 3.35

> **Note:** Instead of taking steps 1–3, you could choose the series RB-Count and then select Minus (only).

Figure 3.36 plots the median per strain plus a range between the 25th and 75th percentiles. You create these error bars as percentile bars by using intermediate calculations, as follows:

1. Calculate the values in B7:D9. The formula in B7 would be: =MEDIAN(B2:B6); in B8: =PERCENTILE(B2:B6,0.25); in B9: =PERCENTILE(B2:B6,0.75).
2. Do the following intermediate calculations in B11:D12:
   - The difference between the 75th percentile and the median is for the plus section of the error bars.
   - The difference between the median and the 25th percentile is for the minus section of the error bars.



	Strain1	Strain2	Strain3
Sample1	0.57	0.96	0.56
Sample2	0.18	0.34	0.44
Sample3	0.65	0.17	0.87
Sample4	0.33	0.62	0.59
Sample5	0.99	0.06	0.69
Median	0.57	0.34	0.59
25% percentile	0.33	0.17	0.56
75% percentile	0.65	0.62	0.69
75-M	0.08	0.28	0.10
M-25	0.24	0.17	0.03

Figure: 3.36

Figure 3.37 uses similar tricks—this time applied to a Stacked Bar graph:

1. Do the intermediate calculations in G8:K11. For example, in cells H8 use the formula `=H2-G2`.
2. For `Series1`, select No Fill and No Line.

> **Note :** If you have to work on an "invisible" component later on, select that component from the top-left drop-down button on the Layout tab and click Format Selection.

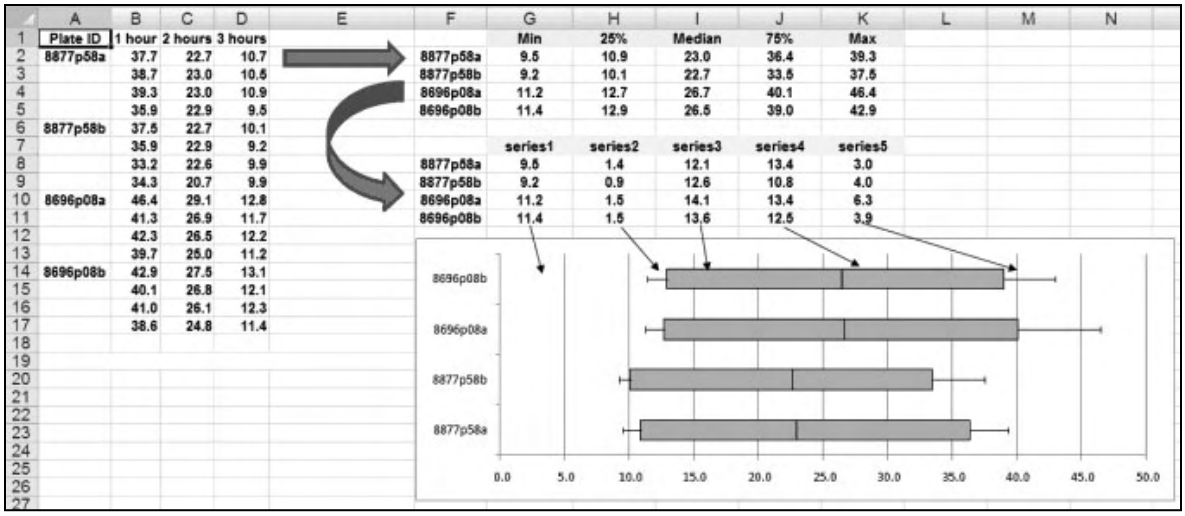


Figure: 3.37

3. For `Series2`, select No Fill and No Line. Also add an error bar with both Minus and Percentage set to 100%.
4. For `Series3` and `Series4`, add a border. Give Series4 the same fill color as `Series3`.
5. For `Series5`, select No Fill and set the Minus Bar to 100%.

> **Note:** Another option for `Series5` is to use `Series4` instead and select Plus Bar and K8:K11.

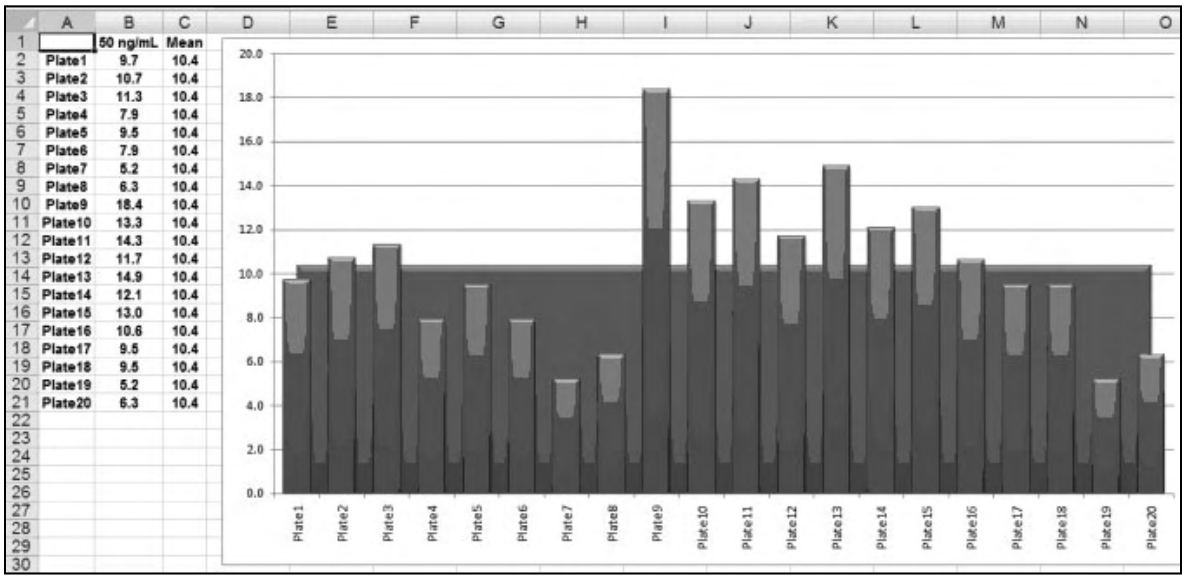
* * *

Chapter 30

Line Markers

You can outfit graphs with extra lines or markers in order to demarcate specific graph sections, locate means, designate quality control limits, and so on. These line markers can dramatically enhance the functionality of your graphs. When you know how to create them, you can make them work to your benefit.

Figure 3.38 includes a dynamic marker for the mean of all the readings. The "secret" series behind this Area graph is located in column C. Because this graph is of the Line and Column type, you need a mean value for each category. That's the price you pay. But it is clear; you could not receive this result with errors bars.



	50 ng/mL	Mean
Plate1	9.7	10.4
Plate2	10.7	10.4
Plate3	11.3	10.4
Plate4	7.9	10.4
Plate5	9.5	10.4
Plate6	7.9	10.4
Plate7	5.2	10.4
Plate8	6.3	10.4
Plate9	18.4	10.4
Plate10	13.3	10.4
Plate11	14.3	10.4
Plate12	11.7	10.4
Plate13	14.9	10.4
Plate14	12.1	10.4
Plate15	13.0	10.4
Plate16	10.6	10.4
Plate17	9.5	10.4
Plate18	9.5	10.4
Plate19	5.2	10.4
Plate20	6.3	10.4

Figure: 3.38

The graph in Figure 3.39, on the other hand, is not of the Line type but of the XY type. Because you are not dealing with categories here, you need pairs of coordinates. Remember that XY graphs work with paired values. You should therefore create a mini-table with a new series of coordinates for the lowest and highest x value paired with the mean of all y values. Here's how:

1. In A14, use the formula `=MIN(A2:A11)`.
2. In A15, use the formula `=MAX(A2:A11)`.
3. In B14 and B15, use the formula `=AVERAGE($B$2:$B$11)`.
4. Add a new series to the graph, with x values from A14:A15 and y values from B14: B15.

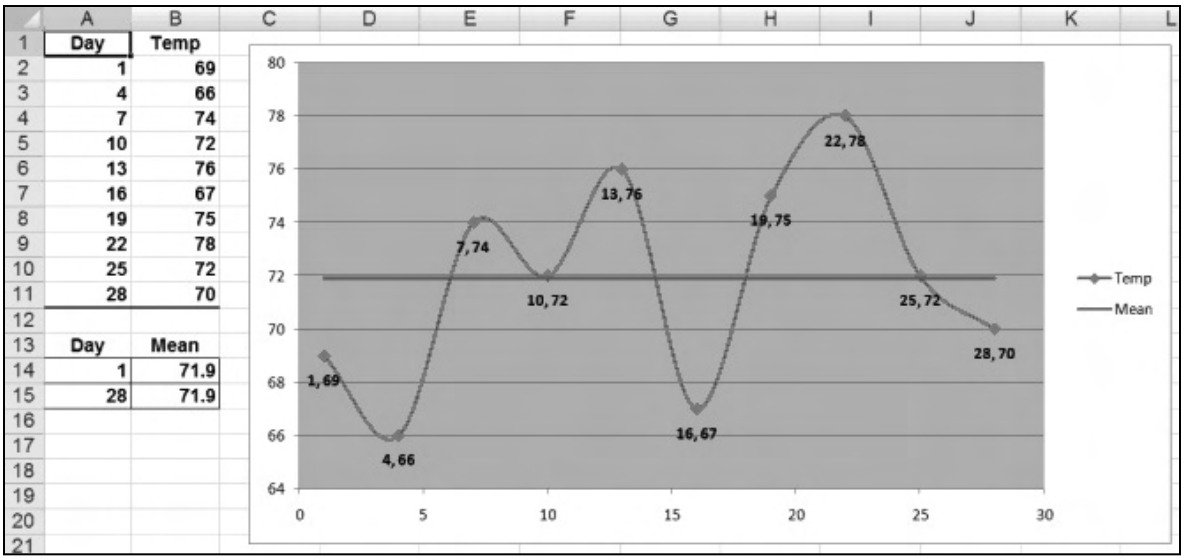


Figure: 3.39

Figure 3.40 shows an example of quality control: Samples should be within the three standard deviations. The graph can be a Line graph or an XY graph, but the line markers are implemented differently in the two types:

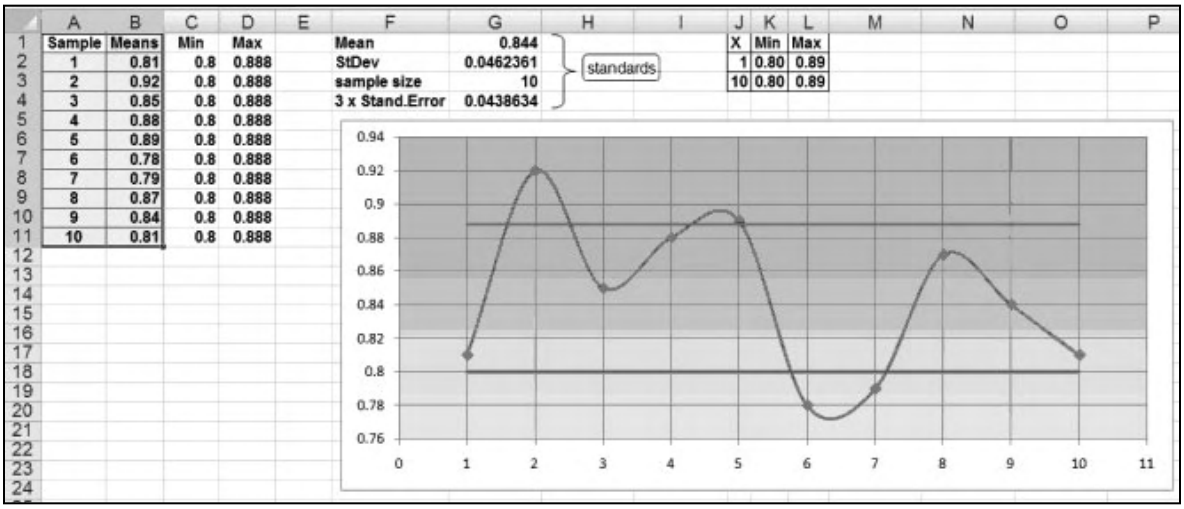


Figure: 3.40

- If this is a Line graph, you need the columns C and D:
  1. In column C: the mean (in G1) minus 3*Standard Error (in G4).
  2. In column D: the mean (in G1) plus 3*Standard Error (in G4).
  3. In the part on Statistics, we will discuss the Standard Error more in detail. The Standard Error is the Standard Deviation (G2) divided by the square root of the sample size (G4).
  4. Add the two columns C and D to the graph as two new series.
- If this is an XY graph, you need a mini-table of coordinates:
  1. Place in columns I and J the coordinates - for instance, in cell I1: =$E$1-$E$4.
  2. Create the Min Series from H2:H3 and I2:I3.
  3. Create the Max Series from H2:H3 and J2:J3.

Figure 3.41 offers one more example of playing with extra columns in a Line graph. It shows a rather unusual situation: One of the units of measurement works with a tiny subscale. Because there is already a secondary axis, you need to find another solution. For example, you could use an extra series of "hidden" x and y values. Here's how:

1. Hide a fake scale in columns F and G.
2. You can place the scale units at any position on the category axis. I chose a (date) value smaller than the lowest value in column A (for instance, 9/8/2001).
3. Add a new series to the XY graph by using x values from F6:F11 and y values from G6:G11.

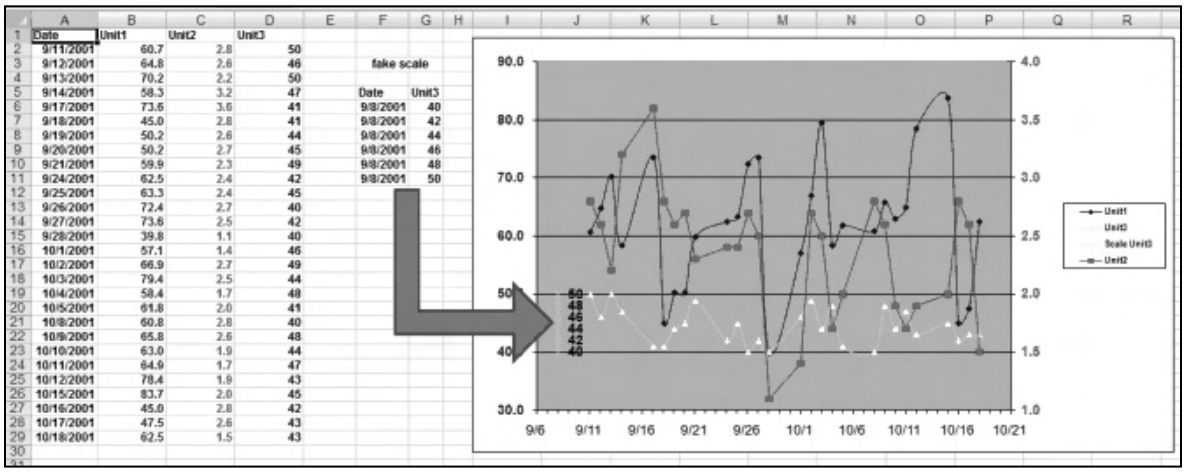


Figure: 3.41

* * *

Chapter 31

INTERPOLATION

Interpolation is a process of estimating a missing value by using existing, observed values. For example, in a graph, you might want to mark a specific point on the curve that may not have been measured; it has to be interpolated. The graph must be of the XY type because interpolation works with values in between—and such values do not exist in graphs carrying a category axis.

Figure 3.42 shows the concept of interpolation.

- The dots lined up along the linear curve are observed pairs of values. An x value of 0.17 was never observed, so you use interpolation to find its corresponding y value—say, 50.
- You have a choice as to what you want marked in the graph:
  - One pair of coordinates: 0.17 and 50
  - Three pairs of coordinates: 0.17,50 and 0,50 and 0.17,0
  - A line that connects the new pairs of coordinates, called an *insert*, or *line*, *marker*

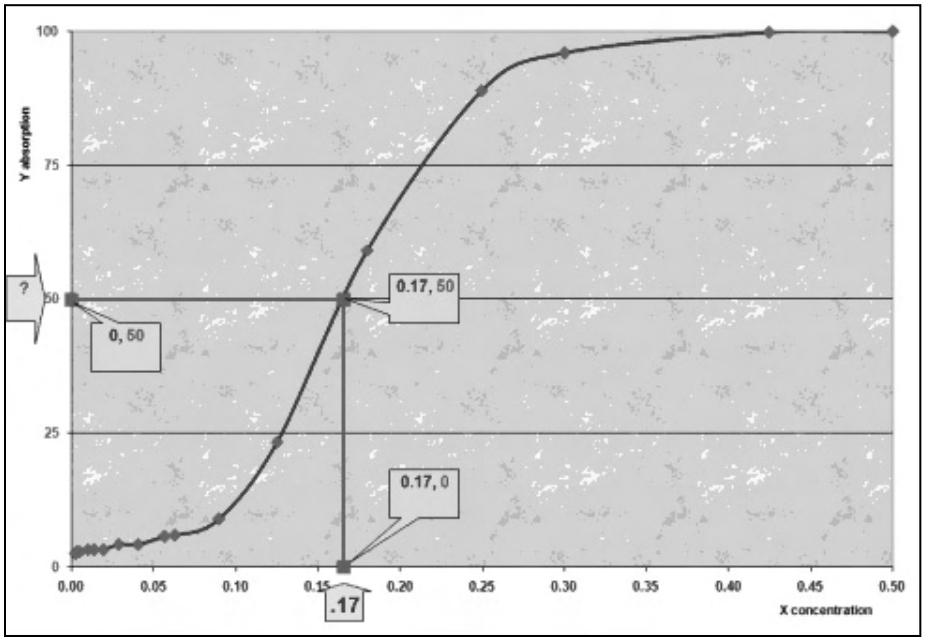


Figure: 3.42

Figure 3.43 uses only one pair of coordinates. The thick interpolated marker on the trendline is based on the pair of coordinates shown in cells G2 and H2. Cell H2 (y) is regulated by cell G2 (x) through a formula based on a linear trend through the formula `Y=1.286*X-43.11`. Cell G2, in turn, is regulated by a control to its left (which runs from 70 to 110). Part 4 discusses how this formula is found and how to implement a control like this. You need to add the two coordinates from G2 and H2 to the graph as a new series. Moving the control makes the interpolated marker move along the trendline.

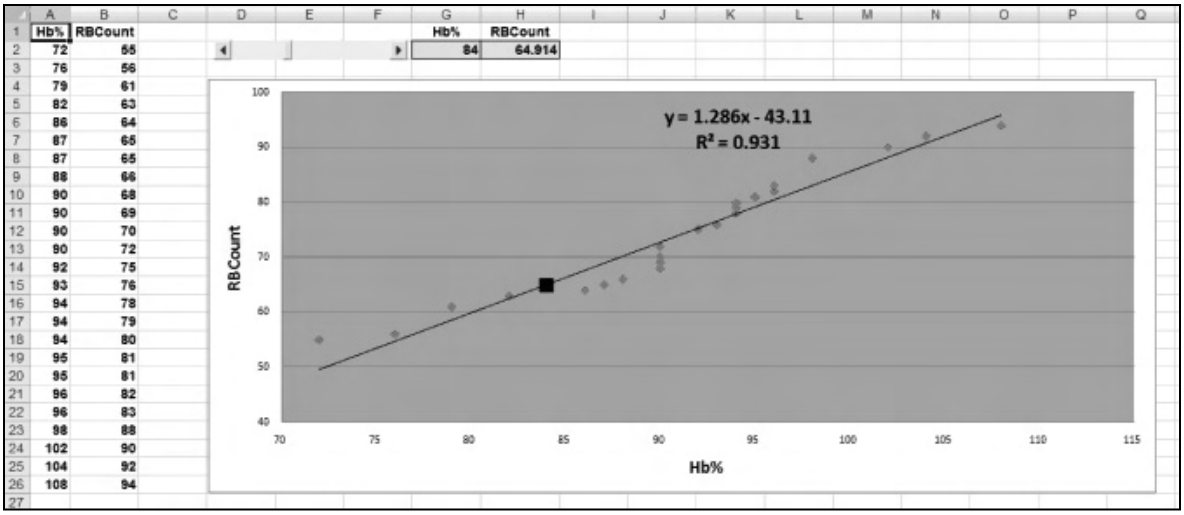


Figure: 3.43

Figure 3.44 uses three pairs of coordinates to mark the mean of the x and of the y values. If you want the line markers to touch the axis, you must use the min and max value of the axes (and not the lowest and highest value in the table of observations). Here's how it works:

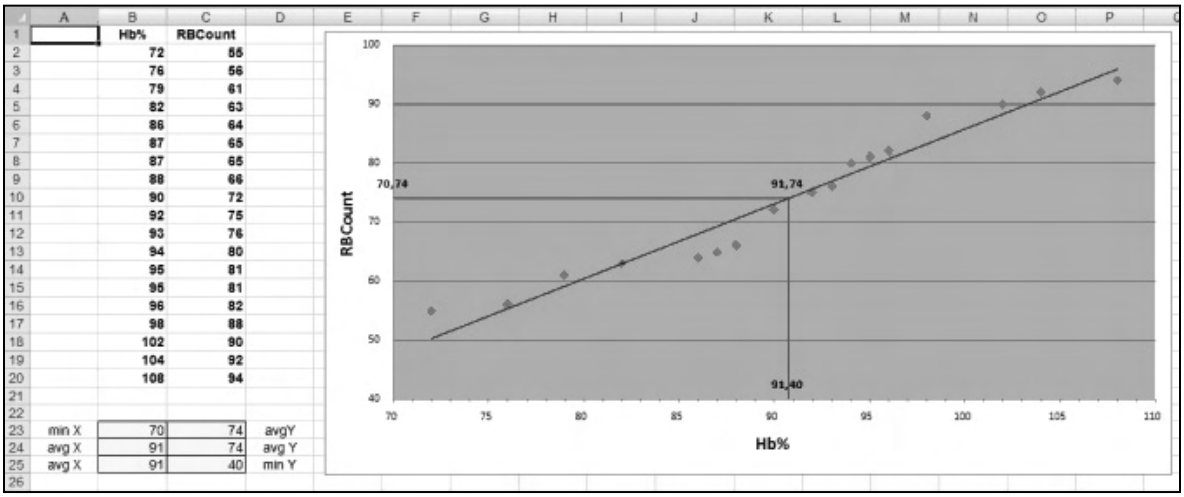


Figure: 3.44

1. Add a new series to the XY graph, with x values from B23:B25 and y values from C23: C25.
2. If you want a connecting line between the new coordinates, change the chart type to Scatter with Straight Lines; otherwise, you get a strangely curved line.
3. If desired, add data labels to the insert.

Figure 3.45 shows another instance of three pairs of coordinates for interpolation. The curve shows the increasing speed of a falling ball. Thanks to the help of a "classic" physics formula, we know the speed in cell B4: `=$B$3*(A4^2)`. Say that you want to interpolate what the speed would be at 4.5 seconds. To find out, you need a mini-table of three pairs of coordinates in A11:B13. Here's how you do it:

1. Ensure that A12 is the only independent cell. In this case, enter `4.5`.
2. Base B12 on A12 by using the formula `=$B$3*(A12^2)`.
3. Set A11 to 0 (the origin of the x axis), A13 to `=A12`, B11 to `=B12`, and B13 to `0` (the origin of y axis)
4. Add the new series to the XY graph and change its type to Scatter with Straight Lines and Markers.

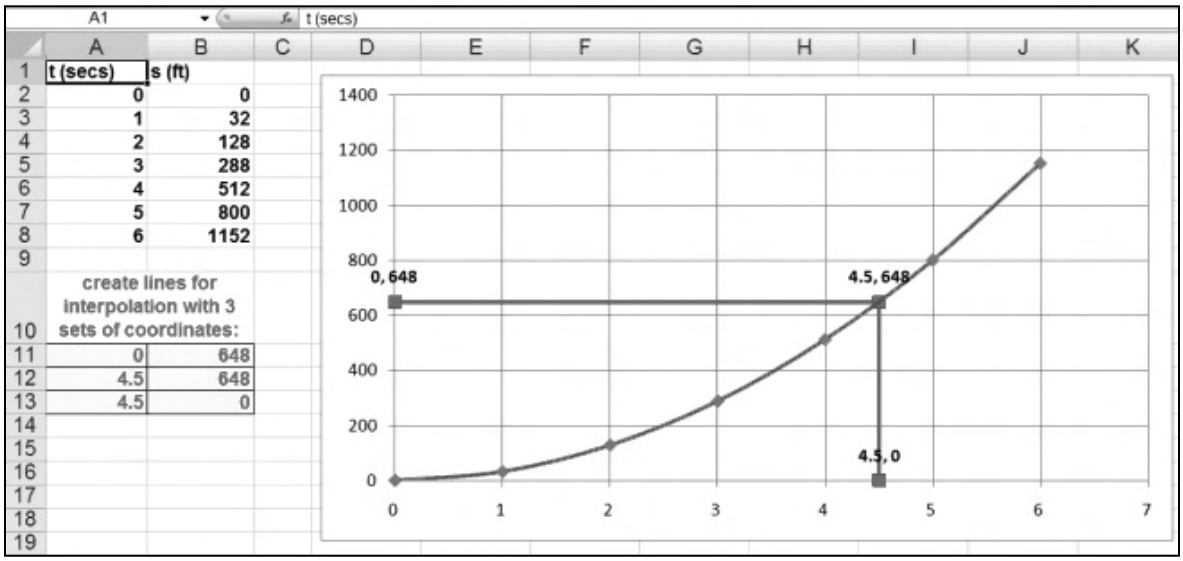
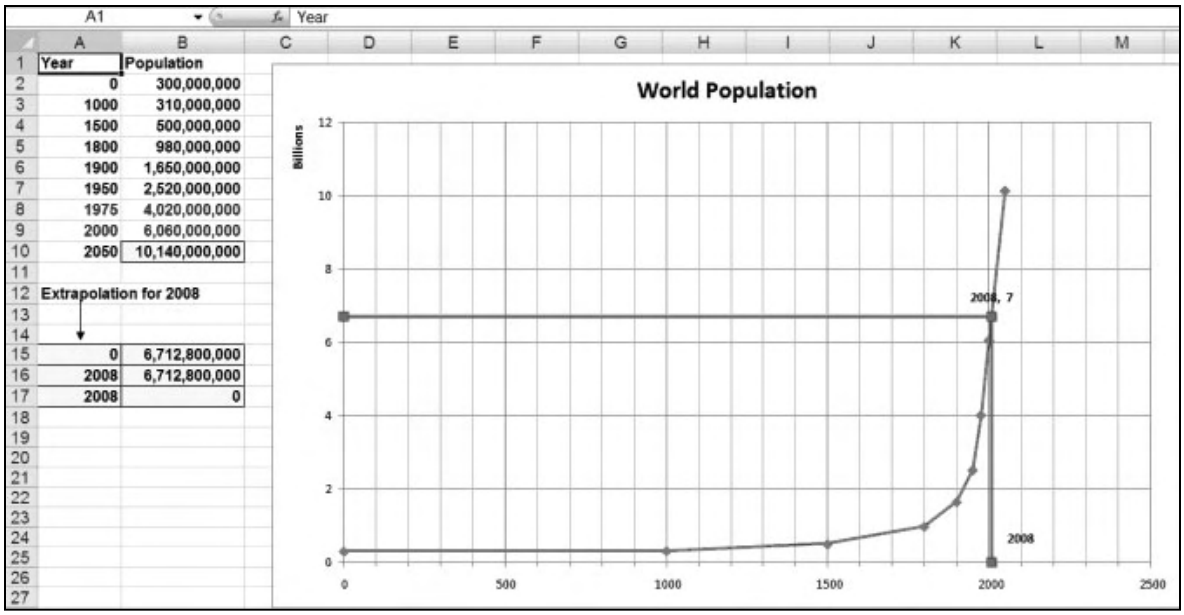


Figure: 3.45

Figure 3.46 applies extrapolation as well as interpolation to data about the world population, but this time without your knowing the formula behind the curve. So you have to somehow estimate the interpolated values:

1. To predict the world population in 2050 (cell B10), apply the `TREND` function to the two closest observations (1975 and 2000).
2. Add 2050 to the graph, either by expanding the series range or by using the data source.
3. Create a line marker for 2008 through the mini-table in A15:B17.
4. Enter the formula `=TREND(B8:B9,A8:A9,A16)` in cell B16; in other words, you use extrapolation here based on the two latest observations rather than use interpolation based on the latest observation and an already extrapolated value.
5. Add the extrapolated coordinates to the graph with straight lines.



Year	Population
0	300,000,000
1000	310,000,000
1500	500,000,000
1800	980,000,000
1900	1,650,000,000
1950	2,520,000,000
1975	4,020,000,000
2000	6,060,000,000
2050	10,140,000,000

Extrapolation for 2008

0	6,712,800,000
2008	6,712,800,000
2008	0

Figure: 3.46

* * *

Chapter 32

Graph Formulas

Excel is a formula program, so it's not surprising that its graphs use formulas as well. Not only do they plot what comes out of formulas, they also use formulas in the background—and you can make them behave differently by using formulas. When you know how to manipulate formulas, the sky is the limit.

You may have never noticed it, but when you select a specific series in a graph, the formula bar displays its formula. Its syntax (which is not available through fx, by the way) is `=SERIES(label,categories,values,order)`. When you click in the formula bar and press F9, you see the formula perform. Then you have a choice: Either press Esc to get the formula back or press Enter to keep these static values (but then the graph is detached from the table and can no longer update).

Thanks to formulas, you can also make graphs automatically expand when the table expands. You do this by using names and two different functions:

- To have a name refer to a dynamic range, you use the `OFFSET` function, whose syntax is `=OFFSET(start,row-offset,col-offset,#rows,#cols)`. Part 1 discusses this issue.
- To have a graph work with dynamic names, you use the `SERIES` function, which has the syntax `=SERIES(label,categories,values,order)`.

Figure 3.47 shows a dynamic graph that expands when more temperature readings are added to the table. You follow these steps to create such an effect:

1. Use the Name Manager to give the ranges A1:A15 and B1:B15 the dynamic name `Weeks`.
2. In the Refers To box, enter `=OFFSET(DynEnd!$A$1,0,0,COUNTA(DynEnd!$A:$A))`.
3. For the Temps range, enter `=OFFSET(DynEnd!$B$1,0,0,COUNT(DynEnd!$B:$B))`.

> **Note**: In steps 2 and 3, you do not use the last argument (`#cols`) because it is not relevant here. Be careful with `COUNT` and `COUNTA`; the first one only counts cells that contain numbers.

4. Use the new dynamic names in the SERIES function:
   - Highlight A1:A15 in the formula bar and replace it with the name `Weeks`.
   - Highlight B1:B15 in the formula bar and replace it with the name `Temps`.

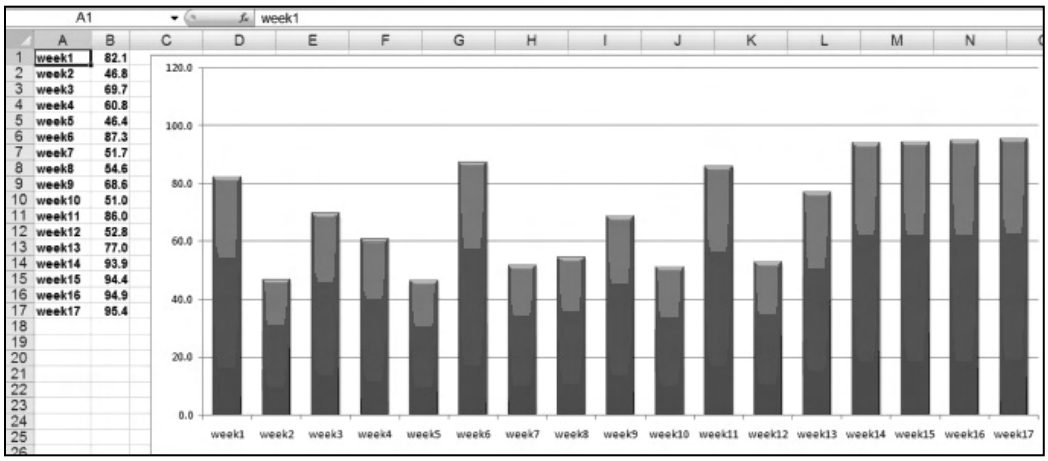


Figure: 3.47

Caution: Do not touch the sheet names in the formula bar!

5. Press Enter, and the sheet references are replaced by book references because these names function at the book level. The end result is `=SERIES(,'BookName.xlsx'!Weeks,'BookName.xlsx'!Temps,1)`.
6. Watch how the ranges get properly highlighted. But this time, they are dynamic and can automatically expand.

When you add new entries to the table, the graph nicely responds.

You saw Figure 3.48 in Chapter 23. This time, the graph should adjust to changes in D1 as well. Here's how you make that happen:

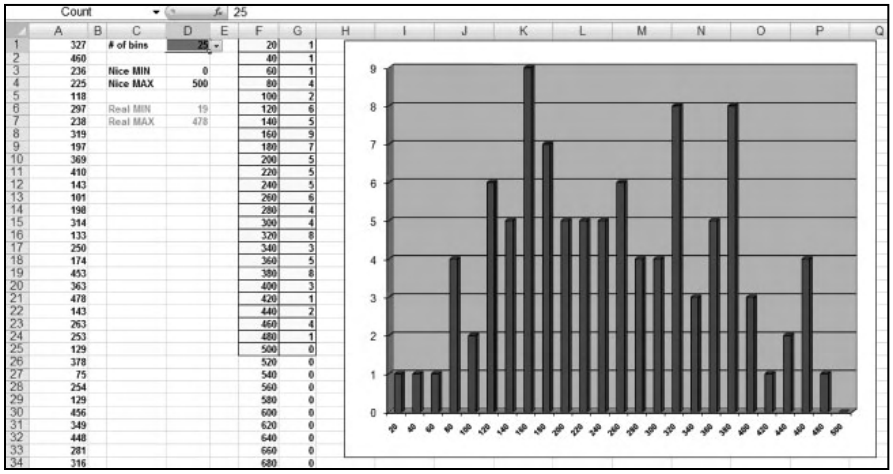


Figure: 3.48

1. Assign the name `Bins` to column F: `=OFFSET(DynBins!$F$1,0,0,DynBins!$D$1)`.
2. Assign the name `Freqs` to column G: `=OFFSET(DynBins!$G$1,0,0,DynBins!$D$1)`.
3. Replace the

ranges (not their sheet names!) with the range names in the graph's `SERIES` formula.

4. Press Enter, and the formula should look like this:
   `=SERIES(,'BookName.xlsx'!Bins,'BookName.xlsx'!Freqs,1)`.

Figure 3.49 has something else going on: In this example, pH readings greater than or equal to 7.1 get flagged. You discovered in Chapter 11 that sheets use conditional formatting. But graphs do not! So what's the secret to the graph in this figure? You add an extra column for a new series:

1. Cell C1 contains a real number, but the number has been formatted with additional text (see Chapter 15 for more on how to do this). If this were not a real number, you could never use comparison operators such as > and <. Now we can create formulas that single out values above the value featured in cell C1.
2. In cell C2, enter the formula `=IF(B2>=$C$1,B2,NA())`, which does not show values below 7.1 (in cell C1). You used the function `NA` earlier (refer to Chapter 22); it does not show up in a graph.
3. Add column C as a new series and fix its format (with no line but some kind of a marker).

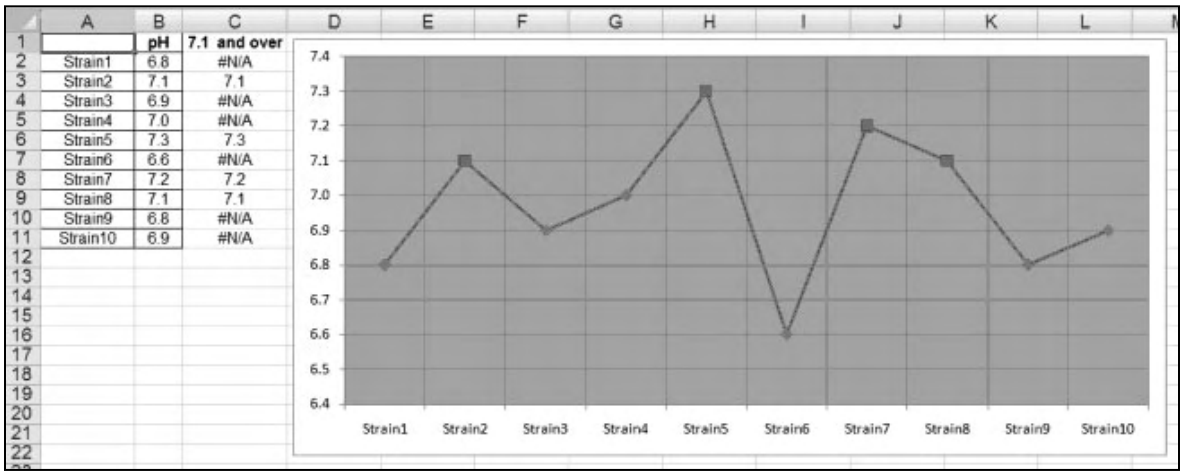
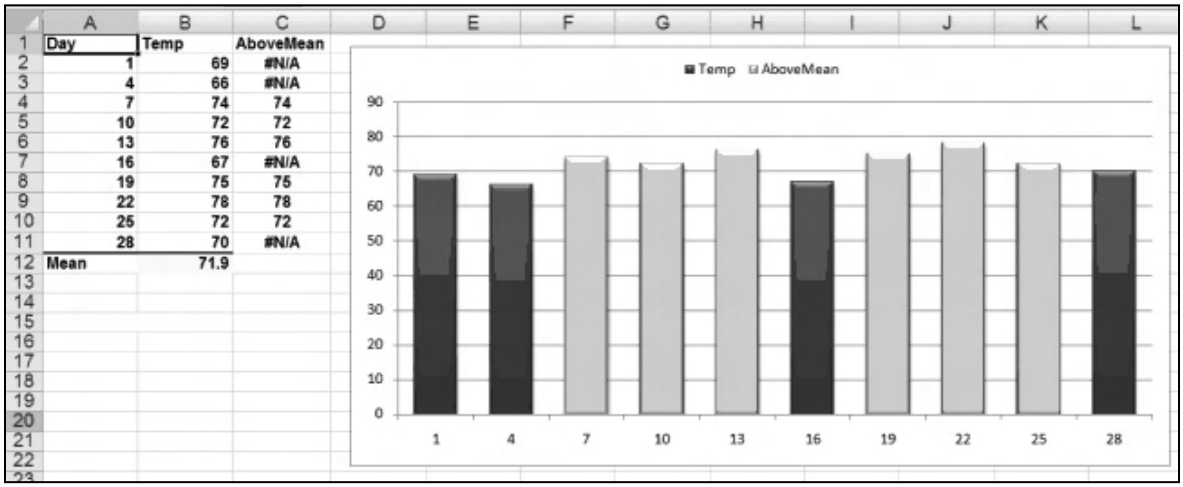


Figure: 3.49

Figure 3.50 shows a similar example: It flags all temperatures above the mean. Again, you need another column for this, so you follow steps 1–3 from the Figure 3.49 example. This time, however, you encounter another problem: The two series do not overlap. You therefore need to change their overlap to 100%. By coincidence, the second series might be hidden

behind the first series. If it is, you change `1` to `2` in the last argument of the `SERIES` formula (or change the series order through the data source).

Excel may not have all the tools needed for graphic representations of your scientific data, but it does have an impressive array—and you can do the rest. This part gives you a number of examples and hints to help you get your creativity going when it comes to graphs.



Day	Temp	AboveMean
1	69	#N/A
4	66	#N/A
7	74	74
10	72	72
13	76	76
16	67	#N/A
19	75	75
22	78	78
25	72	72
28	70	#N/A
Mean	71.9	

Figure: 3.50

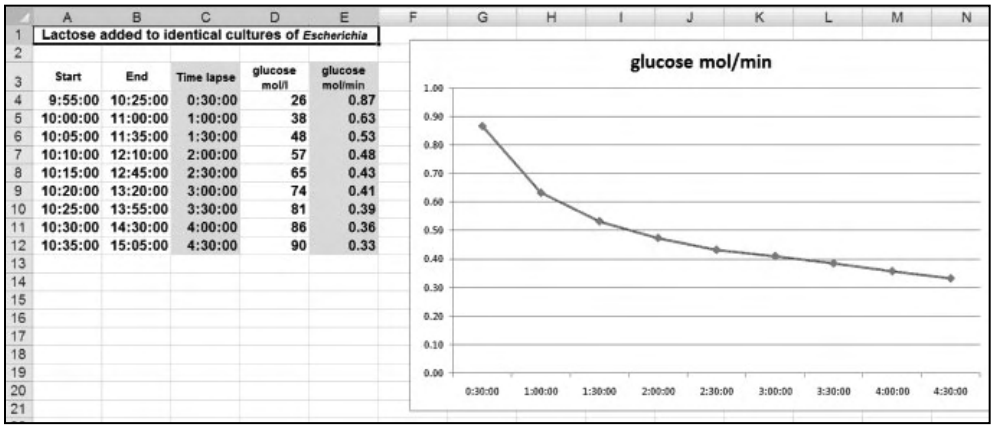
* * *

Excercises - Part 3

You can download all the files used in this book from `www.genesispc.com/Science2007.htm`, where you can find each file in its original version (to work on) and in its finished version (to check your solutions).

Exercise 1

1. Types of Graphs
   1.1. Create a Line graph based only on columns C and E.
   1.2. Change the type from Line to XY, and do the necessary axis work.



Lactose added to identical cultures of *Escherichia*

Start	End	Time lapse	glucose mol/l	glucose mol/min
9:55:00	10:25:00	0:30:00	26	0.87
10:00:00	11:00:00	1:00:00	38	0.63
10:05:00	11:35:00	1:30:00	48	0.53
10:10:00	12:10:00	2:00:00	57	0.48
10:15:00	12:45:00	2:30:00	65	0.43
10:20:00	13:20:00	3:00:00	74	0.41
10:25:00	13:55:00	3:30:00	81	0.39
10:30:00	14:30:00	4:00:00	86	0.36
10:35:00	15:05:00	4:30:00	90	0.33

Figure: Ex-1

Exercise 2

2. Types of Graphs
   2.1. Come up with two graphs that would present this information well in a graphical way.
   2.2. Determine which of your two possible graphs would be best.

	A	B	C	D
1	Activities of 10 Mice in 2 different cages			
2				
3		Sleep	Eat	Run
4	Cage A	47	37	16
5	Cage B	60	21	19

Figure: Ex-2

	A	B	C	D	E
1	Plate ID	50 ng/mL	%CV	25 ng/mL	%CV
2	8877p58a	47.7	2	22.7	0
3	8877p58a	48.7	3	23.0	3
4	8877p58a	49.3	0	23.0	1
5	8877p58a	45.9	2.0	22.9	1.0
6	**8877p58a StdDev**	1.48772757		0.1414214	
7	8877p58b	47.5	0	22.7	2
8	8877p58b	45.9	1	22.9	3
9	8877p58b	43.2	4	22.6	3
10	8877p58b	44.3	3.0	20.7	3.0
11	**8877p58b StdDev**	1.87860764		1.0242884	
12	8696p08a	56.4	12	29.1	11
13	8696p08a	51.3	2	26.9	1
14	8696p08a	52.3	1	26.5	2
15	8696p08a	49.7	2.0	25.0	1.0
16	**8696p08a StdDev**	2.85817541		1.6938615	
17	8696p08b	52.9	3	27.5	6
18	8696p08b	50.1	1	26.8	3
19	8696p08b	51.0	1	26.1	1
20	8696p08b	48.6	1.0	24.8	0.0
21	**8696p08b StdDev**	1.79722008		1.1518102	
22	8697p58b	47.5	0	22.7	2
23	8697p58b	47.5	1	22.9	3
24	8697p58b	43.2	4	22.6	3
25	8697p58b	44.3	3.0	22.6	3.0
26	**8697p58b StdDev**	2.21114601		0.1414214	
27	**Grand StdDev**	3.42779275		2.2181073	
28					

Exercise 3

3. A Graph's Data Source
   - 3.1. Create a Column graph for this table.
   - 3.2. Remove columns C and E from the data source.
   - 3.3. Hide the details in the table.
   - 3.4. Remove the grand standard deviation from the data source—whatever way you prefer.

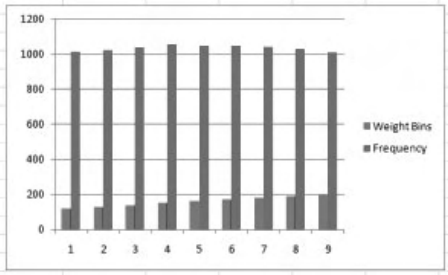
Figure: Ex-3

Exercise 4

4. A Graph's Data Source
   - 4.1. Create a Column graph for the table on the left.
   - 4.2. Fix the trouble that occurs.
   - 4.3. Create a Column graph from the table on the right.
   - 4.4. Why does this second graph cause no trouble?

	A	B
1	Weight Bins	Frequency
2	120	1015
3	130	1022
4	140	1040
5	150	1055
6	160	1050
7	170	1046
8	180	1042
9	190	1028
10	200	1011

	H	I
		Frequency
	120	1015
	130	1022
	140	1040
	150	1055
	160	1050
	170	1046
	180	1042
	190	1028
	200	1011



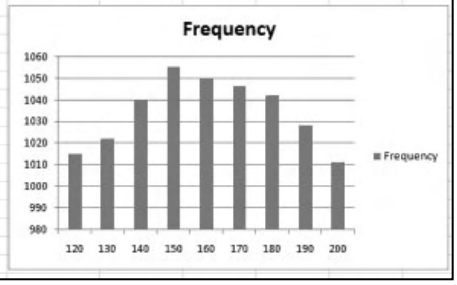
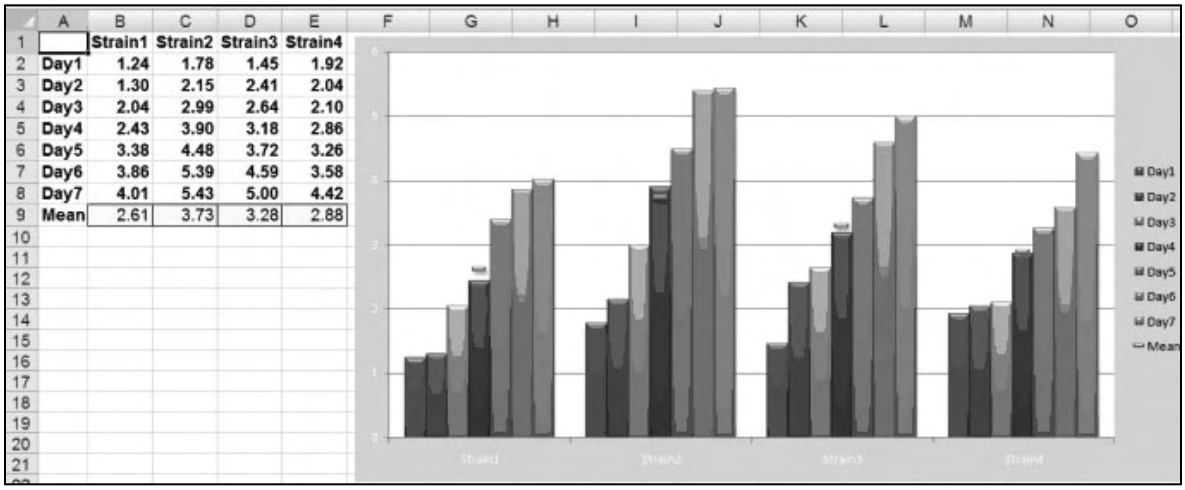


Figure: Ex-4

Exercise 5

5. Combining Graph Types
   - 5.1. Calculate the mean for each strain (in row 9).
   - 5.2. Create a Column graph and select Switch Row/Column.
   - 5.3. Add the means to the graph and fix the format.

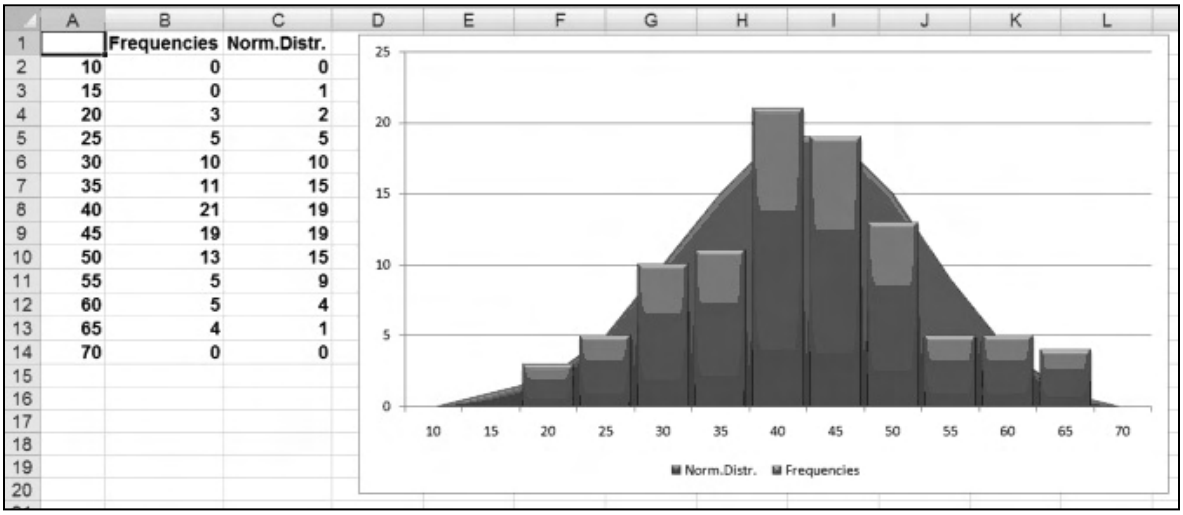


	Strain1	Strain2	Strain3	Strain4
Day1	1.24	1.78	1.45	1.92
Day2	1.30	2.15	2.41	2.04
Day3	2.04	2.99	2.64	2.10
Day4	2.43	3.90	3.18	2.86
Day5	3.38	4.48	3.72	3.26
Day6	3.86	5.39	4.59	3.58
Day7	4.01	5.43	5.00	4.42
Mean	2.61	3.73	3.28	2.88

Figure: Ex-5

Exercise 6

6. Combining Graph Types
   - 6.1. Create a Column graph for this table.
   - 6.2. Change the series of column C into an Area graph.
   - 6.3. Adjust the gap between the columns.



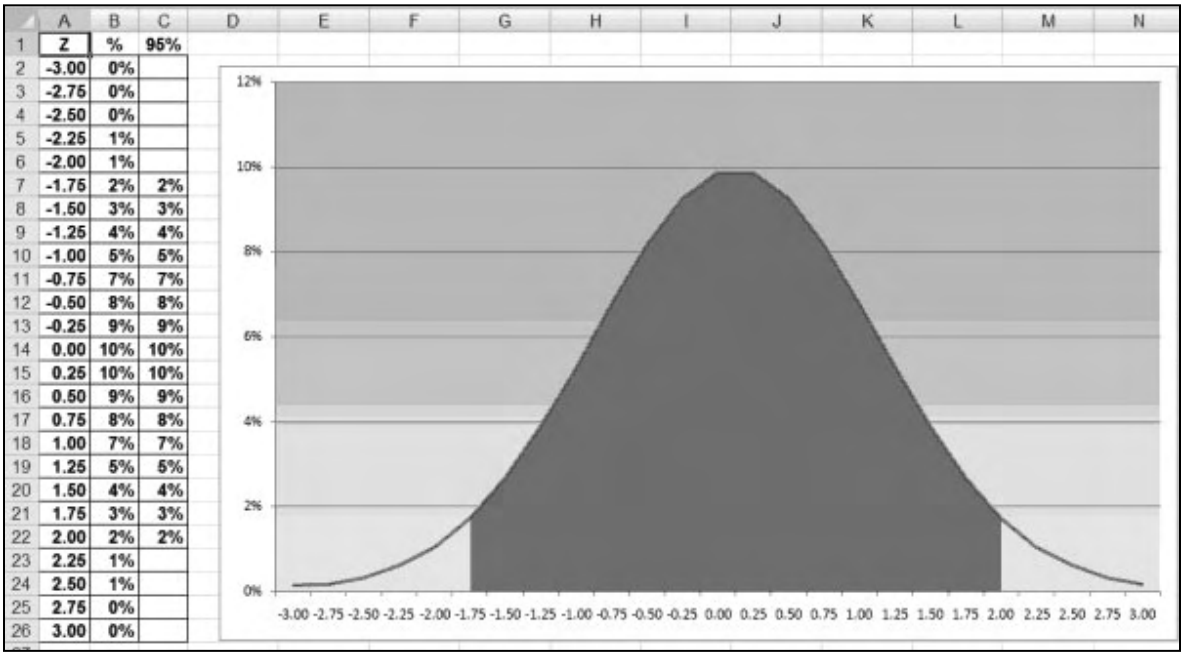
	Frequencies	Norm.Distr.
10	0	0
15	0	1
20	3	2
25	5	5
30	10	10
35	11	15
40	21	19
45	19	19
50	13	15
55	5	9
60	5	4
65	4	1
70	0	0

Figure: Ex-6

Exercise 7

7. Combining Graph Types

7.1. Decide which two graph types have been combined here.



Z	%	95%
-3.00	0%	
-2.75	0%	
-2.50	0%	
-2.25	1%	
-2.00	1%	
-1.75	2%	2%
-1.50	3%	3%
-1.25	4%	4%
-1.00	5%	5%
-0.75	7%	7%
-0.50	8%	8%
-0.25	9%	9%
0.00	10%	10%
0.25	10%	10%
0.50	9%	9%
0.75	8%	8%
1.00	7%	7%
1.25	5%	5%
1.50	4%	4%
1.75	3%	3%
2.00	2%	2%
2.25	1%	
2.50	1%	
2.75	0%	
3.00	0%	

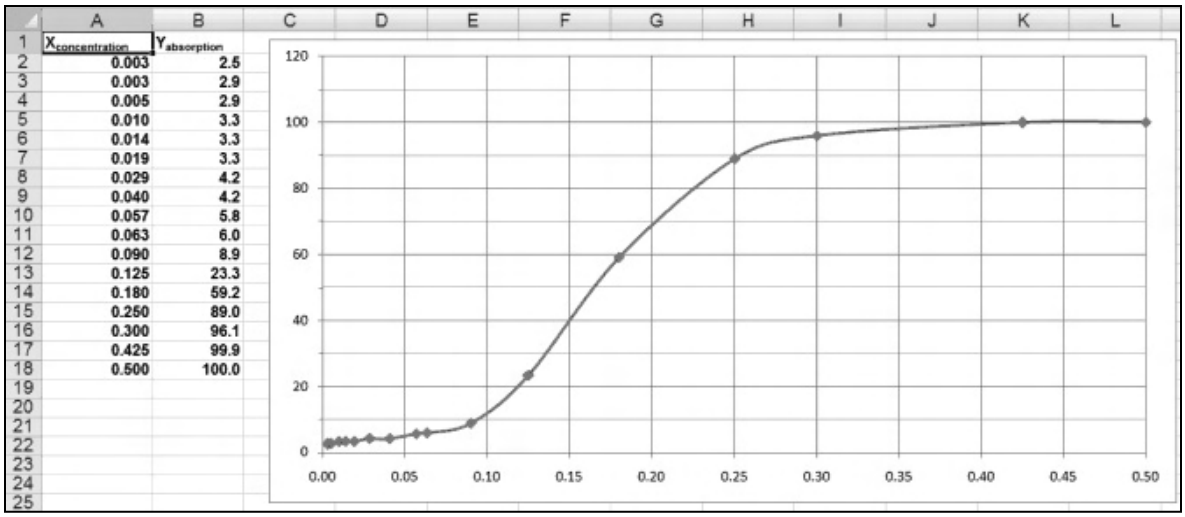
Figure: Ex-7

Exercise 8

8. Axis Scales

8.1. Create the proper scale units, as shown in the figure.

8.2. Add the proper gridlines, as shown in the figure.



$X_{concentration}$	$Y_{absorption}$
0.003	2.5
0.003	2.9
0.005	2.9
0.010	3.3
0.014	3.3
0.019	3.3
0.029	4.2
0.040	4.2
0.057	5.8
0.063	6.0
0.090	8.9
0.125	23.3
0.180	59.2
0.250	89.0
0.300	96.1
0.425	99.9
0.500	100.0

Figure: Ex-8

Exercise 9

9. Axis Scales

9.1. Create an empty chart sheet.

9.2. Move the graph located next to the table into the empty chart sheet twice.

9.3. Create the effect of a broken axis by changing the y axis scales.

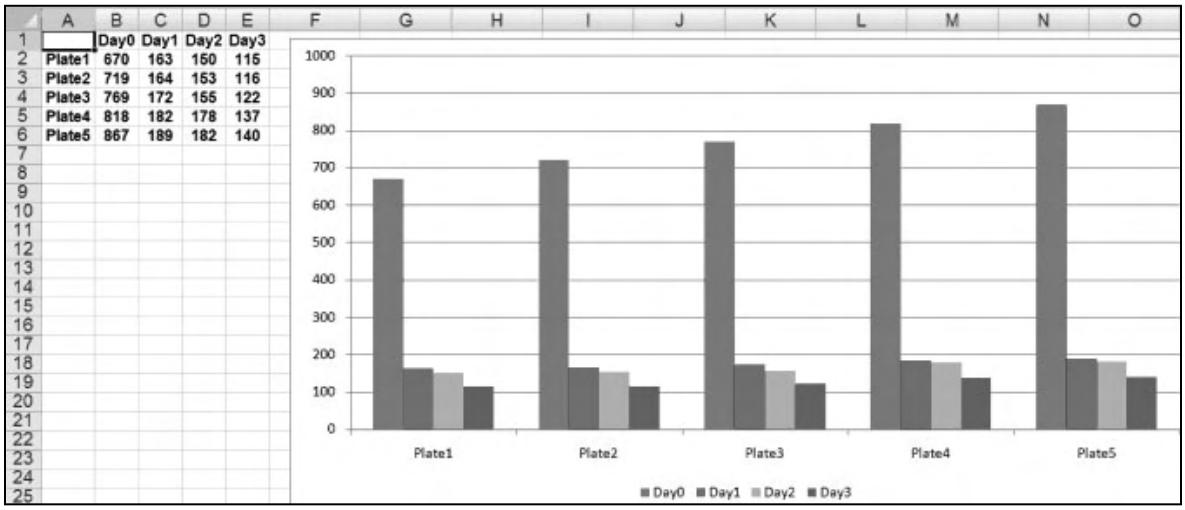


Figure: Ex-9

Exercise10

10. More Axes

10.1. Calculate the cumulative totals in column D.

10.2. Add the cumulative totals to the graph.

10.3. Assign a secondary axis.

10.4. Change the series order, if necessary.

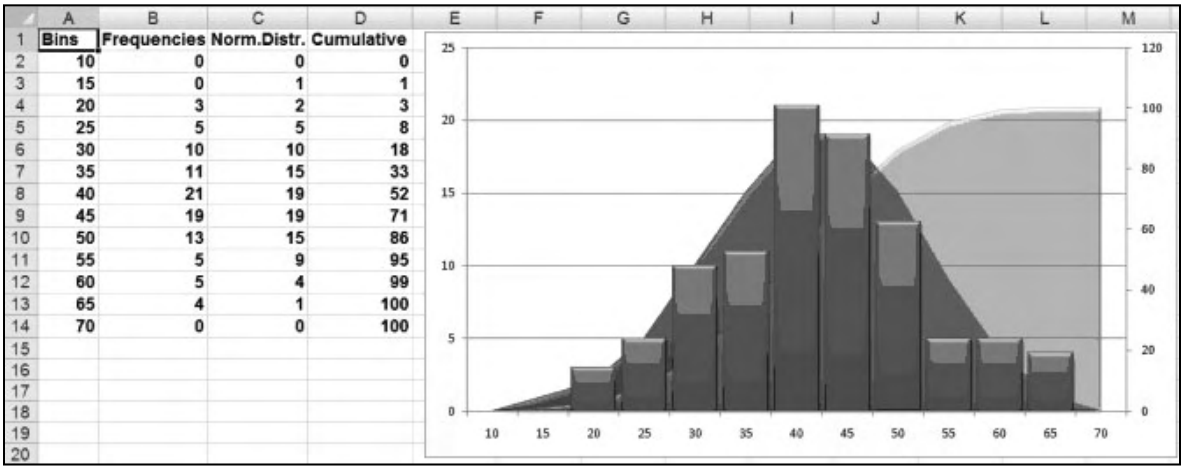


Figure: Ex-10

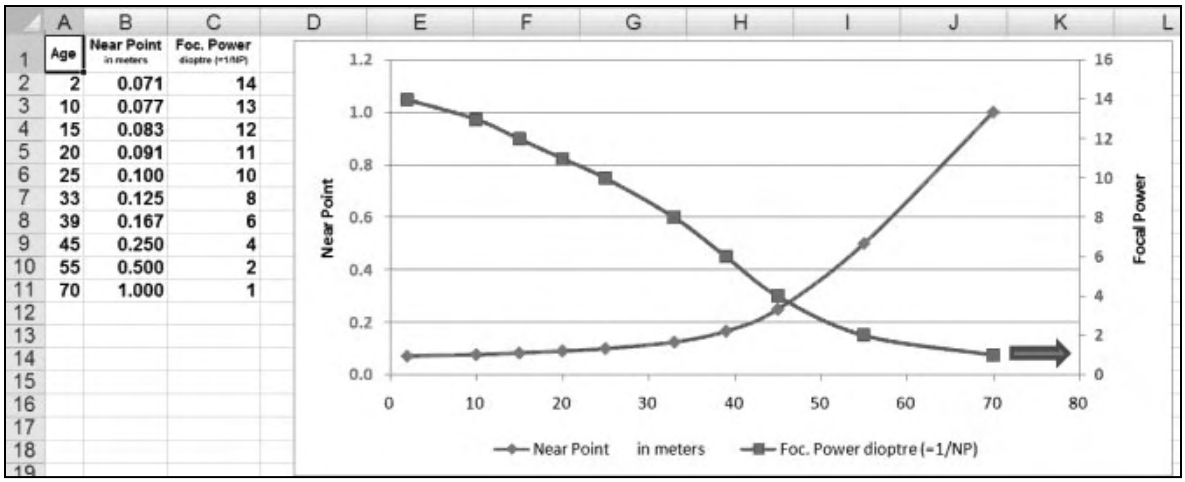
Exercise11

11. More Axes

    11.1. Create an XY graph based on the table on the left.

    11.2. Assign a secondary axis to one of the two curves.

    11.3. Add a label to each axis.



Age	Near Point in meters	Foc. Power dioptre (=1/NP)
2	0.071	14
10	0.077	13
15	0.083	12
20	0.091	11
25	0.100	10
33	0.125	8
39	0.167	6
45	0.250	4
55	0.500	2
70	1.000	1

Figure: Ex-11

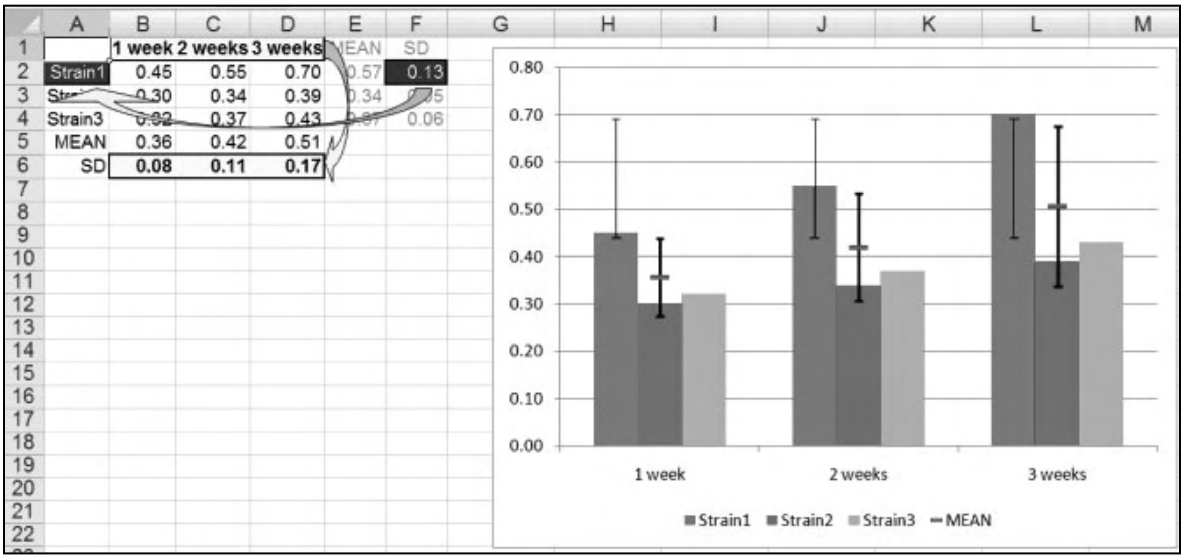
Exercise12

12. Error Bars

    12.1. Add the standard deviation for `Strain1` as errors bars to the column of `Strain1`.

    12.2. Add to the graph the mean for each week (based on row 5).

    12.3. Display in the graph the standard deviation for each week (based on row 6).



	1 week	2 weeks	3 weeks	MEAN	SD
Strain1	0.45	0.55	0.70	0.57	0.13
Strain2	0.30	0.34	0.39	0.34	0.05
Strain3	0.32	0.37	0.43	0.37	0.06
MEAN	0.36	0.42	0.51		
SD	0.08	0.11	0.17		

Figure: Ex-12

Exercise 13

13. Error Bars

13.1. Use the bottom summary table (based on the top table) for this Bar graph.

13.2. Add the standard deviation or standard error values as error bars.

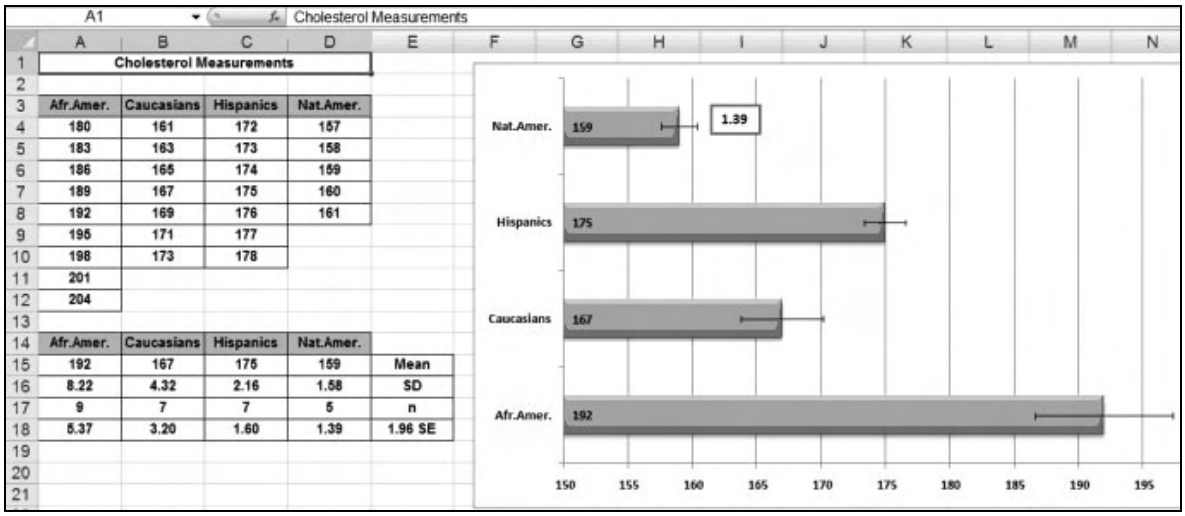


Figure: Ex-13

Exercise 14

14. More Bars

14.1. Add the vertical drop lines as error bars.

14.2. Do something similar for the horizontal drop lines.

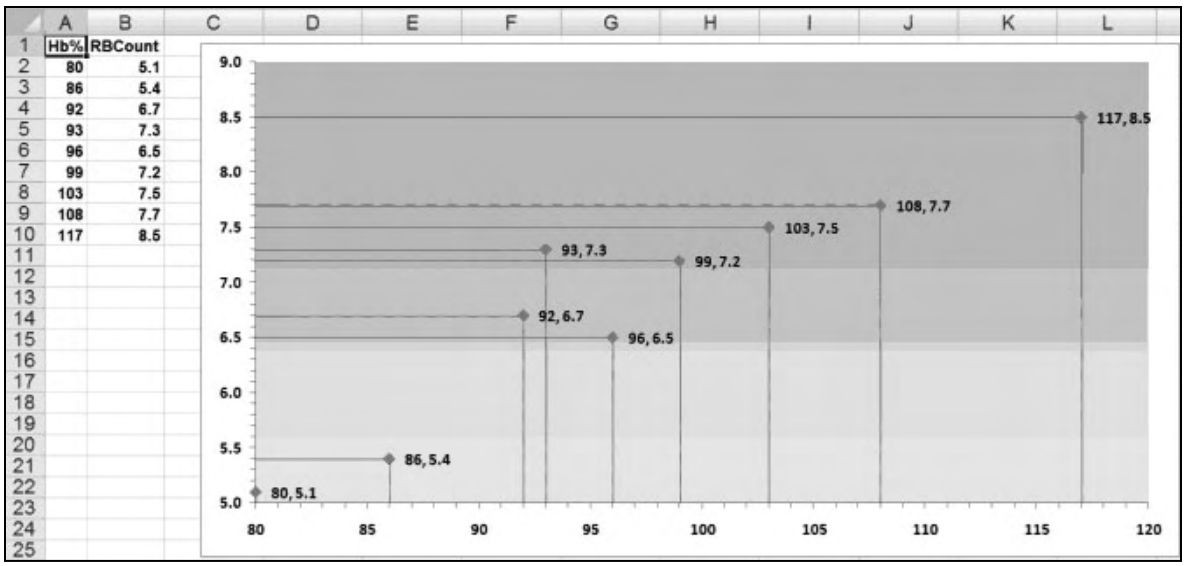


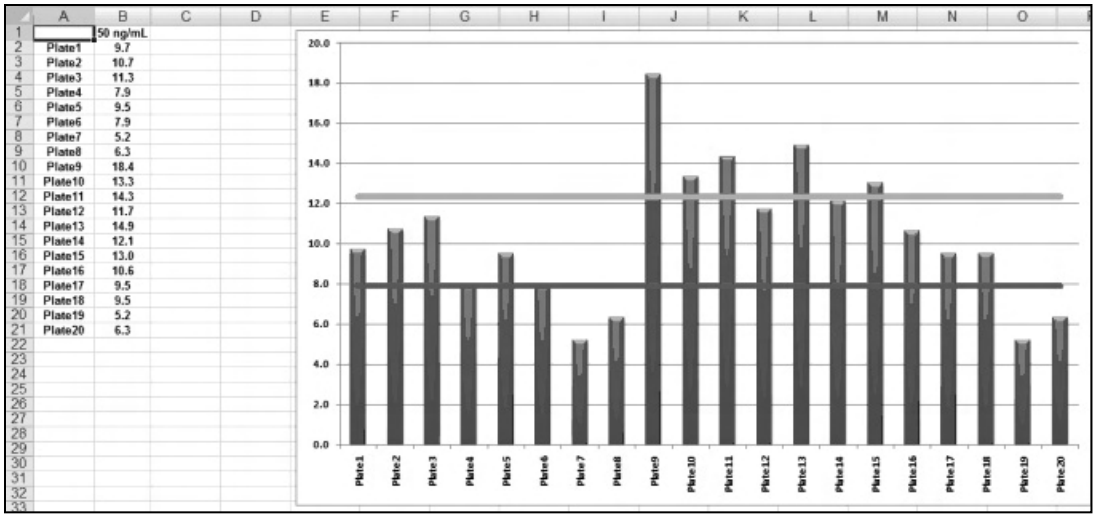
Figure: Ex-14

Exercise 15

15. Line Markers

15.1. Add 25th percentile calculations and make them show up as a line in the graph.

15.2. Add 75th percentile calculations and make them show up as a line in the graph.



	50 ng/mL
Plate1	9.7
Plate2	10.7
Plate3	11.3
Plate4	7.9
Plate5	9.5
Plate6	7.9
Plate7	5.2
Plate8	6.3
Plate9	18.4
Plate10	13.3
Plate11	14.3
Plate12	11.7
Plate13	14.9
Plate14	12.1
Plate15	13.0
Plate16	10.6
Plate17	9.5
Plate18	9.5
Plate19	5.2
Plate20	6.3

Figure: Ex-15

Exercise 16

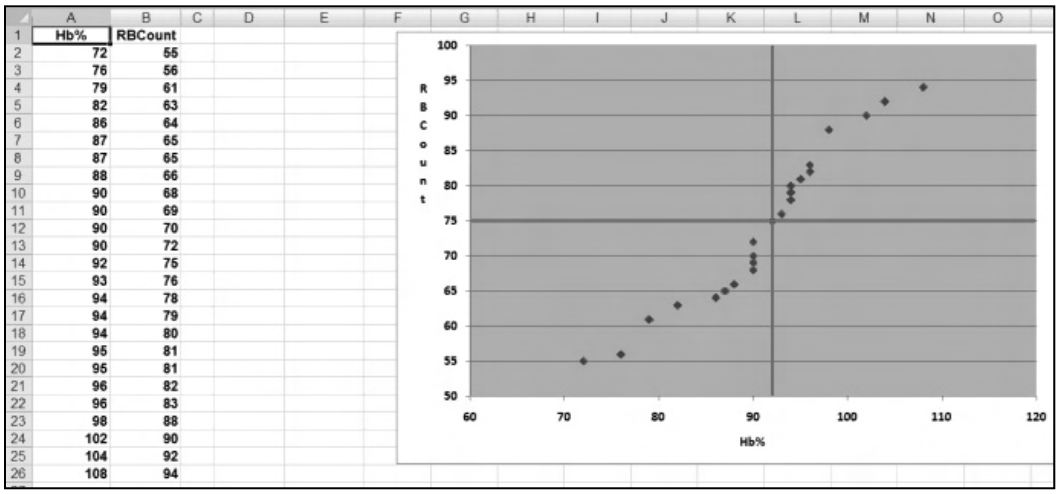
16. Line Markers

16.1. This is an XY graph with fixed maxima and minima on the scales.

16.2. Find the coordinates to draw the vertical median line.

16.3. Find the coordinates to draw the horizontal median line.

16.4. Add both median lines to the graph.



Hb%	RBCount
72	55
76	56
79	61
82	63
86	64
87	65
87	65
88	66
90	68
90	69
90	70
90	72
92	75
93	76
94	78
94	79
94	80
95	81
95	81
96	82
96	83
98	88
102	90
104	92
108	94

Figure: Ex-16

Exercise 17

17. Interpolation

17.1. What type of graph is this? Notice that the vertical axis is logarithmic and thus never reaches 0 (but 1).

17.2. Assuming that the maximum capacity of this population is size 1221, determine the three sets of coordinates for the maximum capacity.

17.3. Assuming that the optimum capacity is at size 195, determine the three sets of coordinates for the optimum capacity.

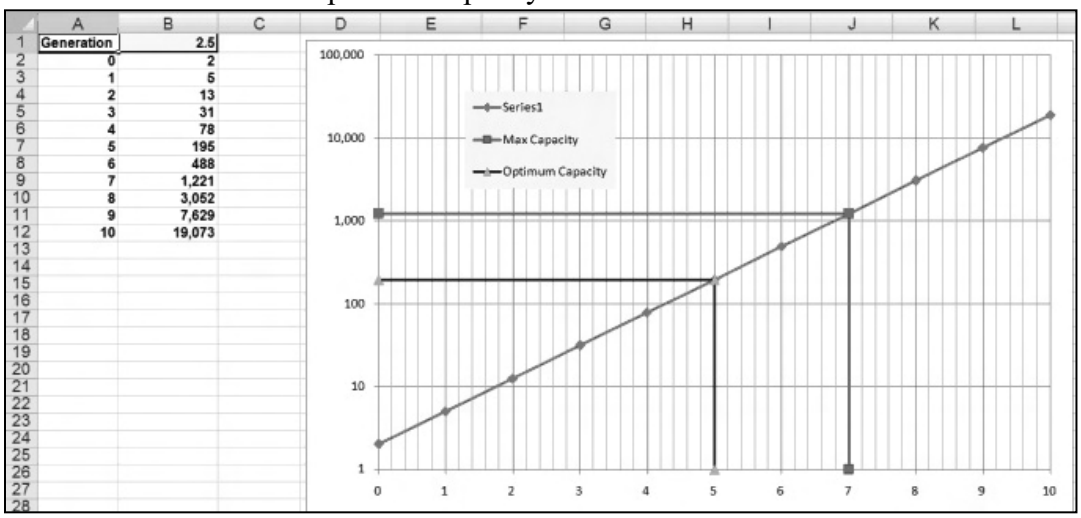


Figure: Ex-17

Exercise 18

18. Interpolation

18.1. Use an insert to mark pKa for acetic acid.

18.2. Mark the range where acetic acid buffers best—which is pKa+1.

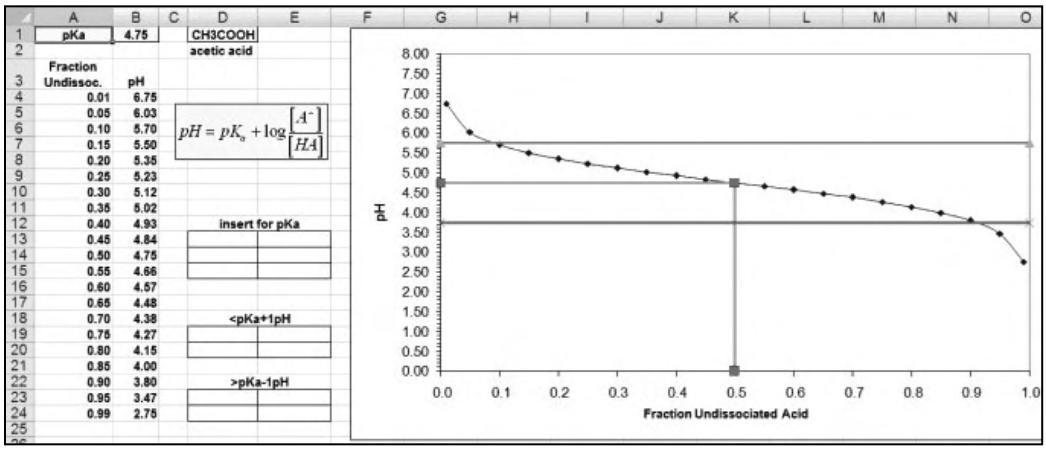


Figure: Ex-18

Exercise 19

19. Interpolation

19.1. Create the insert for the top control by using the proper coordinates in cells D8:E10.

19.2. Create the insert for the bottom control by using the proper coordinates in cells D20:E22.

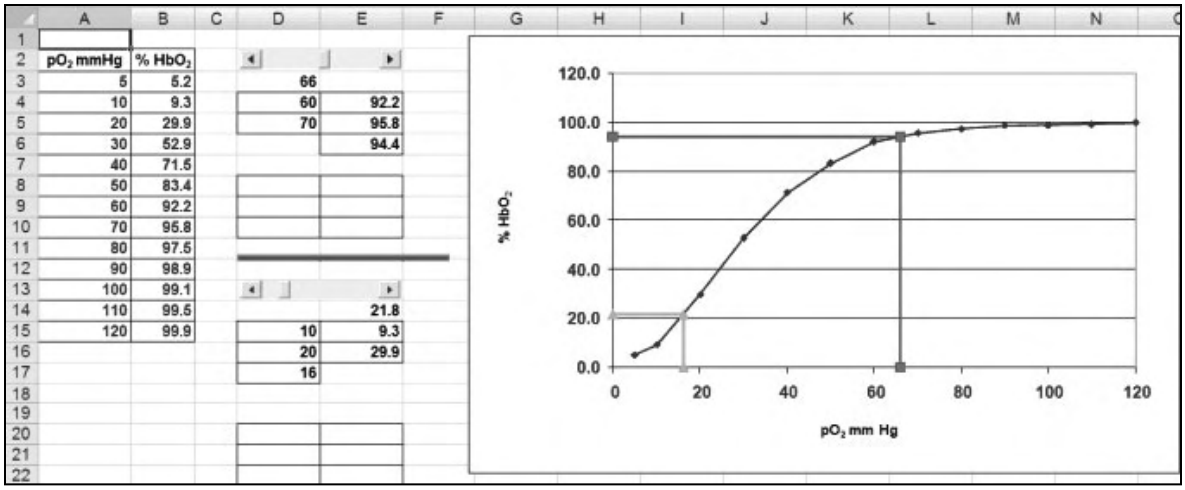


Figure: Ex-19

Exercise 20

20. Graph Formulas

20.1. Create the name `Weeks` for A1:A16 and the name `Temps` for B1:B16.

20.2. Make both names dynamic with the `OFFSET` function and starting with the week mentioned in E1.

20.3. Replace the `SERIES` references with the new names.

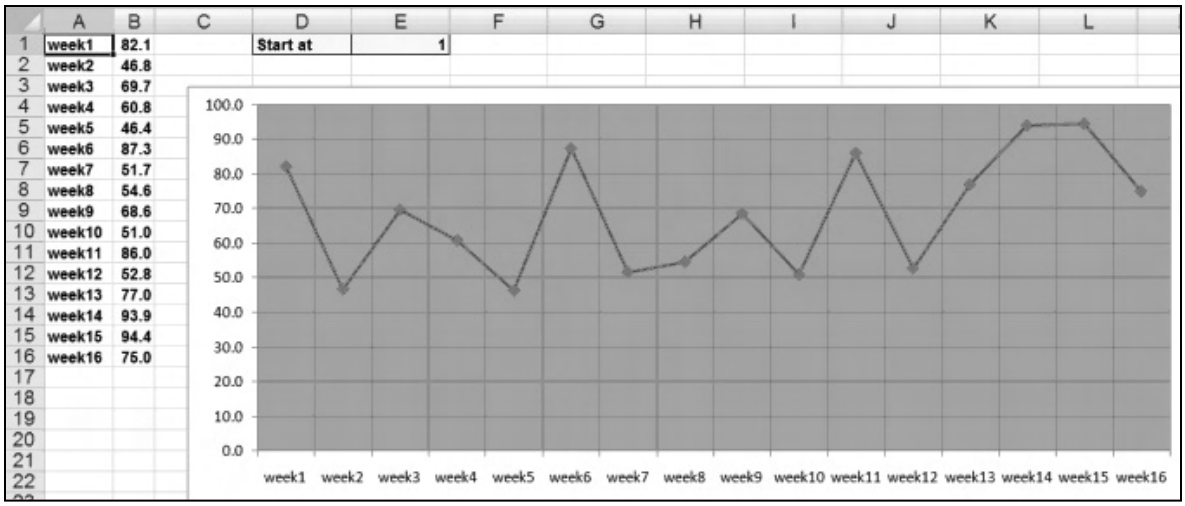


Figure: Ex-20

* * *

PART 4
Regression Analysis

Chapter 33

Linear Regression

Regression analysis is the process of making predictions of some variable, based on the relationship between this *dependent variable* and an *independent variable* (or set of variables). It is a scientist's task to find a model or an equation to make such predictions possible.

Single linear regression assumes a linear relationship between two factors: a dependent factor (y) and an independent factor (x). By using the linear equation $y=a_1x+a_0$, you can derive, estimate, determine, or predict the dependent factor (y) from the independent factor (x).

Figure 4.1 explains a bit of the terminology used in connection with linear regression. Let's assume that there is a linear relationship between hemoglobin percentage and the erythrocyte

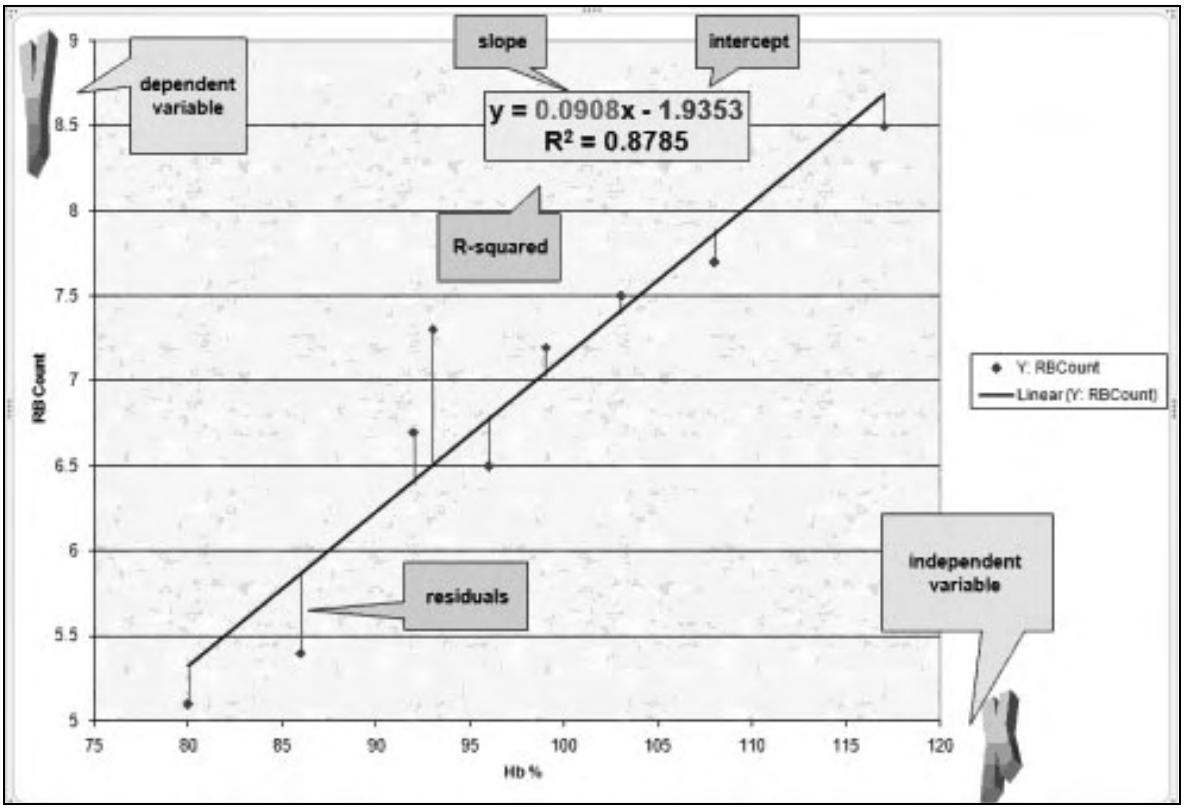


Figure: 4.1

count in human blood. You determine what the independent factor is and plot that variable on the x axis. The linear equation $y=a_1x+a_0$ uses a slope (a_1) and an intercept (a_0). Thanks to this model, you can predict, determine, or estimate y based on x.

The actual, observed values are usually clustered around the linear regression line. RSQ or R^2 is a "measure of scatter" around the regression line; it ranges from 0 to 1. The closer R^2 comes to 1, the better values coincide with the regression line, so the better you can estimate or predict. The differences between the observed values and the predicted values are called *residuals*. The linear regression line is calculated in such a way that the sum of all residuals together is 0 (so they have "evened each other out"). The residuals are basically the unexplained parts of the regression analysis—the "leftovers."

If you want to know more about what is going on in the background, you need to realize that for each point in the graph (or for each data pair in the table), Excel calculates the following sums of squares (SS):

- The squared difference between the estimated y value and the observed y value. The sum of these squared differences is called the *residual sum of squares*, SS_{resid} (which is the unexplained part of the regression).
- The total sum of squares, SS_{total}, which is the sum of the squared differences between the observed y values and the mean of the y values.
- The regression sum of squares, SS_{regr}, or the explained part of the regression: $SS_{regr} = SS_{total} - SS_{resid}$.
- RSQ, or R^2, measures how well the equation explains the relationship among the variables: R^2 equals SS_{regr}/SS_{total}.

The graph in Figure 4.2 plots the relationship between the G-C% of different DNA samples and the thermal denaturation of the double helix. What the independent factor stands for is up to you. If you want to determine, predict, or estimate T_m by using G-C%, then G-C% is the independent factor (x), whereas T_m would be the dependent factor (y). To perform regression analysis on this table, we need to gather some important information:

1. Name the x range `GC` and the y range `Tm`.
2. Use the `SLOPE` function in cell C19: `=SLOPE(Tm,GC)`.
3. Use the `INTERCEPT` function in cell C20: `=INTERCEPT(Tm,GC)`.
4. Use the `RSQ` function in cell C21: `=RSQ(Tm,GC)`.
5. Predict, determine, or estimate in cell C24 what T_m would be if you had a DNA sample containing 50% G-C bonds: `=C19*C23+C20`.
6. Predict or estimate G-C% based on T_m in cell D24: `=D19*D23+D20`.

	A	B	C	D
1			%G-C Pairs	T_m in °C
2	A plot of the G-C % of 15 different DNA samples against the thermal denaturation of the double helix (T_m)		24%	79
3			98%	110
4			55%	94
5			80%	106
6			27%	83
7			9%	73
8			87%	108
9			13%	75
10			94%	109
11			72%	99
12			36%	84
13			79%	103
14			18%	76
15			45%	90
16			62%	97
17				
18			determine T_m by using %G-C	determine %G-C by using T_m
19		Slope (a_1)	42.97705652	0.023070486
20		Intercept (a_0)	69.50755456	-1.59904624
21		R-squared	0.991501581	0.991501581
22				
23		X	50%	95
24		Y	90.99608282	0.59264993

Figure: 4.2

Note: Notice that RSQ is and should be the same after step 6 as after step 5.

Figure 4.3 presents graphically what you have done so far mathematically. To get these regression lines and their equations to show up in a graph, you follow these steps:

1. Right-click the series of data points in the top XY graph and select Add Trendline.

Note: Excel uses the term *trendline* rather than *regression line*.

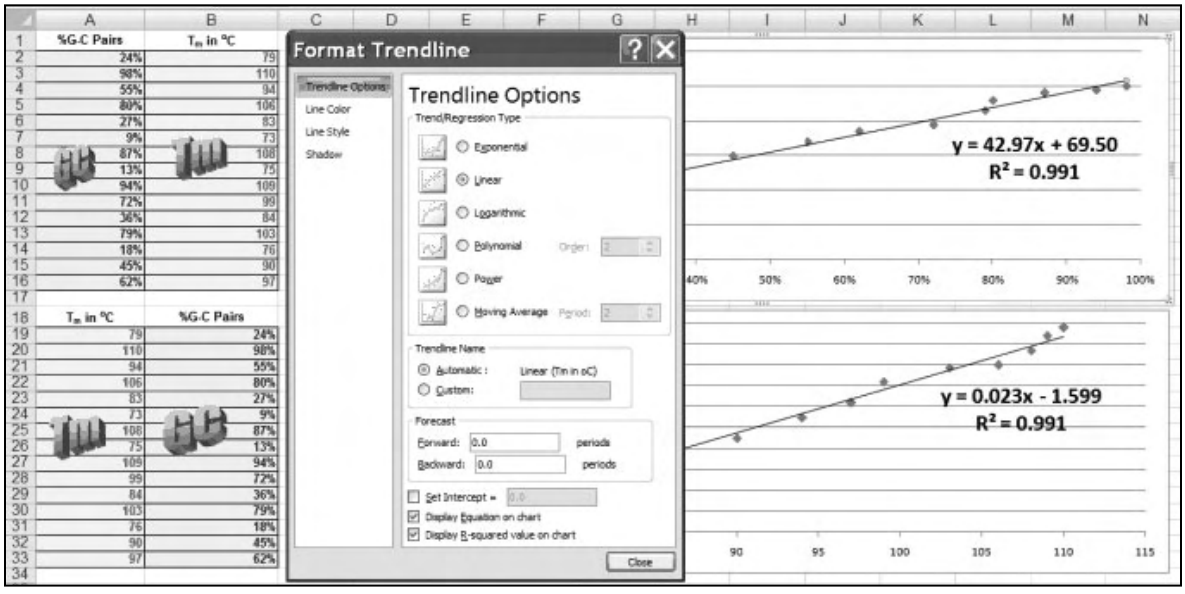


Figure: 4.3

2. By default, Excel always goes for a linear regression line—which is okay here. Make sure you also include RSQ and the equation by marking those two options at the bottom of the dialog box.
3. Notice that slope, intercept, and RSQ are identical to the ones you calculated earlier by using functions. All observed values are very close to the predicted or estimated regression line—that's why RSQ is so high.
4. Repeat steps 2 and 3 for the reversed situation where the x axis and y axis have been interchanged. Because Excel only creates regression lines for values on the x axis, you must switch the axes in order to reverse the situation.
5. Notice that both slope and intercept are different this time, but RSQ has not changed.

If you want both regression lines in the same XY graph, you must create the second one manually. In the graph shown in Figure 4.4, the RSQ was manually lowered; otherwise, both lines would have practically coincided. RSQ is basically a measure of the angle between both regression lines. If the angle were 90 degrees, RSQ would be 0, and the observation points would be scattered all over the graph.

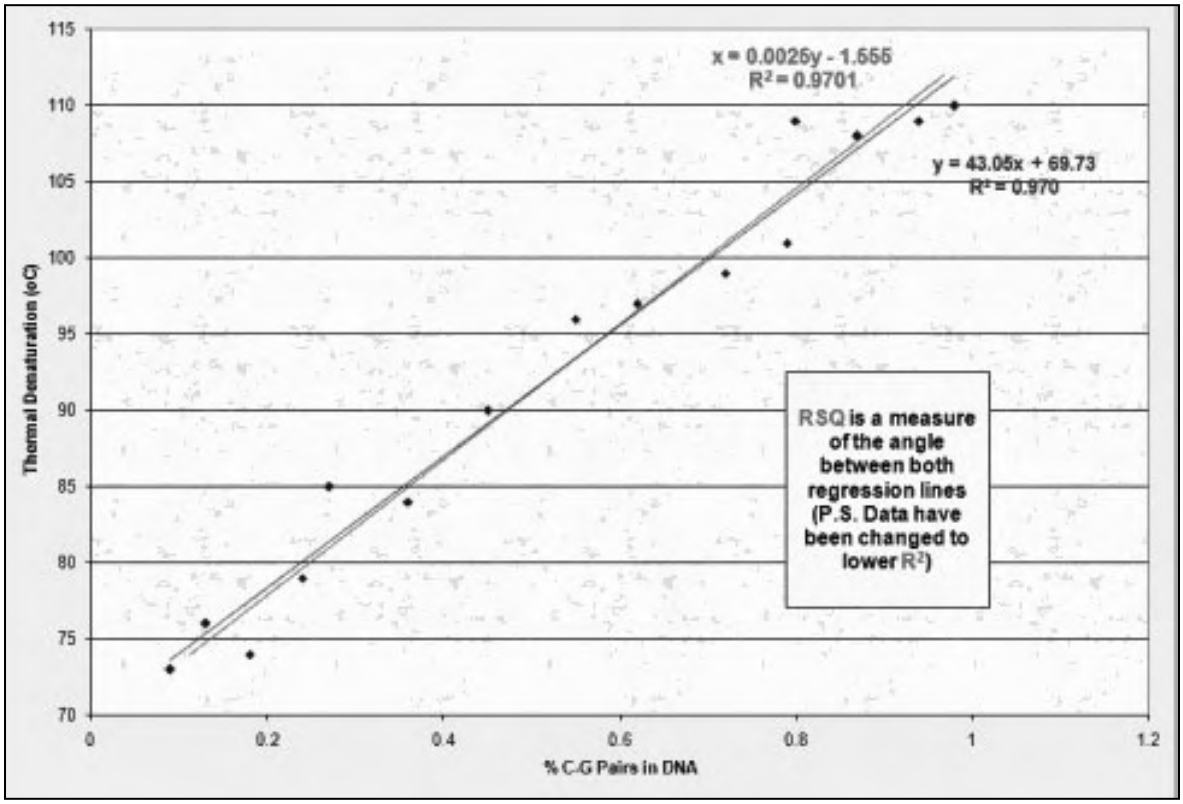
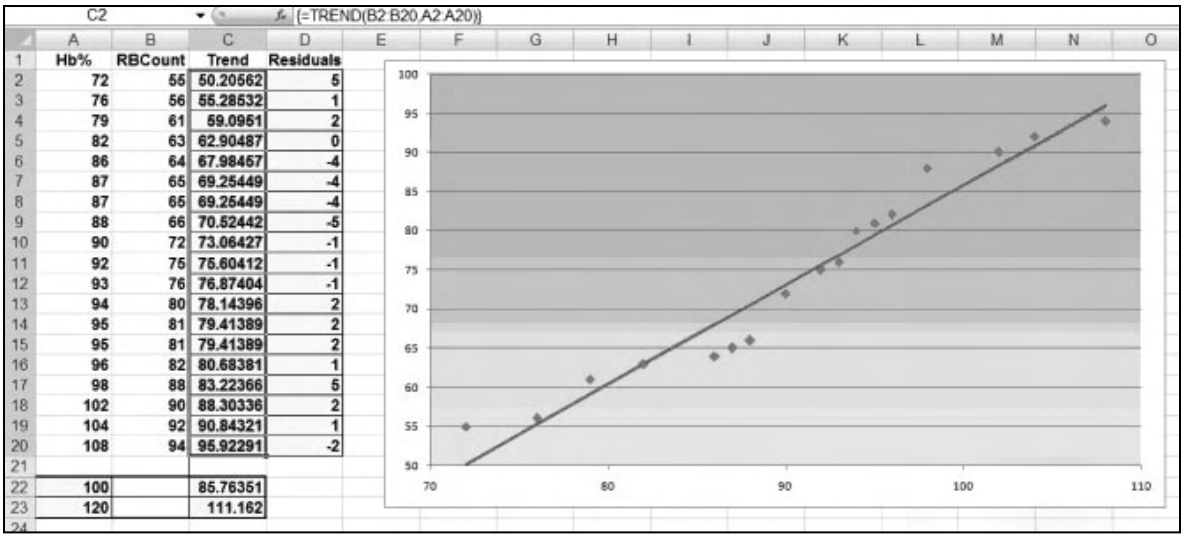


Figure: 4.4

Figure 4.5 shows another way of making predictions or estimates: by using the function TREND. Based on observed x and y values, TREND calculates a series of expected or estimated y values, assuming a linear regression. The advantage of using TREND is that you can calculate the residuals as well—that is, the difference between observed and expected (or reversed, but this book consistently sticks to the first option). Another advantage of using TREND is the possibility of predicting non-observed values. But if you want to use it in that way, you need the third argument. Let us find out what TREND can do for us:

1. Select all cells C2:C20 and apply the formula `=TREND(B2:B20,A2:A20)`.
2. Press Ctrl+Shift+Enter.



	A	B	C	D
1	Hb%	RBCount	Trend	Residuals
2	72	55	50.20562	5
3	76	56	55.28532	1
4	79	61	59.0951	2
5	82	63	62.90487	0
6	86	64	67.98457	-4
7	87	65	69.25449	-4
8	87	65	69.25449	-4
9	88	66	70.52442	-5
10	90	72	73.06427	-1
11	92	75	75.60412	-1
12	93	76	76.87404	-1
13	94	80	78.14396	2
14	95	81	79.41389	2
15	95	81	79.41389	2
16	96	82	80.68381	1
17	98	88	83.22366	5
18	102	90	88.30336	2
19	104	92	90.84321	1
20	108	94	95.92291	-2
21				
22	100		85.76351	
23	120		111.162	
24				

Figure: 4.5

> **Note:** The "manual" regression line (based on the TREND formula results) is exactly identical to the one Excel creates after a right-click on the series of observed data points.

3. Calculate in column D the residuals by entering =B2-C2 in cell D2. Notice that the sum of residuals is—as you would expect—zero. (You can watch the sum on your status bar at the bottom of your screen while the values in column D are selected.)
4. To predict non-observed values, use TREND again, this time with its third argument. The values for the argument New_x's determine what the x values should be in this case. So enter the following formula in C22:C23: `=TREND(B2:B20,A2:A20,A22:A23)`.

Caution: The first value in the TREND formula in step 4 is based on interpolation; the second one is based on extrapolation. Extrapolation is always potentially dangerous; for example, a 120% Hb content may be lethal!

Note: You can do extrapolation graphically by using Excel's regression lines. All you need to do is indicate a forward (or backward) step in the line's dialog box—in this case, 12 periods forward.

Figure 4.6 shows a graph in which you want to predict the erythrocyte count for a specific hemoglobin percentage (regulated by a control). Let's use this example to compare three different methods of interpolation:, , or 85.49 (H2), 85.54 (K2), and 85.54 (N2).

- **Using an equation**: In cell H2, enter `=1.286*G2-43.11` (taken from the equation that comes with the regression line).
- **Using TREND**: In cell K2, enter `=TREND(B2:B26,A2:A26,J2)`.
- **Using a combination of SLOPE and INTERCEPT**: In cell N2, enter `=SLOPE(B2:B26,A2:A26)*M2+INTERCEPT(B2:B26,A2:A26)`.

Only the first cell (H2) differs from the two other cells because it is based on an equation that has been rounded and is therefore less precise than the others.

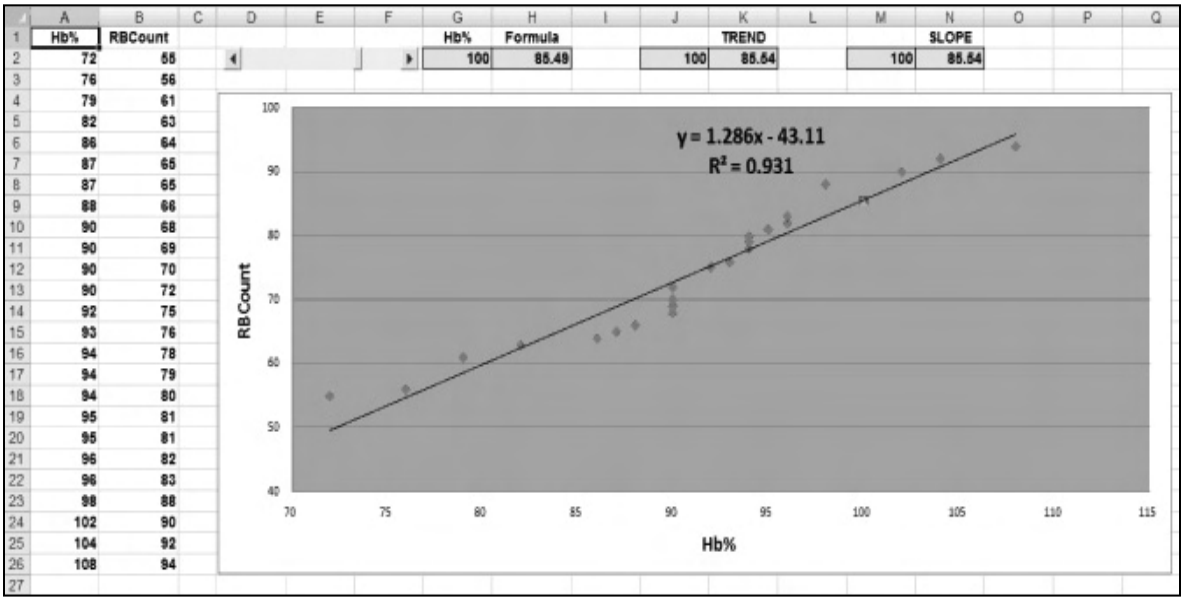


Figure: 4.6

Figure 4.7 plots the pulmonary function FVC (forced vital capacity), per liter, against the age of participants in this project. Although the RSQ is not bad here (0.96), you see many "ups and downs"—even for people of the same age. Why? In general, there are three possible explanations:

- You always deal with errors, inaccuracy, randomness, and just "noise" as part of any measuring procedure; let's disregard such factors here.
- There may be additional variables involved; we'll discuss this issue in Chapter 39.
- The relationship is not really linear; we'll discuss this in Chapter 34.

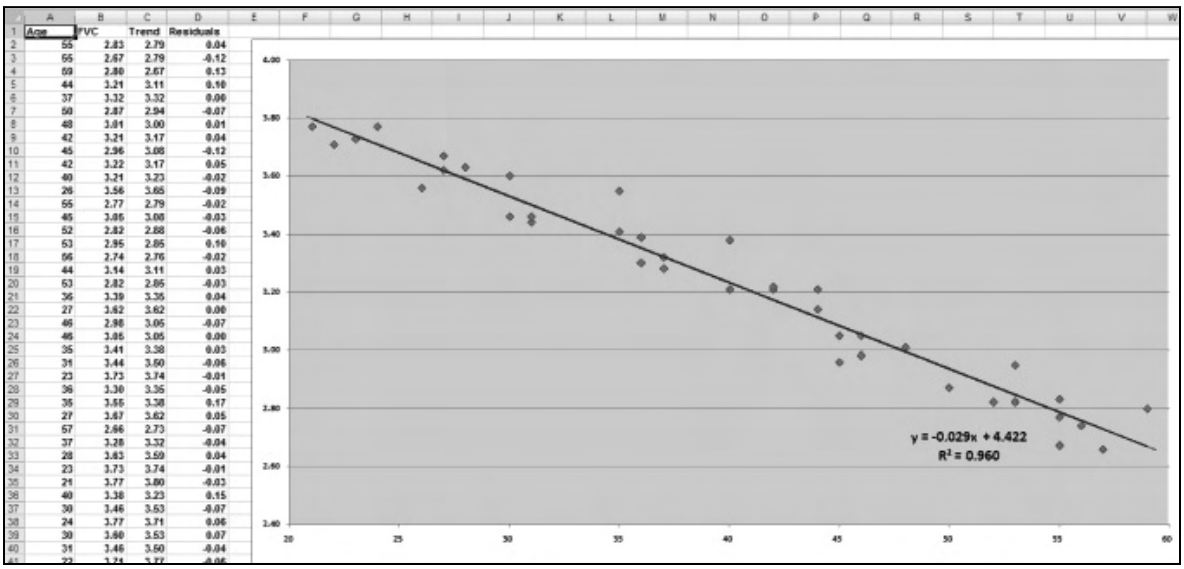


Figure: 4.7

* * *

Chapter 34

Nonlinear Regression

Many, or perhaps most, relationships between paired sets of data are not of the linear type. The most common alternatives are logarithmic, exponential, power, and polynomial. This chapter describes the characteristics of these various relationships.

The most flexible alternative to a linear relationship is a polynomial regression line. This type can grow into an nth degree curve, which can make this type flexible—but also increasingly unmanageable, as you will see. Polynomial curves essentially change direction—from up to down, or from down to up:

- $\mathbf{y = a_2x^2 + a_1x + a_0}$: This is a polynomial curve of the second order. This quadratic curve changes direction once; its pattern is concave downward (u-shaped) when slope a_2 is negative, and it is concave upward (n-shaped) when a_2 is positive.
- $\mathbf{y = a_3x^3 + a_2x^2 + a_1x + a_0}$: This is a polynomial curve of the third order and is also called *cubic*. The cubic or third-degree curve changes direction twice; it is ~-shaped.
- $\mathbf{y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0}$: This is a polynomial curve of the fourth order. The fourth-degree curve changes direction three times; it is w- or m-shaped.
- $\mathbf{y = a_nx^n + \ldots + a_1x + a_0}$: This is an *n*th degree curve—that is, of the *n*th order. The *n*th degree curve changes direction (*n*-1) times.

Figure 4.8 shows a linear regression line with a decent RSQ, but it is not the best regression model. The following are two possible improvements:

- You can add a new line by right-clicking the series and selecting Add Trendline.
- You can replace the current line by right-clicking the line and selecting Format Trendline.

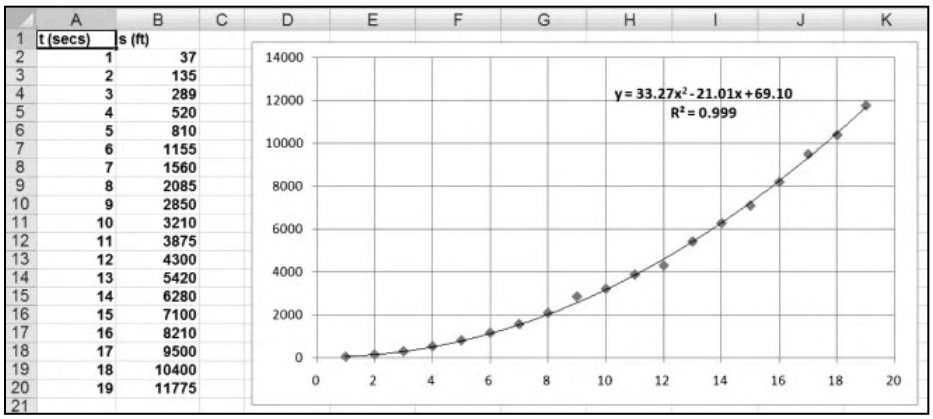


Figure: 4.8

A polynomial regression line of the second order seems to fit better and makes RSQ go up. Yes, this is a trial-and-error method, but it is good enough for now. (You will do better in Chapter 35.)

Figure 4.9 shows the fraction of un-dissociated acetic acid as a function of the pH. Trial and error has suggested a polynomial regression line of the fifth order! As you can see, you can often force your data into a polynomial regression line by adding more and more slopes. The result is often awkward. You could often get the same result with a much simpler formula.

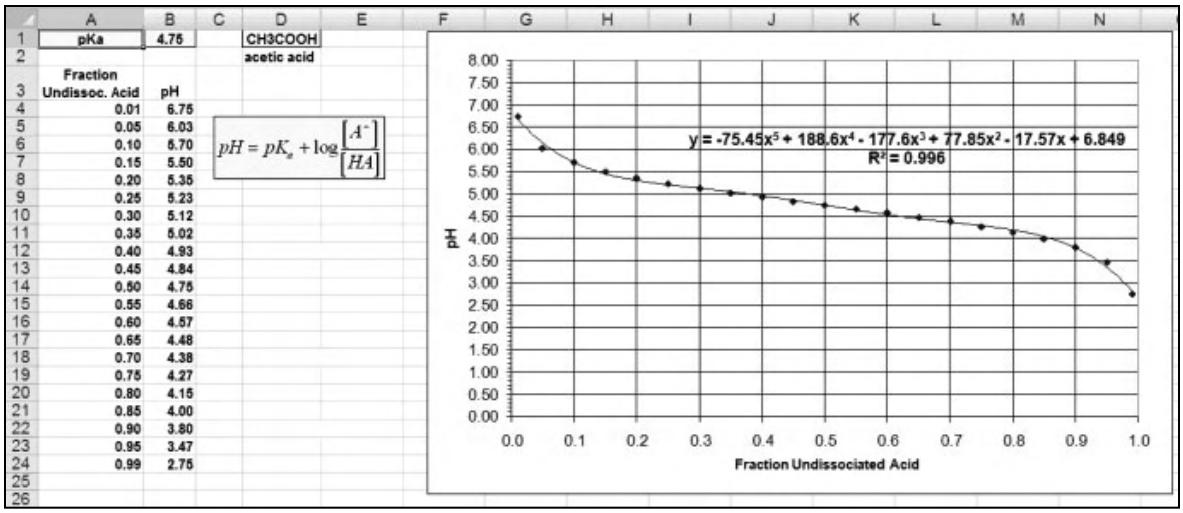


Figure: 4.9

Figure 4.10 shows the increase of population size after each generation. It is definitely a case in which a linear regression line wouldn't fit. You may get far with a polynomial line, but your intuition probably tells you to go for an exponential regression line. Shortly, we will discuss the equation, slope, and intercept of an exponential regression line. For now, remember that an exponential curve is the upper part of a hyperbola—always located above the x axis: It goes up toward infinity when its exponent is positive; it goes down asymptotically toward 0 when its exponent is negative.

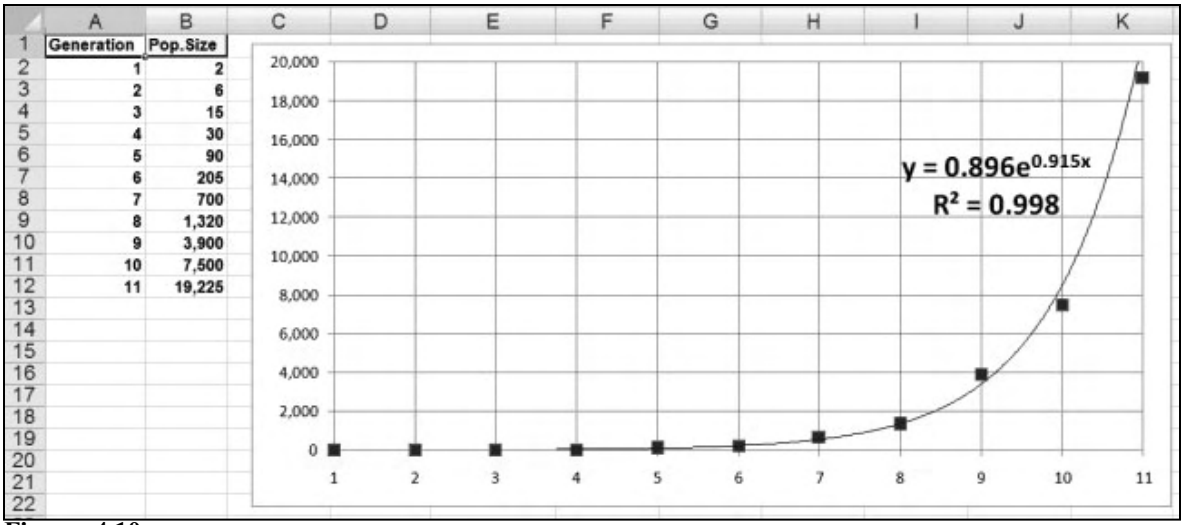


Figure: 4.10

Figure 4.11 shows a similar case—this time for the metabolic rate of breaking down lactose into D-glucose and D-galactose. Again, you could get far with a high-degree polynomial version (the most curved one). But it probably makes more sense to go for a power regression line—although that brings down RSQ a little. Don't forget that you have just a few observations here, so observation errors and randomness have a more prominent impact.

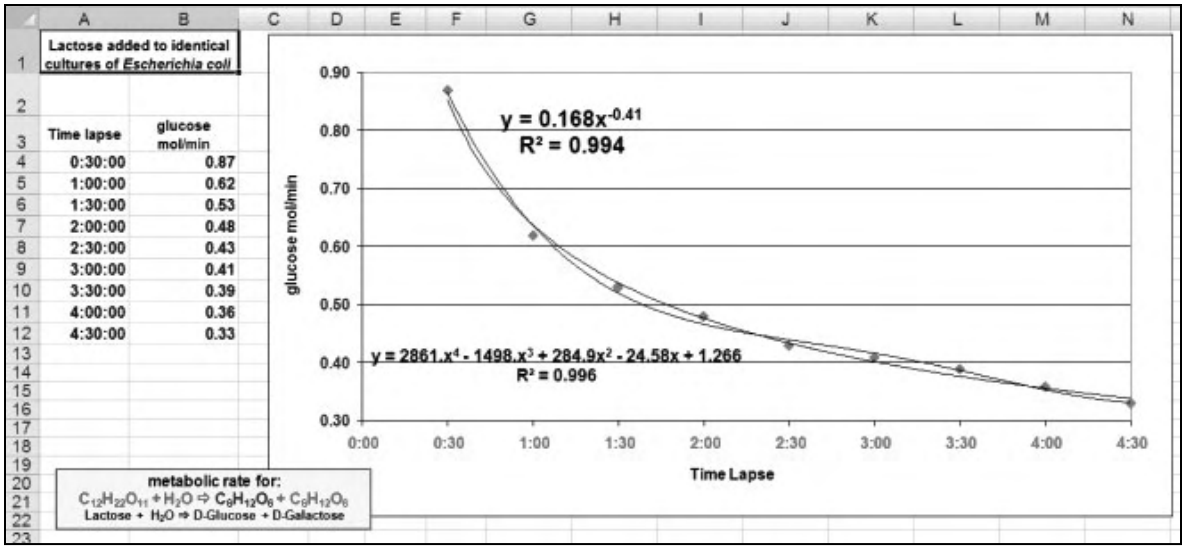


Figure: 4.11

Figure 4.12 shows the facilitated uptake of glucose by erythrocytes. A polynomial regression line would be a stretch in this situation. It probably makes much more sense to try a logarithmic regression line. But even that latter fit is not great because you seem to be dealing here with a sigmoid curve, which you'll learn more about in Chapter 36.

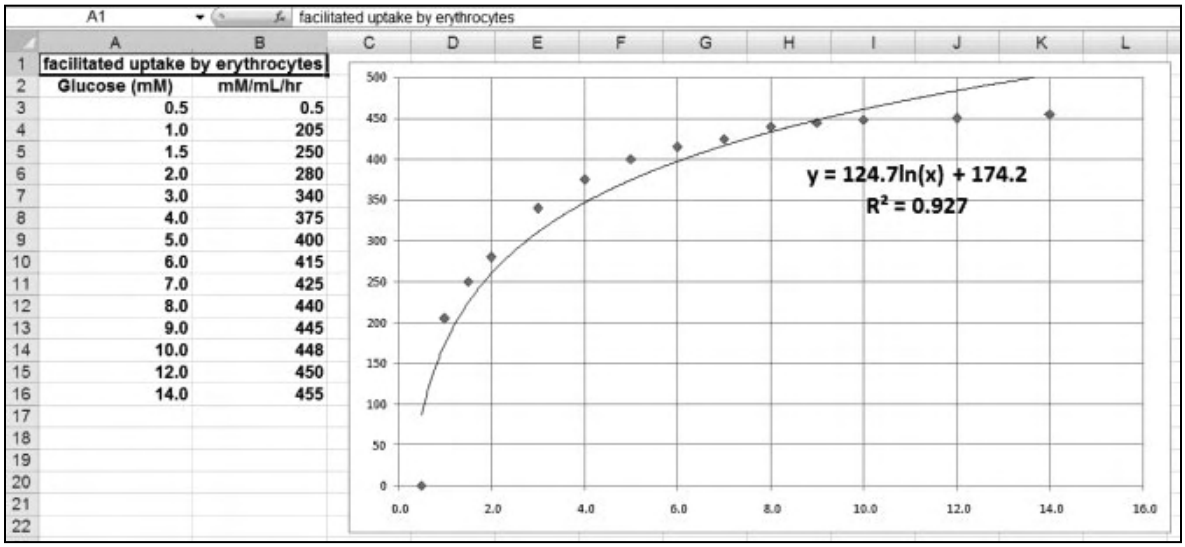


Figure: 4.12

Figure 4.13 provides an overview of the formulas behind the curves we've looked at so far. Excel offers the following functions for use in this context:

- `LN(x)`: `LN` returns the natural logarithm of `X`. Natural logarithms are based on the constant e (that is, 2.71828182845904).
- `EXP(X)`: `EXP` returns e raised to the power of `X`. The constant e equals 2.71828182845904, the base of the natural logarithm. The natural logarithm of e raised to the power of 3 would be `LN(EXP(3))`.
- `LOG(X,base)`: `LOG` returns the logarithm of `X` to the base you specify; the base is 10 by default. So `LOG(X,EXP(1))` is the same as `LN(X)`.
- `POWER(X,n)`: `POWER` returns the result of `X` raised to a power of *n* and is equivalent to `X^n` in Excel.

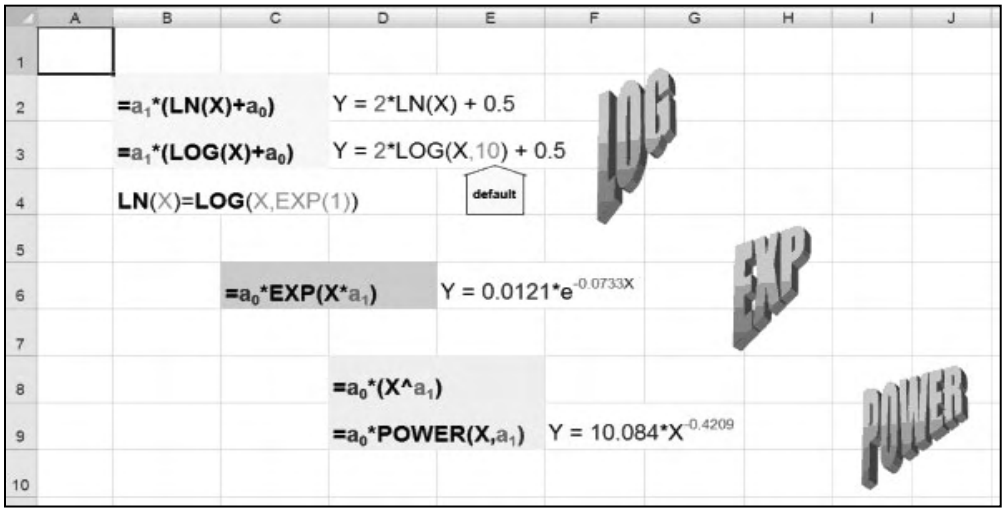


Figure: 4.13

Now that you know the formulas, you are better equipped to understand what is behind each curve. At any time, you can "linearize" values by using the `LN` function. Which values need to be linearized in order to get a linear regression curve? Depending on whether you

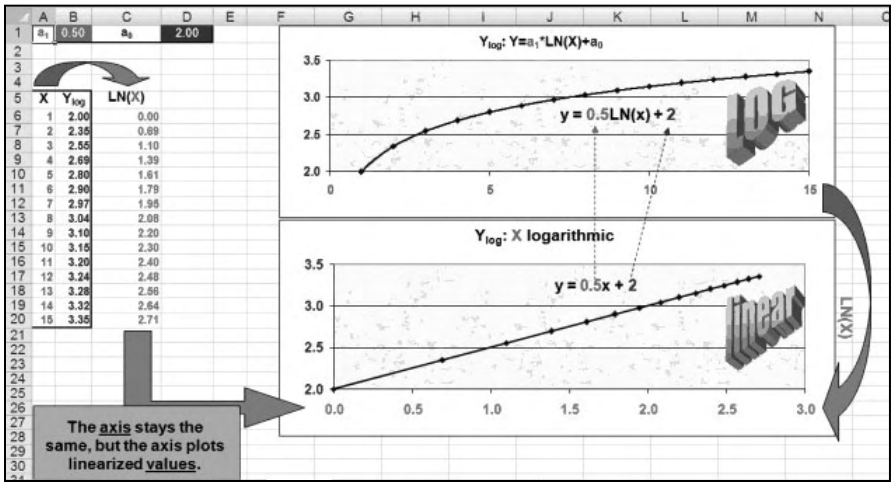


Figure: 4.14

are dealing with a logarithmic, exponential, or power relationship, this question receives a different answer:

- Logarithmic: Figure 4.14 represents a logarithmic case. At any time, you can "linearize" a logarithmic curve by linearizing the x values (in column A) via the `LN` function used in column C. The top graph uses the original values from column A on its x axis. The bottom graph, on the other hand, uses the linearized values of column C on its x axis instead. Notice the linear regression line in the bottom graph versus the logarithmic regression line in the top graph. The logarithmic equation (in the upper graph) has the same slope and intercept as the linear equation (in the lower graph).

- Exponential: Figure 4.15 has an exponential curve in its top graph. In order to linearize an exponential curve, you must apply the `LN` function to the y values from column B—which has been done in column C. The bottom graph uses the linearized values from column C on its y axis. Notice how the slope and intercept match again between the exponential and linear versions of regression, provided that you "unlinearize" the linear slope first by using the function `EXP`.

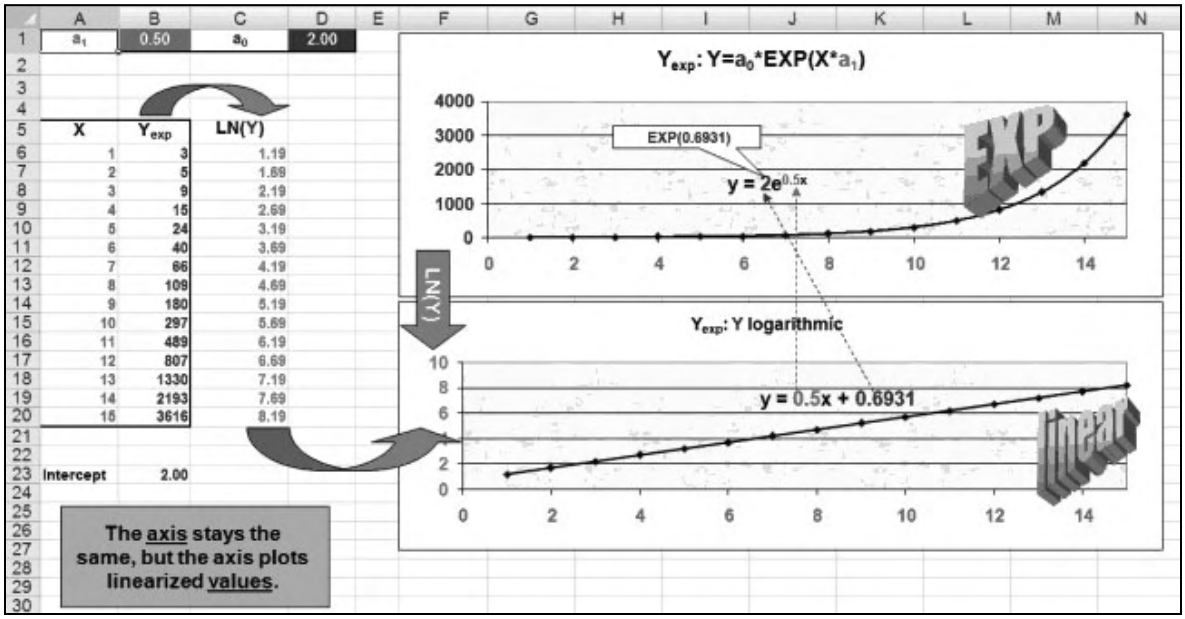


Figure: 4.15

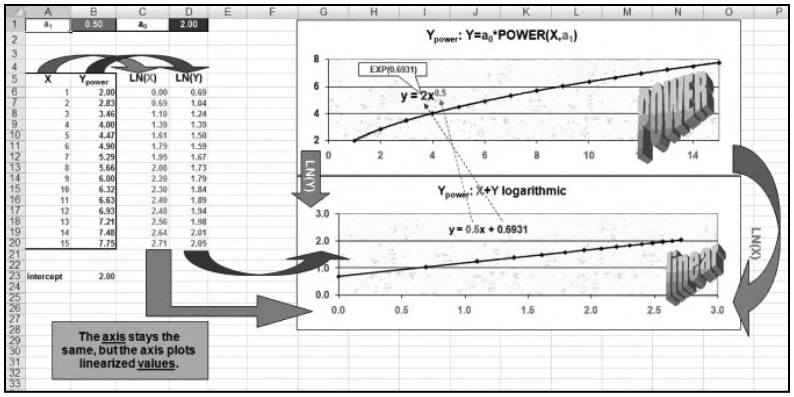


Figure: 4.16

- Power: Figure 4.16 shows a power curve in its top graph. In order to linearize a power curve, you must apply the `LN` function to both x values and y values. In this case, you find linearized values on both axes of the bottom graph.

Instead of linearizing values, you could instead linearize the axis by making its scale logarithmic. Remember that `LN(X)` is the same as `LOG(X,EXP(1))` (see Figure 4.17). Be aware that the graphs to the right use the regular values—not the linearized ones—so each regression line is of the same type as the one used by its partner graph to the left. To put it differently: You linearized the axis here, not the values. Instead of using linearized values, you created a logarithmic scale for a specific axis:

- A logarithmic curve needs the x axis to be logarithmic.
- An exponential curve needs the y axis to be logarithmic.
- A power curve needs both axes to be logarithmic.

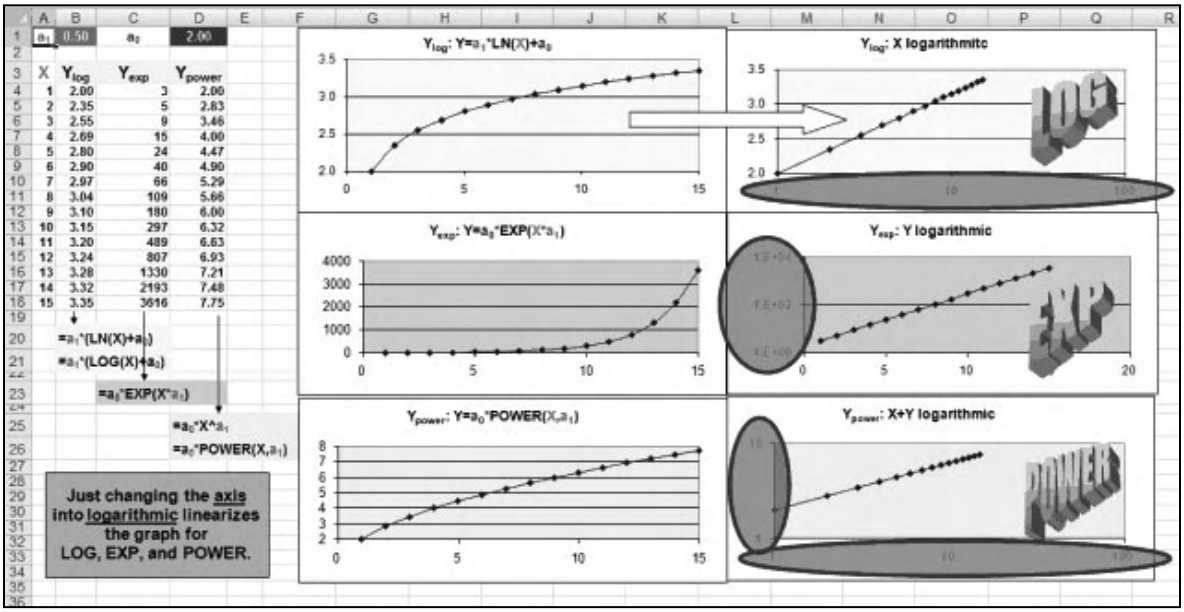


Figure: 4.17

* * *

Chapter 35

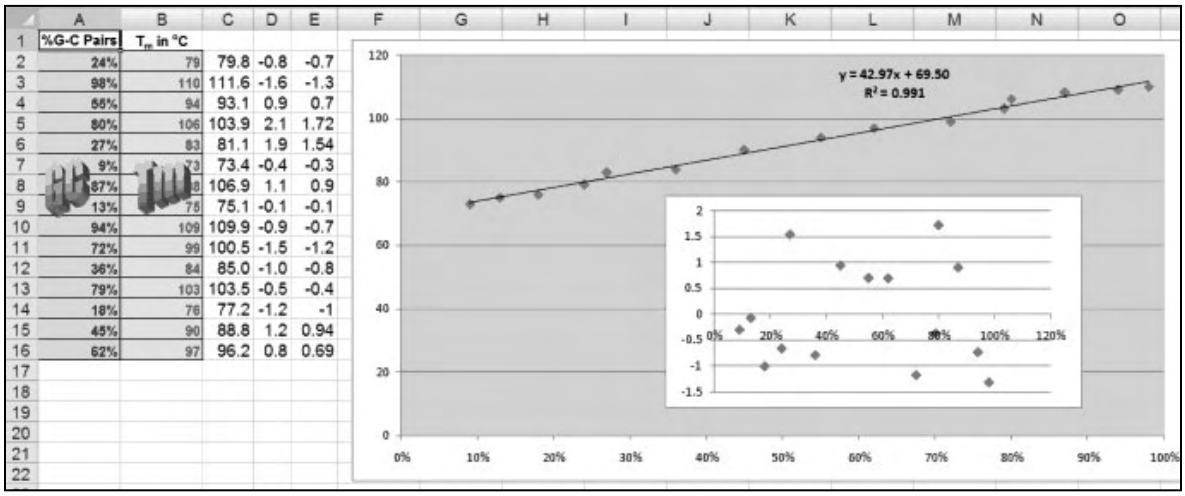
Curve Fitting

Finding the proper regression curve in a methodical way is called *curve fitting*. Curve fitting is the process of trying to find the model or equation that best represents the sample data. So far, you have used intuition and trial-and-error to do this—and that may very often work. But there is also a more methodical way of finding a model and testing whether it fits the data.

You used some clear cases of linear regression earlier in this section. When we speak of a "clear" case, we usually mean that the observed dots seem to nicely coincide with the estimated or predicted linear curve. However, looks may be very deceiving. Changes in the axis scales may change our impressions drastically. Even nonlinear cases may seem linear if you work the scales a bit. So you need additional tools to test the assumed regression model. You have two tools available:

- **The RSQ value**: As discussed before earlier, improving `RSQ` is basically still a matter of trial-and-error.
- **The residuals method**: Residuals should be randomly scattered without showing a particular pattern. Let's now study this method more in detail.

Figure 4.18 shows a nice linear relationship between the percentage of G-C bonds in the DNA helix and the temperature of denaturation. You want to apply the second method here—



	A	B	C	D	E
1	%G-C Pairs	T_m in °C			
2	24%	79	79.8	-0.8	-0.7
3	98%	110	111.6	-1.6	-1.3
4	66%	94	93.1	0.9	0.7
5	80%	106	103.9	2.1	1.72
6	27%	83	81.1	1.9	1.54
7	9%	73	73.4	-0.4	-0.3
8	87%	108	106.9	1.1	0.9
9	13%	75	75.1	-0.1	-0.1
10	94%	109	109.9	-0.9	-0.7
11	72%	99	100.5	-1.5	-1.2
12	36%	84	85.0	-1.0	-0.8
13	79%	103	103.5	-0.5	-0.4
14	18%	76	77.2	-1.2	-1
15	45%	90	88.8	1.2	0.94
16	62%	97	96.2	0.8	0.69

Figure: 4.18

the residuals test:

1. To find the residuals, calculate in column C what the predicted/estimated y values would be if the linear model were correct: `=TREND(B2:B16,A2:A16)`.
2. In column D, calculate the residuals: $Y_{observed} - Y_{predicted}$.

> **Note**: *y* predicted is also called *y* expected and *y* estimated.

3. Plot X versus the residuals in an XY graph (the result is shown in the insert in Figure 4.18). If you are really dealing with a linear relationship, the residuals should be randomly distributed above and below the x axis—and they are in this case!
4. Ideally, go one step further and calculate the standardized version of residuals: residual / $SD_{residuals}$, or `=D2/STDEV($D$2:$D$16)` in column E.

> **Note**: Notice that all the dots nicely occur in the range between -2 and +2 in the standardized residuals. If there are a few standardized residuals beyond +3 or -3, they are called *outliers*; you may want to check whether those extreme values are reliable or find out what else is going on. A quick test for outliers is the function `ZTEST`; it returns the probability of finding a specific value in a population; probabilities below 1% make for outliers.

The bottom line of the residuals test is that the pattern of residuals should be random:

- There shouldn't be any visible pattern.
- Most residual values should be centered around the zero line.
- The farther you get from zero, the fewer of them should occur.
- 95% of them are expected to lie between -2 and +2.

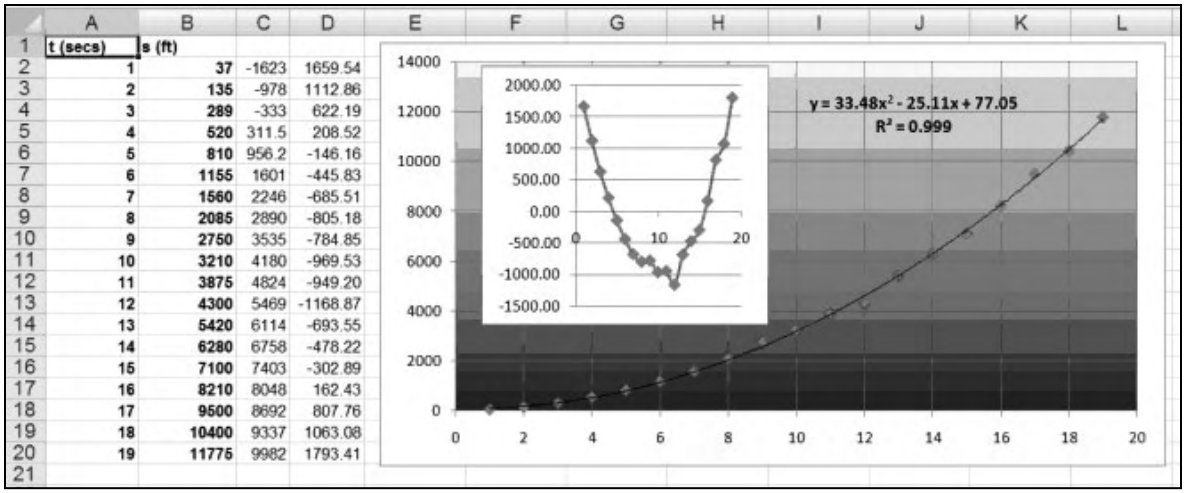
In other words, if the residuals are not randomly distributed above and below the x axis—but show a distinctive pattern instead—you are not dealing with linear regression and need another, nonlinear model.

As you have seen, to determine whether the regression line is linear, you calculate the trend, determine the residuals, and plot the residuals against their y values.

Figure 4.19 shows a case in which you definitely should use the residuals method of curve fitting:

1. Use `TREND` in column C.
2. Calculate residuals in D: `=B-C`.
3. Plot the residuals against the y values (see the insert in Figure 4.19). The pattern of the residuals is far from randomly scattered. It looks more like a parabola.

A parabola is a curve type that is symmetric on both sides of a vertical axis, which runs parallel with the y axis. When the pattern of residuals comes close to being a parabola, you should at least consider using a polynomial type of regression.



t (secs)	s (ft)		
1	37	-1623	1659.54
2	135	-978	1112.86
3	289	-333	622.19
4	520	311.5	208.52
5	810	956.2	-146.16
6	1155	1601	-445.83
7	1560	2246	-685.51
8	2085	2890	-805.18
9	2750	3535	-784.85
10	3210	4180	-969.53
11	3875	4824	-949.20
12	4300	5469	-1168.87
13	5420	6114	-693.55
14	6280	6758	-478.22
15	7100	7403	-302.89
16	8210	8048	162.43
17	9500	8692	807.76
18	10400	9337	1063.08
19	11775	9982	1793.41

Figure: 4.19

Each of the four main nonlinear regression types shows its own distinctive pattern of residuals. How do they differ, then, in their residuals patterns? Figure 4.20 systematizes their differences:

- Only the polynomial type has a symmetrical, vertical residuals pattern—a parabola. This pattern is concave downward when the slope, a_2, is negative, and it is concave upward when a_2 is positive.
- The three other types are not symmetrical, and their central axis is not parallel to the y axis. How do they differ? There are some general tendencies in their residuals pattern, as

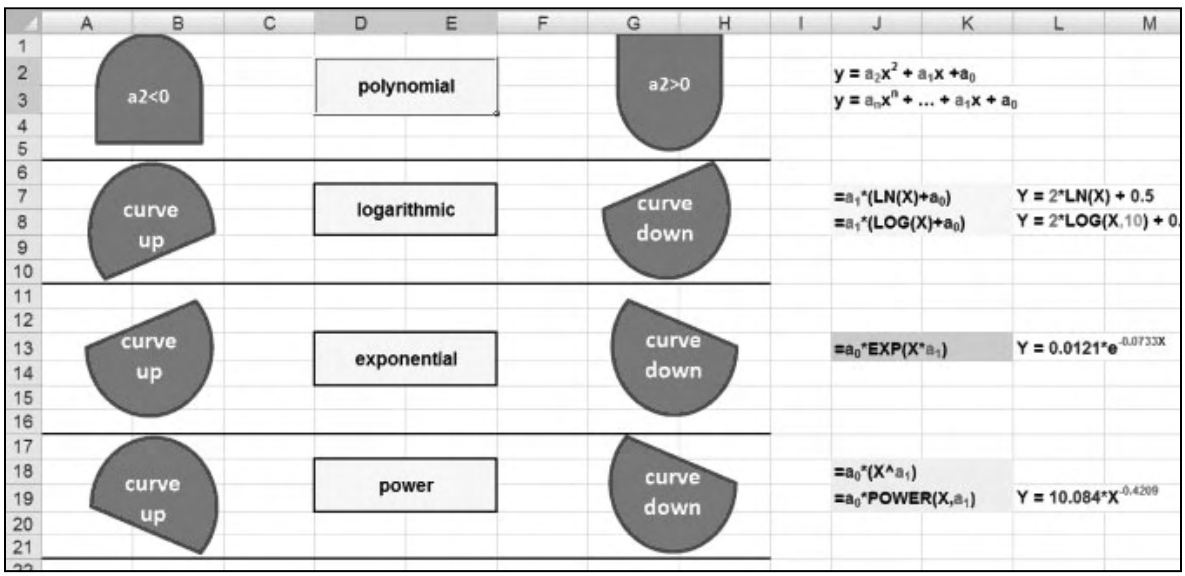


Figure: 4.20

summarized in Figure 4.20. This pattern also depends on whether the original curve went upward or downward.

In many cases, the residuals pattern based on linear regression may not clearly fit into one of the particular categories described here. So you may still end up with trial-and-error by checking whether the RSQ value went up or down. However, when you have come up with a "final" model, you should not forget to test this nonlinear model—no matter what type it is—with a new residuals test, but this time based on the nonlinear model. You should check your nonlinear model by using its new regression equation and then plotting the residuals pattern again. The residuals pattern should be randomly scattered, without showing any distinctive pattern. If the pattern is not randomly scattered, you have to keep revising your latest model until you come up with one that does have a random pattern in its residuals test.

Figure 4.21 presents a case in which several nonlinear models would qualify—including a linear model. To reject the linear model quickly, you might want to go for a less time-consuming alternative: using the Analysis Toolpak, which has a tool for regression analysis that calculates all the related statistics plus a series of plots, including a residuals plot. Based on this information, you can decide whether linear regression is an acceptable model or has to be replaced by a nonlinear model.

The Analysis Toolpak comes with Excel as an add-in, but it is not automatically active. In order to activate it, you need to install it:

1. Click the Office icon.
2. Choose Excel Options.
3. Select Add-Ins.
4. Under Manage, select Excel Add-Ins.
5. Click Go.

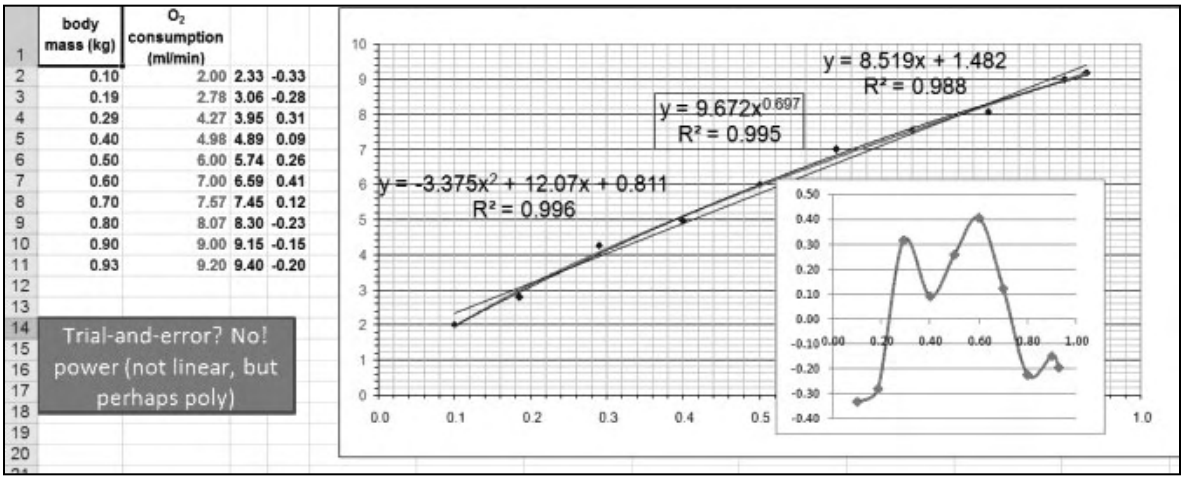
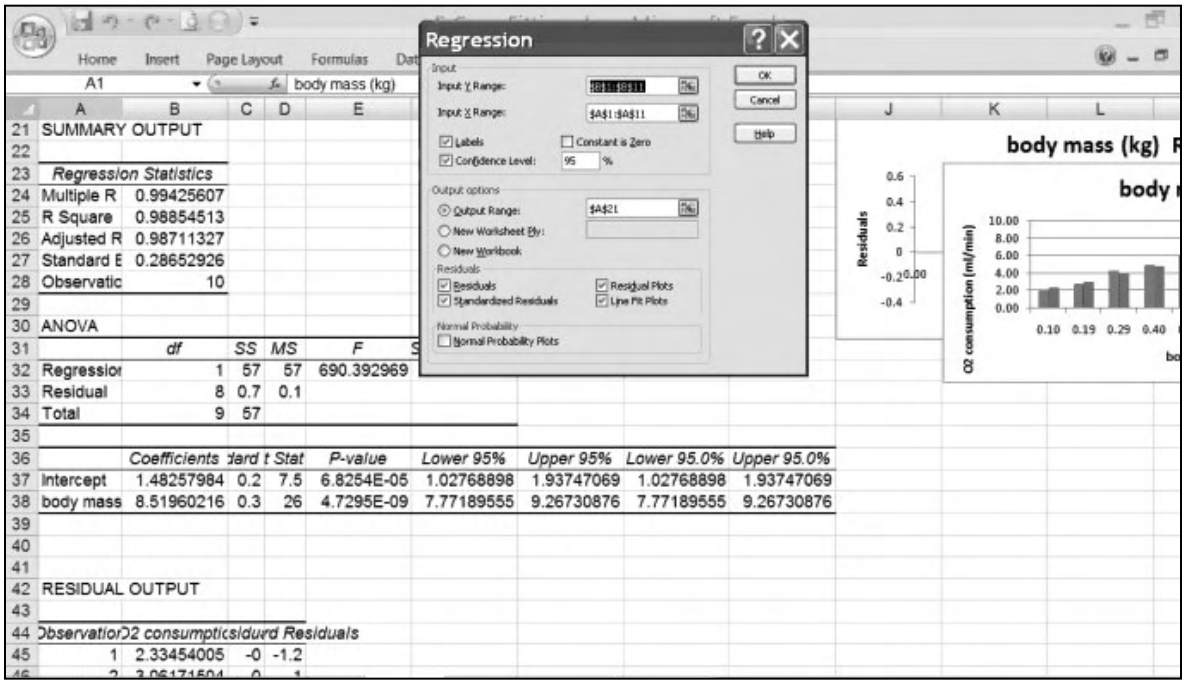


Figure: 4.21

When the Analysis Toolpak is active, it is available through the Data tab. You can use it for several purposes, including regression analysis.

Figure 4.22 shows a residuals plot, based on linear regression, of the data from Figure 4.21. It shows a rather distinctive pattern, so you should reject the linear model and test the residuals pattern with alternative models. Unfortunately, the Analysis Toolpak cannot do the latter part for you; that's still a manual task.



	A	B	C	D	E
21	SUMMARY OUTPUT				
22					
23	*Regression Statistics*				
24	Multiple R	0.99425607			
25	R Square	0.98854513			
26	Adjusted R	0.98711327			
27	Standard E	0.28652926			
28	Observatic	10			
29					
30	ANOVA				
31		*df*	*SS*	*MS*	*F*
32	Regressior	1	57	57	690.392969
33	Residual	8	0.7	0.1	
34	Total	9	57		
35					
36		*Coefficients*	*dard t Stat*	*P-value*	*Lower 95%*
37	Intercept	1.48257984	0.2	7.5	6.8254E-05
38	body mass	8.51960216	0.3	26	4.7295E-09
39					
40					
41					
42	RESIDUAL OUTPUT				
43					
44	*Observatior*	*O2 consumptic*	*sidurd Residuals*		
45	1	2.33454005	-0	-1.2	

	J	K	L	M	N	O	P	Q	R
36	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*						
37	1.93747069	1.02768898	1.93747069						
38	9.26730876	7.77189555	9.26730876						

Figure: 4.22

* * *

Chapter 36

Sigmoid Curves

Many regression curves used in research (especially in the life sciences) are S-shaped; they are also called *sigmoid*, or *logistic*, *curves*. None of the nonlinear types discussed so far cover these cases (although a high-level polynomial model may come close sometimes). Sigmoid curves require much more work than the common nonlinear curves.

A general characteristic of a sigmoid curve is that, in the beginning, the curve grows exponentially, but its growth gets increasingly inhibited up to an asymptotic saturation point. So, in general, a sigmoid curve has two important landmarks or markers:

- **The inflection point**: This is where a concave upward curve changes into a concave downward curve (or vice versa); this point is also called the *saddle*.
- **The saturation point**: This is the maximum value and is approached asymptotically.

Figure 4.23 offers a typical example of a sigmoid curve: It has an inflection point (*N*+100) and a saturation point (*N*+200). The formula shows that the rate of increase depends on the relationship between the value reached (*N*) and the value of the saturation point (Max). Closer to Max, *n*/Max comes closer to 1, so (1 - *n*/Max) comes closer to 0 and thus reduces the rate more and more. This is why column C in Figure 4.23 initially goes up and then

	A	B	C	D
1	t	N	$Growth_{predicted}$	$Total_{predicted}$
2	1	10	10	10
3	2	15	5	15
4	3	22	7	22
5	4	32	10	31
6	5	45	13	45
7	6	63	17	62
8	7	80	21	84
9	8	111	24	107
10	9	133	25	132
11	10	153	22	154
12	11	177	18	172
13	12	185	10	182
14	13	189	7	189
15	14	194	5	194
16	15	197	3	197
17	16	198	1	199
18	17	199	1	200
19	18	200	0	200
20		Maximum	200	
21		Rate	0.497193474	
22				

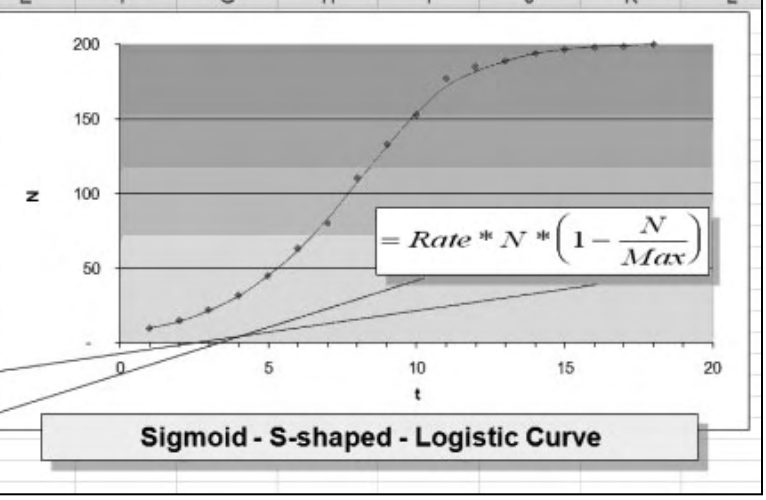


Figure: 4.23

gradually slows down. But to get the formula behind column D, you need mathematical tools such as differentiation and integration. Integration would make the distance between the x values infinitely small.

Many scientists don't feel comfortable with this kind of math, especially with integration. Excel can help you with the math, but it does not do the math for you. So next we'll look at some more friendly alternatives. If you know how to integrate painlessly and feel comfortable doing it, go ahead and skip the rest of this chapter. Otherwise, you might discover an alternative way of dealing with sigmoid curves. Here are two suggestions:

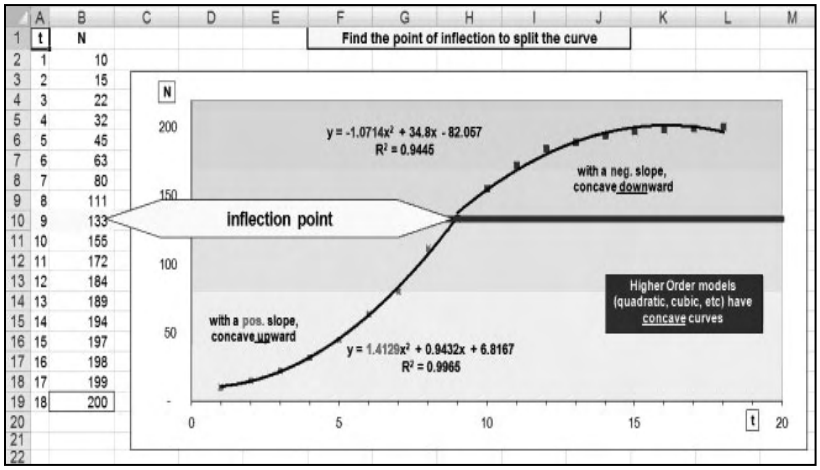


Figure: 4.24

- Figure 4.24 shows a simple, but rather primitive, solution. You locate the inflection point (by estimating visually) and split the series of data points into two sections—each one with its own polynomial curve. This solution is very unsatisfying because you need two different equations to predict a single regression line.

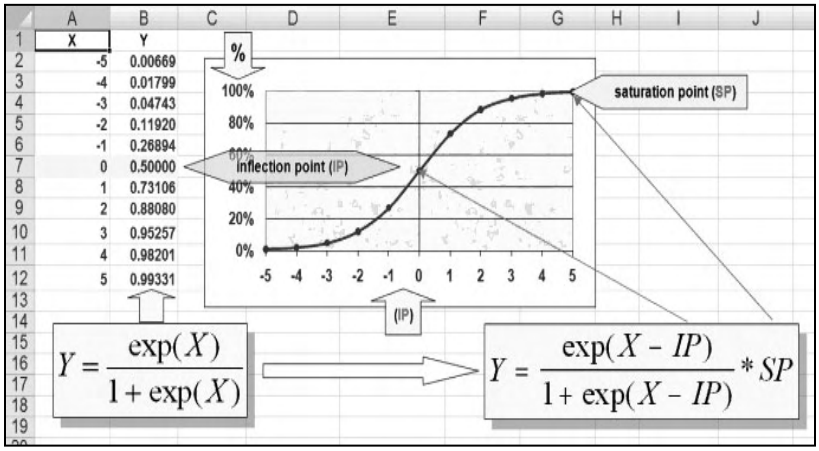


Figure: 4.25

- Figure 4.25 offers a more sophisticated approach (without requiring the use of integration). The formula on the left creates a sigmoid curve, but it returns values between 0 and 1 (%). So you could transform it into the formula on the right by including the inflection point (IP) and the saturation point (SP).

For the rest of this chapter, you will use the second approach, with an improved formula. Figure 4.26 shows the addition of a slope (a_1) to the equation. Now three variables in the equation help you regulate the sigmoid curve:

- SP determines the asymptotic top of the curve.
- IP allows you to shift the curve to the left or to the right.
- a_1 determines the steepness of the curve.

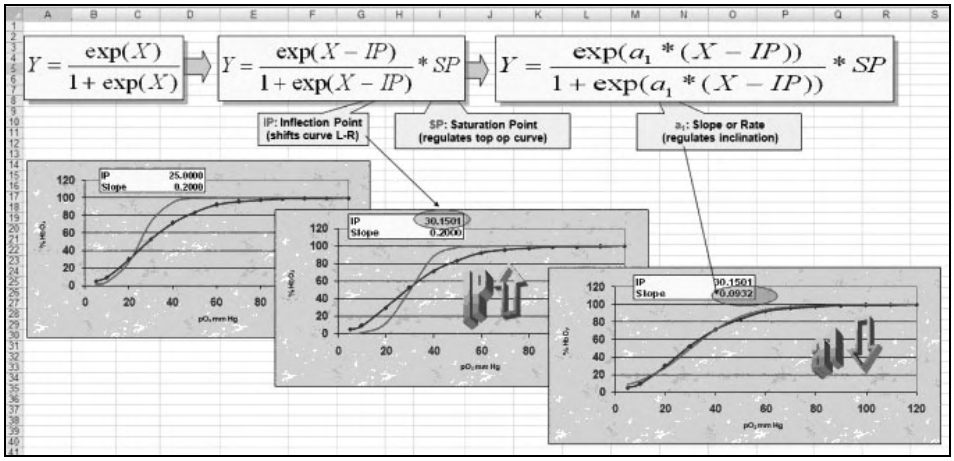


Figure: 4.26

Figure 4.27 has two controls to change settings for IP and slope independently (see Chapter 42). Moving the controls makes the sigmoid regression curve change in response, which in turn affects the sum of the squared residuals in cell D21. The first control regulates IP and thus moves the curve to the left or to the right. The second control regulates the slope and thus changes the steepness of the curve. Using trial-and-error on the controls, I have come up with the settings IP=7.5388 and slope=0.5032. Perhaps you can do better! When you have proper settings for slope, IP, and SP, you can apply the equation to column C in order to create a sigmoid regression line—without using integration!

Chapter 41 shows another way of finding the best settings for IP and slope—without the use of trial-and-error controls.

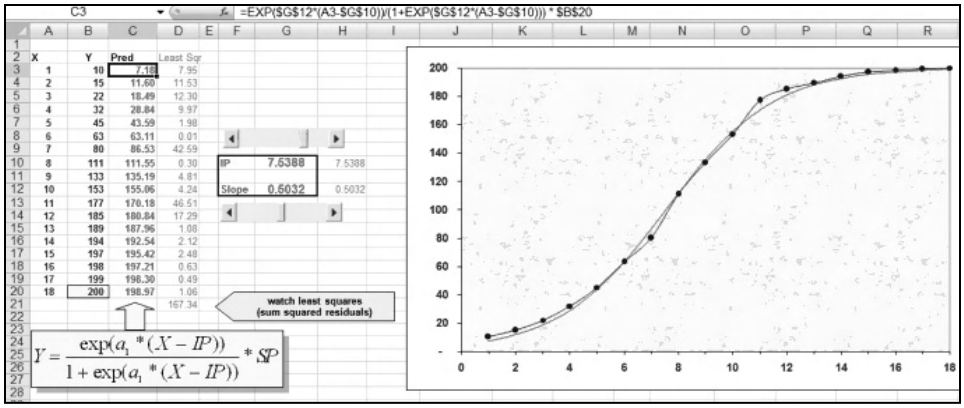


Figure: 4.27

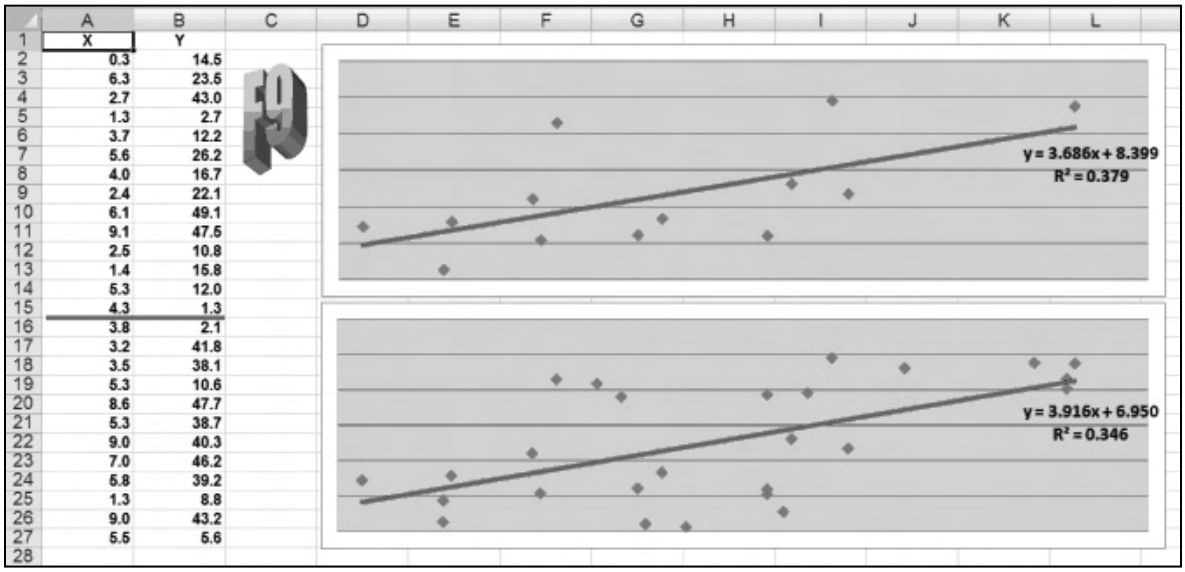
* * *

Chapter 37

PREDICTABILITY

Regression analysis is aimed at predictability. But predictability is a very ambiguous concept. Does it mean how close (or far off) your predictions are in relationship to your observations? RSQ is a good measure of that closeness. It measures how well you can predict in a particular sample. But there is another dimension to predictability: How well can you replicate the results you have found so far? In other words, how well can you predict in the population from which this particular sample comes? This chapter tackles the latter question, which takes us deeper into statistics.

The fact that you have found a linear regression line in your data—even with a "reasonable" RSQ value—doesn't mean you have hit on a "real" connection. Figure 4.28 illustrates this phenomenon. All the "observations" plotted in these two curves are based on random numbers in columns A and B, generated by the RAND function. The top graph covers all data up to row 15; the bottom graph covers all data up to row 27. If you just keep pressing F9, you may get some terrific (but unscientific) results! A high RSQ means only that a relatively high proportion of y variance can be credited to x variance. But variance can be a random



	X	Y
2	0.3	14.5
3	6.3	23.5
4	2.7	43.0
5	1.3	2.7
6	3.7	12.2
7	5.6	26.2
8	4.0	16.7
9	2.4	22.1
10	6.1	49.1
11	9.1	47.5
12	2.5	10.8
13	1.4	15.8
14	5.3	12.0
15	4.3	1.3
16	3.8	2.1
17	3.2	41.8
18	3.5	38.1
19	5.3	10.6
20	8.6	47.7
21	5.3	38.7
22	9.0	40.3
23	7.0	46.2
24	5.8	39.2
25	1.3	8.8
26	9.0	43.2
27	5.5	5.6

Figure: 4.28

effect—as it is here. The more cases you have (as in the lower curve), the more unlikely it is that both variances correlate by chance—but the possibility is still there. The problem is that RSQ may be strong in the particular sample under investigation here, but this doesn't mean that the sample is representative for the entire population. Would you get similar results if you were to test other samples of the same size? As they say, "Results may vary."

The concept of predictability covers two very different issues:

- How close can you get in your predictions, estimations, or expectations as far as a particular sample is concerned? This is a matter of correlation—RSQ is the correlation coefficient that measures the proportion of the variance in y attributable to the variance in x. Go for high correlations!
- How well can you repeat your observations regarding the same population? This is a matter of probability; the more probable the results, the more randomness has interfered. Go for low probabilities!

Figure 4.29 offers an assessment as to how repeatable the linear relationship between HbA1C readings and glucose levels in humans would be—given the fact that only 15 pairs of observations are available:

- **The outer boundaries mark the 95% prediction interval**: 95% of the y values to be found for a certain x value will be within this interval range around the linear regression line.
- **The inner boundaries mark the 95% confidence interval**: 95% of the y means to be found for a certain x value will be within this interval range around the linear regression line.

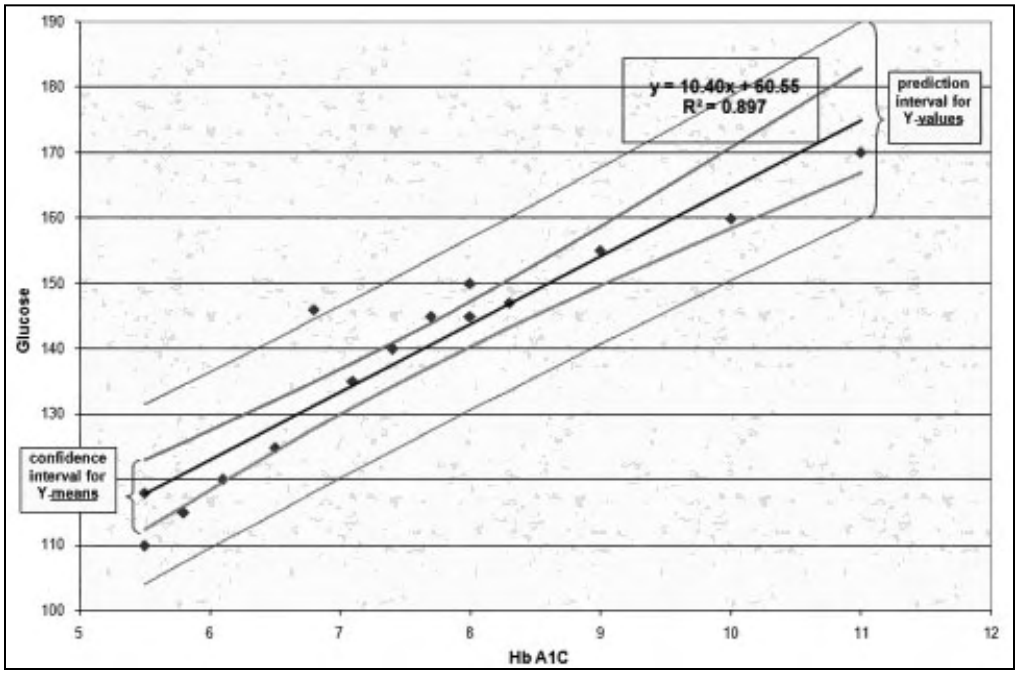


Figure: 4.29

Consequently, when you find several y values for a specific x value, their mean will be within a range narrower than the prediction range—which is the confidence interval.
If you increased the number of cases in your sample, both interval ranges would get narrower and narrower. So although the results may vary, they will vary less and less so.
How do you calculate these intervals? Many calculations have to be performed! This chapter doesn't go into all the details, but one of the tools you need is the function `LINEST`, one of the multi-cell array functions mentioned in Chapter 17. `LINEST` returns several statistical values for linear regression lines, and it does so in the following order:

- Both slope and intercept
- The standard error (SE) of both
- `RSQ` and the SE of y values
- *F* and *df* (which are explained in Part 5)
- Two sums of squares

Figure 4.30 applies `LINEST` to the linear relationship between age and FVC. Here's how you create it:

1. Select F2:G6.
2. Call `LINEST` and supply its arguments: `=LINEST(B9:B17,A9:A17,1,1)`.
3. Make sure the last argument is set to `1` or `TRUE`; otherwise, you won't get the rows 4–6.
4. Press Ctrl+Shift+Enter (not just Enter). Notice that the slope is negative (-0.027) and `RSQ` s reasonably high (0.93).
5. Sometimes, you want only one or two of these statistics; at other times, you want all of them. To get only one, apply `INDEX` to `LINEST`—in other words, nest `LINEST` inside `INDEX`. For `RSQ` (in row 3 of column 1), the formula would be `=INDEX(LINEST(B9:B17,A9:A17,1,1),3,1)`.

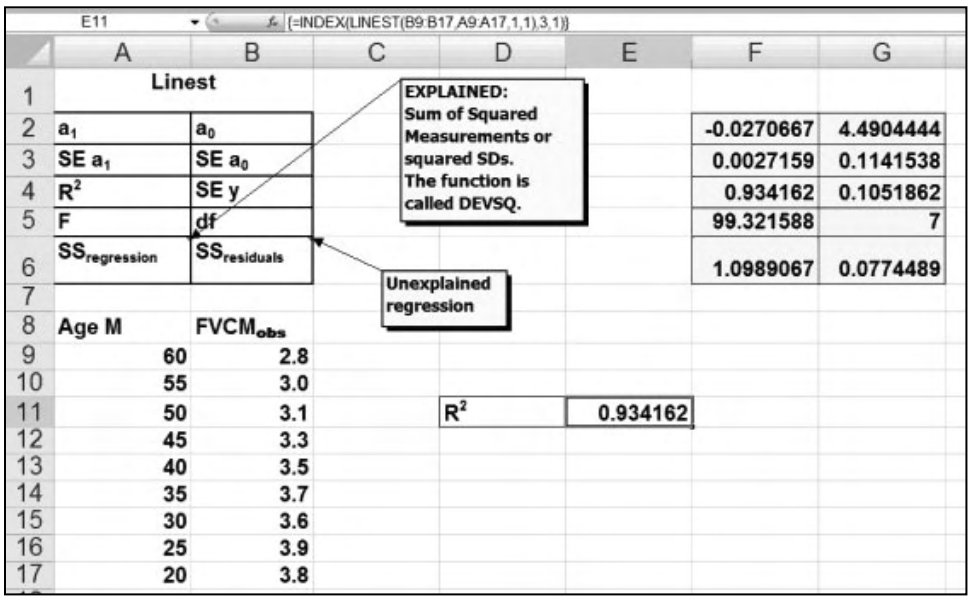


Figure: 4.30

Let's consider again the confidence intervals in Figure 4.31. The sheet has a `LINEST` calculation in the cells G20:H24, and it has the output of the Analysis Toolpak with the regression tool in the columns K:S. What is the difference between the `LINEST` results and the Analysis Toolpak results? The latter ones may be more detailed and faster, but they are "dead." Changing data later on does not affect those results anymore, unless you run the Toolpak again. But their advantage is that they display more information—including some probabilities in O15:O16. These probabilities are very low, so there is a good chance that you would get similar results if you tested more samples of the same size. Let's leave it at that for now. (You'll learn more about probabilities in Part 5.)

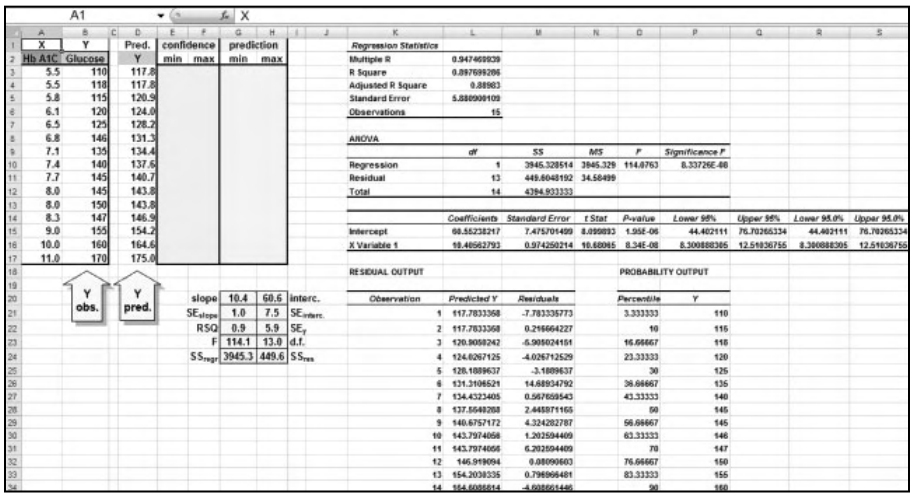


Figure: 4.31

Even all the information so far is not enough to get the final results for confidence intervals. Figure 4.32 shows a few more intermediate steps, which are not discussed here. All this is old-fashioned manual work. After you have implemented the calculations in columns E:H, you can create the graph in Figure 4.29.

The information discussed in this chapter is highly statistical in nature. We will dwell on this issue much more intensely in Part 5.

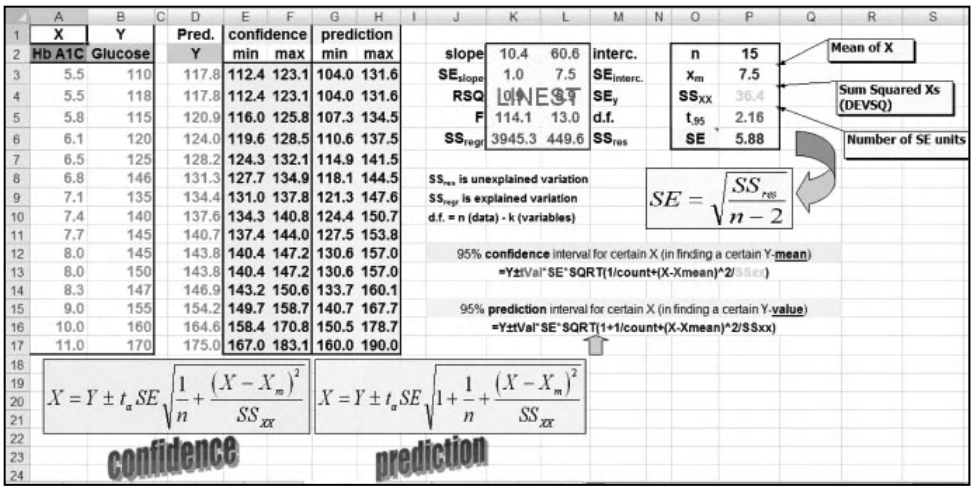


Figure: 4.32

* * *

Chapter 38

CORRELATION

The correlation coefficient (r) is the ratio of the explained variation to the unexplained variation. The better x variations explain y variations, the smaller the unexplained variation is, which makes r come closer to +1. Its squared version is `RSQ` (r^2).

Excel has two different functions you can use to determine the correlation coefficient: `CORREL` and `PEARSON`. Both of these functions use the same formula (see Figure 4.33).

$$\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Figure: 4.33

Note: I don't know why there are two formulas to do exactly the same thing. I was once told that `PEARSON` is for samples and `CORREL` for populations; that may have been the case in older versions of Excel, but when you test the two formulas in Excel 2007, the results are completely identical. So you can just take your pick.

Figure 4.34 takes us back to an earlier example: the relationship between a person's age and their FVC, in liters (see Chapter 33). You assume that variations in age determine variations in FVC. There is some unexplained variation here—for instance, three 55-year-olds have different FVC readings. Where does this unexplained variation come from? There are three theoretical answers:

- There is always inaccuracy and randomness in the observations. Let's skip this possibility for now.
- The linear model is incorrect, as discussed in Chapter 34.
- Additional factors are involved. This is the main topic in this chapter.

Let's think about the possibility that other independent variables, such as weight and height, affect variation in FVC. This idea suggests some kind of multiple regression. So far, we have only discussed single regression—with only one factor affecting the dependent factor. Because there may be many factors of potential impact, you need a tool to assess their individual effects—and that tool is the correlation coefficient. One of the reasons for using r is the fact that you can test the usefulness of independent variables in a pilot study first. You don't need to spend money on variables that have no impact! Here's how you calculate r for

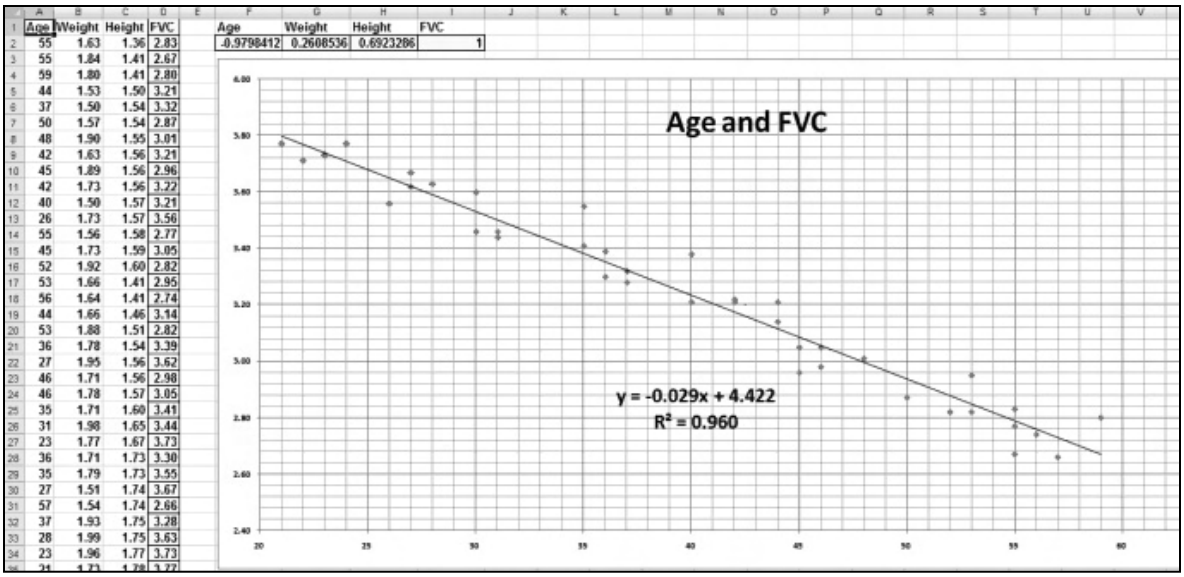


Figure: 4.34

each of three independent factors in relationship to the dependent factor, FVC:

1. Select F2:I2.
2. Call `CORREL` or `PEARSON` and provide the proper arguments. In cell F2, enter `=CORREL(A2:A41,$D$2:$D$41)`, and then use Ctr+Enter (not Ctr+Shift+Enter, because this is not an array function).
3. The order of diminishing impact appears to be age (-0.98), height (0.69), and then weight (0.26). (Later we will discuss whether weight has enough impact to even be included in a multiple regression formula.)

Figure 4.35 shows a similar case: Someone studied the impact of five different independent factors on the dependent factor systolic blood pressure. You can test the individual impact of each factor on the dependent variable by calculating correlations.

There is another reason it is prudent to assess the correlation between factors: A very high correlation between the independent variables themselves can cause trouble. This phenomenon is called *colinearity*. You test for colinearity as follows:

1. In cells I8:O8, perform a correlation test for each independent factor in relation to the dependent factor systolic blood pressure. The formula in cell I8 is: =PEARSON($A$2:$A$51,A2:A51). The order of impact turns out to be exercise, drink, smoke, weight, age, parents.
2. Test for colinearity by using the Correlation option of the Analysis Toolpak in cells I16: P23. The section K18:P23 shows us that there is no correlation beyond +0.8, so colinearity does not seem to interfere here.

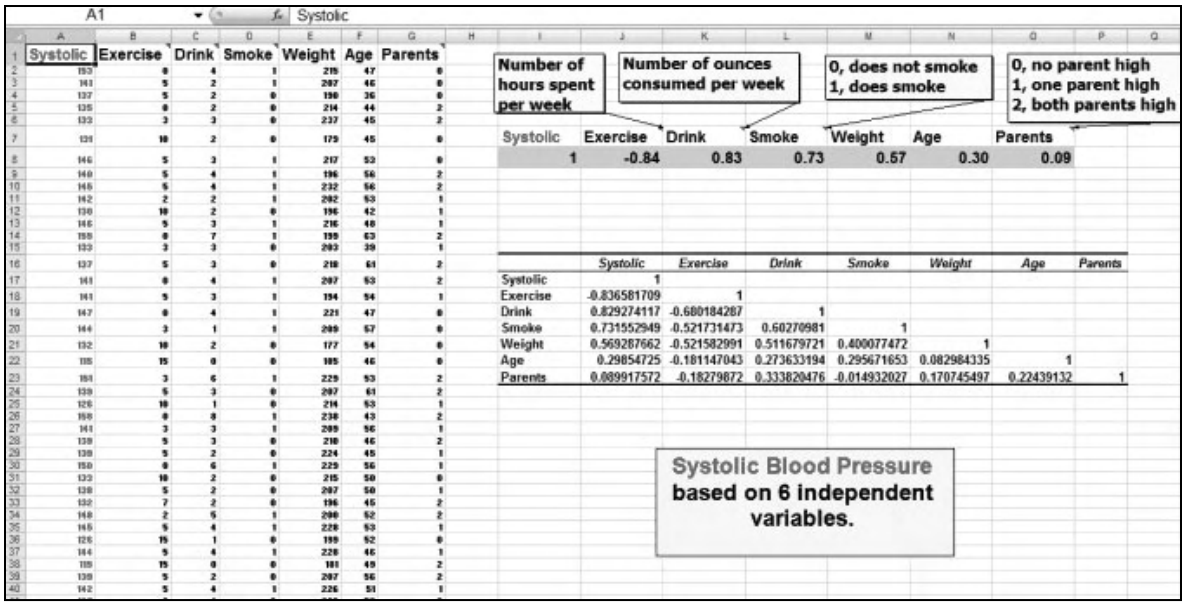


Figure: 4.35

> **Note**: An alternative way of testing for colinearity involves range names and the use of `INDIRECT`, as discussed in Chapter 3. You can also test for colinearity by using range names (as discussed in Chapter 4) in combination with the function `INDIRECT` (as was done in one of the exercises at the end of Part 1). The advantage of doing things this way is that these results would update, whereas the Analysis Toolpak delivers static results.

Figure 4.36 requires a different treatment. You need to determine whether there is a correlation between the number of drinks per week and the diastolic blood pressure. Unfortunately, the correlation coefficient can only be used when the observations are normally distributed. But here you find most observations clustered in the left-lower corner of the graph. In other words, the correlation coefficient, as found here with `CORREL` or `PEARSON` in cell B13, is not reliable. So you need a *distribution-free* test (see Chapter 53). In this case, you can apply the Spearman's rank test first, by using the function `RANK`. Here's how it works:

1. Create ranks in columns D and E by entering the formula `=RANK(A2,A$2:A$11,1)` in cell D2.
2. Because you cannot use `PEARSON` or `CORREL` until the many ties in the ranks have been

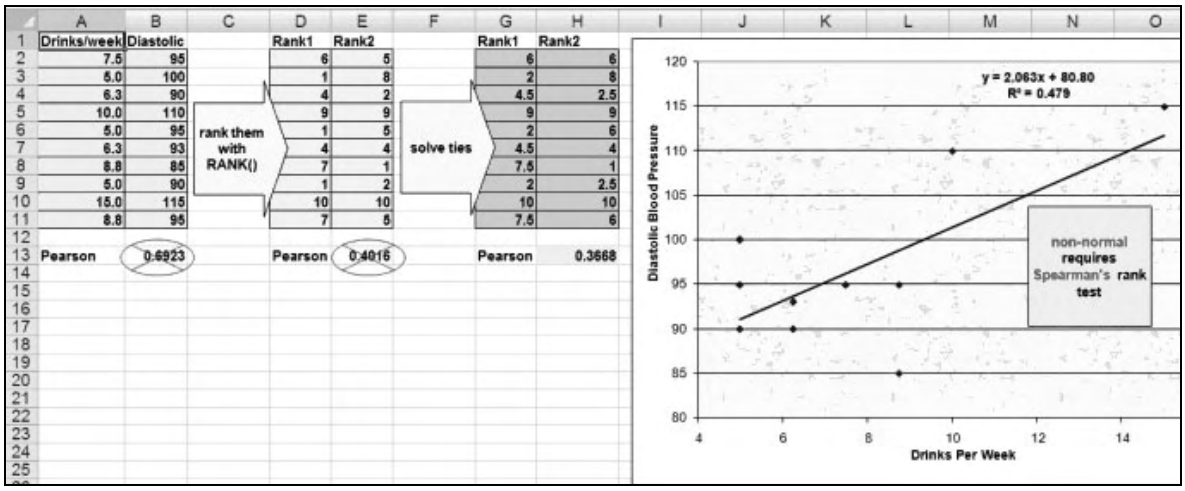


Figure: 4.36

adjusted, fix the ties. You can do this manually—1-1-1 ⇨ 2-2-2 ; 4-4 ⇨ 4.5-4.5 ; and so on—or by using a formula. You should use a formula if the columns are much longer than they are here. To use a formula to fix ties, enter `=D2+(COUNTIF(D$2:D$11,D2)-1)/2` in cell G2.

3. Apply `PEARSON` or `CORREL` in cell H13.

Notice that the spuriously high initial assessment goes down dramatically: from 0.69 to 0.37

* * *

Chapter 39

Multiple Regression: Linear Estimates

Most variables in science do not depend on one other single variable; most variables in real life depend on several variables. You should therefore be more able to predict or estimate some particular factor if you have several other factors available that impact the factor under investigation. When you decide on multiple (or multifactorial) regression analysis, you can often go for linear estimates because there are several factors involved, which makes linearity more acceptable. (This chapter does not discuss nonlinear versions.)

To perform multiple regression analysis, you use `LINEST`, which can handle multiple factors in its second argument. The formula for multiple regression has a slope for each factor (also called a *coefficient*) plus an intercept, with the following syntax: $a_n x_n + \ldots + a_2 x_2 + a_1 x_1 + a_0$. So when `LINEST` displays the slopes and standard errors, it does so in this syntactical order: $a_n x_n + \ldots + a_2 x_2 + a_1 x_1 + a_0$ (that is, the last factor first).

`LINEST` is used in Figure 4.37. The top-left section shows what `LINEST` returns when dealing with up to *n* multiple factors. This case shows only three independent factors in addition to the dependent factor FVC. In cells F13:I17, you use the formula `=LINEST(D9:D17,A9:C17,TRUE,TRUE)`. Be aware that the slopes (or coefficients) appear in reversed order, starting at factor n. The standard errors appear in the next row down.

A1 The order of regression statistics (4th argument set to True)

	A	B	C	D	E	F	G	H	I
1	The order of regression statistics (4th argument set to *True*)								
2	a_n	a_{n-1}	…	a_2	a_1	a_0			
3	SE a_n	SE a_{n-1}	…	SE a_2	SE a_1	SE a_0			
4	R^2	SE y							
5	F	df							
6	$SS_{regression}$	$SS_{residuals}$							
7									
8	**Age M**	**Height M**	**Weight M**	**$FVCM_{obs}$**					
9	**60**	**164.0**	**60.6**	**2.8**					
10	**55**	**165.0**	**52.0**	**3.0**					
11	**50**	**167.6**	**55.8**	**3.1**					
12	**45**	**168.0**	**60.9**	**3.3**			weight	height	age	intercept
13	**40**	**169.0**	**57.1**	**3.5**			-0.0090317	0.06738713	-0.0101895	-7.0772263
14	**35**	**172.0**	**59.0**	**3.7**			0.010849	0.05417439	0.01379816	9.61428952
15	**30**	**171.0**	**55.5**	**3.6**			0.95240303	0.10582151	#N/A	#N/A
16	**25**	**173.0**	**52.0**	**3.9**			33.3495709	5	#N/A	#N/A
17	**20**	**174.0**	**60.6**	**3.8**			1.1203646	0.05599096	#N/A	#N/A

Figure: 4.37

Figure 4.38 applies multiple regression to a case you studied in Chapter 38. You have six independent factors here in addition to the dependent factor systolic blood pressure. Some independent factors may not score high for correlation with the dependent factor, but you still use all of them in the multiple regression model. (You'll learn more on this issue later.) Here's how you create Figure 4.38:

1. Enter the observations in columns A:G.
2. Use cells K2:Q2 to calculate the correlations between the independent factors and the dependent factor. In K2, enter `=PEARSON($A$2:$A$51,A2:A51)`.
3. Use cells K9:Q13 to return `LINEST` results: `=LINEST(A2:A51,B2:G51,1,1)`. Smoking seems to have the highest coefficient. Coefficients that are very close to zero are basically useless. (You'll learn more on this issue later.)
4. The section K17:R24 shows that there is no colinearity interfering, so predict or estimate systolic blood pressure in column I, using either of these methods:
   - Use the coefficients from `LINEST`: `=$K$9*G2+$L$9*F2+$M$9*E2+$N$9*D2+$O$9*C2+$P$9*B2+$Q$9` (in I2).
   - Use the `TREND` function: `=TREND(A2:A51,B2:G51)`.

This example shows all factors included, which is not a good practice. It is often difficult to decide which factors to use. Here are some general rules:

- Eliminate the factors with a low correlation.
- Eliminate the factors with near-to-zero coefficients.
- Eliminate one of the two factors that show colinearity.

Next, let's look at one more rule.

Systolic	Exercise	Drink	Smoke	Weight	Age	Parents	SBP
153	0	4	1	215	47	0	150.0
141	5	2	1	207	46	0	140.2
137	5	2	0	190	36	0	134.4
135	0	2	0	214	44	2	138.0
133	3	3	0	237	45	2	138.5
131	10	2	0	179	45	0	130.1
146	5	3	1	217	53	0	143.7
140	5	4	1	196	56	2	142.0
145	5	4	1	232	56	2	143.7
142	2	2	1	202	53	1	142.0
130	10	2	0	196	42	1	128.9
146	5	3	1	216	48	1	141.4
155	0	7	1	199	63	2	154.5
133	3	3	0	203	39	1	137.9
137	5	3	0	218	61	2	137.6
141	0	4	1	207	53	2	147.0
141	5	3	1	194	54	1	141.1
147	0	4	1	221	47	0	150.3
144	3	1	1	209	57	0	141.3
132	10	2	0	177	54	0	131.0
115	15	0	0	185	46	0	121.1
151	3	6	1	229	53	2	149.6
139	5	3	0	207	61	2	137.0
126	10	1	0	214	53	1	128.8
158	0	8	1	238	43	2	156.2
141	3	3	1	209	56	1	144.0

Systolic	Exercise	Drink	Smoke	Weight	Age	Parents
1	-0.84	0.83	0.73	0.57	0.30	0.09

Parents	Age	Weight	Smoke	Drink	Exercise	
-1.66	0.12	0.05	3.83	2.24	-0.97	121.52
0.62	0.08	0.04	1.21	0.44	0.16	9.06
0.90	3.11	#N/A	#N/A	#N/A	#N/A	#N/A
62.70	43.00	#N/A	#N/A	#N/A	#N/A	#N/A
3629.61	414.87	#N/A	#N/A	#N/A	#N/A	#N/A

	Systolic	Exercise	Drink	Smoke	Weight	Age	Par
Systolic	1						
Exercise	-0.836581709	1					
Drink	0.829274117	-0.680184287	1				
Smoke	0.731552949	-0.521731473	0.60270981	1			
Weight	0.569287662	-0.521582991	0.511679721	0.400877472	1		
Age	0.29854725	-0.181147043	0.273633194	0.295871653	0.082984335	1	
Parents	0.089917572	-0.16279872	0.333820476	-0.014932027	0.170745497	0.22439132	

Figure: 4.38

Figure 4.39 shows that when you eliminate factors step by step (the poorest-correlated one first), your prediction/estimation power (RSQ) goes down—or, reversed, it does go up when you add more factors. But don't get fooled by RSQ: It will always go up or down when you add or eliminate additional factors. Every extra factor helps, of course. That's why you need an adjusted RSQ, as shown in the formula on the sheet in Figure 4.39. The adjusted RSQ takes into account how many values are in the model already; it considers the number of cases (n) plus the number of variables (df+1). Adding another variable may counteract the effect of its added values. Notice in column I that adding variables does not always increase RSQ_{adj}. Its value actually goes down after you add the factors weight and age.

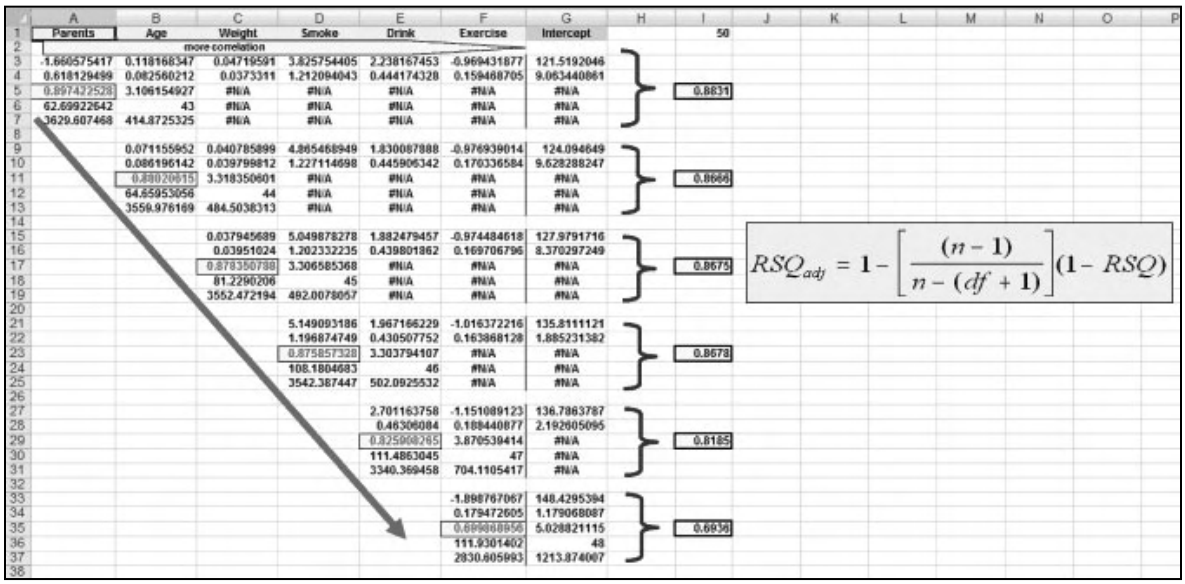


Figure: 4.39

Once again, you could do much of this work by using the Analysis Toolpak, as shown in Figure 4.40. Its regression tool gives an overview with some extra information:

- It automatically calculates the adjusted RSQ.
- It also shows the probability for each coefficient. Remember that a high probability indicates a great deal of randomness. A low probability means that testing another sample of the same size would not greatly sway the results. Notice that the two factors with a relatively high probability (weight and age) are the very same ones that didn't improve the value of RSQ_{adj}.

You now have two ways of testing multiple regression: Either add the factors with the lowest probabilities one by one or eliminate the factors with the highest probabilities first.

To summarize, there are some good tests for multiple regression. You should use them together and combine them with common sense. Here are the rules in a nutshell:

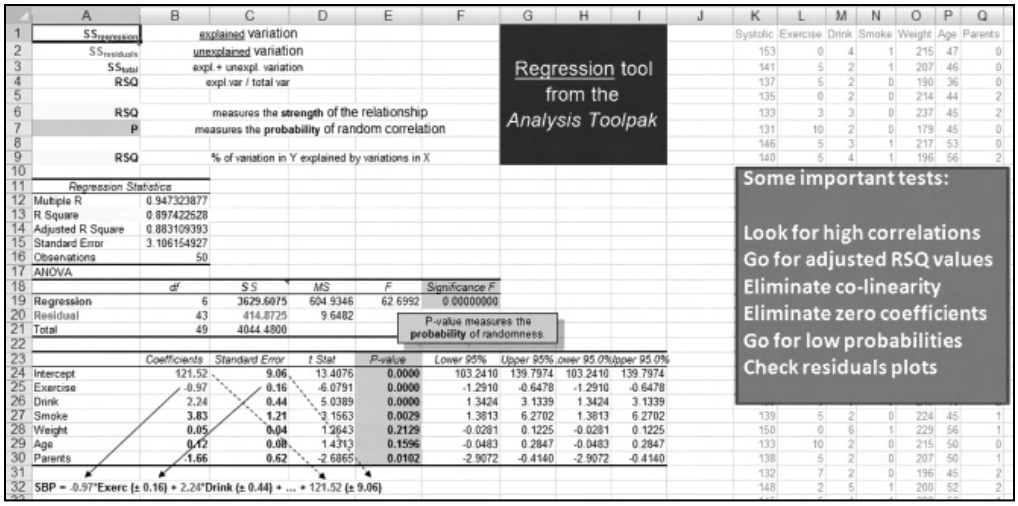


Figure: 4.40

- Look for high correlations.
- Go for adjusted RSQ values.
- Avoid colinearity.
- Eliminate zero coefficients.
- Go for low probabilities.
- Check the residual plots for random scatter.

There is one more issue that needs attention: the interaction between factors. For example, in Figure 4.41, say that bone growth depends on Ca and P intake, yet `RSQ` is rather poor, so the prediction is poor. Why? There is interaction between these independent factors: Either more Ca inhibits P or the reverse is true. What is the solution? You add another "factor" and base it on the interaction of the two original factors: Ca * P. Here's how:

1. In cell I2, enter `=G2*H2`.
2. In F11:I15, enter `=LINEST(F2:F9,G2:I9,1,1)`. Notice that `RSQ` goes up dramatically—from a low of 0.32 to a high of 0.99.
3. Make your prediction in column J in one of the following ways:
   - Use `TREND`: `=TREND(F2:F9,G2:I9)`.
   - Use coefficients: `=$F$11*I2+$G$11*H2+$H$11*G2+$I$11`.

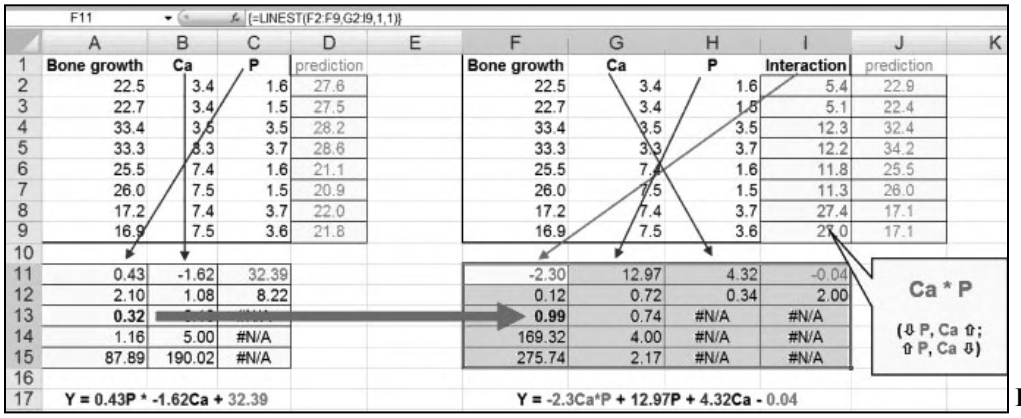


Figure: 4.41

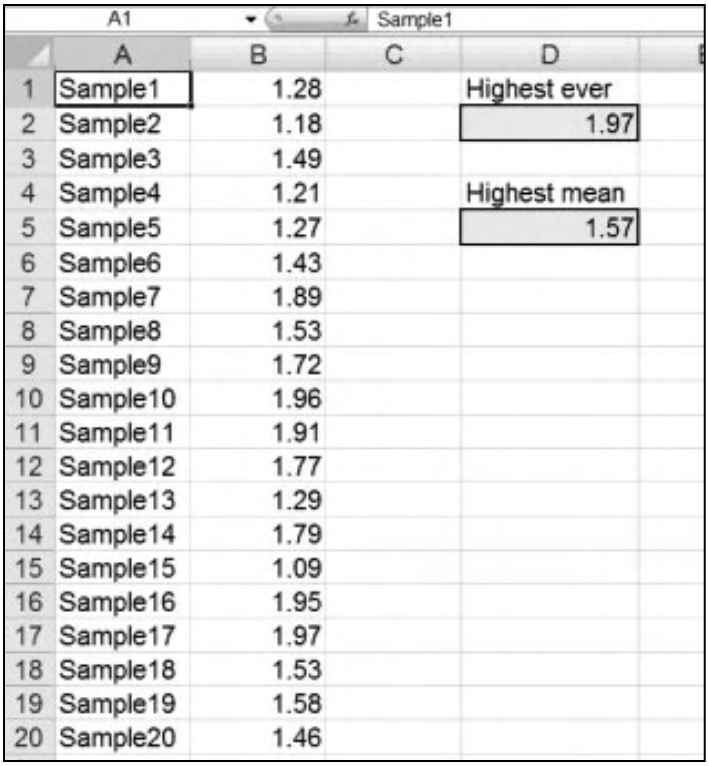
* * *

Chapter 40

Reiterations and Matrixes

One of the many tasks scientists have to perform in their work is solving complex equations. One possible way of solving equations is to have a computer go through a series of iterations (trials and errors) until it finds the correct solution. When the cell containing the equation is itself part of the equation, we speak of *circular reference*.

The two marked cells in Figure 4.42 (cells D2 and D5) have a circular reference problem. They display the highest value from column B or their own value—whichever is the highest. So they have some kind of memory stored inside them. A formula that refers to its own results creates a circular reference; Excel usually rejects circular references because very often those formulas were created by mistake. But there are times when you want to create them intentionally. Here's how you do it:



	A	B	C	D
1	Sample1	1.28		Highest ever
2	Sample2	1.18		1.97
3	Sample3	1.49		
4	Sample4	1.21		Highest mean
5	Sample5	1.27		1.57
6	Sample6	1.43		
7	Sample7	1.89		
8	Sample8	1.53		
9	Sample9	1.72		
10	Sample10	1.96		
11	Sample11	1.91		
12	Sample12	1.77		
13	Sample13	1.29		
14	Sample14	1.79		
15	Sample15	1.09		
16	Sample16	1.95		
17	Sample17	1.97		
18	Sample18	1.53		
19	Sample19	1.58		
20	Sample20	1.46		

1. Enter the following formula in cell D2: `=MAX(B:B,D2)`. (Notice the reference to D2 in D2's formula.)
2. Make the formula final. Excel gives you a circular reference warning. Click Cancel.
3. Excel does not give you any results until you go to Excel Options and enter the Formulas section, where you can enable Iterative Calculation. By default, Excel performs a maximum of 1,000 iterations.
4. Click OK. The calculation in cell D2 works and always shows the highest value.
5. Enter the following formula for the highest mean in cell D5: `=MAX(AVERAGE(B:B),D5)`. When you close this file and open it again, you won't be alerted until you turn off iterations again.

Figure: 4.42

Figure 4.43 presents a similar situation. Column A holds 100 random numbers. In column D and E, you would like to create 10 equal bins between the minimum and maximum values. Column F shows the frequencies at this point. Here's what you do:

D1 =MIN(A1:A100)

	A	B	C	D	E	F
1	0.035		min	0.0020	0.0020	1
2	0.719			0.1125	0.1123	15
3	0.854			0.2226	0.2224	11
4	0.010			0.3324	0.3322	5
5	0.033			0.4422	0.4421	10
6	0.057			0.5522	0.5521	10
7	0.542			0.6625	0.6624	9
8	0.337			0.7723	0.7723	13
9	0.703			0.8814	0.8814	18
10	0.357		max	0.9900	0.9900	8
11	0.775					
12	0.546					
13	0.365					
14	0.834					
15	0.218					
16	0.715					
17	0.150					
18	0.975					

1. Enter the following formula in cell D2: `=AVERAGE(D1:D3)`. This formula contains circular reference.
2. Enter the following formula in cell E2: `=AVERAGE(E1,E3)`. This formula does not contain a circular reference.

   Results may differ between column D and E, but the formula with circular reference is more accurate.

Figure: 4.43

Iterations are a great tool for simulations, as Figure 4.44 illustrates. Say that you want to simulate what happens in between a few points (marked at the four corners of A4:E8) that you know following a gradient. This is the kind of situation you may encounter when dealing with gradients in pressure, temperature, concentration, gene flow, and so on. Create a Surface type of graph to better visualize the situation and follow these steps:

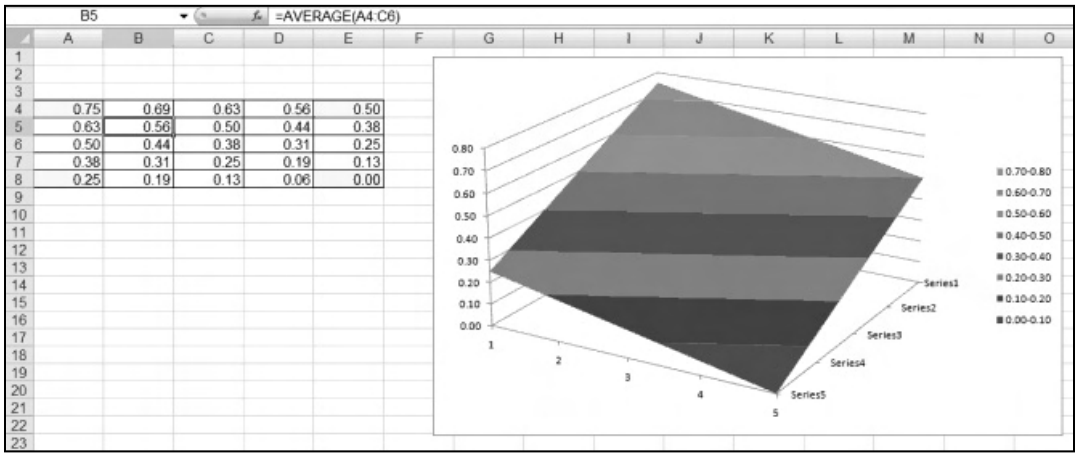


Figure: 4.44

1. Mark the known points (at the four corners).
2. Determine the four boundary areas: B4:D4, B8:D8, A5:A7, and E5:E7.
3. Enter the following formula in cell B4: `=AVERAGE(A4:C4)`. (There is a circular reference again.)
4. Fill the areas in between. For B5, you enter `=AVERAGE(A4:C6)`.
5. Make sure 1,000 iterations are enough for Excel to reach a stable matrix, you press F9 until stability is reached. The finer the grid or matrix system, the longer it takes to get final results.

You can sometimes use matrixes to solve equations. For example, Figure 4.45 has three equations with three unknown x values:

- Each equation uses three different coefficients for A, as shown in matrix [A].
- The three equations should equate to the y values, as shown in matrix [Y].
- You need to determine what the x values are supposed to be.

Here's what you do:

1. Invert matrix [A] by using the multiple-cell array function `MINVERSE` in cells C14:E16: `=MINVERSE(C6:E8)`.
2. Multiply the matrix `Inv[A]` with the matrix `[Y]` by using the array function `MMULT` in cells C18:C20: `=MMULT(C14:E16,C10:C12)`. Thanks to this multiplication, the cells C18:C20 contain the three x values that we were looking for to solve 3 equations with 3 unknown x values. The 3 x values we have found make the three equations, based on the A values specified in the first matrix equate to the y values specified in the second matrix.

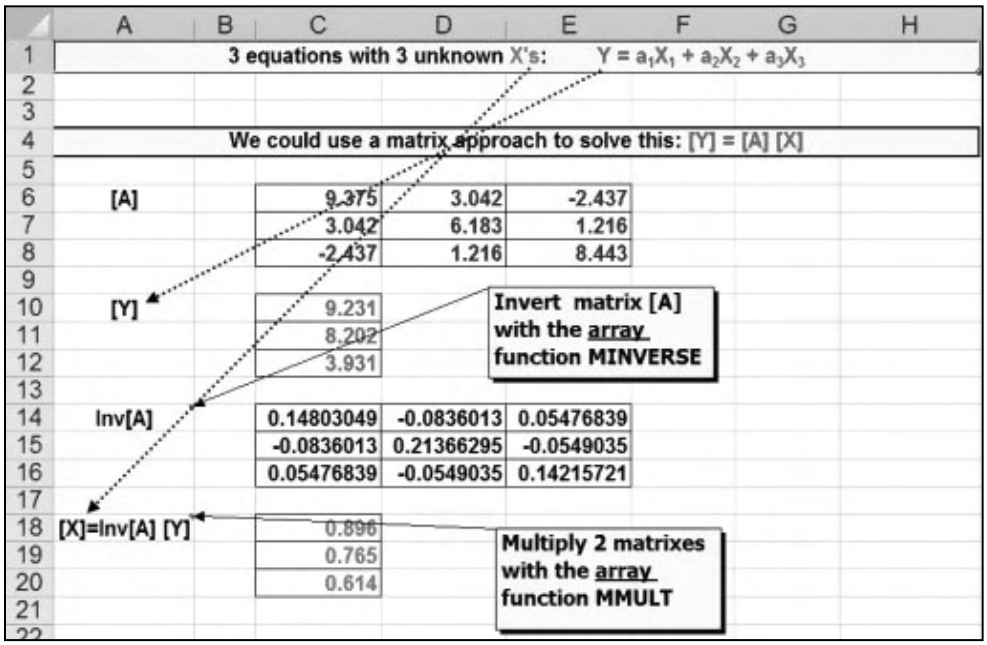


Figure: 4.45

* * *

Chapter 41

Solving Equations

In Chapter 40, you solved some mathematical problems manually. Fortunately, Excel has also two dedicated solving tools to offer: Goal Seek and Solver. Goal Seek is for simple situations, and Solver is for more complicated mathematical problems. Both use iterative processes.

You can use Figure 4.46 to explore both of Excel's solving tools. Notice that the dark gray cells on the sheet contain formulas; the formulas are displayed in the lighter gray cells above them.

Let's start with GoalSeek, the simpler of the two solving tools. To open it, you go to the Data tab and select What-If Analysis, then Goal Seek. Say that you want to find the root of the quadratic equation, as shown in row 8—in other words, you want to solve for the x at which the equation equates to 0. GoalSeek can find the answer, but you must observe a few rules:

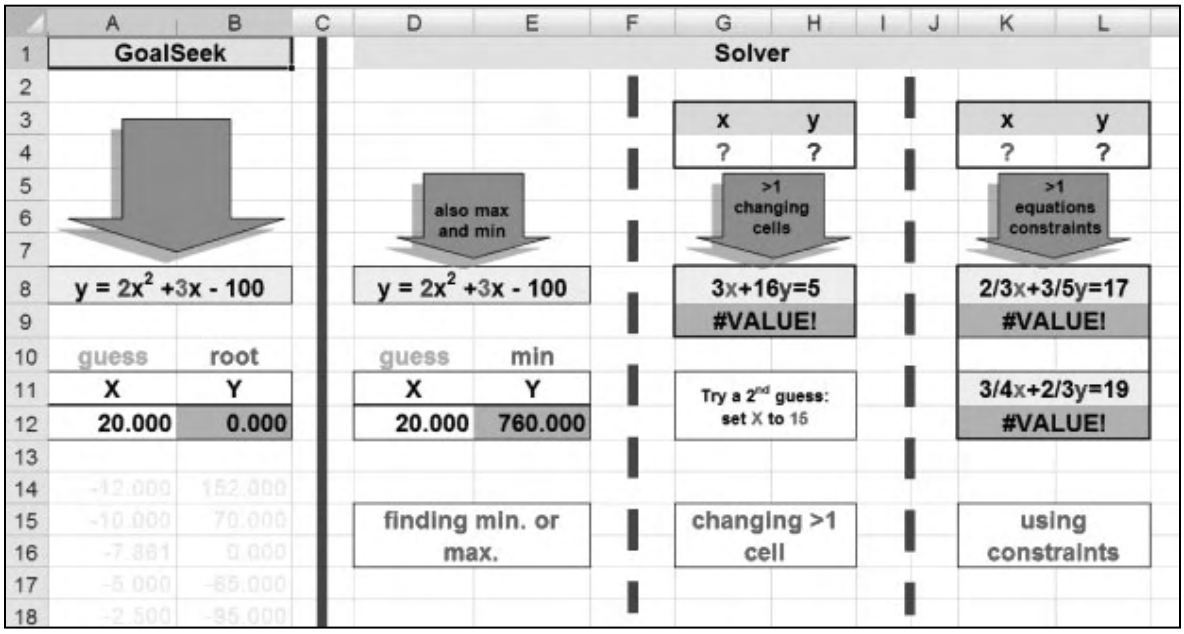


Figure: 4.46

- The set cell must contain a formula.
- The changing cell must contain a value that has been used in the previous formula.

When you apply GoalSeek to the problem just described, it uses the initial value for x (20) as a starting point and goes through a series of iterations to come up with a solution: x is 6.4.

When you look at the graph behind the equation in Figure 4.47, you can see why GoalSeek comes up with 6.4 (and not -7.8). It searches in the direction that comes closer to a solution. A good starter for -7.8 would be anything down from -1 (so definitely not 0). Very often, GoalSeek finds the solution you are looking for, but if you have a wrong starter or ask for the impossible (for example, when is y -120?), it does not give the correct solution. And then there are certain questions GoalSeek could never solve, such as: What is the lowest y value? That's where GoalSeek has reached it limits, and Solver comes to the rescue because Solver can also find maximum and minimum values.

Solver is not installed by default. If you haven't already done so, you need to activate this add-by clicking the Office icon and selecting Excel Options, Add-Ins. Once Solver has been

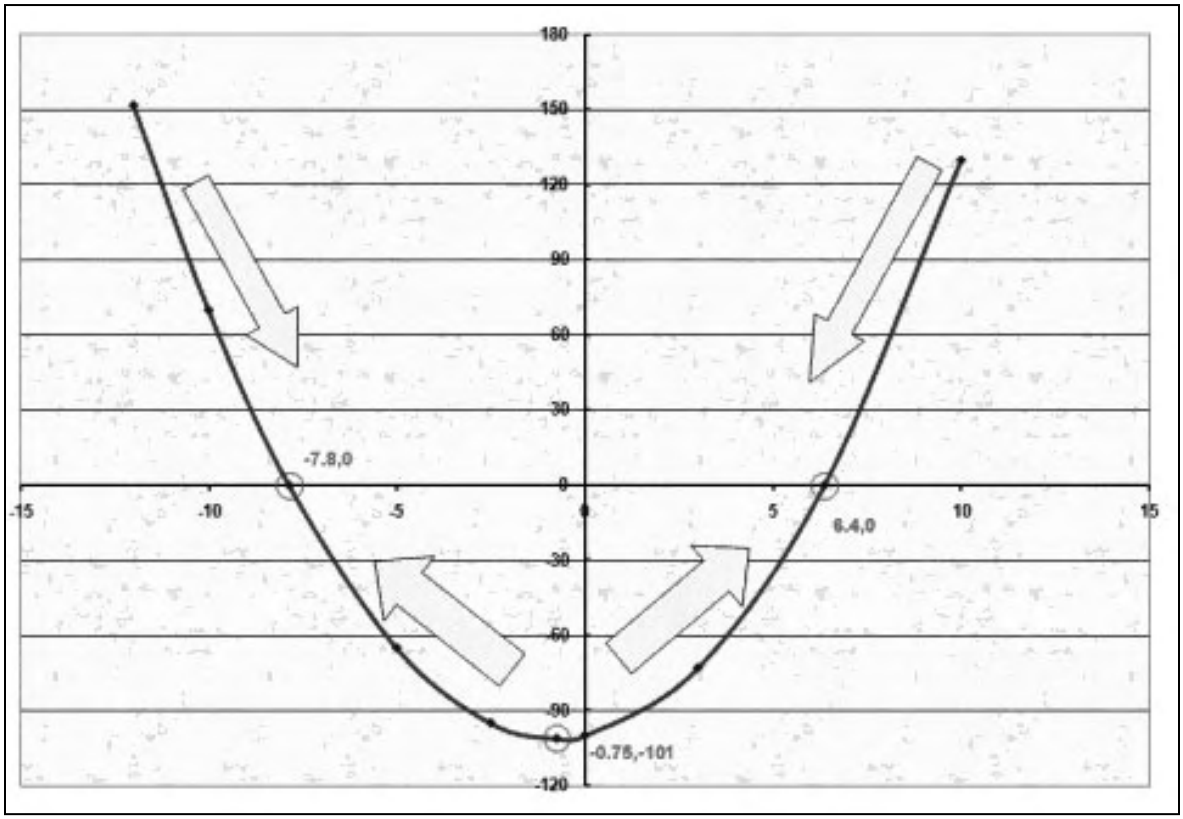


Figure: 4.47

activated, it is available from the Data tab, in a new Analysis group. When using Solver, you need to follow these rules:

- The target cell is always a formula—but only a single formula. (This is the same as in GoalSeek.)
- Besides value, there are two more options: Min and Max. (This is not like GoalSeek.)
- Unlike in GoalSeek, Solver's changing cells can consist of more than 1 cell. (The Guess button finds all the value cells the formula is based on, which is sometimes what you want.)
- Unlike GoalSeek, Solver accepts constraints.
- When Solver finds a solution, you can either keep or reject the solution.

Figure 4.46 shows how you can apply Solver to some common kinds of mathematical problems. For example, in columns D:E, you can use Solver to find the minimum value of the equation $y = 2x^2 + 3x - 100$:

1. Set the target cell E12 to Minimum.
2. Set D12 as the changing cell or click the Guess button. Once you click OK, Solver finds the following solution: x=-0.75 and y=-101.125.
3. To see if there are more solutions, try other starting values.

In columns G:H, you can set the equation 3x+16y to a specific value (5) by changing more than one value cell (both x in G4 and y in H4). You cannot do this with GoalSeek, but here's how it works with Solver:

1. Set the target cell G9 to `5`.
2. Set G4:H4 as the changing cells or click the Guess button. Solver finds x=0.0566 and y=0.302 as one of the many solutions.

You can also use Solver to solve several equations at the same time, as is done in columns K:L. Because Solver can only handle one equation in its target cell, you must treat the other equation(s) as constraints. Here's what you do:

1. Set the target cell K9 to `17`.
2. Set K4:L4 as the changing cells or click the Guess button.
3. Add a constraint to set the second equation to `19`. Solver finds this solution based on the starting settings: x=12 and y=15.

Figure 4.48 looks like another candidate for Solver. It shows a sigmoid-like equation (as discussed in Chapter 36) based on an estimated slope (in cell B2) and IP (in cell B1). You may be able to improve the estimated slope and IP by applying Solver to either the sum of residuals (in cell C23) or the sum of the least squares (in cell C24). Here's what you do:

1. In cell C23, enter the following array formula for the sum of residuals: `=SUM((B5:B22)-(C5:C22))`.
2. In cell C24, enter the following array formula for the sum of the least squares: `=SUM(((B5:B22)-(C5:C22))^2)`. You could instead use Excel's function `SUMXMY2`.
3. Try Solver in three different ways:
   - **Set C23 to 0**: Solver finds 7.4421 for IP and 0.5252 for slope.
   - **Set C24 to Minimum**: Solver finds 7.5388 and 0.5032.
   - **Set C24 to Minimum and add a constraint of 0 for C23**: Solver finds 7.4336 and 0.5038.

If you click the Options button in Solver's dialog box, you find several intricate options. One of them is the precision with which you want Solver to work. You can test these options on your own.

	A	B	C
1	IP	7.5000	
2	Slope	0.5000	
3			
4	X	Y	Predict
5	1	10	7.47
6	2	15	12.02
7	3	22	19.07
8	4	32	29.61
9	5	45	44.54
10	6	63	64.16
11	7	80	87.56
12	8	111	112.44
13	9	133	135.84
14	10	153	155.46
15	11	177	170.39
16	12	185	180.93
17	13	189	187.98
18	14	194	192.53
19	15	197	195.40
20	16	198	197.19
21	17	199	198.28
22	18	200	198.96
23	Sum Residuals		13.17
24	Least Squares		172.80

	0	
0	IP	7.4421
232.91	Slope	0.5252
	Min	
20.41	IP	7.5388
167.34	Slope	0.5032
	Both	
0	IP	7.4336
204.53	Slope	0.5038

Figure: 4.48

* * *

Chapter 42

What-If Controls

Besides using manual methods, GoalSeek, and Solver, you can solve equations by using controls that allow you to regulate values used in formulas. You can use controls to apply what-if analysis. Although they're a form of trial-and-error, controls can be great for simulations because they beautifully imitate the impact of certain changes. However, controls may not always be the ideal tools for finding the very best settings; Solver is better at that

Let's consider Figure 4.49 as a starting point. To create a control like the one used there for interpolation, you follow these steps:

1. In cell H2, enter the linear regression formula shown in the graph: =1.286*G2-43.11. When the value in G2 changes, the values in H2 will change. Now we will use a control to regulate the value in G2 – and thus indirectly in H2.
2. To create the control itself, activate the Design mode on the Developer tab. If the Developer tab is not available, click the Office icon and select Excel Options, Popular and select Show Developer Tab. From now on, you can insert controls on your sheets—for instance, a scrollbar—by using the second section (called Active X Controls) of the drop-down button.

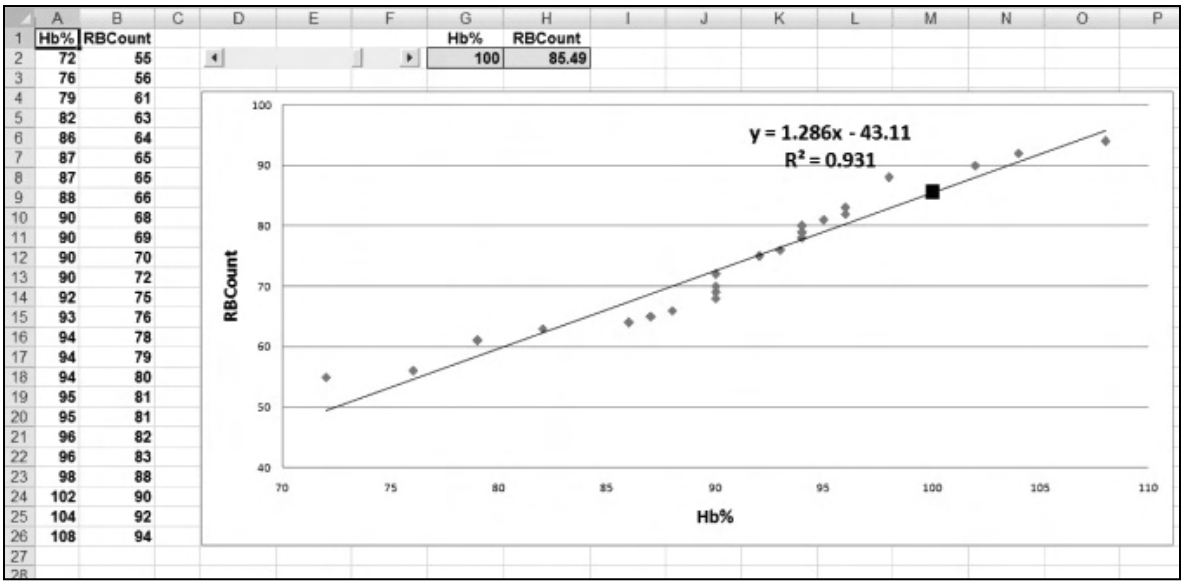


Figure: 4.49

3. After the control is drawn, right-click the control, select Properties, and then set at least these properties:
   - **Min**: 72.
   - **Max**: 102.
   - **Linked Cell**: G2. You must type this address; do not select the cell; ignore error messages.
   - **SmallChange and LargeChange**: Set these properties to integer values.
4. Turn off Design mode; otherwise, you continue to work on the scrollbar object, which is just floating on the sheet.
5. When you are back on the sheet, make the value in cell H2 go up and down and watch the line marker in the graph walk along the linear regression line.

Because controls can use only integers, the minimum change is always 1. This could be a problem in a case like the one shown in Figure 4.50, where cell F17 requires decimals. Situations like these force you to store a control's integer in an intermediate cell and then use a division formula in the real target cell. Here's how you do that:

1. Create a scrollbar.
2. Link the scrollbar to cell F18.
3. Set **Min** to **20** and **Max** to **180**.
4. Stop Design mode.
5. Enter the following formula in cell F17: =F18/1000.
6. Hide the intermediate cell by either using a white font or placing it behind the control.

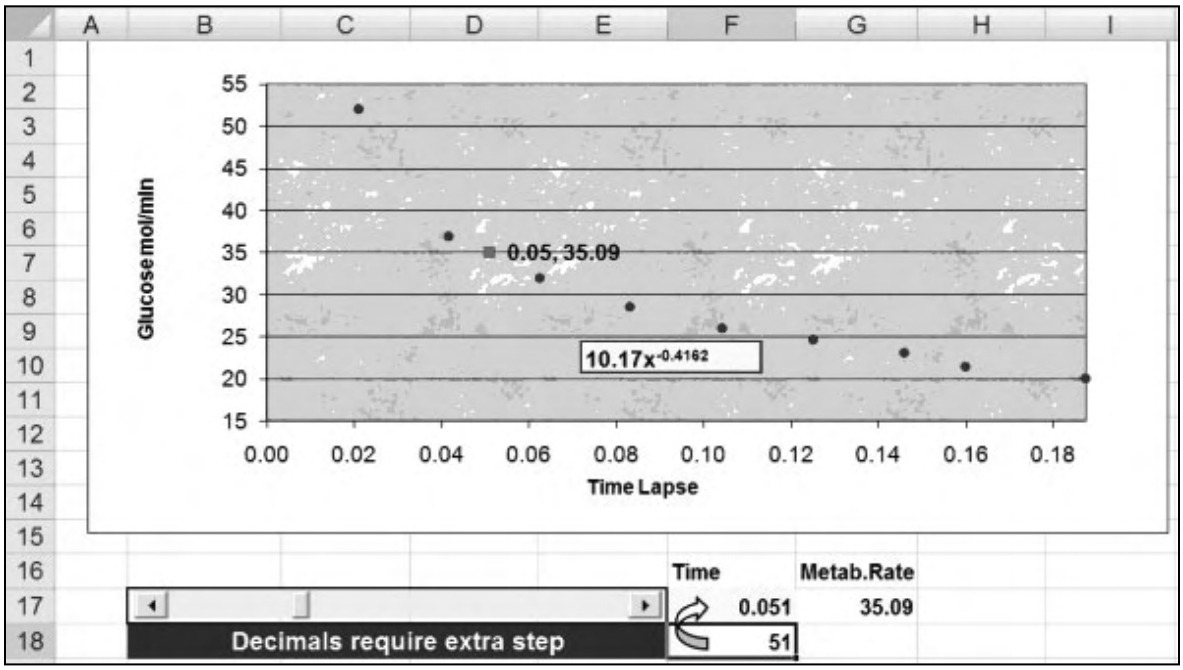


Figure: 4.50

* * *

Chapter 43

Syntax of Functions

Although Excel provides numerous built-in functions, you may sometimes need to construct your own customized functions for the particular time-consuming calculations you perform regularly. Excel allows you to create customized functions, which opens a world of unlimited power.

All the functions you have encountered so far return something—usually a number—but other types of return values are possible as well; but even numbers can be of different types. Figure 4.51 lists the most important Data Types returned by common Excel functions. For numbers, you have a choice between long and integer (both have no decimals, but they differ in maximum value) and between single and double (both accept decimals, but they differ in their precision). Variant is the most comprehensive type; it can hold any of the other types, but it requires with more computer memory and slows the processing speed.

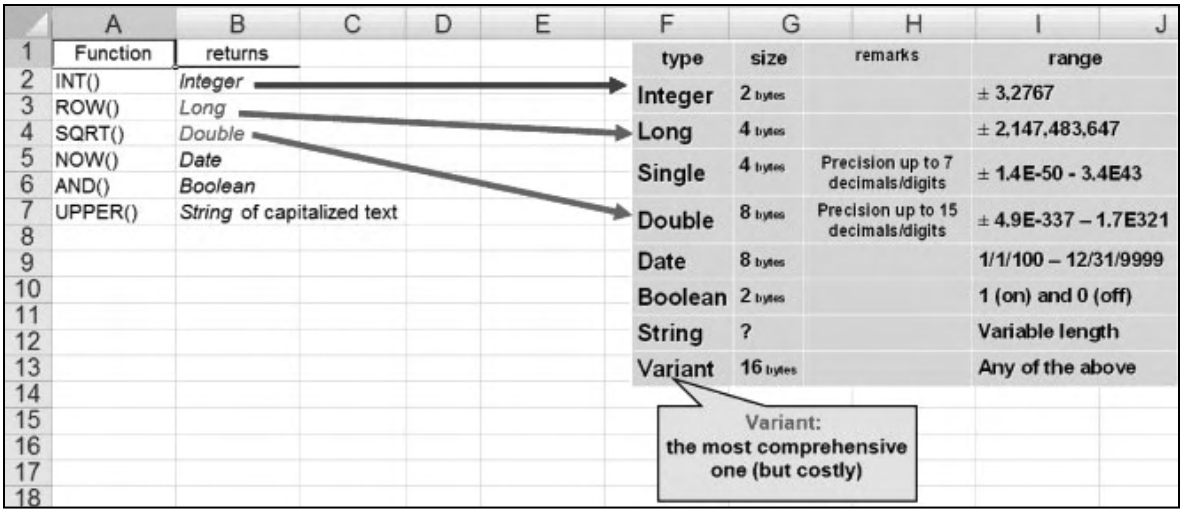


Figure: 4.51

Most functions have *arguments*—that is, information that the functions work on. Here are some examples:

- `SQRT` has one argument—the number whose square root you want.
- `COUNTIF` has two arguments—the range in which to count and the criteria that define when to count.

- IF has three arguments—the criteria plus what to do when TRUE and what to do when FALSE.

Some arguments are optional; in the dialog box, they show up as non-bold:

- TREND and LINEST, for instance, have two optional arguments at the end; when you don't specify them, a default setting kicks in.
- IF has two optional arguments, but if you leave them blank, the function returns either TRUE or FALSE.
- SUMIF has the syntax SUMIF(*range,criteria,sum_range*), with the last argument being optional; however, the last argument is not optional when *range* and *criteria* are non-numeric.

Excel provides numerous functions. However, in some situations, Excel does not offer a function that meets your needs. For example, Excel has a function to calculate the square root of a number (SQRT), but it has no function to calculate the cube root. In this situation, you can create a function of your own.

You create functions in Visual Basic or VBA. VBA stands for Visual Basic for Applications; in this case, the application is Excel. This chapter does not delve into VBA programming; it shows just the tip of the iceberg, and it describes only functions in VBA.

> **Note**: To delve more deeply into VBA, see the *Excel 2007 VBA* CD-ROM from MrExcel's *Visual Learning series* at www.mrexcel.com.

1. Click the Visual Basic button on the Developer tab. If the Developer tab is not available, click the Office icon and select Excel Options, Popular and select Show Developer Tab.
2. When you are in Visual Basic, select Insert, Module.
3. On the new Module sheet that appears, enter the following:

> **Note**: On a Module sheet, you can type just as you would on a piece of paper, but you cannot just type whatever you want. You need to enter text in the correct sequence.

   - Start a function by typing the word Function. (Be sure to spell it correctly.)
   - Assign the name CubeRoot to the function. (Note that a function name cannot have a space in it.)
   - Type (.
   - Type a creative name, such as the word number, for an argument.
   - Specify the type of the argument: As Double. (A list pops up when you type the d in Double).

- Type).
- Specify the type that the function returns: `As Double`.
- Altogether, you type `Function CubeRoot(number As Double) As Double`. Press Enter, and VBA encapsulates the function.

4. Inside the function, type what the function should achieve:
   `CubeRoot = number ^ (1 / 3)`.

This is probably not the function you have been dreaming of for years, but it is a good example of what is possible. Now you need to test it in Excel.

1. The new function should now feature in the regular listing of functions under fx , where you can find it in both the categories All and User Defined.
2. Type on a new sheet in cell A1 the number 4. Let us find in cell B1 the cube root of this number by using our new function: =CubeRoot(A1). And the result should be 1.587 (or more decimals).

The new function seems to work fine, but it doesn't provide user-friendly information until you do the following:

1. In Excel, click the Macros button on the Developer tab.
2. Because the function is not a macro, it's not listed under Macros, so type its name in the top box.
3. Click the Options button.
4. Type a description or an explanation and click OK.
5. Close the left box (but do not click Run because this is not a macro). Now the function box should look a bit more professional. Adding a real Help feature, though, is beyond the scope of this chapter.

> **Note**: If you ever run into trouble because you make a mistake in your VBA code, you can click the Reset button on the VBA toolbar and then make the needed corrections.

The new function you have created here is rather limited because it only calculates cube roots. Here's how you create a function for any kind of root:

1. Start a new function in VBA—this time, with two arguments:
   `Function AnyRoot(number As Double, root As Integer)` As Double. Press Enter, and VBA encapsulates the function
2. Inside the function, type `AnyRoot = number ^ (1 / root)`.
3. Test the new function in Excel.

Now you're ready to create a function that is a more realistic and useful function—one that calculates the standard error, as Excel does not provide a function for this. Most scientists calculate the standard error manually: = SD / √n. In VBA, it goes like this:

1. Go back to Visual Basic if it is still open, or open it from the Developer tab.
2. Type after all previous functions:
   `Function StError(SD As Double, size As Double) As Double.`
3. Inside the function, type `StError = SD / AnyRoot(size,2)`.

> **Note**: Notice that here you use the function AnyRoot that you just created. If you want to use `Sqr(size)` instead, be aware that the VBA version is `SQR`, whereas Excel uses `SQRT`.

4. If you use your `AnyRoot` function, specify the second argument because it is not optional until you do make it optional. To do so, add to the second argument the word `Optional` and set it to `2` (thus making it the default value). This is what the first line looks like now:
   `Function AnyRoot(number As Double, Optional root As Integer = 2) As Double.`

> **Note**: Optional arguments should always be the last ones in the list of arguments. Consequently, using `AnyRoot` anywhere in Excel does not always require a second argument now. If you don't specify the second argument, the function assumes that you want the square root. In `StError`, for instance, you could use this line: `StError = SD / AnyRoot(size)`. But the original line is still acceptable as well: `StError = SD / AnyRoot(size,2)`.

Figure 4.52 shows how you can apply the custom-made function `StError` in a regular spreadsheet. Cell B2 holds the following formula: `=StError($A2,B$1)`. Part 4 discusses the statistical implications of this. For now, just know that increasing sample sizes reduces the standard error for a given standard deviation.

There is one more issue we need to address. When you close an Excel file, you save it as a Macro-enabled file with the .xlsm extension. When you open it again, you are warned that there are macros in this file. (There are not really macros, but there is VBA code.) If you don't enable the VBA code by clicking the Options button just below the menus or ribbons and enabling the content, you won't be able to recalculate or apply the new functions. The existing calculations remain, but you cannot reapply them.

When you close the file in which you made custom functions and open another file, you won't find your custom functions because they are stored in the original file only. You can solve this problem in two different ways. The first way is using a VBAproject called `FUNCRES.XLAM`. This solution may not always be possible, but if it is, do the following:

1. Open Visual Basic (again).
2. Locate the Solution Explorer in the left panel of the VBA screen; it shows one or more VBAProjects, including their modules. If that panel is missing, click the Project Explorer button on the toolbar.

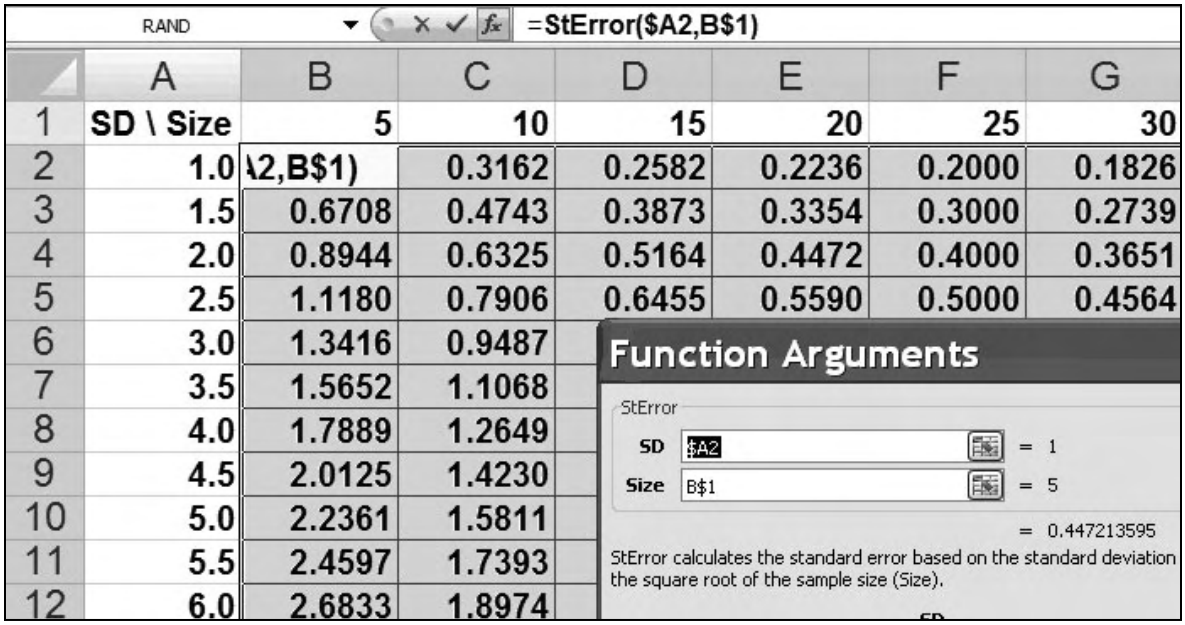


Figure: 4.52

3. If you see the project FUNCRES.XLAM, open it and select its Module section.
4. Select Insert, Module.
5. Double-click the original module (in the left panel) that you have worked on before and copy its entire function code.
6. Paste the code into the new module.

You can now use all functions in any `.xlsx` file on your machine.

Another solution to the problem of having no access to custom functions outside the file we created them in would be to create a personal macro workbook by recording a fake macro first. Here's how:

1. Click the Record Macro button on the Developer tab.
2. Accept the macro name that is already there.
3. Indicate to store the macro in the personal macro workbook and click OK.
4. Record anything—for example, Ctrl+Home.
5. Stop recording by again clicking the button that was Record Macro. Now there is a new section in the left panel of your VBA screen, and it includes a module that contains your macro.
6. Replace the macro code with a copy of your function code.

Any Excel file on your machine has access to the personal macro workbook, including its customized function(s).

* * *

Chapter 44

WORKSHEET FUNCTIONS

When you create functions on your own, you can base them on preexisting Excel worksheet functions. You don't want to keep reinventing the wheel, so you can use existing Excel functions in your own customized functions in order to get exactly the functionality you need.

Figure 4.53 uses a customized function that works with cell ranges to incorporate existing Excel functions that use cell ranges as well. The function here works with the preexisting function AVERAGE. Here's how you create it:

1. To declare variables as being of the Range type, enter the following line: `Function MeanChange(Before As Range, After As Range) As Double`.
2. Inside the function, we want to find the difference between the mean of the `After` range and the mean of the `Before` range. So type inside the function skeleton this line: `MeanChange = WorksheetFunction.Average(After) - WorksheetFunction.Average(Before)`.
3. Switch back to Excel. The function `MeanChange` should now work correctly in cell C19: `=MeanChange(B2:B17,C2:C17)`.

C19 =MeanChange(B2:B17,C2:C17)

	A	B	C	D	E
1		Before	After		
2	Patient1	213.4	200.1		
3	Patient2	225.0	216.4		
4	Patient3	217.0	195.6		
5	Patient4	183.7	175.0		
6	Patient5	197.2	202.3		
7	Patient6	223.6	214.8		
8	Patient7	224.2	215.7		
9	Patient8	215.2	200.7		
10	Patient9	202.4	211.7		
11	Patient10	217.7	216.1		
12	Patient11	221.0	208.5		
13	Patient12	219.9	188.4		
14	Patient13	205.4	211.4		
15	Patient14	195.1	180.9		
16	Patient15	218.0	184.1		
17	Patient16	207.6	202.3		
18					
19	Mean Change		-10.15		
20					

Figure: 4.53

Note: In Chapter 18, you got the same result by using a single-cell array formula. However, creating a customized function makes more sense if you do this kind of calculation frequently.

Figure 4.54 presents a similar case. In Chapter 43, you created a function for the standard error, based on a standard deviation and a count. This time, you need to create one that uses the original list of readings—which is a cell range—instead of using intermediate calculations. Here's what you do:

1. For the first line, enter `Function SERange(Series As Range) As Double`.
2. Inside this function, type the three existing Excel functions `STDEV`, `SQRT`, and `COUNT` so that the second line looks like this: `SERange = WorksheetFunction.StDev(Series) / Sqr(WorksheetFunction.Count(Series))`.

Note: This example uses the shorter VBA version `Sqr(…)` because `WorksheetFunction. Sqrt(…) does not exist`. However, you can use your own customized `AnyRoot` function from Chapter 43, provided that it is available to you in this file.

In Excel, cell E5 uses the new function: `=SERange(B1:B18)`. The advantage of using this function is that you don't have to do the calculations of cells E1 and E2 first.

E5 =SERange(B1:B18)

	A	B	C	D	E
1	abc	1.71		SD	0.28
2	abc	1.11		Count	18
3	abc	1.84		SE	0.07
4	abc	1.24			
5	abc	1.94		SE	0.07
6	klm	1.97			
7	klm	1.18			
8	klm	1.54			
9	klm	1.69			
10	klm	1.76			
11	mno	1.50			
12	mno	1.78			
13	mno	1.78			
14	mno	1.40			
15	xyz	1.97			
16	xyz	1.29			
17	xyz	1.80			
18	xyz	1.45			

Figure: 4.54

Figure 4.55 shows a rather complicated scenario. As you know, Excel provides functions such as `COUNTIF`, `SUMIF`, and `AVERAGEIF`, but it does not have anything like `STDEVIF`. In other words, there is no built-in Excel function for calculating the standard deviation for specific values in a list. This last example, which is a pretty tough one, shows how much you can do with VBA. Follow these steps:

1. On the first line of the VBA Code Window, enter `Function StDevIf(SDRange As Range, CritRange As Range, Crit As String) As Double`.
2. Declare a temporary variable of the `Variant` type on the second line: `Dim TempRange As Variant`.
3. The variable `TempRange` is just going to hold a copy of `SDRange`, so you can work on it and remove values that do not qualify to be included in the standard error calculations. To create this copy, fill `TempRange` with the values from `SDRange` on the third line, so you

G2 =StDevIf($B$1:$B$18,$A$1:$A$18,D2)

	A	B	C	D	E	F	G
1	abc	1.71			Count	Mean	SD
2	abc	1.11		abc	5	1.63	0.32
3	abc	1.84		klm	5	1.54	0.31
4	abc	1.56		mno	4	1.62	0.19
5	abc	1.94		xyz	4	1.63	0.31
6	klm	1.97					
7	klm	1.18					
8	klm	1.54					
9	klm	1.69					
10	klm	1.34					
11	mno	1.50					
12	mno	1.78					
13	mno	1.78					
14	mno	1.40					
15	xyz	1.97					
16	xyz	1.29					
17	xyz	1.80					
18	xyz	1.45					

Figure: 4.55

can adjust the values that do not qualify to be in the standard deviation: `TempRange = SDRange`.

4. On the fourth line, loop through all the cells of `CritRange` to see which cells do not match the criterion (a specific analyst). Keep looping until you reach the last cell in the range (by means of `Cells.Count`). In this loop, you use the variable `i` as a counter. Start the loop like this: `For i = 1 To CritRange.Cells.Count`.
5. Inside the loop on the next line, empty each value in `TempRange` if that value is not (<>) identical to `Crit`: `If CritRange.Cells(i, 1) <> Crit Then TempRange(i, 1) = ""`.
6. Close the loop on a closing line: `Next i`.
7. Using a new line after the loop, return the standard deviation of all values that are left in `TempRange`: `StDevIf = WorksheetFunction.StDev(TempRange)`.
8. Use the new function in cell G2: `=StDevIf($B$1:$B$18,$A$1:$A$18,D2)`.

Figure 4.56 shows all the code for this example. If you don't understand what is going on in the background, don't worry about it; all you need to know for now is that VBA is a very powerful tool that you can use to customize your Excel experience.

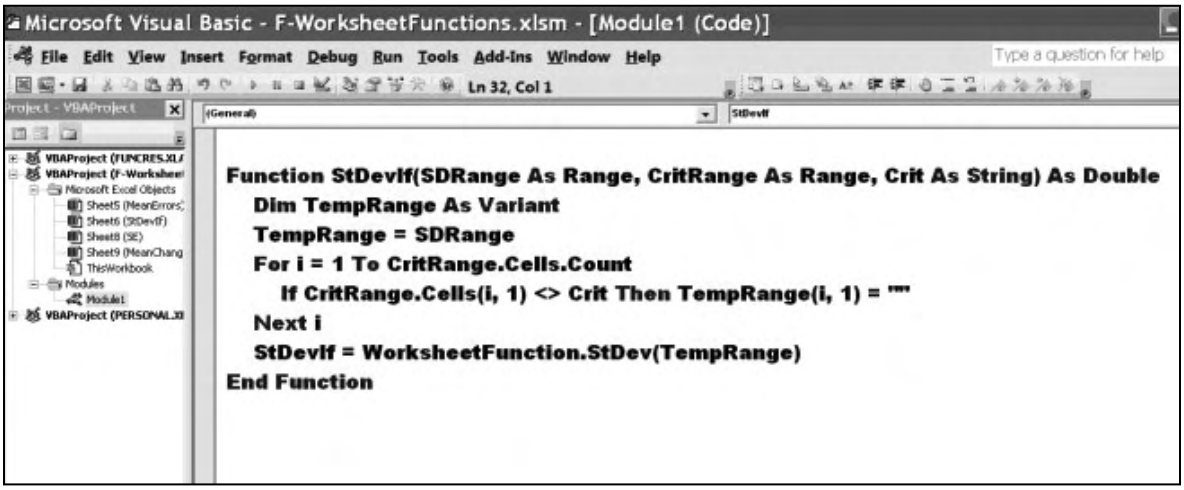


Figure: 4.56

> **Note**: To delve more deeply into VBA, see the *Excel 2007 VBA* CD-ROM from MrExcel's *Visual Learning series* at `www.mrexcel.com`.

* * *

Excercises - Part 4

You can download all the files used in this book from `www.genesispc.com/Science2007.htm`, where you can find each file in its original version (to work on) and in its finished version (to check your solutions).

Exercise 1

1. Linear Regression
   1.1. Create an "automatic" linear regression line.
   1.2. Extrapolate a 20% mutation rate on the automatic linear regression line.
   1.3. Create a linear regression line manually by using column C.
   1.4. Extrapolate a 20% mutation rate on the linear regression line by using the `TREND` function.

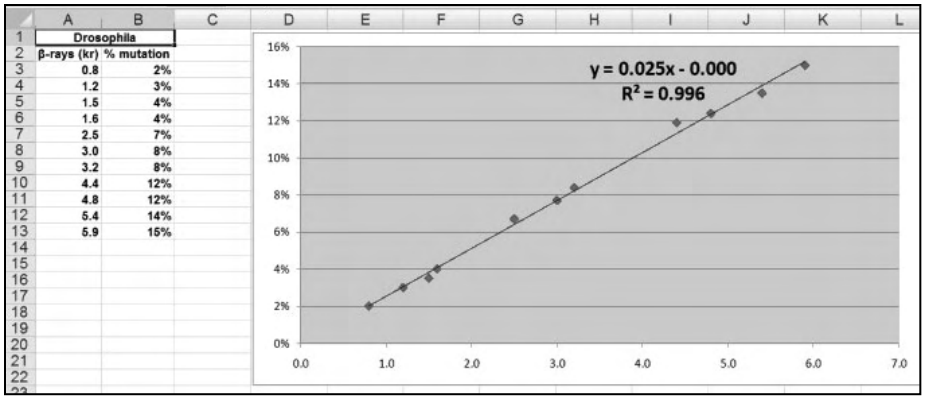
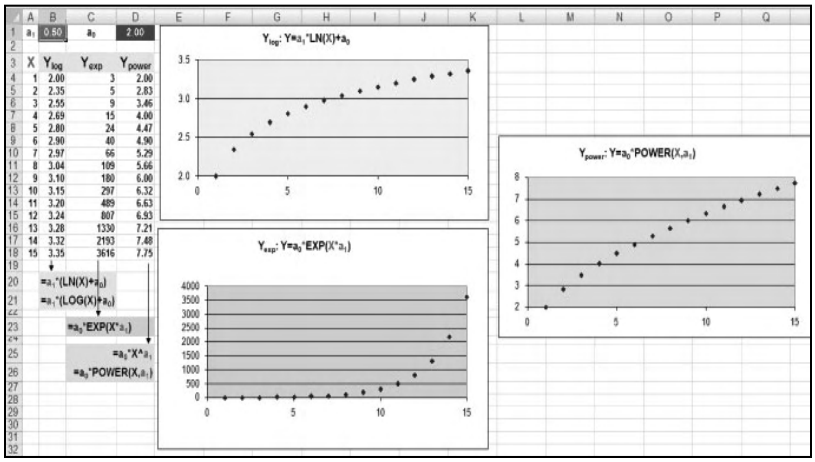


Figure: Ex-1



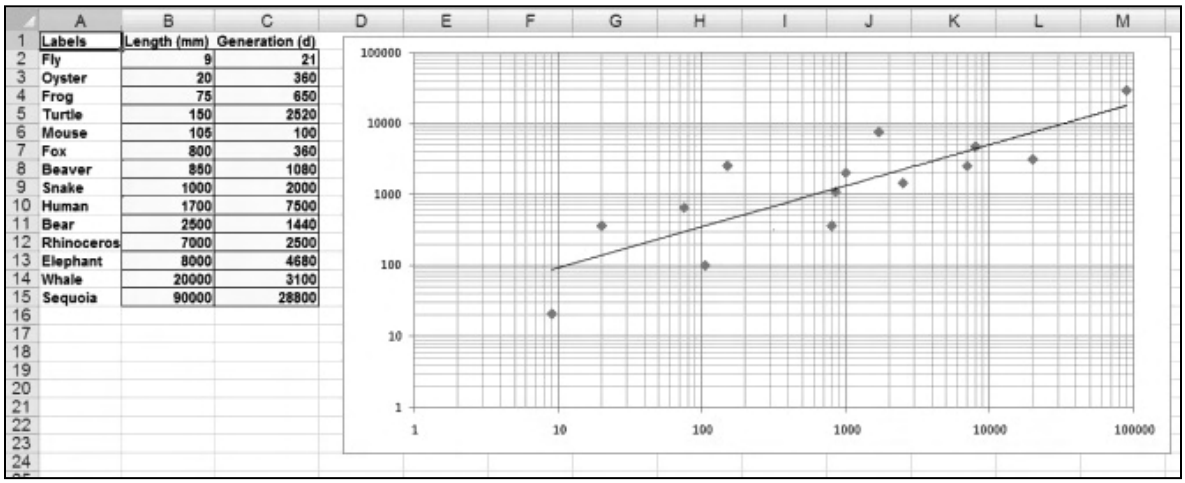
Exercise 2

2. Nonlinear Regression
   2.1. For each graph shown here, make the proper axis or axes logarithmic.
   2.2. Add the proper regression line to each curve.

Figure: Ex-2

Exercise 3

3. Nonlinear Regression
    3.1. Create a graph that looks like the one shown here.
    3.2. Apply the correct type of regression line.

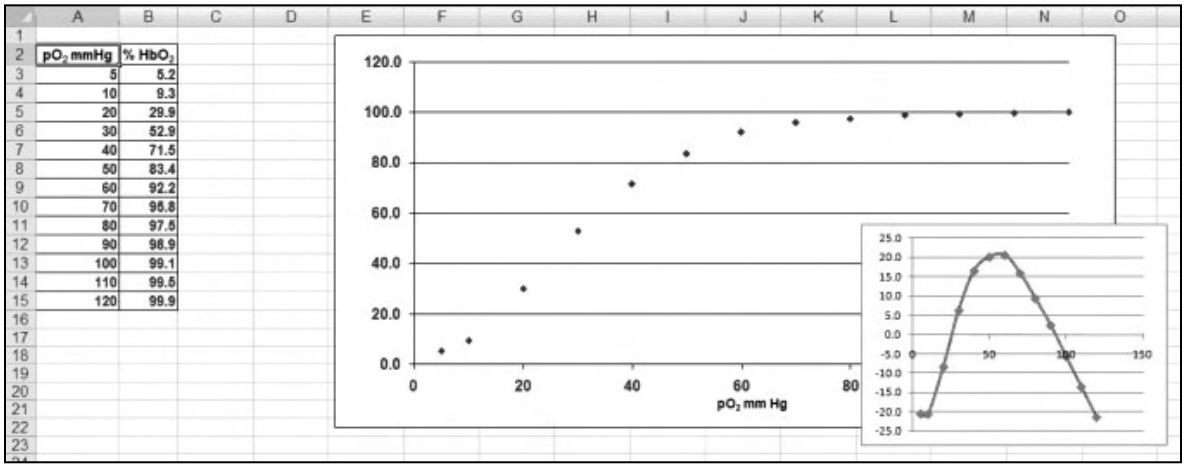


Labels	Length (mm)	Generation (d)
Fly	9	21
Oyster	20	360
Frog	75	650
Turtle	150	2520
Mouse	105	100
Fox	800	360
Beaver	850	1080
Snake	1000	2000
Human	1700	7500
Bear	2500	1440
Rhinoceros	7000	2500
Elephant	8000	4680
Whale	20000	3100
Sequoia	90000	28800

Figure: Ex-3

Exercise 4

4. Curve Fitting
    4.1. Create a residuals plot as shown in the insert.
    4.2. Decide on the proper regression line for the original curve.
    4.3. Apply a final curve-fitting test.



pO_2 mmHg	% HbO_2
5	5.2
10	9.3
20	29.9
30	52.9
40	71.5
50	83.4
60	92.2
70	95.8
80	97.5
90	98.9
100	99.1
110	99.5
120	99.9

Figure: Ex-4

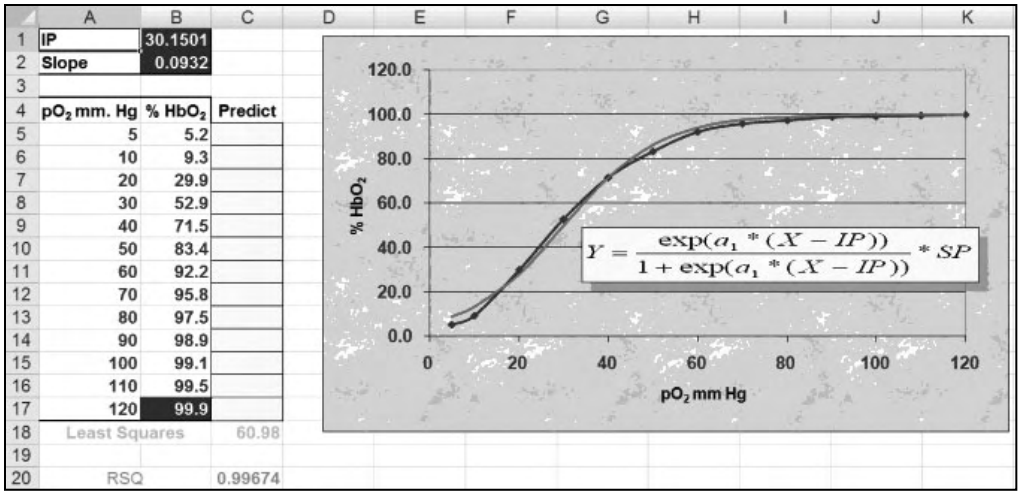
Exercise 5

5. Sigmoid Curves

5.1. Predict or estimate the percentage of hemoglobin oxygenation based on partial oxygen pressure if the relationship is of the sigmoid type.

5.2. Add the sigmoid regression curve to the graph.

5.3. Do the necessary curve fitting.



	A	B	C
1	IP	30.1501	
2	Slope	0.0932	
3			
4	pO_2 mm. Hg	% HbO_2	Predict
5	5	5.2	
6	10	9.3	
7	20	29.9	
8	30	52.9	
9	40	71.5	
10	50	83.4	
11	60	92.2	
12	70	95.8	
13	80	97.5	
14	90	98.9	
15	100	99.1	
16	110	99.5	
17	120	99.9	
18	Least Squares		60.98
19			
20	RSQ		0.99674

Figure: Ex-5

Exercise 6

6. Linear Estimates

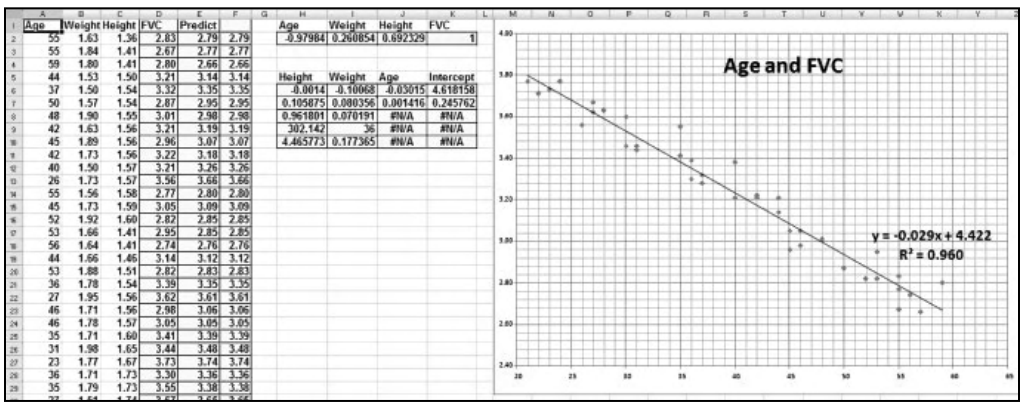
6.1. Calculate the correlations in H2:K2.

6.2. Find the multiple regression coefficients in H6:K10.

6.3. Predict or estimate FVC in column F by using coefficients.

6.4. Predict or estimate FVC in column G by using the `TREND` function.

6.5. Add a linear regression line to the graph.



Age	Weight	Height	FVC	Predict	
59	1.63	1.36	2.83	2.79	2.79
55	1.84	1.41	2.67	2.77	2.77
59	1.80	1.41	2.80	2.66	2.66
44	1.53	1.50	3.21	3.14	3.14
37	1.50	1.54	3.32	3.35	3.35
50	1.57	1.54	2.87	2.95	2.95
48	1.90	1.55	3.01	2.98	2.98
42	1.63	1.56	3.21	3.19	3.19
45	1.89	1.56	2.96	3.07	3.07
42	1.73	1.56	3.22	3.18	3.18
40	1.50	1.57	3.21	3.26	3.26
26	1.73	1.57	3.56	3.66	3.66
55	1.56	1.58	2.77	2.80	2.80
45	1.73	1.59	3.05	3.09	3.09
52	1.92	1.60	2.82	2.85	2.85
53	1.66	1.41	2.95	2.85	2.85
56	1.64	1.41	2.74	2.76	2.76
44	1.66	1.46	3.14	3.12	3.12
53	1.88	1.51	2.82	2.83	2.83
36	1.78	1.54	3.39	3.35	3.35
27	1.95	1.56	3.62	3.61	3.61
46	1.71	1.56	2.98	3.06	3.06
46	1.78	1.57	3.05	3.05	3.05
35	1.71	1.60	3.41	3.39	3.39
31	1.98	1.65	3.44	3.48	3.48
23	1.77	1.67	3.73	3.74	3.74
36	1.71	1.73	3.30	3.36	3.36
35	1.79	1.73	3.55	3.38	3.38

Age	Weight	Height	FVC
-0.97984	0.260854	0.692329	1

Height	Weight	Age	Intercept
-.0014	-0.10068	-0.03015	4.618158
0.105875	0.080356	0.001416	0.245762
0.961801	0.070191	#N/A	#N/A
302.142	36	#N/A	#N/A
4.465773	0.177365	#N/A	#N/A

Figure: Ex-6

	A	B	C	D	E
1	Clearance mL/min	Weight	Age	Serum Creatinine mg/dL	
2	206.49	170	36	1.07	218.58
3	257.79	172	57	0.81	252.46
4	311.36	178	24	0.93	321.23
5	283.94	181	44	0.81	306.65
6	463.83	184	37	0.58	420.44
7	433.02	148	26	0.58	405.94
8	307.42	169	32	0.83	320.99
9	401.69	135	29	0.53	397.27
10	153.85	130	52	0.93	164.92
11	293.05	165	28	0.94	286.61
12	222.17	186	35	1.21	190.87
13	314.42	175	33	0.78	345.03
14	192.16	152	67	0.89	162.36
15	310.61	184	71	0.55	322.65
16	287.92	159	41	0.71	323.74

Exercise 7

7. Linear Estimates

7.1. Check for colinearity between weight, age, and serum creatinine levels in predicting creatinine clearance.

7.2. Find the coefficients for this multiple regression by using the Analysis Toolpak.

7.3. Predict or estimate creatinine clearance in column E.

7.4. Check RSQ and determine whether any factor should be eliminated.

Figure: Ex-7

Exercise 8

8. Reiterations

8.1. Activate iterations in this workbook.

8.2. Fill in the entire table, keeping A4, A10, D7, G4, and G10 fixed and calculating the gradient values in between.

8.3. Check whether all values have been stabilized.

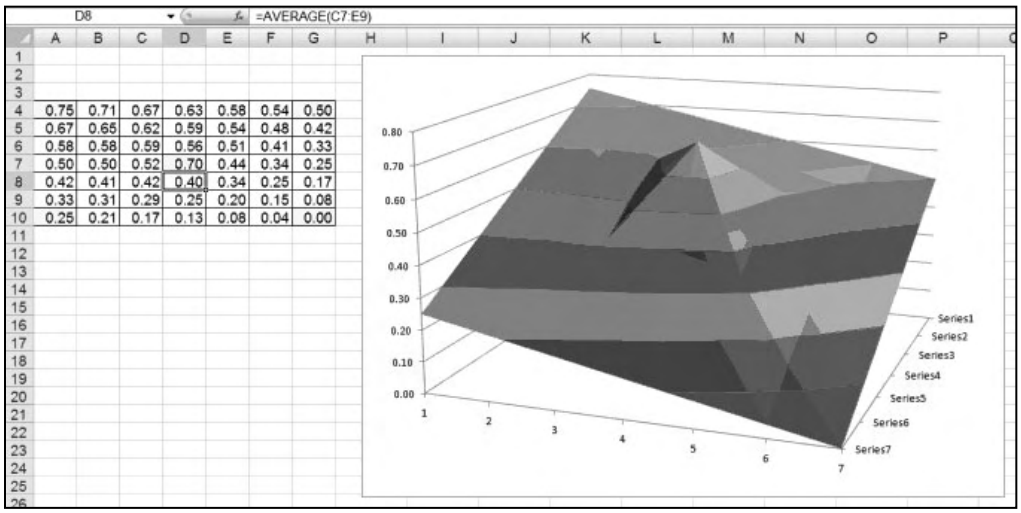


Figure: Ex-8

Exercise 9

9. Solving Tools

9.1. Use Solver to solve three equations with three unknown x values (in A4:A6) by setting H4:H6 to the values in G4:G6.

9.2. Use Solver again, but this time minimize the sum of the squared residuals in cell I15.

9.3. In both cases, ensure that the three unknown x values are the same as the ones found with the matrix system in cells I24:I26.

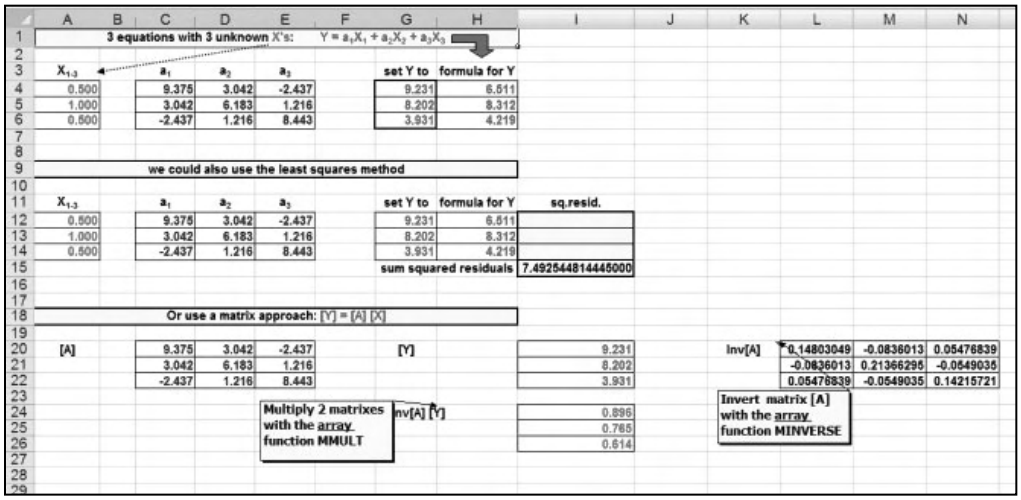


Figure: Ex-9

Exercise 10

10. What-if Controls

10.1. Create a control that regulates IP (used in the formula of column C) with a precision of two decimals.

10.2. Create a control that regulates the slope with a precision of two decimals.

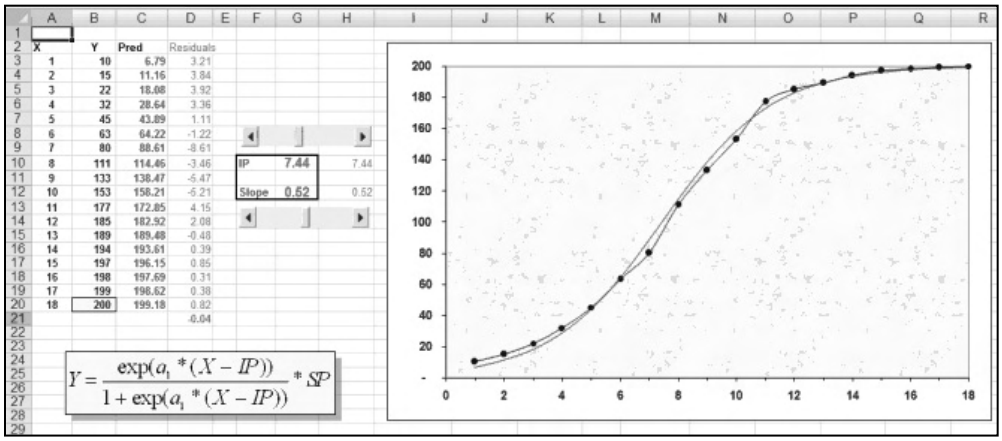


Figure: Ex-10

Exercise 11

11. Syntax of Functions

11.1. Create a customized function that calculates the offspring of y organisms after x generations, with a growth rate of z per generation.

11.2. Add some explanation to the function.

11.3. Use the new function in the table.

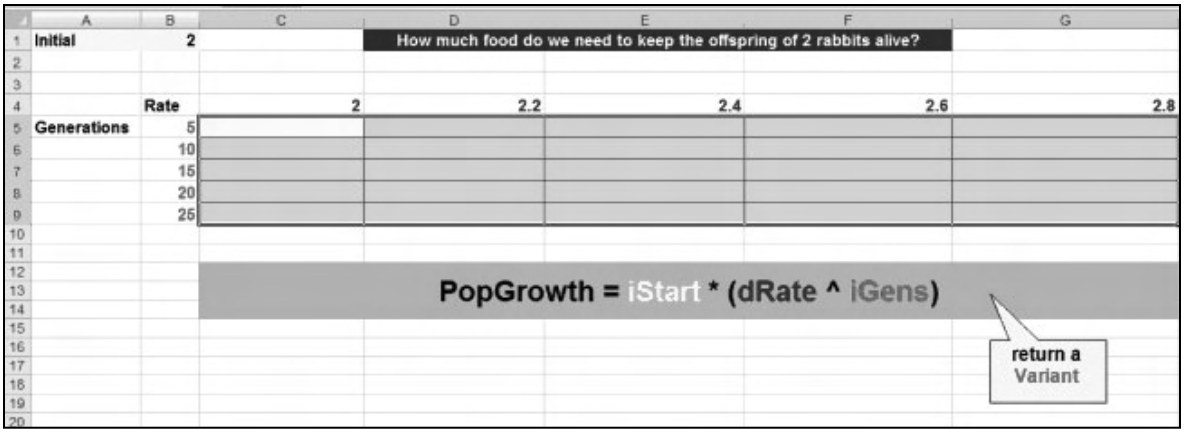


Figure: Ex-11

Exercise 12

12. Syntax of Functions

12.1. Create a custom function that calculates the minimum sample size based on a certain mean and SD plus a requested margin of error. Use the formula shown.

12.2. Make the variable 1.96 optional; it stands for a confidence level of 95% but should be adjustable.

12.3. Apply the new function to E2:J2 and so on.

	A	B	C	D	E	F	G	H	I	J
1	Mean	4.15		Margin	0.05	0.1	0.15	0.2	0.25	0.3
2	SD	0.32		Size	157	39	17	10	6	4
3										
4	Mean	6.80		Margin	0.01	0.02	0.03	0.04	0.05	0.06
5	SD	0.05		Size	96	24	11	6	4	3
6										
7	Mean	175.00		Margin	5	10	15	20	25	30
8	SD	25		Size	96	24	11	6	4	3

$$n = \left(\frac{1.96}{margin / \mu} \right)^2 \left(\frac{SD}{\mu} \right)^2$$

Figure: Ex-12

Exercise 13

13. Worksheet Functions

13.1. Create a custom function that calculates the mean but skips error values. Do this by using existing functions such as `AVERAGE` and `IFERROR`.

13.2. Use the new function in cell E20.

13.3. Use a single-cell array formula to compare the results.

E20 =MeanErrors(B2:B19)

	A	B	C	D	E	F
1					SD	
2	abc	1.71	1.46	1.25	0.23	
3	abc	1.11	1.90	1.85	0.44	
4	abc	1.84	1.46	1.57	0.20	
5	abc				#DIV/0!	
6	abc	1.94	1.82	1.18	0.41	
7	klm	1.97	1.34	1.99	0.37	
8	klm	1.18	1.47	1.38	0.15	
9	klm	1.54		1.39	0.11	
10	klm	1.69	1.28	1.34	0.22	
11	klm			1.86	#DIV/0!	
12	mno	1.50	1.07	1.47	0.24	
13	mno	1.78	1.72	1.22	0.31	
14	mno	1.78	1.76	1.46	0.18	
15	mno	1.40	1.74	1.94	0.27	
16	xyz	1.97	1.03	1.91	0.53	
17	xyz	1.29	1.33	1.14	0.10	
18	xyz	1.80	1.66	1.89	0.12	
19	xyz	1.45	1.30	1.17	0.14	
20		Mean			1.44	
21						

Figure: Ex-13

* * *

PART 5
Statistical Analysis

Chapter 45

Why Statistics?

Because scientists usually work with samples taken from huge populations, they need to deal with the fact that samples are never exact replicas of the population they represent. In order to assess how much this fact may impact results, you need statistics. This chapter doesn't provide a crash course on statistics, but you'll learn the basics of how to use Excel in your statistical work.

Figure 5.1 illustrates what happens in sampling. Let's pretend you want to study the "infinite population" of random numbers between 0 and 10 by taking 20 different samples with size 10. Each row represents one of those 20 samples. You calculate for each sample the mean of all 10 random numbers drawn from this population. Notice that those means may vary quite a bit; the extreme low and high ones are displayed in a different font. What you try to simulate here is drawing new samples from the same population—and yet the mean found in the sample keeps changing. That's what happens in research! As they say, "Results may vary." However, there is one value that doesn't change as widely: the mean of all these 20 means, featured in cell L22.

Whereas the individual means may reach 3 or 7 rather easily, it is unlikely that the mean of means ends up outside the range of 4.5 to 5.5. A sample's mean is often symbolized as $\overline{x}$, x_m, or m. The mean of means is often symbolized as μ_m.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	sample	each sample holds 10 random numbers between 0 and 10										mean			
2	1	9	10	3	7	10	5	9	10	9	0	7.2			
3	2	0	0	2	2	4	10	6	9	7	7	4.7			
4	3	9	2	9	0	8	8	2	9	10	1	5.8			
5	4	7	7	4	2	5	10	5	7	10	10	6.7			
6	5	2	5	10	9	0	10	9	10	0	0	5.5			
7	6	4	10	0	9	6	0	7	5	2	3	4.6			
8	7	8	3	5	0	6	4	0	4	1	3	3.4			
9	8	7	6	1	7	10	0	9	7	4	4	5.5			
10	9	2	9	8	1	8	4	4	6	9	10	6.1		each sample holds	
11	10	2	8	1	5	6	2	2	5	1	5	3.7		10 random numbers	
12	11	7	4	1	0	2	5	4	5	10	9	4.7		between 0 and 10	
13	12	7	1	2	6	8	8	0	7	7	4	5.0			
14	13	2	7	9	10	6	6	6	6	0	6	5.8			
15	14	5	6	9	3	2	4	10	10	8	0	5.7			
16	15	2	2	5	8	0	10	2	8	8	3	4.8			
17	16	4	6	8	6	5	5	2	2	7	1	4.6			
18	17	0	3	6	4	9	5	10	2	3	2	4.4			
19	18	0	5	10	3	9	4	8	5	3	3	5.0			
20	19	0	3	3	5	10	9	4	2	6	0	4.2			
21	20	2	6	10	8	7	3	1	0	2	7	4.6			
22									mean of 20 sample means			5.1			

Figure: 5.1

Figure 5.2 shows another aspect of this phenomenon. To create a table like this, you first create a frequency table of individual means (on the right); then, you plot those frequencies in a graph (in the center). The vertical line in the graph represents the current mean of the means; it is the mean of all 20 sample means combined. The range of the mean of the means is rather narrow. The curve representing the frequency of means usually has some kind of a bell shape. A common technique in research is to find the range of means based on the mean of a particular sample.

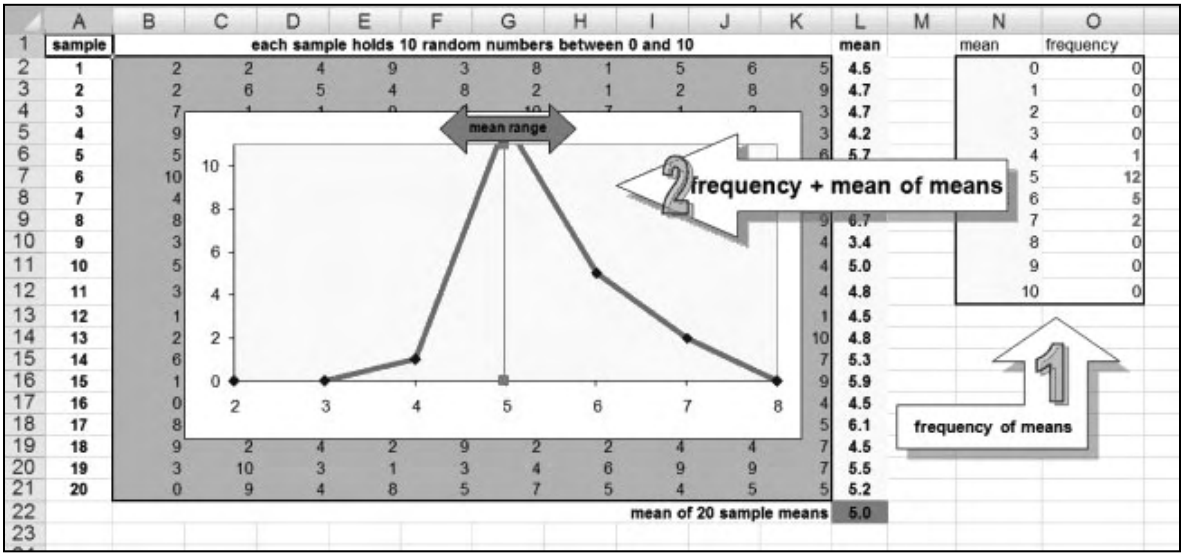


Figure: 5.2

Figure 5.3 shows this same idea in another way. Again, you do two things to create a table like this: First, you create cumulative means for increasing sample sizes; second, you plot those means in a graph. The sample size in columns N:O therefore keeps going up. Very often, the means are rather scattered at the left side of the graph (for small sample sizes), but they tend

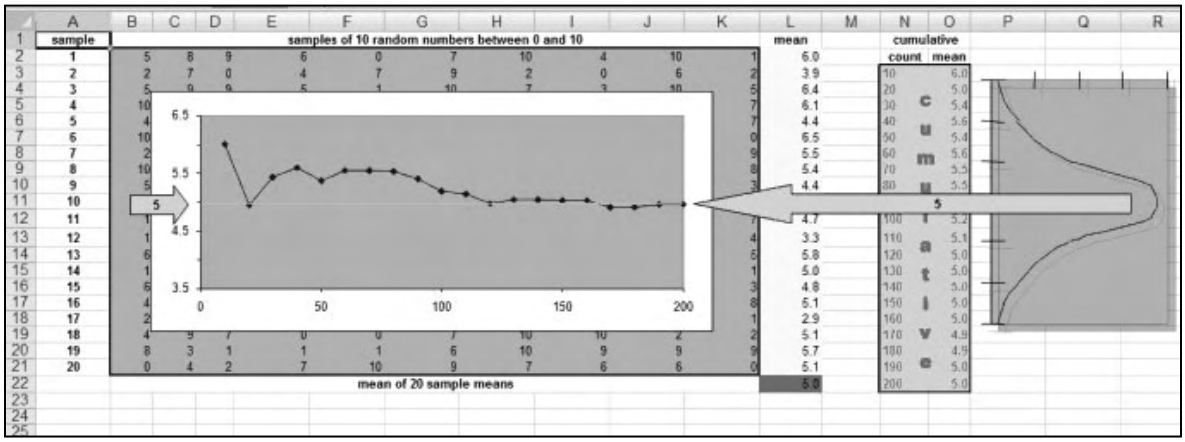


Figure: 5.3

to converge to "the real mean" toward the right (for large sample sizes). This phenomenon is based on the law of large numbers: With increasing sample size, the mean becomes more reliable—that is, it becomes representative for the entire population. The rotated normal distribution curve (bell shape) on the right shows that low and high means have a lower frequency than means in the center of the range of means.

Figure 5.4 brings us closer to the central issue. It shows two frequency distributions: The one on the left is for frequencies of values (as found in sample 5, for instance); it is called a `sample distribution`. The distribution on the right is for frequencies of means (all samples taken together), and it is called a `sampling distribution`. In statistics, the normal distribution curve is the graph on the right. Basically, research is about the graph on the left, and statistics is about the graph on the right.

Another key concept in this context is standard deviation (SD); the standard deviation is a measure of how widely values are dispersed from their mean. Figure 5.4 calculates the standard deviation twice (in row 14): The left one is the SD of observations; the right one is the standard deviation of means. How are these two SDs related? The SD of means is always smaller than the SD of observations; the SD of means decreases with increasing sample size.

Because research is basically about the left part of this screen, you often do not know about the mean of the means and the standard deviation of the means to the right. All you have on hand are the mean and SD as found in the sample to the left. So you have to obtain an estimate of the SD of the means based on the SD of the observations. You do this by calculating the standard error (SE), also called the relative SD. The formula for the SE is $SE = SD / \sqrt{n}$. As you can gather from this formula, the SE decreases when the sample size increases.

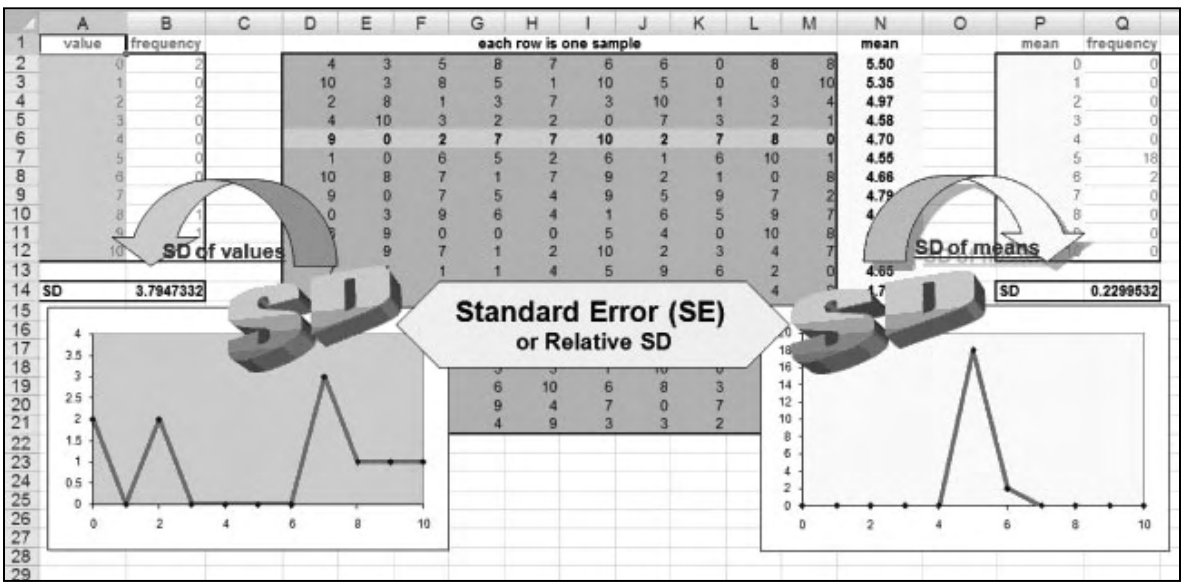


Figure: 5.4

Figure 5.5 shows the normal distribution, which is at the heart of statistics. The normal distribution is a sampling distribution of means. The mean of means is located at 0 on the x axis. The units plotted on this axis are units of SE. The SE units on the horizontal axis are unit free—that is, they don't change when the units of measurement change. It doesn't matter whether you are dealing with degrees Celsius or Fahrenheit or with ounces or kilograms—these SE units are unit free.

Means that are more than 1.96 SE units away from the mean of means are extremely rare—actually, only 5% of the means occur in those two outer ranges. The other 95% of means are located in the center shaded range. The units of SE around the mean of means determine how much of the curve's surface has been covered: 70% for +1 SE; 90% for +1.65 SE; 95% for +1.96 SE; and 97.5% for +2.25 SE. You'll learn more on this later.

Imagine that a population has a mean of 15 (whether °C, ng/ml, mol, or whatever), and a sample of 16 cases has a mean of 14 and a SD of 2. How many SE units are these two means apart from each other? The answer is (14-15)/SE = (14-15)/(SD/√16) = -0.5 SE units.

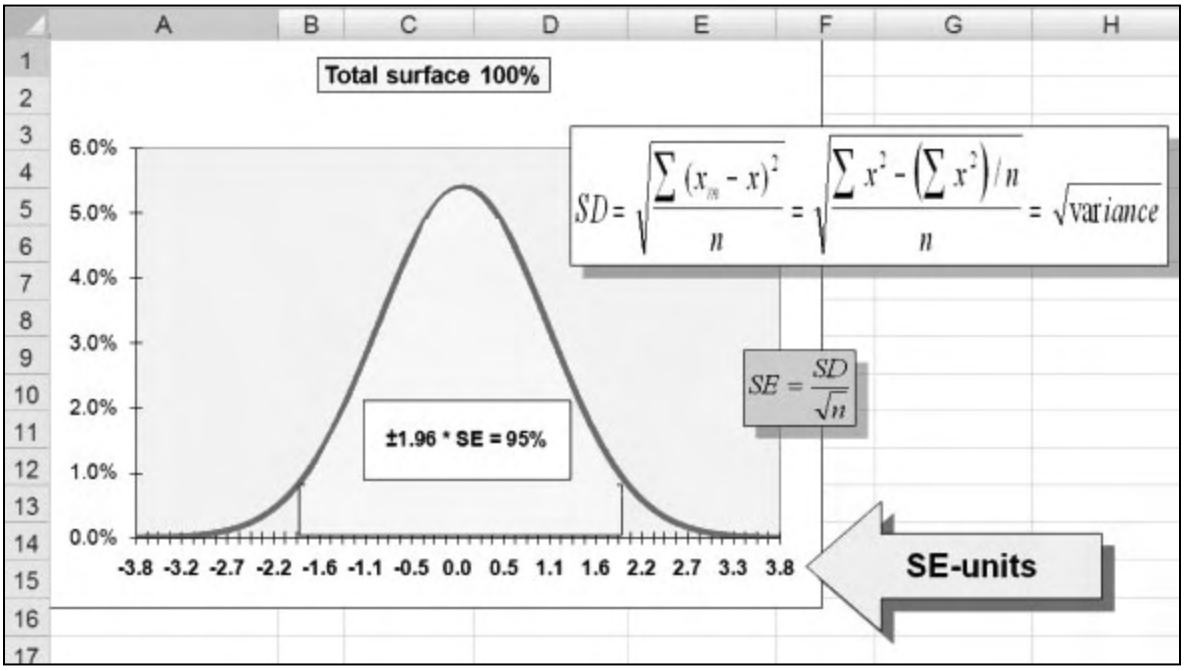


Figure: 5.5

* * *

Chapter 46

Types of Distributions

The normal distribution is only one of the several types of sampling distributions used in statistics. This chapter discusses the key distributions, their characteristics, and when to use each one. Distributions are sampling distributions that are meant to help evaluate sample distributions.

The normal distribution has two versions, as Figure 5.6 demonstrates: the noncumulative version (to the left) and the cumulative version (to the right). The cumulative graph shows that the area to the left of 0 in the noncumulative graph covers 50% of all cases. It also shows that a mean being +2 SE units away from the mean of means covers up to 97.5% of all cases. As mentioned earlier, the x axis features units of SE. These are "universal" units that can be applied to means of any magnitude (pH, °C, ng/ml, mol, volts, and so on). In case of a normal distribution, these units are called *z-values*. They can be positive or negative because the normal distribution is symmetrical.

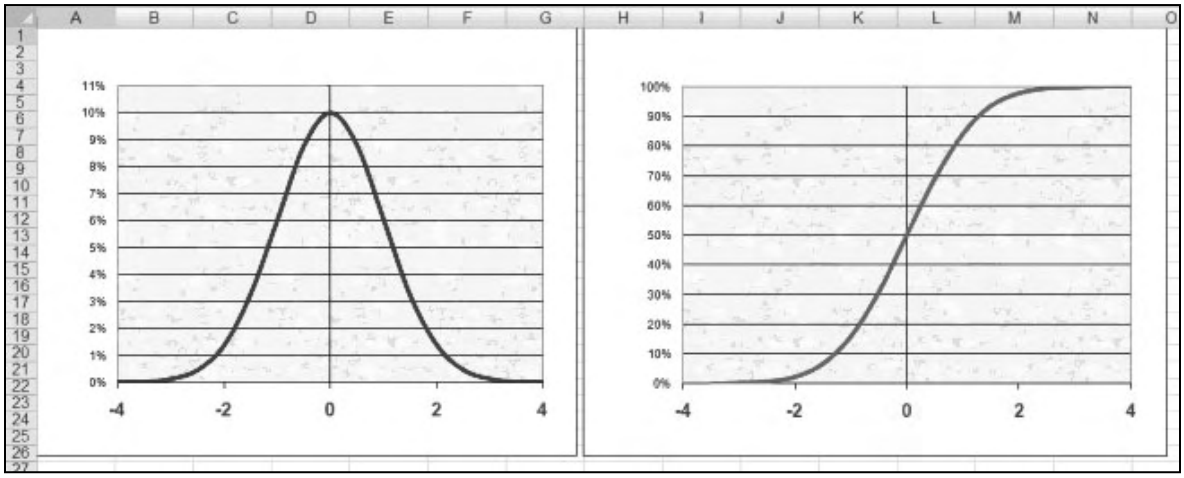


Figure: 5.6

But there are additional types of distributions. For example, Figure 5.7 shows the *t*-distribution, the chi-distribution, the binomial distribution, and the *F*-distribution. There are a few more distributions, but these are the ones discussed in this part of the book. The four curves shown in Figure 5.7 are all nonsymmetrical, so the x axis has only positive units (unlike the normal bell-shaped curve), and they happen to be of the cumulative type, so the vertical axis has a scale up to 50% or 100%. And again, the x axis has units of SE. These units of SE have been named after the distributions they come with:

- *t*-values are for means.
- Chi-values are for frequencies.
- *p*-values are for proportions.
- *F*-values are for variances.
- *z*-values are for means.

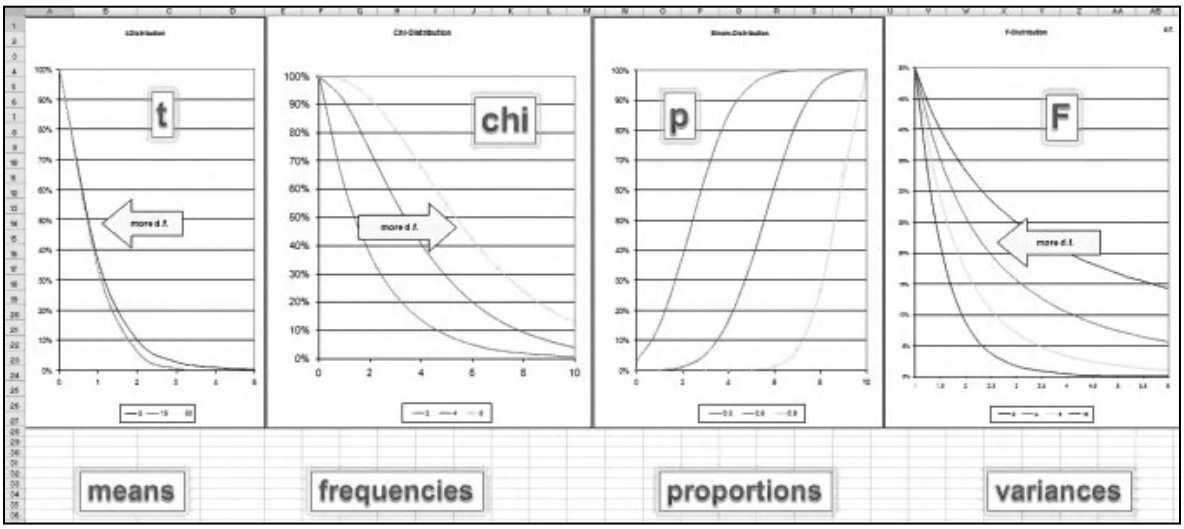


Figure: 5.7

How do you use Excel functions to calculate the values on these axes? Figure 5.8 offers an overview:

- The x axis holds the units of SE (*z*-values, *t*-values, and so on). To calculate SE values, Excel uses functions whose names end with `INV`. `NORMSINV`, for instance, would find the *z*-value for a certain probability, and `TINV` would find the *t*-value for that probability.
- The y axis plots the probabilities with which those values occur (either noncumulative, or cumulative). To calculate probabilities, Excel uses functions whose names end with `DIST`. `NORMSDIST`, for instance, would find the probability of a certain *z*-value.

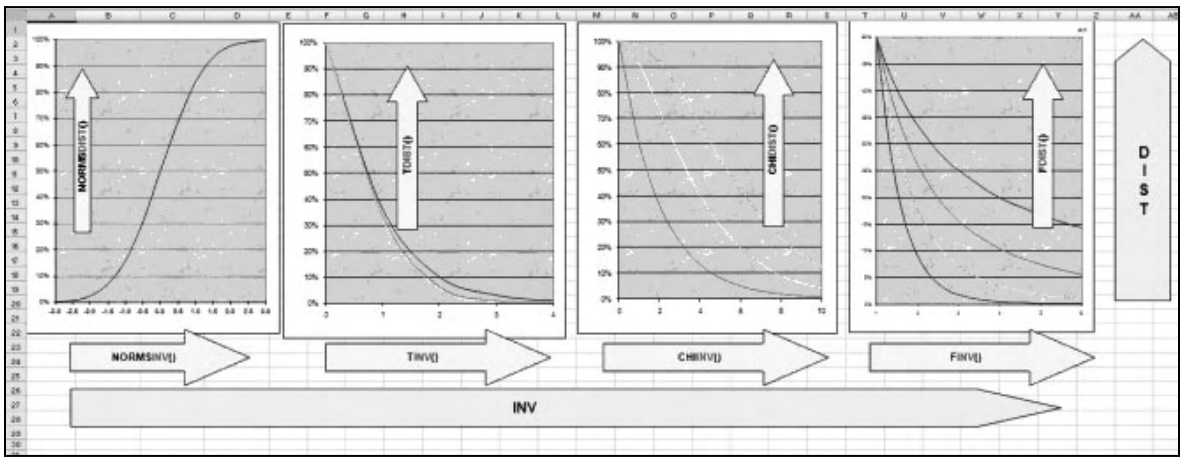
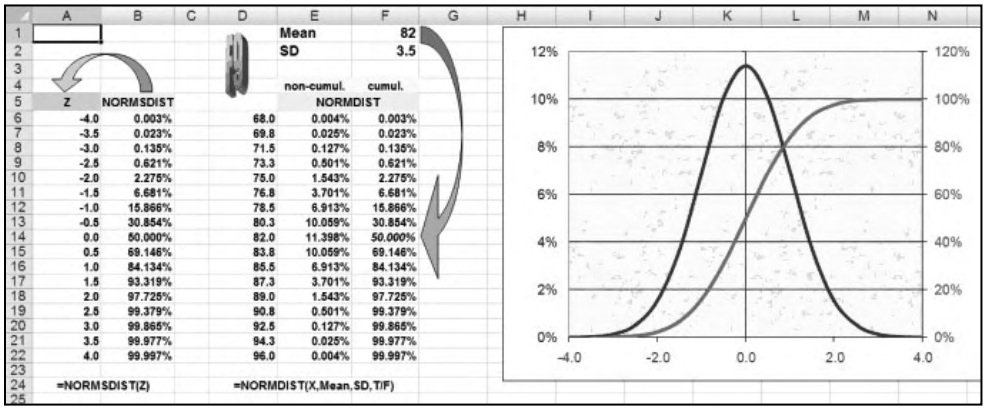


Figure: 5.8

Let's look at the normal distribution. As you can see in Figure 5.9, normal distributions work with two different `DIST` functions: `NORMDIST` and `NORMSDIST`. Both of these functions are designed to find probabilities:

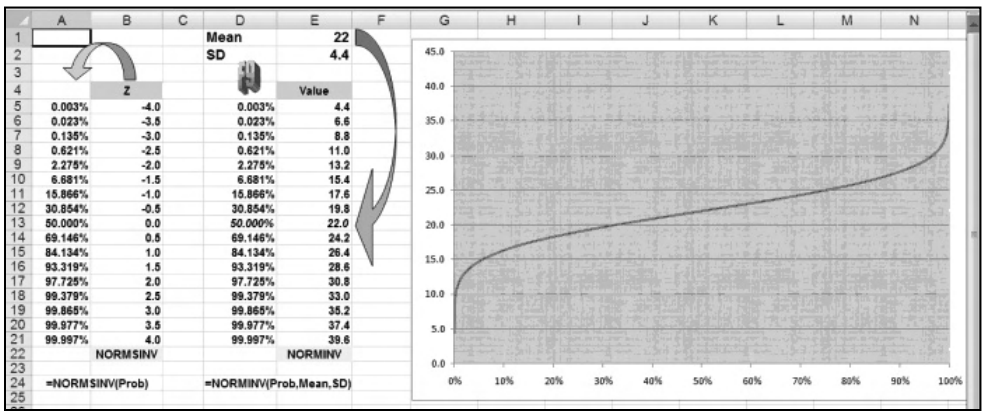
- **NORMSDIST**: This function is *z*-related and cumulative: It uses *z* to find its probability (see column B). Cell B6 uses it: `=NORMSDIST(A6)`. `NORMSDIST` returns 50% for a *z*-value of `0` because it is cumulative.
- **NORMDIST**: This function (without an S) is not *z*-related: It uses a specific mean (of any magnitude) to find its probability. In addition, `NORMDIST` has two versions:
  - Noncumulative version (in column E); cell E6: `=NORMDIST(D6,Mean,SD,FALSE)`. The noncumulative version of NORMDIST never returns 50%; the last argument makes it noncumulative.
  - Cumulative version (in column F); cell F6: `=NORMDIST(D6,Mean,SD,TRUE)`. The cumulative version of `NORMDIST` returns 50% for the mean used in cell F1.



	A	B	D	E	F
1			Mean		82
2			SD		3.5
4				non-cumul.	cumul.
5	Z	NORMSDIST		NORMDIST	
6	-4.0	0.003%	68.0	0.004%	0.003%
7	-3.5	0.023%	69.8	0.025%	0.023%
8	-3.0	0.135%	71.5	0.127%	0.135%
9	-2.5	0.621%	73.3	0.501%	0.621%
10	-2.0	2.275%	75.0	1.543%	2.275%
11	-1.5	6.681%	76.8	3.701%	6.681%
12	-1.0	15.866%	78.5	6.913%	15.866%
13	-0.5	30.854%	80.3	10.059%	30.854%
14	0.0	50.000%	82.0	11.398%	*50.000%*
15	0.5	69.146%	83.8	10.059%	69.146%
16	1.0	84.134%	85.5	6.913%	84.134%
17	1.5	93.319%	87.3	3.701%	93.319%
18	2.0	97.725%	89.0	1.543%	97.725%
19	2.5	99.379%	90.8	0.501%	99.379%
20	3.0	99.865%	92.5	0.127%	99.865%
21	3.5	99.977%	94.3	0.025%	99.977%
22	4.0	99.997%	96.0	0.004%	99.997%
24	=NORMSDIST(Z)		=NORMDIST(X,Mean,SD,T/F)		

Figure: 5.9

Figure 5.10 shows the use of two `INV` functions. The graph may look somewhat unusual to you because the two axes have been interchanged. This is how you would read the columns: In this case, a mean of 11 (cell E8) is -2.5 *z*-values (cell B8) away from the mean in cell E1, and it has a 0.625% probability (cell A8 or D8) of occurring. But again, there are two



	A	B	D	E
1			Mean	22
2			SD	4.4
4		Z		Value
5	0.003%	-4.0	0.003%	4.4
6	0.023%	-3.5	0.023%	6.6
7	0.135%	-3.0	0.135%	8.8
8	0.621%	-2.5	0.621%	11.0
9	2.275%	-2.0	2.275%	13.2
10	6.681%	-1.5	6.681%	15.4
11	15.866%	-1.0	15.866%	17.6
12	30.854%	-0.5	30.854%	19.8
13	50.000%	0.0	*50.000%*	22.0
14	69.146%	0.5	69.146%	24.2
15	84.134%	1.0	84.134%	26.4
16	93.319%	1.5	93.319%	28.6
17	97.725%	2.0	97.725%	30.8
18	99.379%	2.5	99.379%	33.0
19	99.865%	3.0	99.865%	35.2
20	99.977%	3.5	99.977%	37.4
21	99.997%	4.0	99.997%	39.6
22		NORMSINV		NORMINV
24	=NORMSINV(Prob)		=NORMINV(Prob,Mean,SD)	

Figure: 5.10

INV functions for a normal distribution:

- **NORMSINV**: This function is *z*-related and cumulative: It finds *z* for a certain probability. For a 50% probability, *z* would be 0. Cell B5 uses it: `=NORMSINV(A5)`.
- **NORMINV**: This function (without an S) is not *z*-related and is cumulative: It finds a value on either side of the mean at a certain probability level. A 50% probability comes with the mean featured in cell E1. Cell C5 uses this function: `=NORMINV(D5,Mean,SD)`.

Figure 5.11 shows another type of sampling distribution. The Student's *t*-distribution works with *t*-values instead of *z*-values. The curve on the right is cumulative (notice 100% on the y axis) but not symmetrical—in other words, *t*-values (unlike *z*-values) are always only positive. Another major difference from the normal distribution is that the *t*-distribution takes into account the size of the sample; it does this by using degrees of freedom—which is the sample size minus 1. When the sample size increases, the *t*-curve becomes steeper and high *t*-values become more unlikely. To put it differently, the x axis of *t*-values extends farther to the right when the samples become smaller; it stretches like chewing gum.

The table on the left in Figure 5.11 is based on a series of *t*-values (in column A) and a series of degrees of freedom (in row 1). The intersection uses `TDIST`, with cell B2 set to `=TDIST($A2,B$1,2)`. The last argument determines whether you work with one or two tails. The function in this example is two-tailed—which comes closest to the two tails in a symmetrical, cumulative normal distribution (5% two-tailed is the same as 2.5% one-tailed). You'll learn more on tails later. Notice that higher *t*-values are more likely in smaller samples. When the sample size comes closer to 30, the scale of *t*-values becomes shorter; above 30, the *t*-scale is almost identical to the *z*-scale. In other words, in samples over 30, you can use either distribution, but under 30, you should use the *t*-distribution because it acknowledges the relatively small size of the sample (which *z* plainly ignores).

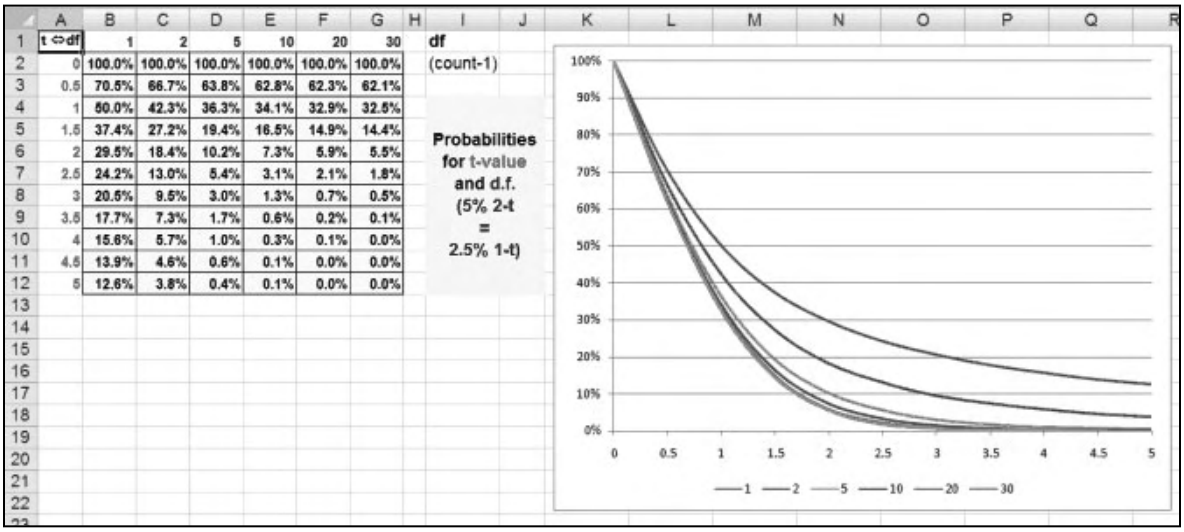
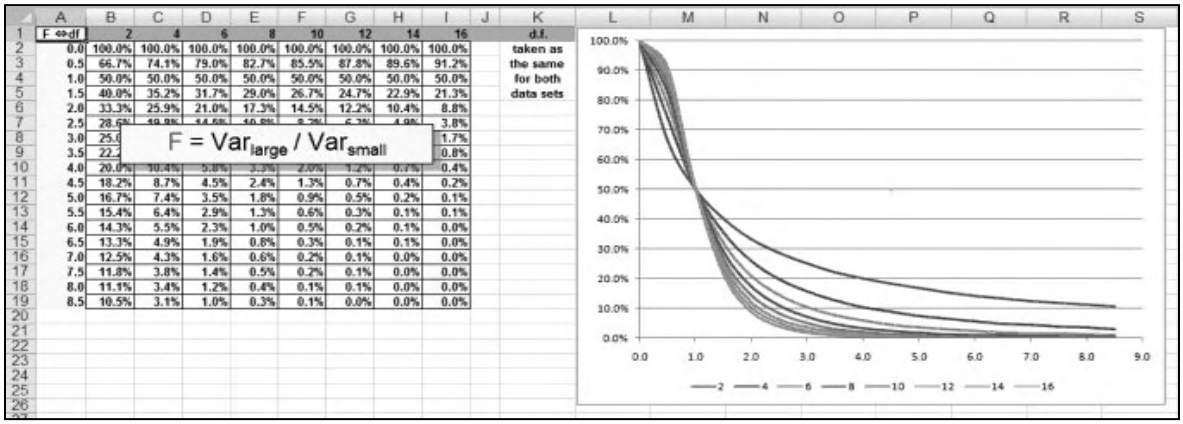


Figure: 5.11

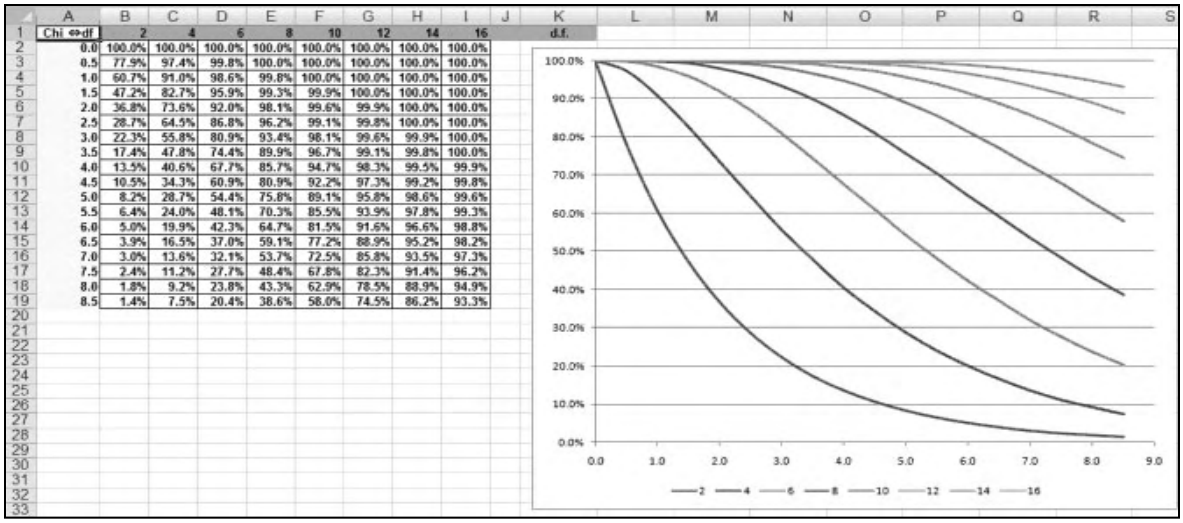
The *F*-distribution, which is plotted in Figure 5.12, is used for variances, not for means. The variance is the squared SD (so SD is the square root of variance). The *F*-value (or *F*-ratio) compares the variances of two data sets: the larger variance divided by the smaller variance. *F* works with two degrees of freedom—the one that comes with the larger variance first. The table's formula uses the same degrees of freedom for both, so the formula in cell B2 is `=FDIST($A2,B$1,B$1)`. As you can gather from the plot, `FDIST` is two-tailed and becomes steeper for larger samples.



F ⇔df	2	4	6	8	10	12	14	16
0.0	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
0.5	66.7%	74.1%	79.0%	82.7%	85.5%	87.8%	89.6%	91.2%
1.0	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%
1.5	40.0%	35.2%	31.7%	29.0%	26.7%	24.7%	22.9%	21.3%
2.0	33.3%	25.9%	21.0%	17.3%	14.5%	12.2%	10.4%	8.8%
2.5	28.6%	19.8%	14.6%	10.8%	8.2%	6.3%	4.9%	3.8%
3.0	25.0%	15.5%	10.2%	7.0%	4.8%	3.3%	2.3%	1.7%
3.5	22.2%	12.4%	7.4%	4.5%	2.8%	1.8%	1.2%	0.8%
4.0	20.0%	10.4%	5.8%	3.3%	2.0%	1.2%	0.7%	0.4%
4.5	18.2%	8.7%	4.5%	2.4%	1.3%	0.7%	0.4%	0.2%
5.0	16.7%	7.4%	3.5%	1.8%	0.9%	0.5%	0.2%	0.1%
5.5	15.4%	6.4%	2.9%	1.3%	0.6%	0.3%	0.1%	0.1%
6.0	14.3%	5.5%	2.3%	1.0%	0.5%	0.2%	0.1%	0.0%
6.5	13.3%	4.9%	1.9%	0.8%	0.3%	0.1%	0.1%	0.0%
7.0	12.5%	4.3%	1.6%	0.6%	0.2%	0.1%	0.0%	0.0%
7.5	11.8%	3.8%	1.4%	0.5%	0.2%	0.1%	0.0%	0.0%
8.0	11.1%	3.4%	1.2%	0.4%	0.1%	0.1%	0.0%	0.0%
8.5	10.5%	3.1%	1.0%	0.3%	0.1%	0.0%	0.0%	0.0%

Figure: 5.12

Figure 5.13 shows another type of distribution: the chi-distribution (chi-squared, or χ^2). This is the perfect distribution for frequencies. With this distribution, the degrees of freedom are related to the number of bins, or categories, you have used. The formula in cell B2 is `=CHIDIST($A2,B$1)`. Notice that more bins make higher chi-values more likely—the opposite of what you see for a *t*-distribution.

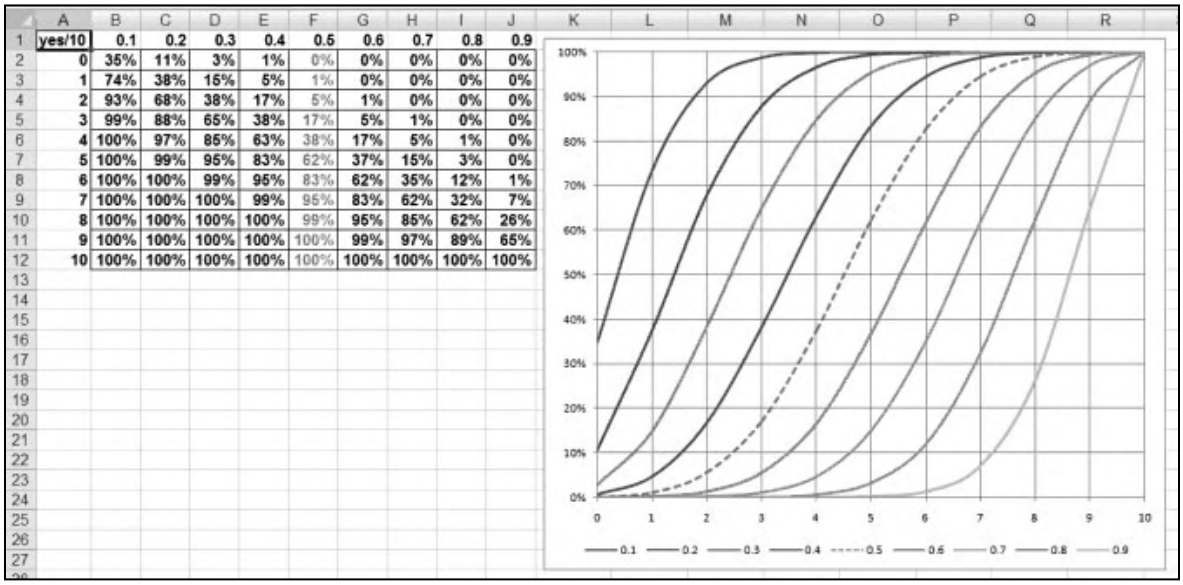


Chi ⇔df	2	4	6	8	10	12	14	16
0.0	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
0.5	77.9%	97.4%	99.8%	100.0%	100.0%	100.0%	100.0%	100.0%
1.0	60.7%	91.0%	98.6%	99.8%	100.0%	100.0%	100.0%	100.0%
1.5	47.2%	82.7%	95.9%	99.3%	99.9%	100.0%	100.0%	100.0%
2.0	36.8%	73.6%	92.0%	98.1%	99.6%	99.9%	100.0%	100.0%
2.5	28.7%	64.5%	86.8%	96.2%	99.1%	99.8%	100.0%	100.0%
3.0	22.3%	55.8%	80.9%	93.4%	98.1%	99.6%	99.9%	100.0%
3.5	17.4%	47.8%	74.4%	89.9%	96.7%	99.1%	99.8%	100.0%
4.0	13.5%	40.6%	67.7%	85.7%	94.7%	98.3%	99.5%	99.9%
4.5	10.5%	34.3%	60.9%	80.9%	92.2%	97.3%	99.2%	99.8%
5.0	8.2%	28.7%	54.4%	75.8%	89.1%	95.8%	98.6%	99.6%
5.5	6.4%	24.0%	48.1%	70.3%	85.5%	93.9%	97.8%	99.3%
6.0	5.0%	19.9%	42.3%	64.7%	81.5%	91.6%	96.6%	98.8%
6.5	3.9%	16.5%	37.0%	59.1%	77.2%	88.9%	95.2%	98.2%
7.0	3.0%	13.6%	32.1%	53.7%	72.5%	85.8%	93.5%	97.3%
7.5	2.4%	11.2%	27.7%	48.4%	67.8%	82.3%	91.4%	96.2%
8.0	1.8%	9.2%	23.8%	43.3%	62.9%	78.5%	88.9%	94.9%
8.5	1.4%	7.5%	20.4%	38.6%	58.0%	74.5%	86.2%	93.3%

Figure: 5.13

Figure 5.14 shows a fourth kind of distribution: the binomial distribution. Binomial distributions deal with *proportions*, such as the proportion of finding one of two. The classic example is head/tails, but anything dual or binary qualifies for a binomial distribution: yes/no, success/failure, sick/healthy, defect/non-defect, immunized/nonimmunized, male/female, increased/decreased, or true/false.

Let's consider an example: How often will you find up to 4 females in a group of 10 individuals if the proportion of females is 0.4? The answer is 63%, as cell E6 shows: `=BINOMDIST($A6,$A$12,E$1,1)`. `BINOMDIST` has four arguments: `number_s (4)`, `trials (10)`, `probability (0.4)`, `cumulative` (`1` or `TRUE`). The proportion of success (or yes) is *p*, whereas the proportion of failure (or no) is (1-*p*).



yes/10	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	35%	11%	3%	1%	0%	0%	0%	0%	0%
1	74%	38%	15%	5%	1%	0%	0%	0%	0%
2	93%	68%	38%	17%	5%	1%	0%	0%	0%
3	99%	88%	65%	38%	17%	5%	1%	0%	0%
4	100%	97%	85%	63%	38%	17%	5%	1%	0%
5	100%	99%	95%	83%	62%	37%	15%	3%	0%
6	100%	100%	99%	95%	83%	62%	35%	12%	1%
7	100%	100%	100%	99%	95%	83%	62%	32%	7%
8	100%	100%	100%	100%	99%	95%	85%	62%	26%
9	100%	100%	100%	100%	100%	99%	97%	89%	65%
10	100%	100%	100%	100%	100%	100%	100%	100%	100%

Figure: 5.14

> **Note**: This chapter uses the word probability only for what DIST functions return, and it uses the term *proportion* for the probability of "success."

The last argument is set to cumulative here. But because the table is cumulative, you cannot find what the chance is of finding exactly 4 females. That would be only a 25% chance (vs. 63%). Figure 5.15 presents the table and graph for a noncumulative binomial distribution. The noncumulative version has a curve in the middle that resembles a normal distribution. But

the other curves are more "squeezed" to either side of the graph. For "extremely squeezed" situations, there is an alternative: the Poisson distribution.

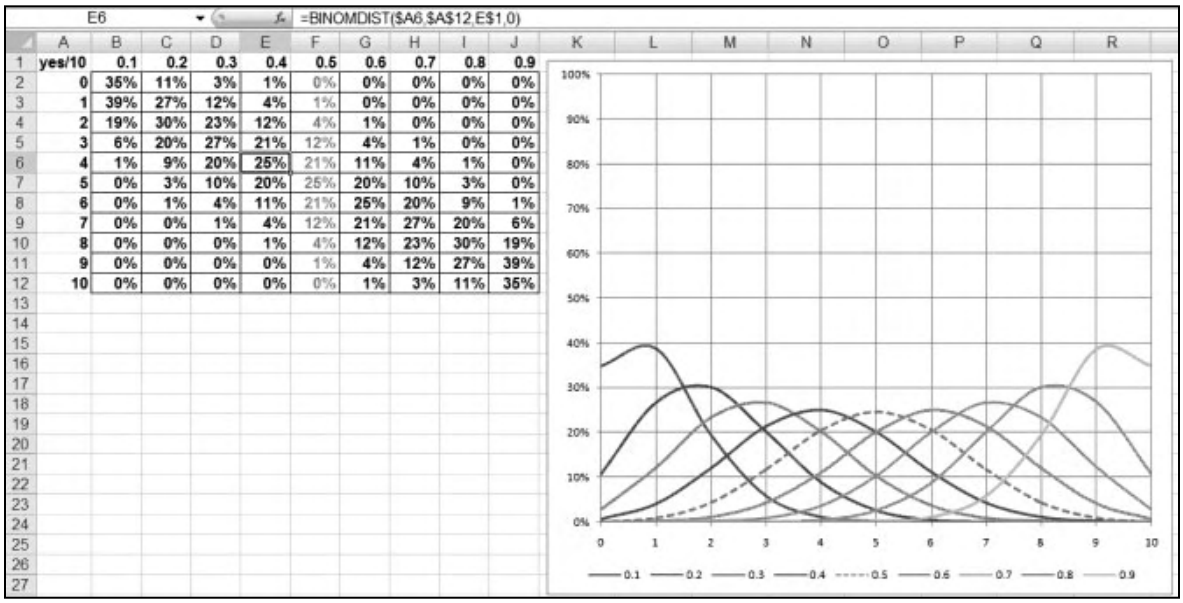


Figure: 5.15

The left part of Figure 5.16 shows calculations for a Poisson distribution—noncumulative in column B and cumulative in column C. This type of distribution is for very rare occasions: low proportions (*p*) among numerous cases (*N*). The general rule is *Np*<5. (Np is also called the *mean*.) Notice how the results for the Poisson (on the left) and binomial (on the right) distributions come pretty close to each other. Poisson distributions deal with the probability

	A	B	C	D	E	F	G	H	I
1	POISSON					BINOMIAL			
2									
3							2%		
4	0	30.1%	30.1%		(should be >50)	0	29.8%	29.8%	
5	1	36.1%	66.3%		size sample	1	36.4%	66.2%	
6	2	21.7%	87.9%		60	2	21.9%	88.1%	
7	3	8.7%	96.6%			3	8.7%	96.8%	
8	4	2.6%	99.2%		% affected	4	2.5%	99.3%	
9	5	0.6%	99.8%		2%	5	0.6%	99.9%	
10	6	0.1%	100.0%			6	0.1%	100.0%	
11	7	0.0%	100.0%		Np=lambda=mean	7	0.0%	100.0%	
12	8	0.0%	100.0%		1.2	8	0.0%	100.0%	
13	9	0.0%	100.0%		(should be <5)	9	0.0%	100.0%	
14	10	0.0%	100.0%			10	0.0%	100.0%	
15									

Figure: 5.16

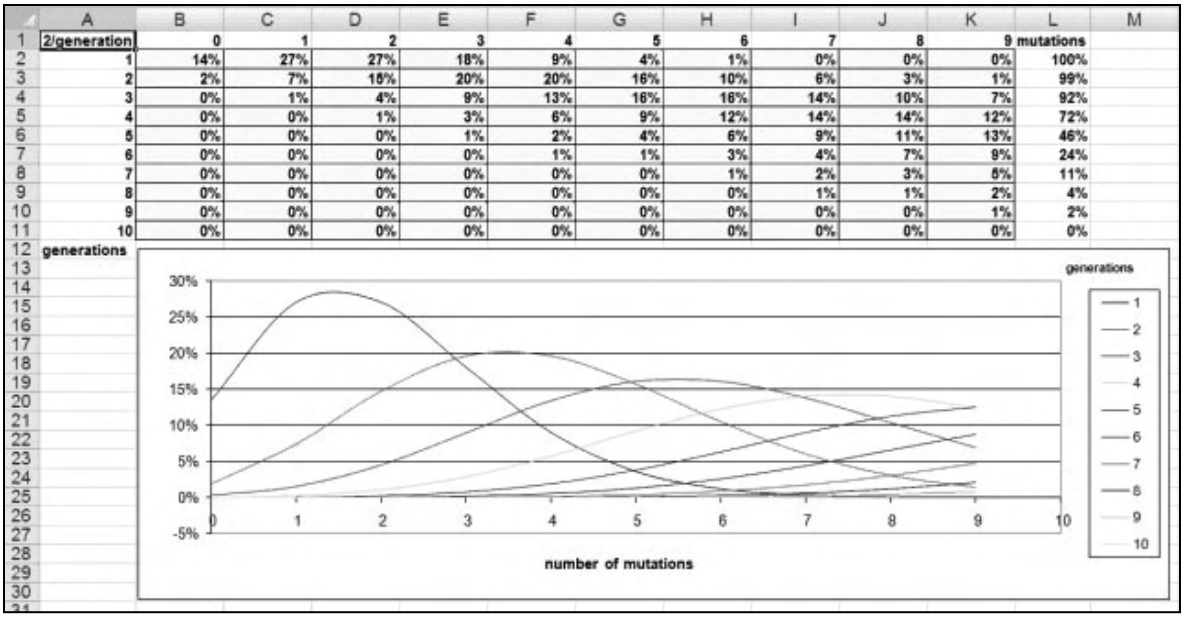
of rare events—under two conditions: their occurrences are independent of each other; their probability is proportional to the length of time interval. In science, Poisson distributions are used for situations such as the number of defects in production, the number of electrons emitted by a heated cathode, the number of mutations occurring in bacteria over time, and the number of atoms disintegrating per second in radioactive materials.

Figure 5.17 shows a typical example of a Poisson distribution. For each generation, you expect two mutations to occur in a large population of microorganisms. `POISSON` has three arguments: `x` (that is, the number of events), *mean*, and *cumulative*. In cell B2, the formula is `=POISSON(B$1,$A$1*$A2,0)`. It may look strange that the total of mutations (in column L) decreases after each generation, but don't forget that the table shows only up to nine mutations.

In the following chapters, you'll learn more about these five types of distributions:

- The normal distribution for means (but large samples)
- The *t*-distribution for means
- The chi-distribution for frequencies
- The binomial (and Poisson) distribution for proportions
- The *F*-distribution for variances

You can learn about even more distributions in specialized books on statistics.



2/generation	0	1	2	3	4	5	6	7	8	9	mutations
1	14%	27%	27%	18%	9%	4%	1%	0%	0%	0%	100%
2	2%	7%	18%	20%	20%	16%	10%	6%	3%	1%	99%
3	0%	1%	4%	9%	13%	16%	16%	14%	10%	7%	92%
4	0%	0%	1%	3%	6%	9%	12%	14%	14%	12%	72%
5	0%	0%	0%	1%	2%	4%	6%	9%	11%	13%	46%
6	0%	0%	0%	0%	1%	1%	3%	4%	7%	9%	24%
7	0%	0%	0%	0%	0%	0%	1%	2%	3%	5%	11%
8	0%	0%	0%	0%	0%	0%	0%	1%	1%	2%	4%
9	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%
10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figure: 5.17

* * *

Chapter 47

Simulating Distributions

It is possible to simulate the types of distributions discussed in Chapter 46 and then show them on an Excel spreadsheet. You might want to do so, for example, to study their characteristics or to compare an empirical distribution with an ideal one. You can make simulations from the option Data Analysis on the Data tab. If that option is not available, you must install the Analysis Toolpak first as an add-in (see Chapter 35). Through the Data Analysis option, you can activate the option Random Number Generation.

Figure 5.18 shows the simulation of a normal distribution; it is characterized by a specific mean and a specific standard deviation. The settings for this graph are specified in cell A1. If you leave the option Random Seed empty, Excel uses its own preset seeds. Otherwise, you need to enter an optional value from which to generate random numbers. You can reuse this value later to produce the same set of random numbers. Then, when you tell it to, Excel generates these 100 random numbers. When you calculate their frequencies in column D, you get a simulated normal distribution.

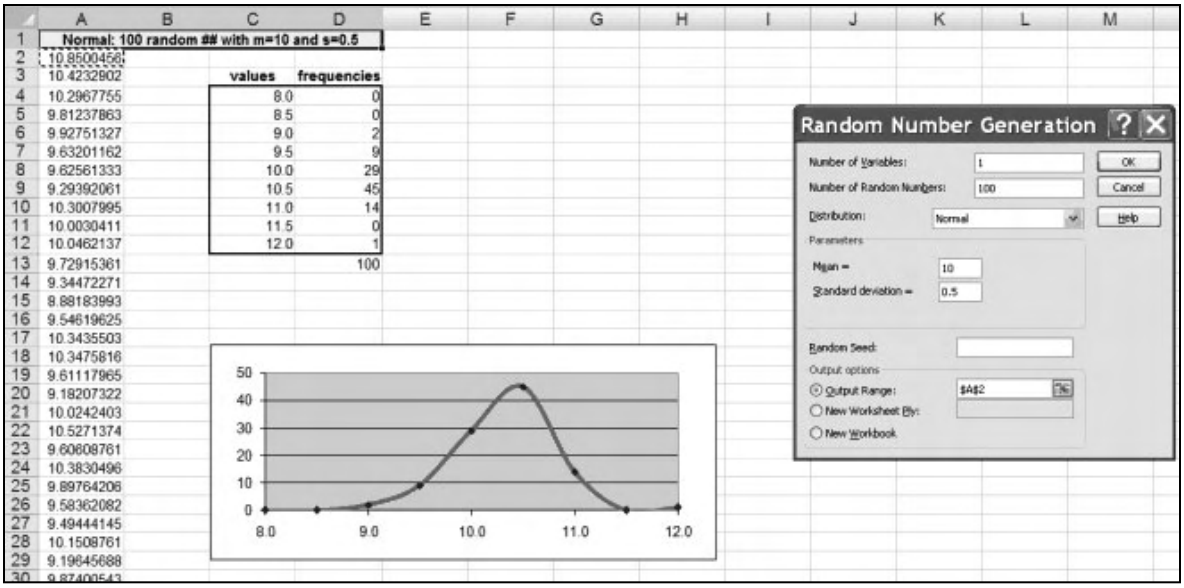
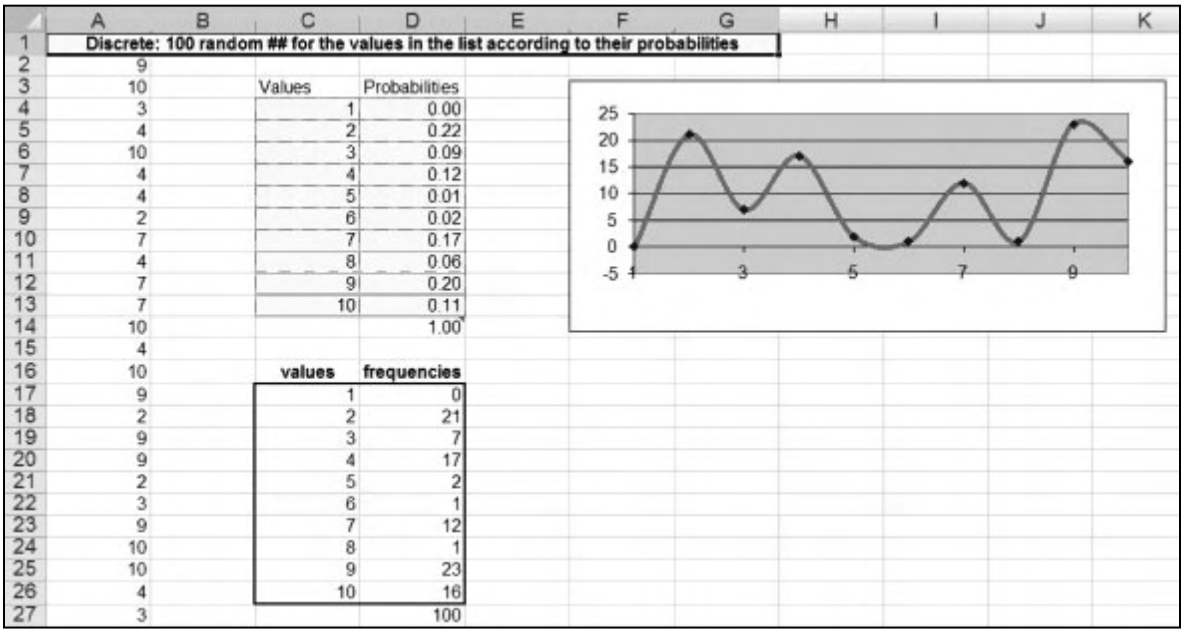


Figure: 5.18

Figure 5.19 simulates a situation in which certain values occur with specific probabilities (for example, multiple alleles of a gene, with each one having its own frequency in the population). This is a *discrete distribution*. It is characterized by a value and the associated probability. First you need to make a table of values and their probabilities. The table must contain two ranges: range C4:C13 for values and range D4:D13 for probabilities associated with the values in that row; the sum of the probabilities must be 1. The random number generator takes care of the rest, so you can create a frequency table and plot the results.



	A	B	C	D	E	F	G	H	I	J	K
1	Discrete: 100 random ## for the values in the list according to their probabilities										
2	9										
3	10		Values	Probabilities							
4	3		1	0.00							
5	4		2	0.22							
6	10		3	0.09							
7	4		4	0.12							
8	4		5	0.01							
9	2		6	0.02							
10	7		7	0.17							
11	4		8	0.06							
12	7		9	0.20							
13	7		10	0.11							
14	10			1.00							
15	4										
16	10		values	frequencies							
17	9		1	0							
18	2		2	21							
19	9		3	7							
20	9		4	17							
21	2		5	2							
22	3		6	1							
23	9		7	12							
24	10		8	1							
25	10		9	23							
26	4		10	16							
27	3			100							

Figure: 5.19

Figure 5.20 simulates a binomial distribution for 100 random numbers with p=0.5 and 10 trials. The curve happens to look like a normal distribution because p equals 0.5 in this case. The curve would get more squeezed to the left or right if p were closer to 0 or 1.

The random number generator offers a few more distributions, which are only briefly described here:

- **Poisson**: As mentioned in Chapter 46, this distribution is characterized by a value of mean occurrences. A Poisson distribution is often used to characterize the number of events that occur per unit of time (for example, the mean mutation rate in a population).
- **Uniform**: This distribution is characterized by lower and upper bounds. Variables are drawn with equal probability from all values in the range. A common application uses a uniform distribution in the range 0 to 1.
- **Bernoulli**: This distribution is characterized by a probability of success (p-value) on a

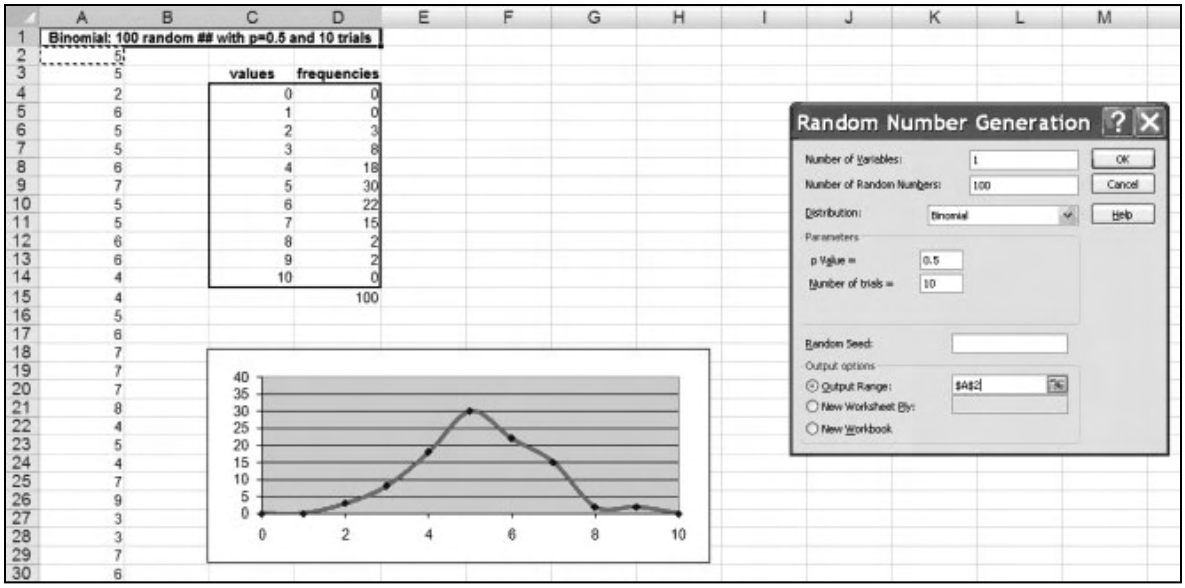


Figure: 5.20

given trial. A Bernoulli random variable has the value 0 or 1. For example, you can draw a uniform random variable in the range 0 to 1. If a Bernoulli random variable is less than or equal to the probability of success, the variable is assigned the value 1; otherwise, it is assigned the value 0.

- **Patterned**: This distribution is characterized by lower and upper bounds, a step, a repetition rate for values, and a repetition rate for the sequence.

Let's examine the concept of simulation in greater depth by using Figure 5.21:

1. Use the Analysis Toolpak in column A to create 100 random numbers according to a normal distribution. Assume that you want to simulate a sample that has a mean of 10 and a SD of 0.5. (You do not need to provide a random seed.) Excel creates the simulated output as shown.
2. Calculate the actual mean in cell D1 and the SD in cell D2. The randomness of 100 cases makes the mean and SD slightly deviate from the targets you set.
3. Use the function `SKEW` in cell G2. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values; negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.
4. Use the function `KURT` to characterize the relative peakedness or flatness (that is, kurtosis) of a distribution compared with the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

	A	B	C	D	E	F	G
1	9.8498839		Mean	9.98		skewness	0.0907014
2	9.3611584		StDev	0.54		kurtosis	-0.4757128
3	10.122129						
4	10.638237						
5	10.599175						
6	10.866567			X	Frequency	non-cum.	cumul.
7	8.9082062		X to 8.5	8.50	0	0.018	0.003
8	9.8829094		8.5 to 8.75	8.75	1	0.056	0.012
9	10.547511		8.75 to 9	9.00	2	0.144	0.036
10	9.4566497		9 to 9.25	9.25	7	0.298	0.089
11	9.6548979		9.25 to 9.5	9.50	9	0.497	0.188
12	9.1547838		9.5 to 9.75	9.75	16	0.672	0.336
13	9.0765446		9.75 to 10	10.00	18	0.734	0.515
14	9.5111853		10 to 10.25	10.25	14	0.649	0.691
15	9.6132465		10.25 to 10.5	10.50	16	0.464	0.831
16	8.9410344		10.5 to 10.75	10.75	9	0.269	0.922
17	9.7160376		10.75 to 11	11.00	5	0.126	0.970
18	9.7979762		11 to 11.25	11.25	3	0.048	0.990
19	10.067427		11.25 to 11.5	11.50	0	0.015	0.997
20	9.8172535				0		
21	9.8365047						
22	9.8148797						
23	10.671321						
24	9.9573578						
25	9.9069212						

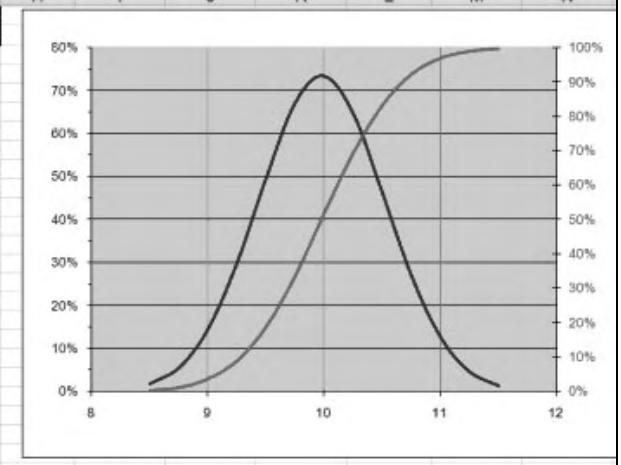


Figure: 5.21

5. Create the frequency bins in column D.
6. Calculate the frequency for each bin in column E by using the function `FREQUENCY` (see Chapter 17).
7. Use `NORMDIST` in columns F:G.
8. Plot the probabilities in an XY graph with a secondary axis.

Let's say that the research figures you are trying to simulate with a normal distribution have a precision of two decimal places. What should you do to the values in Figure 5.21? Figure 5.22 shows the answer:

1. Select cell A1.
2. In the Name box, type `A100`.
3. Press Shift+Enter to select A1:A100.
4. Open fx for the `ROUND` function.
5. Accept the function settings by pressing Ctrl+Enter.
6. To check whether rounding affects the main statistics, calculate the basic statistics again in cells D1:D2 and H1:H2.
7. Compare the basic statistics in cells D1:D2 and H1:H2 with the original values in cells E1:E2 and I1:I2. (The figure uses bold for rounded data and italics for nonrounded values.)

A1 =ROUND(SimulNorm!A1:A100,2)

	A	B	C	D	E	F	G	H	I
1	9.85		**mean**	**9.979800**	*9.979758*		**skewness**	**0.090123**	*0.090701*
2	9.36		**stdev**	**0.543158**	*0.542814*		**kurtosis**	**-0.476373**	*-0.475713*
3	10.12								
4	10.64								
5	10.6		**Display Descriptive Statistics**						
6	10.87		***Column1***						
7	8.91								
8	9.88		**Mean**	**9.9798**					
9	10.55		**Standard Error**	**0.05431576**					
10	9.46		**Median**	**9.96**					
11	9.65		**Mode**	**10.07**					
12	9.15		**Standard Deviation**	**0.54315758**					
13	9.08		**Sample Variance**	**0.29502016**					
14	9.51		**Kurtosis**	**-0.47637327**					
15	9.61		**Skewness**	**0.09012332**					
16	8.94		**Range**	**2.48**					
17	9.72		**Minimum**	**8.71**					
18	9.8		**Maximum**	**11.19**					
19	10.07		**Sum**	**997.98**					
20	9.82		**Count**	**100**					
21	9.84		**Confidence Level(95.0%)**	**0.10777425**					

Figure: 5.22

Instead of taking these steps, if you want a quick overview of your main statistics, you can use the Analysis Toolpak's Descriptive Statistics option. Be aware that these results do not update when values in column A change unless you apply the tool again.

* * *

Chapter 48

Sampling Techniques

The validity of research depends on good samples. Good samples must have the proper size in order to be representative. In addition, good samples need items that had an equal chance to be chosen. In biased samples, some items are more likely to be chosen than others—and that's not good research. Unfortunately, the mind's eye is not a good guide in selecting items for a sample. You need the unbiased verdict of a mathematical tool.

Scientists have available many sampling tools. The simplest one is the RAND function, as illustrated in Figure 5.23. Here's how you use it:

1. In column A, apply the function RAND.
2. Change the formula results into values: Copy and paste values.
3. Sort the values by random number.

	A	B	C	D	E	F	G
1	Biased Samples: Some more likely to be chosen than others						
2							
3	Top 5	Plate ID	Analyst		25%	Plate ID	Analyst
4	0.0532	8696p08a	ksm			8696p08a	ksm
5	0.1405	8696p08a	ksm		+	8696p08a	ksm
6	0.2648	8877p58a	gmv			8696p08a	gmv
7	0.2907	8696p08a	gmv			8696p08a	gmv
8	0.3521	8697p58b	tkm		+	8696p08b	bdo
9	0.4092	8877p58b	bdo			8696p08b	bdo
10	0.4140	8697p58b	tjk			8696p08b	ksm
11	0.4171	8697p58b	tjk		+	8696p08b	ksm
12	0.4379	8696p08b	ksm			8697p58b	tjk
13	0.4419	8877p58a	tjk			8697p58b	tjk
14	0.4722	8877p58a	tjk		+	8697p58b	tkm
15	0.5087	8877p58b	tkm			8697p58b	tkm
16	0.5522	8696p08b	bdo		+	8877p58a	gmv
17	0.5571	8696p08b	bdo			8877p58a	gmv
18	0.5762	8877p58b	tkm		+	8877p58a	tjk
19	0.6572	8877p58b	bdo			8877p58a	tjk
20	0.6765	8697p58b	tkm			8877p58b	tkm
21	0.7962	8696p08b	ksm			8877p58b	tkm
22	0.8132	8696p08a	gmv			8877p58b	bdo
23	0.9707	8877p58a	gmv			8877p58b	bdo
24							

Figure: 5.23

4. Select the first *n* items for your sample. (Later in this chapter, we'll discuss what the magnitude of *n* should be.)

If you want a certain percentage of cases (see cell E3), and you want even this percentage to be randomly fluctuating, you can use `RAND` again but this time nest it inside an `IF` function, as is done in cell E4: `=IF(RAND()<$E$3,"+","")`. Each time you press F9, you get a different selection of cases of varying sizes.

The Analysis Toolpak also has a simple sampling tool, as Figure 5.24 demonstrates. An ideal sampling technique puts a chosen item back into the population so it can be chosen again. The Analysis Toolpak uses this technique – in spite of that fact that this is not a common practice in Science. Figure 5.24 shows the use of the sampling tool from the Analysis Toolpak.

- Column C has the result of choosing Random for a sample of 20. Notice that there happens to be a duplicate set in the outcome—but you may not have this.
- Another draw of 20 items in column E happens to include another set of duplicates—at least in this example.
- Choosing the option Periodic for in column G eliminates duplicates if the original list doesn't have any. But the original list should not have some hidden pattern because Periodic will follow that pattern as well.

	A	B	C	D	E	F	G
1	1		Using the Sampling tool in the Analysis Toolpak				
2	92						
3	6		Col A: 100 unique numbers between 1 and 100				
4	63						
5	61		sample of 20		one more of 20		periodic (5)
6	38		64		19		61
7	93		27		49		75
8	39		34		100		76
9	12		72		64		56
10	75		74		99		52
11	45		20		36		57
12	33		16		43		21
13	68		94		93		65
14	69		35		29		32
15	76		81		42		83
16	51		1		72		74
17	95		2		77		17
18	43		40		76		30
19	90		19		70		50
20	56		32		87		13
21	88		40		5		37
22	73		81		44		18
23	27		38		9		41
24	94		22		77		86
25	52		36		89		48
26	81						

Figure: 5.24

A second important sampling rule says that the sample must have a proper size to be representative for the population. One of the considerations is that the larger the SD is in proportion to the mean, the larger the sample should be. Another consideration is that the closer you want to stay to the mean (called the *margin* in this case), the larger the sample should be. This technique and its formula are shown in Figure 5.25:

1. In cell D4, enter the formula `=$C4^2*($C$3/D$3)^2`. After you apply the formula, you find out what the minimum sample size should be under the given conditions.
2. In the section below the table, apply parts of the formula as follows:
   - **Cell D19**: =C19/C20
   - **Cell D21**: =C21/C20
   - **Cell D24**: =D19^2*(C22/D21)^2

These results say that you need a sample size of at least 33 to find a mean between 4.04 and 4.26.

	Z or t	margin/mean				
	2	0.01	0.02	0.05	0.1	0.25
SD/mean	0.1	384	96	15	4	1
	0.2	1537	384	61	15	2
	0.3	3457	864	138	35	6
	0.4	6147	1537	246	61	10
	0.5	9604	2401	384	96	15
	0.6	13830	3457	553	138	22
	0.7	18824	4706	753	188	30
	0.8	24586	6147	983	246	39
	0.9	31117	7779	1245	311	50
	1	38416	9604	1537	384	61
	2	153664	38416	6147	1537	246
	3	345744	86436	13830	3457	553

SD	0.3	0.0771084
Mean	4.2	
Margin	0.1	0.026506
Z or t	2	
Size		32.510731

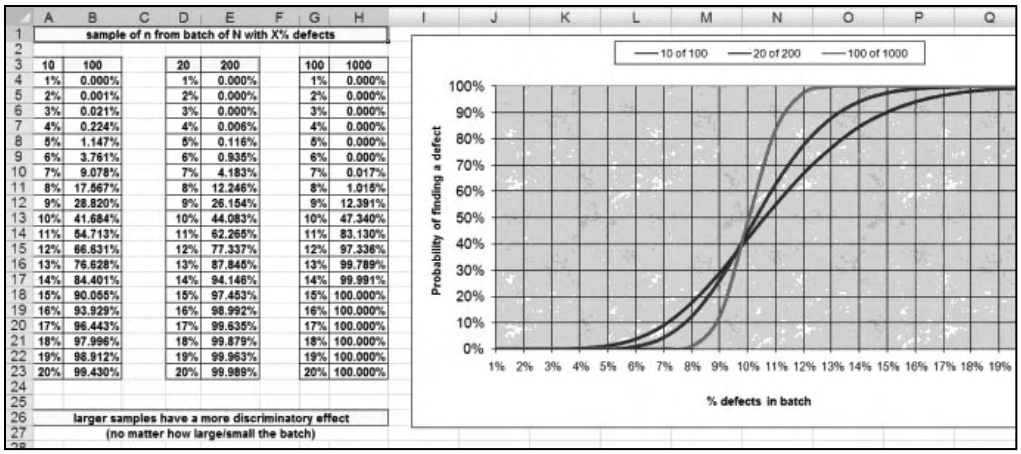
$$n = \left(\frac{1.96}{margin / \mu}\right)^2 \left(\frac{SD}{\mu}\right)^2$$

Figure: 5.25

Figure 5.26 tests how often you detect defects in samples of different sizes from batches with different sizes.

- Cell B4 holds the following formula: `=1-BINOMDIST($A$3,$B$3,A4,1)`. To determine the chances of finding defects, you use `1-BINOMDIST` because the chance of rejecting the batch is 1 minus the chance of accepting the batch.
- Now we can find out, for instance, what the chance would be of detecting a defect in a sample of 10 from a batch of 100 that has 7% defects. The answer is: a 9% chance.

- Based on the data, we can conclude that larger samples have a more discriminatory effect.
- Apparently the ratio of sample-size to batch-size doesn't matter at all. Only sample size matters.



sample of n from batch of N with X% defects						
10	100		20	200	100	1000
1%	0.000%		1%	0.000%	1%	0.000%
2%	0.001%		2%	0.000%	2%	0.000%
3%	0.021%		3%	0.000%	3%	0.000%
4%	0.224%		4%	0.006%	4%	0.000%
5%	1.147%		5%	0.116%	5%	0.000%
6%	3.761%		6%	0.935%	6%	0.000%
7%	9.078%		7%	4.183%	7%	0.017%
8%	17.567%		8%	12.246%	8%	1.015%
9%	28.820%		9%	26.154%	9%	12.391%
10%	41.684%		10%	44.083%	10%	47.340%
11%	54.713%		11%	62.265%	11%	83.130%
12%	66.631%		12%	77.337%	12%	97.336%
13%	76.628%		13%	87.845%	13%	99.789%
14%	84.401%		14%	94.146%	14%	99.991%
15%	90.055%		15%	97.453%	15%	100.000%
16%	93.929%		16%	98.992%	16%	100.000%
17%	96.443%		17%	99.635%	17%	100.000%
18%	97.996%		18%	99.879%	18%	100.000%
19%	98.912%		19%	99.963%	19%	100.000%
20%	99.430%		20%	99.989%	20%	100.000%

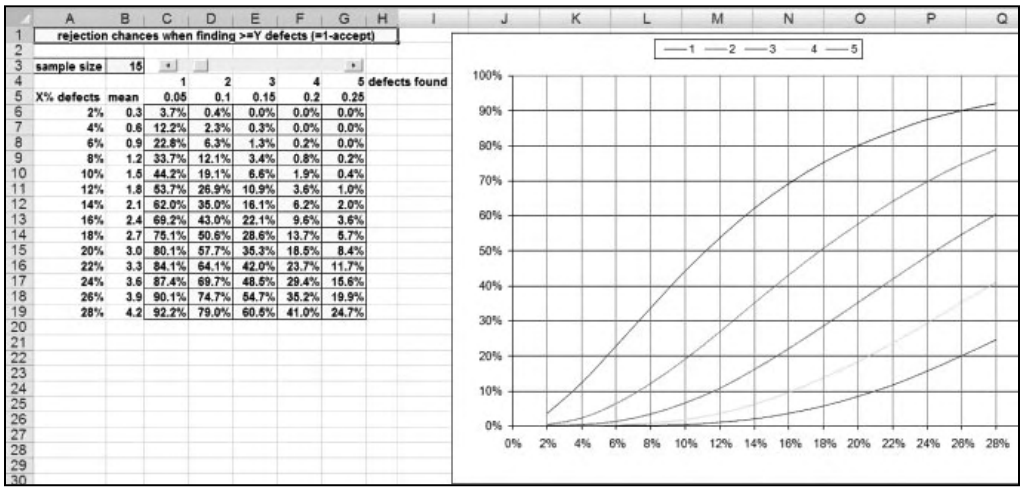
larger samples have a more discriminatory effect
(no matter how large/small the batch)

Figure: 5.26

Figure 5.27 shows all this in action. By moving the control located above the table, you can see the effect of increasing sample sizes:

- Because the size of the batch doesn't matter, this example uses only the size of the sample (in cell B3).
- Because you use the POISSON function in this example, you also need the mean, so you enter `=Np` in column B.
- The formula in cell C6 is `=1-POISSON(C$4,$B6,1)`.
- This example calculates the chances of rejecting a batch when greater than or equal to y defects are found.

In the chapters to come, you will study this phenomenon in more depth.



rejection chances when finding >=Y defects (=1-accept)						
sample size	15					
		1	2	3	4	5 defects found
X% defects	mean	0.05	0.1	0.15	0.2	0.25
2%	0.3	3.7%	0.4%	0.0%	0.0%	0.0%
4%	0.6	12.2%	2.3%	0.3%	0.0%	0.0%
6%	0.9	22.8%	6.3%	1.3%	0.2%	0.0%
8%	1.2	33.7%	12.1%	3.4%	0.8%	0.2%
10%	1.5	44.2%	19.1%	6.6%	1.9%	0.4%
12%	1.8	53.7%	26.9%	10.9%	3.6%	1.0%
14%	2.1	62.0%	35.0%	16.1%	6.2%	2.0%
16%	2.4	69.2%	43.0%	22.1%	9.6%	3.6%
18%	2.7	75.1%	50.6%	28.6%	13.7%	5.7%
20%	3.0	80.1%	57.7%	35.3%	18.5%	8.4%
22%	3.3	84.1%	64.1%	42.0%	23.7%	11.7%
24%	3.6	87.4%	69.7%	48.5%	29.4%	15.6%
26%	3.9	90.1%	74.7%	54.7%	35.2%	19.9%
28%	4.2	92.2%	79.0%	60.5%	41.0%	24.7%

Figure: 5.27

* * *

Chapter 49

Each Test Has Its Own Conditions

When should you use which sampling distribution? We have discussed some general considerations—such as using the binomial distribution for proportions. However, most distributions are subject to some extra conditions. The most frequent condition is that a sample—and thus the population it is taken from—must be normally distributed. And this is not always the case, as you know.

Figure 5.28 shows with a few plots what can go "wrong" with a bell-shaped distribution:

- The first curve is actually composed of two subsets—which could be a subset of males and a subset of females—each of which has its own bell shape. The means of the samples taken from this population would vary equally to either side of the mean of the population—that's why the composite curve is still normally distributed.
- If the means of the subsets were farther apart, the curve would become bimodal. (A *mode* is a peak in a curve.) This is the case with the two curves on the right.
- IF the SDs of the subsets were different, that would definitely affect the skewness of the curve. The lower-left curve is an example of this situation. In other words, the means of the samples taken from this population would not vary equally to either side of the mean of the population.

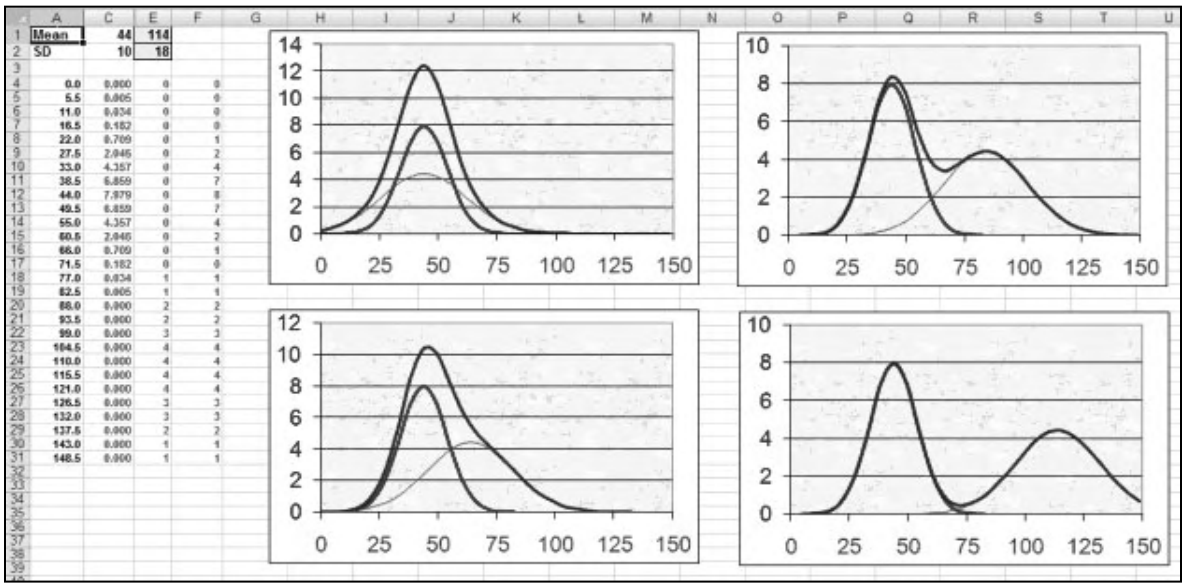


Figure: 5.28

For “abnormal” situations like these, you could not use the normal sampling distribution to test the sample distribution for issues such as estimating, significance, and confidence. What alternatives do you have? Figure 5.29 provides an overview of some conditions.

- Cases 2 through 4 are heavily skewed, whereas case 5 has two very different variances, or SDs.
- The *z*-test and the *t*-test (based on their respective sampling distributions) can be used only when the samples or populations are not too skewed and do not vary too much in their variances.
- An added condition for *z* versus *t* is that *z* can only be used for samples greater than 30, as mentioned earlier.
- The *F*-test requires a normally distributed sample and/or population.
- The chi-test does not have these extra conditions and can be applied even in the absence of a normally distributed sample and/or population, but it has another requirement, which is discussed in Chapter 54.
- The *p*-distribution applies only to binomial situations. But sometimes the binomial sampling distribution can be a solution that works where the previous distributions have failed. See Chapter 53.

You end up with the following series of alternatives: When *z* is not possible because the sample is too small, *t* may be feasible. When *t* is out of the question because the variances are too different, you can try *F*. If all the previous ones are unacceptable because the sample and/or population is not normally distributed, you may have to reshuffle the data and try *p* and/or chi. You will learn more about these issues in the chapters to come.

Now that you know the basics of some sampling distributions and their limitations, you should be ready to estimate margins based on samples (see Chapters 50 and 51) and test for significance based on samples (see Chapters 52 to 54).

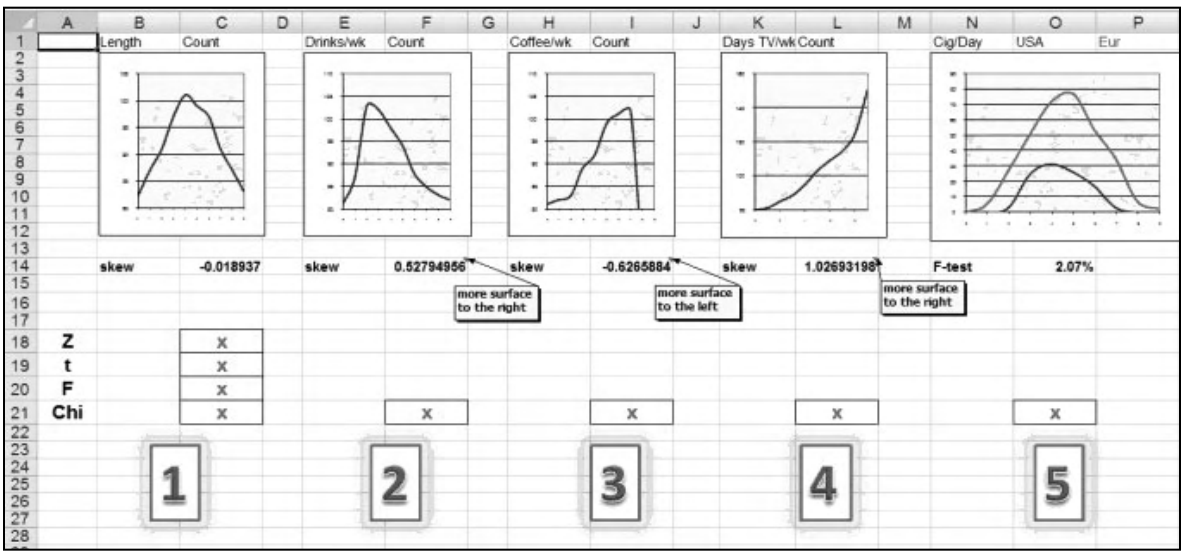


Figure: 5.29

* * *

Chapter 50

Estimating Means

Finding a specific mean in a sample does not imply that other samples taken from the same population will have the same mean—nor will the population. Remember the slogan "Results may vary"? In other words, the mean found in the sample stands for a much wider range of means. A scientist must estimate the margins around the mean found in a specific sample. These margins are often called the "margins of error."

Say that in a normally distributed sample, you have measured a mean of 4.15 (pH, °C, ng/ml, mol, or whatever). Now you need to estimate what range of means you should/could expect in other samples taken from that same population. Let's go for 95% of the normal curve, so the $mean_{exp}$ ranges from a minimum value (at 1.96 * SE to the left of $mean_{obs}$) to a maximum value (at 1.96 * SE to the right of $mean_{obs}$). The distance to the left and the right of the observed mean is called a *confidence limit*, *confidence margin*, or *margin of error.* Figure 5.30 depicts this scenario for a normal distribution based on a *z*-test.

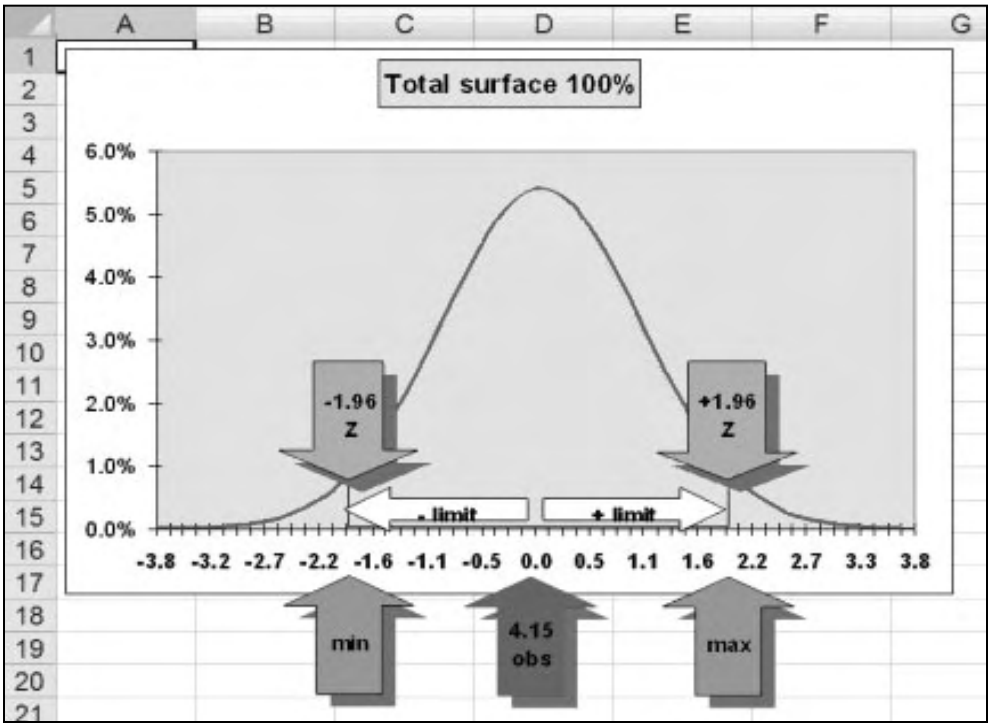


Figure: 5.30

The 95% area under the curve that you usually choose represents the area you want to be covered. So you say that you have 95% confidence that the mean will be found inside the stated range. The area under the curve to the left of the minimum value amounts to 2.5% (or half of the area outside the 95% confidence range). The area under the curve to the right of the maximum value amounts also to 2.5% (or half of the area outside the 95% confidence range). Because you are checking the lower range as well as the upper range, you are dealing with a two-tailed situation:

- The cumulative probability of -1.96 *z* (at the left tail) is 2.5%.
- The probability of +1.96 *z* (at the right tail) is 97.5% because you are dealing with cumulative probabilities.

Be aware, though, that the situation is a bit different for the *t*-distribution because that distribution is not symmetrical (so there are no negative *t*-values). You therefore need to treat *t*-values differently:

- In a two-tailed test, you should choose 5%.
- In a one-tailed test, you should choose 2.5%.

Figure 5.31 shows an example of estimating the margin (of error) around the mean. The first situation is as follows: In a sample of 35 cases, you have measured a mean of 4.15 and a SD of 0.32. But, again, "results may vary." In other words, you shouldn't come up with a single value but with a range of values. What are the lowest mean and the highest mean you could encounter with 95% confidence? In order to find out, you need the *z*-value that comes with 2.5%, plus the SE of this sample mean. Here's how you find it:

1. In cell I8, calculate the *z*-value that comes with 2.5%: `=NORMSINV(H8)`.
2. In cell J8, find the standard error: `=C8/SQRT(D8)`.
3. In cell K8, calculate the margin (of error) or confidence limit on either side of the mean (using ABS): `=ABS(I8)*J8`.
4. In cell L8, use the formula `=B8-K8` to find the lowest mean you expect to find (with 95% confidence) in another sample of the same size and the same population. And use the formula `=B8+K8` in cell M8 to find the highest mean.
5. If you find this margin too wide, either reduce the level of confidence (which is not wise to do!) or increase the size of the sample (which costs you or your organization time and money!). You can determine the best option by using GoalSeek (see Chapter 41).

If you find these steps too involved, you can use a shortcut function called `CONFIDENCE`. However, this function works a bit differently:

- It uses one tail (so 5% instead of 2.5%).

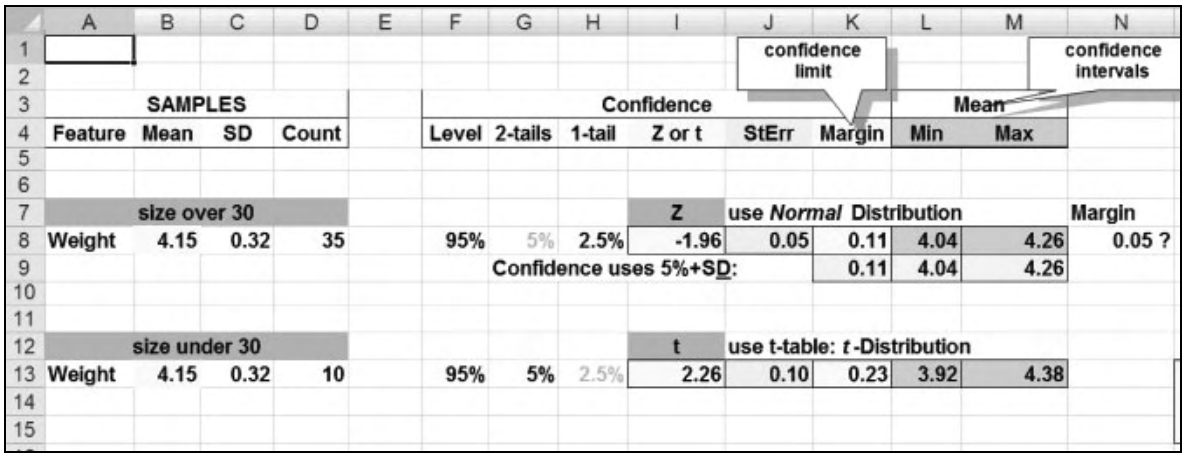


Figure: 5.31

- It calls for the SD instead of the SE; it does the calculation for you!
- The formula in cell K9 would be `=CONFIDENCE(G8,C8,D8)`.

Next let's tackle the same sample again, but this time based on a smaller sample size (10 vs. 35). Because of the sample size, you cannot use *z*-values but only *t*-values. Because there is no `CONFIDENCE` function for the *t*-distribution, you need to do all the statistical work manually, as follows:

1. Find the *t*-value in cell I13: `=TINV(G13,D13-1)`. Because `TINV` uses two tails, you must look for the *t*-value that comes with the 5% area outside the confidence area of 95%.
2. Use the same formulas in J13:M13 as in the previous scenario (J8:M8).

Obviously, the margins are quite different for the first sample and the second sample because *t*-distributions are more cautious when the sample size becomes smaller. Smaller samples are more susceptible to random effects. The *t*-distribution would give you results similar to the *z*-distribution if you increased the sample size to 35, as in the first sample.

It is common policy to use a 5% error range, as explained in Chapter 52. However, you or your organization may decide for some reason to change this number someday or in certain cases. If so, you have to change all your formulas and/or cell entries. So it might be prudent to use a name that represents the current standard of 5%. To do this, follow these steps:

1. In the Formulas tab, select Name Manager.
2. Click the New button and type its new Name: `signif` (or whatever name you want).
3. Set Refers To to `=0.05`.
4. Instead of typing 5% in your formulas, use the name `signif` (or whatever name you chose) from now on. When you change what the name refers to, all cells using that name will automatically apply the new value.

Chapter 37 discusses a regression graph like the one depicted in Figure 5.32. This chapter doesn't go through all the details again, but now that you've learned more about statistics, it might be easier to understand what's happening in this graph. The regression line is based on sample information, but samples represent a range of possible values that fluctuate per sample. `RSQ` may be strong in this particular sample, but it is an altogether different issue whether this sample is representative for future samples of the same population. Therefore, you also want to find a margin of error based on 95% confidence. Because the `CONFIDENCE` function cannot achieve this, you need to do manual work again—which is even more involved than what you did earlier in this chapter. After finishing all necessary calculations, you end up with a 95% confidence interval for each x (in finding a certain y mean) plus a 95% prediction interval for each x (in finding a certain y value).

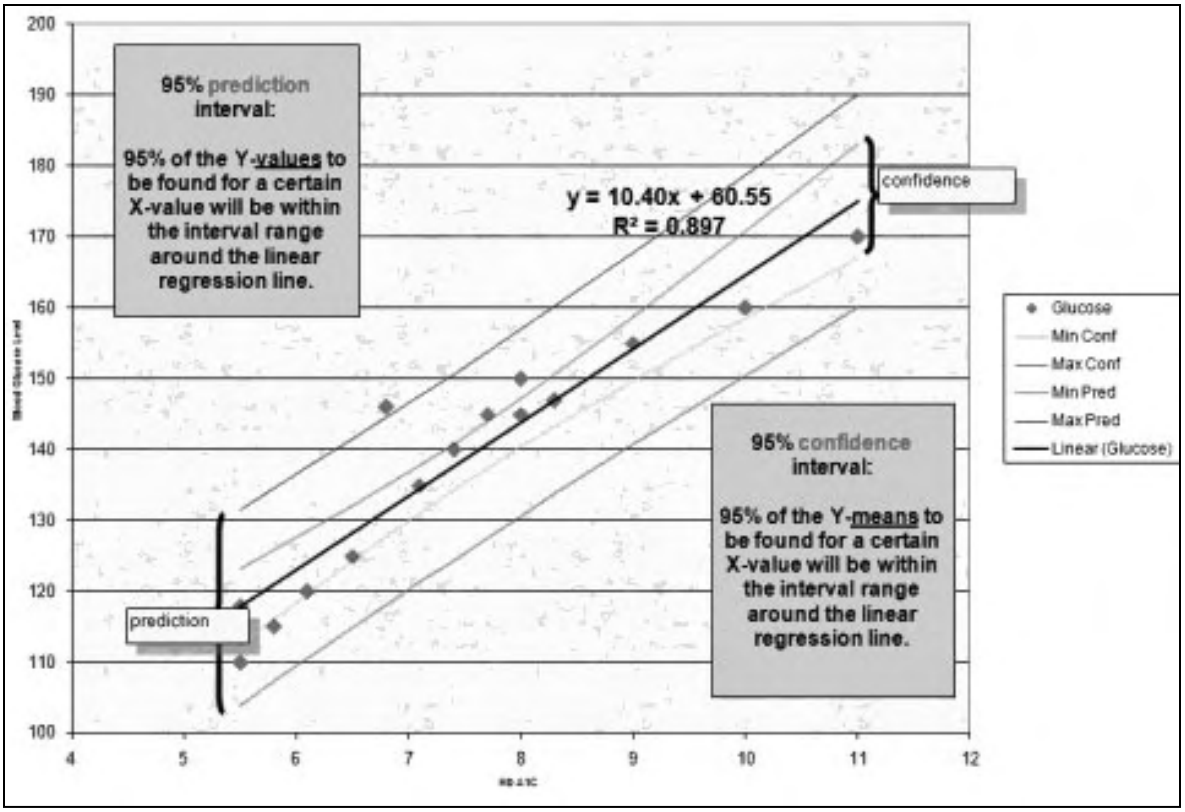


Figure: 5.32

* * *

Chapter 51

Estimating Proportions

When finding a specific proportion in a sample, you realize now that you should expect a wider range when taking other samples from the same population. Although those proportions don't completely follow a normal distribution pattern, they come close.

Before going into confidence limits and intervals for proportions, you should review the basics of proportions by using Figure 5.33. `BINOMDIST` has the following syntax: `BINOMDIST(Yes,Trials,pYes,F/T)`. Here's what it means

- If you find 24 defects (`Yes`) in a sample of 100 items, the proportion of "yes," or "success," is 0.24, or 24% (cell E2).
- If you take 10 trials from this batch (with 24% defects), what is the chance of finding 10 defects? `BINOMDIST` tells the following:
  - The chance of finding exactly 10 defects is almost 0%:
    `=BINOMDIST($H2,$G2,$E2,0)`
  - The chance of finding up to 10 defects is 100% - cumulative:
    `=BINOMDIST($H2,$G2,$E2,1)`
- You can do something similar for the other cases.

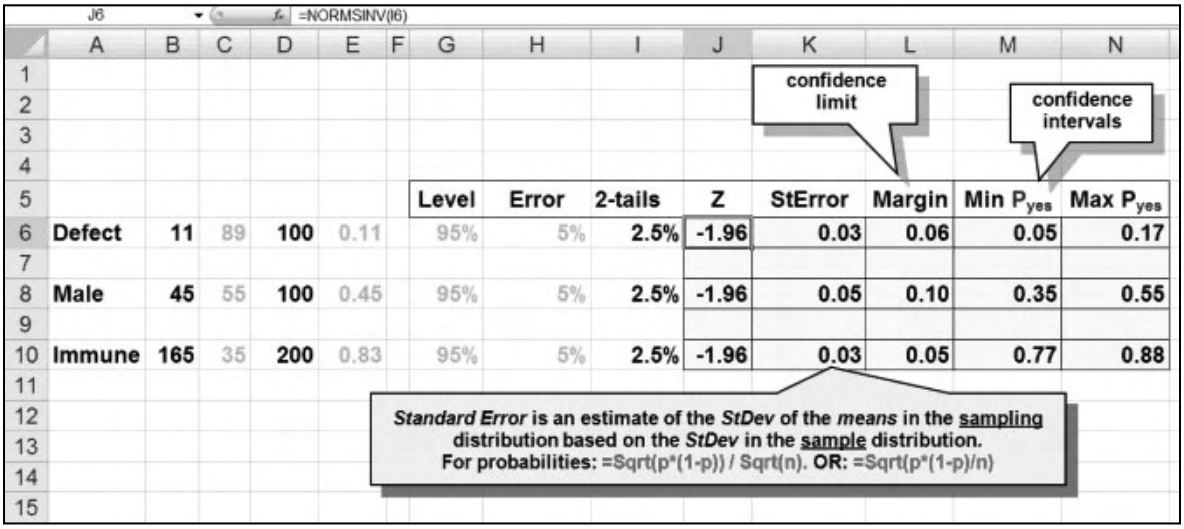
	A	B	C	D	E	F	G	H	I	J
1	Trait	Yes	No	Total	p_{yes}		Trials	Find Yes	Exactly	Up to
2	Defect	24	76	100	0.24		10	10	0.00%	100.00%
3										
4	Male	45	55	100	0.45		10	6	15.96%	89.80%
5										
6	Immune	165	35	200	0.83		10	3	0.03%	0.04%
7										

Figure: 5.33

Now you are ready to get back to estimating 95% confidence intervals in Figure 5.34. Be aware that SE is calculated differently for binomial distributions: `=SQRT(p*(1-p)/n)`. Here's how you estimate the 95% confidence intervals:

1. Determine z for a 2.5% two-tailed error margin in cell J6: `=NORMSINV(I6)`.
2. Determine the standard error in cell K6: `=SQRT(E6*(1-E6)/D6)`.
3. Determine the margin of error, *z**SE, using `=ABS(J6*K6)`.
4. Determine the confidence limits and confidence intervals like you did before.

In the last case scenario (row 10), you have 95% confidence of finding between 77% and 88% immunized cases in samples of size 200.



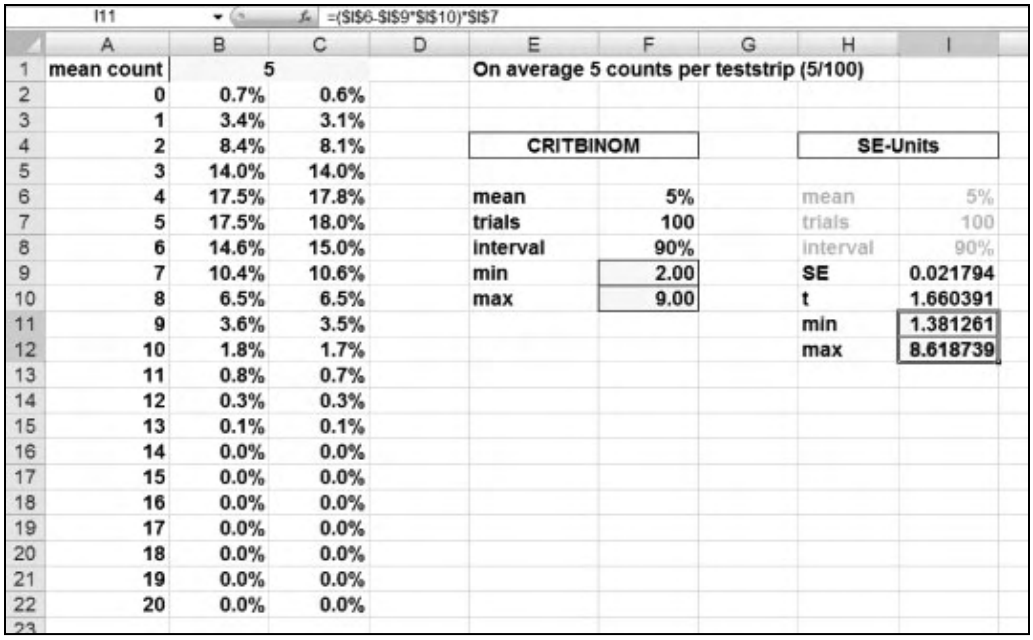
						Level	Error	2-tails	Z	StError	Margin	Min P_{yes}	Max P_{yes}
6	Defect	11	89	100	0.11	95%	5%	2.5%	-1.96	0.03	0.06	0.05	0.17
8	Male	45	55	100	0.45	95%	5%	2.5%	-1.96	0.05	0.10	0.35	0.55
10	Immune	165	35	200	0.83	95%	5%	2.5%	-1.96	0.03	0.05	0.77	0.88

Figure: 5.34

Figure 5.35 demonstrates another binomial function, `CRITBINOM`, that can find a border value (instead of a percentage) for, let's say, 5% and 95% (making for a confidence interval of 90%). If you were told that certain test plates usually have 5 colonies per plate, you could find out which minimum and maximum count to expect:

1. In cell F9, enter `=CRITBINOM($I$7,$I$6,(1-$I$8)/2)`.
2. In cell F10, enter `=CRITBINOM($I$7,$I$6,$I$8+(1-$I$8)/2)`.

The results tell you that you can expect between 2 and 9 colonies in 90 out of 100 plates; the other 10 may be outside this range. Had you done this as you did before (with SE and *z* or *t*), you would have gotten slightly different results in I11:I12. Part of the explanation is that the left calculation works with (rounded) values, whereas the right calculation works with percentages.



I11 =($I$6-$I$9*$I$10)*$I$7

	A	B	C	D	E	F	G	H	I
1	mean count	5			On average 5 counts per teststrip (5/100)				
2	0	0.7%	0.6%						
3	1	3.4%	3.1%						
4	2	8.4%	8.1%		CRITBINOM			SE-Units	
5	3	14.0%	14.0%						
6	4	17.5%	17.8%		mean	5%		mean	5%
7	5	17.5%	18.0%		trials	100		trials	100
8	6	14.6%	15.0%		interval	90%		interval	90%
9	7	10.4%	10.6%		min	2.00		SE	0.021794
10	8	6.5%	6.5%		max	9.00		t	1.660391
11	9	3.6%	3.5%					min	1.381261
12	10	1.8%	1.7%					max	8.618739
13	11	0.8%	0.7%						
14	12	0.3%	0.3%						
15	13	0.1%	0.1%						
16	14	0.0%	0.0%						
17	15	0.0%	0.0%						
18	16	0.0%	0.0%						
19	17	0.0%	0.0%						
20	18	0.0%	0.0%						
21	19	0.0%	0.0%						
22	20	0.0%	0.0%						
23									

Figure: 5.36

Figure 5.36 shows very clearly that confidence margins for binomial distributions depend heavily on sample sizes:

- Cell D2: `=TINV(signif,A2-1)`. Cell E2: `=SQRT(B2*(1-B2))/SQRT(A2)`. Cell F2: `=ABS(E2*D2)`. Cell G2: `=B2-F2`. Cell H2: `=B2+F2`.
- You read the graph or table this way: A proportion of .7, for instance, in a sample of 50 has a 95% range between .55 and .85
- You could have obtained similar results by using `CRITBINOM`. But for small sizes, the curves would show some rounding effects.

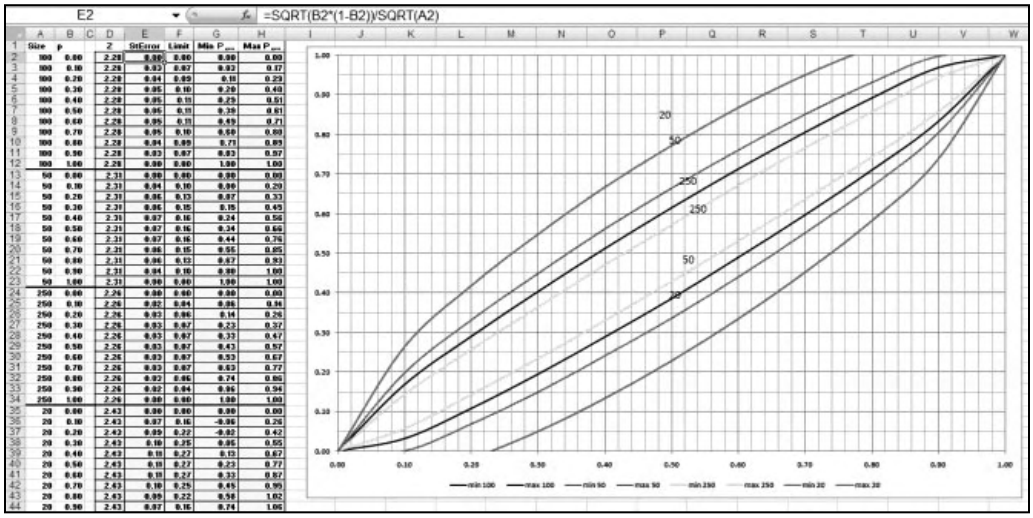


Figure: 5.36

* * *

Chapter 52

Significant Means

When you find a mean outside the range or margin you had expected for samples from a specific population, you may wonder whether that mean is really coming from the same population. This is considered to be an issue of testing for significance—which is the topic of this chapter.

What does testing for significance entail? Say that you had expected a mean of 33 but in fact observed or measured a mean of 35.3. Is this difference significant? In other words, is this difference likely to be the mere result of random sampling? Or is the actual difference (measured in SE units as *z*- or *t*-values) beyond the critical difference that you take as a borderline case for being random? If the latter is the case, you would consider this sample to be from a different population, which usually means that some specific treatment affected the sample and had a significant impact.

When dealing with testing for significance, the term *hypothesis* comes into play:

- The null hypothesis states that the difference between observed and expected is the outcome of randomness: "Results may vary."
- The alternative hypothesis states that the difference is caused by a real difference in the underlying sample (caused by the factor under investigation).
- There are two possible outcomes in testing for significance:
  - When the actual *z*- or *t*-value is less than the critical *z*- or *t*-value, the null hypothesis is accepted. Conclusion: The difference is (very) likely a matter of randomness.
  - When the actual *z*- or *t*-value is greater than the critical *z*- or *t*-value, the alternative hypothesis is accepted. Conclusion: The difference is most likely caused by the factor under investigation.

Where are the critical *z*- and *t*-values located? Usually you place them at the border(s) where 95% of the potential means are covered. So 5% is left out in the critical area—which is 2.5% to the left and 2.5% to the right of the margin of error on a symmetrical curve. Figure 5.37 shows the critical values:

- At 2.5% and 97.5% for *z*-values if you test for both tails
- At 5% if you test for *t*-values in a two-tailed test
- At 2.5% if you test for t-values in a one-tailed test

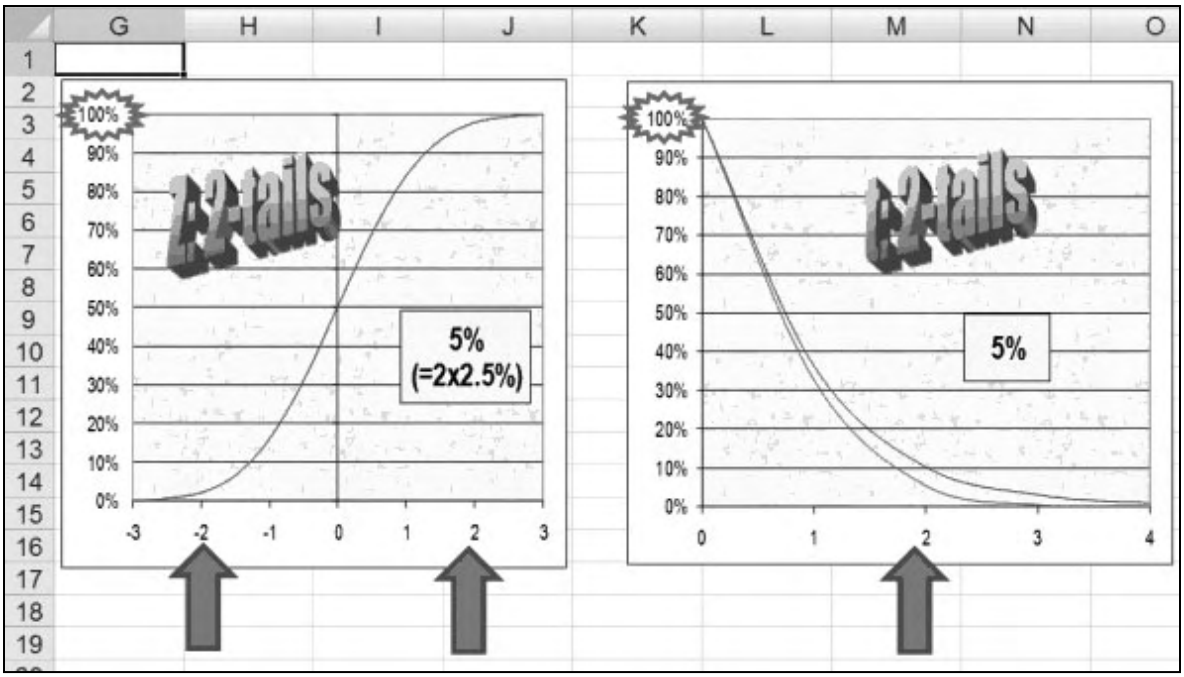


Figure: 5.37

The area outside the 95% range is called the *significance* area; it is beyond the critical *z*- or *t*-values. Why this magic 95% versus 5%? It is basically a strategic decision. You say it is so unlikely that values found in the significance area can be attributed to mere randomness. But what is "so unlikely"? Why decide on 5% and not on 10% or 2.5%?

Let's find out why by examining Figure 5.38. In a significance range of 5%, you accept the alternative hypothesis, but you take a 5% chance of rejecting a true null hypothesis; in other words, the difference between observed and expected could still be random. This value of 5% is also called an *alpha error*. The *n*-shaped curve shows the (alpha) chances of accepting a true hypothesis for a 5% significance limit. Notice that the chances of accepting a true null hypothesis dramatically decline when you get farther away from the center.

Whereas *alpha* designates the error chance of rejecting a true null hypothesis, there is also a beta chance—alpha's mirror image. *Beta* designates the error chance of accepting a false null hypothesis. The u-shaped curve plots beta errors for 5% alpha error settings. So you have a dilemma here: A smaller risk of rejecting a true null hypothesis results in a larger risk of not recognizing the null hypothesis as false. In other words, you have to make a compromise: A small value of alpha is certainly desirable, but making it too small may result in a beta so large that you seldom recognize a false null hypothesis. That's why most scientists have settled on a 5% (alpha) error chance; it is the point where the two curves cross each other.

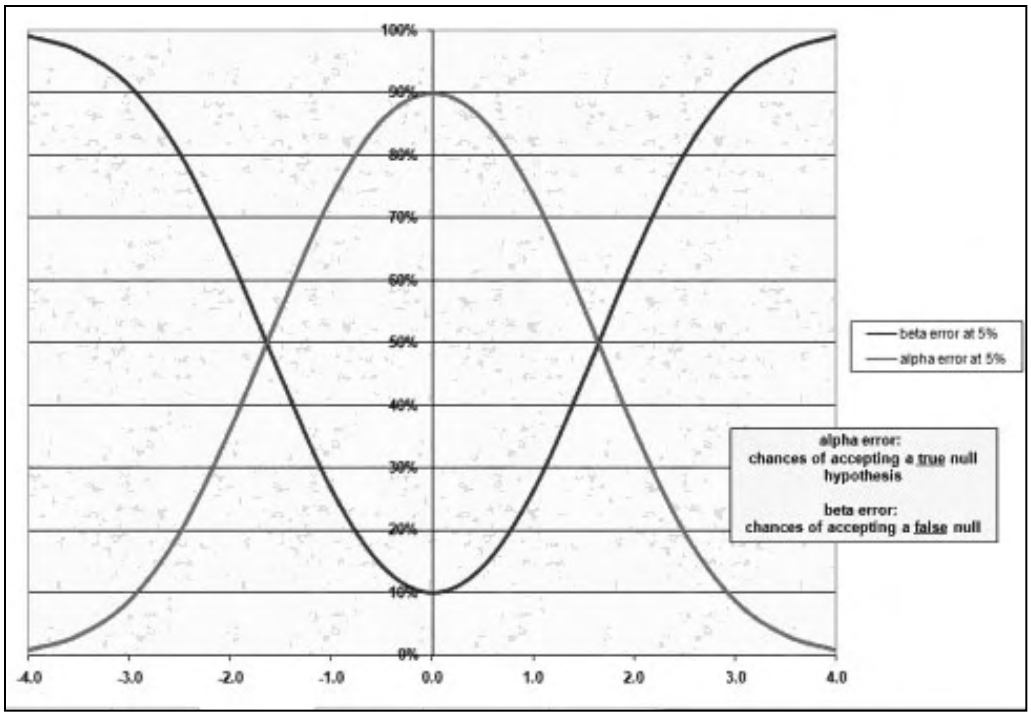


Figure: 5.38

In Figure 5.39, you want to test whether the mean of a set of sample measurements (in column A) is significantly different from the mean you had expected (`33` in cell E5). The null hypothesis states that the difference is a random outcome. Here's what you do:

1. Before applying a *z*- or *t*-test, check whether the distribution is more or less normally distributed by using `SKEW` in cell E2. This is one of the conditions for using a *z*- or *t*-test (see Chapter 49).
2. Calculate the mean, SD, and count of all measurements in D5:D7.
3. Assign a name for the 5% significance level and use it in cell D8 (see Chapter 50).
4. Calculate the SE in cell D9.
5. Calculate the actual *t*-value in D10: `=ABS(D5-E5)/D9`. *t*-values are SE units, so divide the actual distance by the SE.
6. Find the critical *t*-value in D11: `=TINV(D8,D7-1)`. Don't test the SE units against a normal distribution, because *z*-values would be too optimistic, given a sample size below 30.
7. If you rather want to know what the probability is of finding a mean that is 1.46 SE units (cell D10) away from what was expected (cell E5), use TDIST and find in D13 that there is a 19.39% chance.

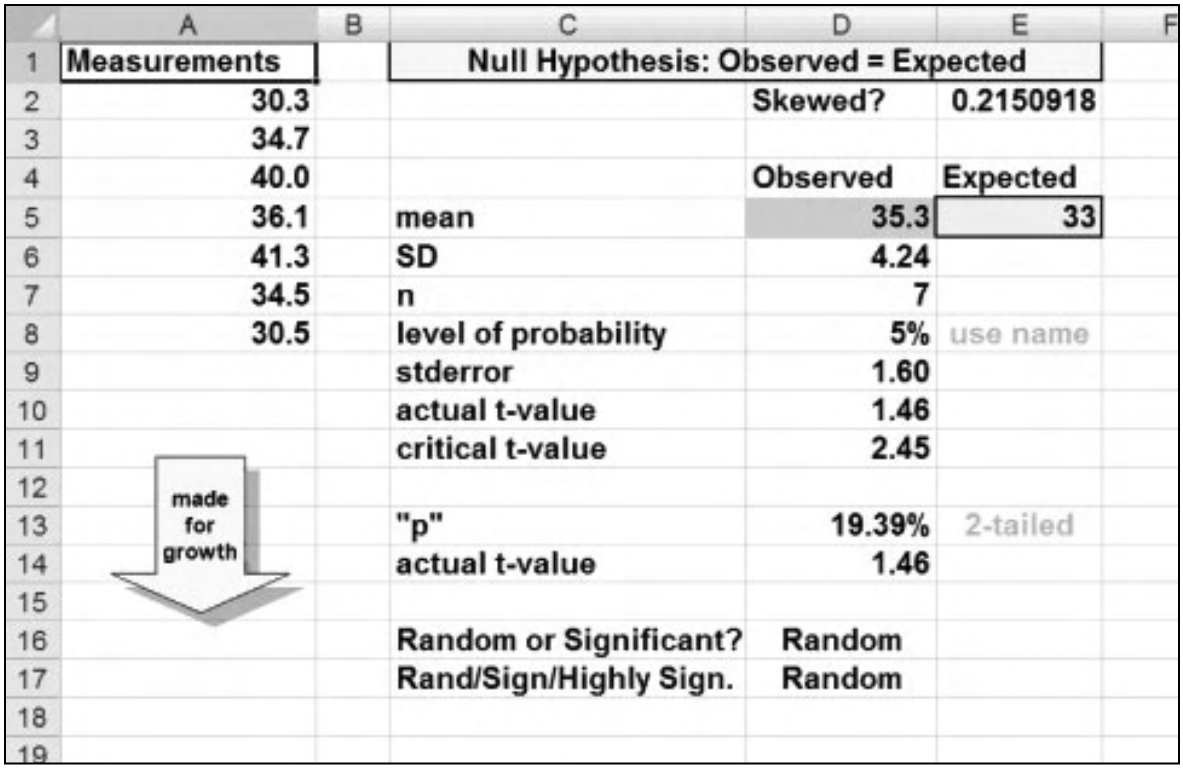


Figure: 5.39

> **Note**: Especially in medical literature, this chance is often called *p*, but this book reserves *p* for binomial proportions.

8. As a test, use cell D14 to find the actual *t*-value based on *p*: `=TINV(D13,D7-1)`.
9. Enter the following: `=IF(D10<D11,"Random","Significant")`.
10. For a three-tiered verdict, use a nested `IF` function:
    `=IF(D10<D11,"Random",IF(D10<TINV(D8/2,D7-1),"Significant","Highly Sign."))`.
    (Chapter 5 describes how to create this formula.)

Had you expected a mean of 30 (instead of 33) in cell E5, the second verdict would become highly significant.

Figure 5.40 shows a similar case in which you test whether a weight-loss pill actually works. This time, you check whether the difference before and after is significantly different from a zero weight loss—which is the null hypothesis here. Here's what you do:

1. In cell C2, type =A2-B2.
2. Check for skewness in F2:H2. Since 0.22 is at the low end, you can continue with a *t*-test.
3. Calculate the basic statistics on the differences in F5:H12. You could use two different tests:

	A	B	C	D	E	F	G	H
1	Before	After	Bef-Aft		Null Hypothesis: weight loss is random			
2	219.0	211.0	8.0		skewness	-0.36	-0.25	-0.03
3	215.0	209.0	6.0					
4	194.0	191.0	3.0			2-tail		1-tail
5	222.0	211.0	11.0		mean of diff.	7.13		7.13
6	217.0	220.0	-3.0		stdev of diff.	7.35		7.35
7	204.0	200.0	4.0		n	16		16.00
8	192.0	175.0	17.0		level of probability	5%		10%
9	180.0	178.0	2.0		stderror	1.84		1.84
10	223.0	224.0	-1.0		actual t-value	3.88		3.88
11	219.0	202.0	17.0		critical t-value	2.13		1.75
12	187.0	169.0	18.0		random or significant	significant		significant
13	205.0	193.0	12.0					
14	213.0	200.0	13.0					
15	193.0	198.0	-5.0					
16	190.0	179.0	11.0					
17	222.0	221.0	1.0					
18					Pill causes >5 pounds weight loss?			
19								
20	estimating				min. loss on average	3.21		
21					max. loss on average	11.04		
22								

Figure: 5.40

- Testing at two tails is proper when testing for any significant weight change—increase or decrease (*t* at 5%).
- Testing at one tail is appropriate when testing for a significant weight loss only (*t* at 10%).

Apparently, the pill causes a significant weight loss—both in the one-tailed test and the two-tailed test.

Can you claim that the pill causes a weight loss of at least 5 pounds on average? No, such a claim is not justified; 7.13 was just a sample value, but "results may vary" in the next sample. You therefore need to find the 95% confidence margin of this sample mean (see Chapter 50). Cell F20 finds the lowest weight loss based on a 95% confidence: `=F5-(F9*F11)`. All you could claim with 95% confidence is that the pill causes an average weight loss of at least 3 pounds. Anything higher is not statistically sound.

If you don't like all the in-between calculations you've had to do so far, you can use the function `TTEST`, which nicely combines many tedious calculations. `TTEST` does all the work for you, but it returns a probability (or *p* for some). In the example shown in Figure 5.40, a one-tailed `TTEST` would look like this: `=TTEST(A2:A17,B2:B17,1,1)`. The next-to-last argument determines the number of tails. (The last argument will be discussed in the next paragraph.) So `TTEST` would come up with a very low probability of 0.0007—which makes it very unlikely that this combination of figures would be a random outcome. With a two-tailed test, the probability would be a little higher: 0.0015.

Not only does TTEST do all the work for you internally, you can also use it for more complicated situations, thanks to its last argument. The last argument allows you to specify the type of test:

- two paired samples
- two samples with equal variances
- two samples with unequal variances

This last argument is especially helpful when you are not dealing with paired samples, as Figure 5.41 demonstrates. Using the manual calculations in this unpaired case would be very cumbersome because you would have to deal with pooled standard deviations and what comes with it. TTEST eliminates all this work, but you must decide whether you are dealing with equal or unequal variances. You can do this with the help of FTEST, which is used in cell E17: `=FTEST(A:A,B:B)`. Because the combination of these two variances turns out to be very likely (61%), you may decide on an equal-variance test in E16: `=TTEST(A:A,B:B,2,2)`. In this case, the null hypothesis wins, with a high 99% probability. This conclusion is that the effect of a specific treatment on the sample in column A cannot be substantiated.

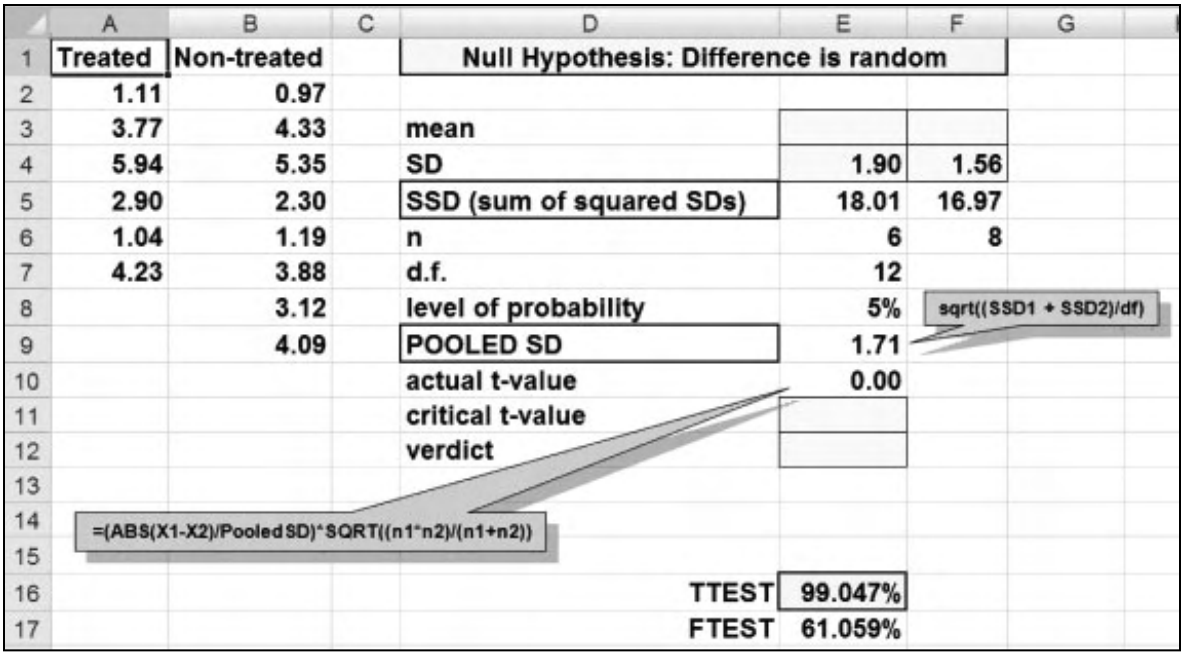


Figure: 5.41

* * *

Chapter 53

Significant Proportions

What you did for means in the previous chapter you can also do for proportions. When finding a proportion different from an expected proportion, you want to test whether the proportion found is significantly different from the proportion expected.

Figure 5.42 helps explain this concept. You perform three different tests—a one-tailed test at the lower end, a one-tailed test at the higher end, and a two-tailed test:

- In a sample of 60 vaccinated cases, you found 16 anthrax infections ("yes," or "success") versus 44 non-infected cases—so p=27%:
  - The null hypothesis claims no effect from vaccination, so p=50% in cell I4. It claims that 27% is a random deviation from 50% due to sample size.
  - The alternative hypothesis claims that vaccination has a lowering effect. Therefore, you need to test only at one tail—the lower tail (p<5%). It claims that 27% is significantly below 50%.
  - You use `CRITBINOM` to test the null hypothesis: 60 trials with a proportion of

L15 =IF(AND(B14>K15,B14<K16),"Random","Significant")

	A	B	C	D	E	F	G	H	I	J	K	L
1					Among n vaccinated cases							
2										95% cumul.		
3	Occurrence	Yes	No	Count	p_{yes}		Alternative Hypo		pNull	pSign	Min.L.	Verdict
4	Anthrax	16	44	60	0.27		vaccine lowers p		50%	5%	24	Significant
5												
6												
7										95% cumul.		
8	Diet	Yes	No	Count	p_{yes}		Alternative Hypo		Null: p	pSign	Max.R.	Verdict
9	Caffeine	29	21	50	0.58		caffeine raises p		50%	95%	31	Random
10												
11												
12												
13	Gender	Yes	No	Count	p_{yes}		Alternative Hypo		Null: p	pSign	Min+Max	Verdict
14	Female	35	45	80	0.44		gender has an impact					
15							fewer females	$<p_{null}$	50%	2.5%	31	Random
16							more females	$>p_{null}$	50%	97.5%	49	

Figure: 5.42

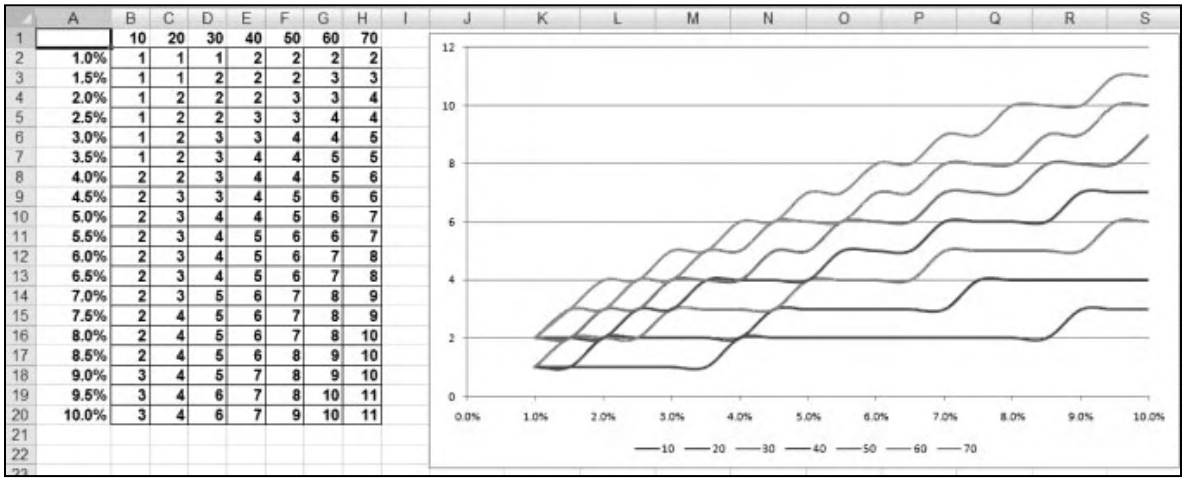
50% at the lower end's significance level (one-tail: 5%). In cell K4, you use `=CRITBINOM(D4,I4,0.05)`.

- Because 16 cases is below the 5% level of 24 cases (if random), the verdict is: significant. Cell L4 contains the formula: `=IF(B4<K4,"Significant","Random")`.

- In the second case, you test whether drinking caffeine increases the proportion of high systolic blood pressure:
  - You test the null hypothesis (p=50%) at one tail: >95%.
  - In cell K9, you type `=CRITBINOM(D9,I9,J9)`.
  - The verdict is: `=IF(B9<K9,"Random","Significant")`.
- In the third case, you test whether gender has any impact on the proportion of diabetics (up or down):
  - The test is two-tailed; with 2.5% in cell J4 and 97.5% in cell J5.
  - In cell K15, you check the lower tail: `=CRITBINOM($D$14,I15,J15)`.
  - In cell K16, you test for the upper tail: `=CRITBINOM($D$14,I16,J16)`.
  - The verdict requires a nested `AND` function: `=IF(AND(B14>K15,B14<K16),"Random","Significant")`.

Figure 5.43 examines how many defects you can accept at the most in samples of 10, 20, and so on before you reject the whole batch if that batch is supposed to have 1%, 1.5%, and so on defects. Here's how it works:

- Cell B2 holds the formula `=CRITBINOM(B$1,$A2,95%)`.
- By setting the last argument to 95%, you take a 5% risk of rejecting a batch that should not have been rejected.
- Each curve seems jagged; the reason is that `CRITBINOM` rounds to numbers of the integer type.

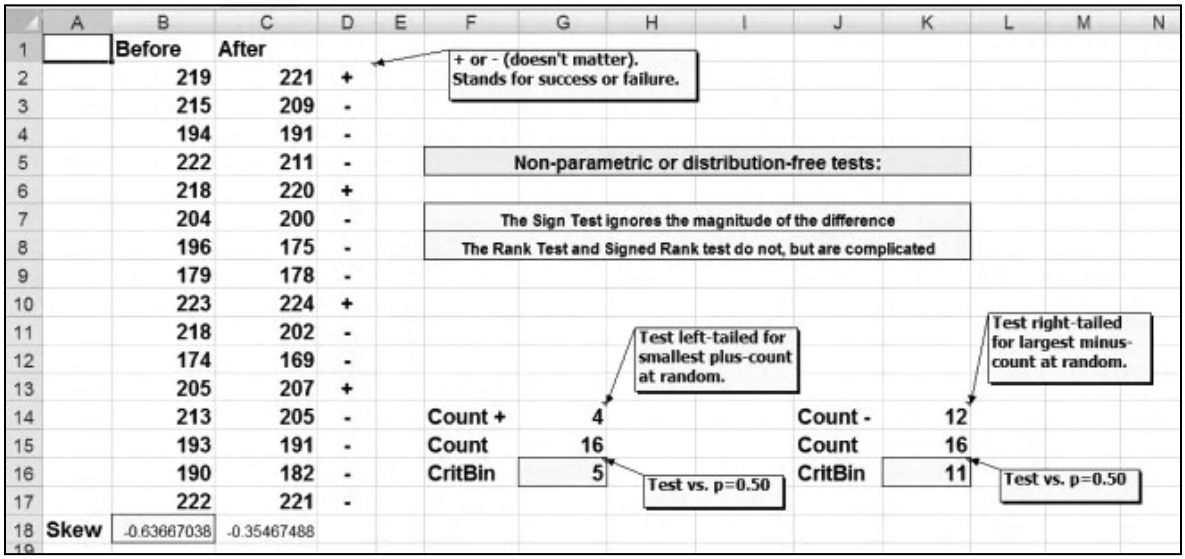


	A	B	C	D	E	F	G	H
1		10	20	30	40	50	60	70
2	1.0%	1	1	1	2	2	2	2
3	1.5%	1	1	2	2	2	3	3
4	2.0%	1	2	2	2	3	3	4
5	2.5%	1	2	2	3	3	4	4
6	3.0%	1	2	3	3	4	4	5
7	3.5%	1	2	3	4	4	5	5
8	4.0%	2	2	3	4	4	5	6
9	4.5%	2	3	3	4	5	6	6
10	5.0%	2	3	4	4	5	6	7
11	5.5%	2	3	4	5	6	6	7
12	6.0%	2	3	4	5	6	7	8
13	6.5%	2	3	4	5	6	7	8
14	7.0%	2	3	5	6	7	8	9
15	7.5%	2	4	5	6	7	8	9
16	8.0%	2	4	5	6	7	8	10
17	8.5%	2	4	5	6	8	9	10
18	9.0%	3	4	5	7	8	9	10
19	9.5%	3	4	6	7	8	10	11
20	10.0%	3	4	6	7	9	10	11

Figure: 5.43

Figure 5.44 shows a case in which you might never consider using a binomial test. And yet you should because the observations before and after a specific treatment seem to be rather skewed (see cells B18 and C18). The *t*-test may therefore not be appropriate here (see Chapter 49). In these kinds of situations, you should go for a *non-parametric*, or *distribution-free*, *test*. One such test is the sign test: It compares the plus cases ("successes") with the minus cases, so you end up actually dealing with a binomial distribution. Here's how it works:

> **Note**: The sign test ignores the magnitude of the differences. If you want to consider the magnitude, you need more sophisticated tests, such as the rank test and the signed rank test—which are beyond the scope of this book.



	A	B	C	D
1		Before	After	
2		219	221	+
3		215	209	-
4		194	191	-
5		222	211	-
6		218	220	+
7		204	200	-
8		196	175	-
9		179	178	-
10		223	224	+
11		218	202	-
12		174	169	-
13		205	207	+
14		213	205	-
15		193	191	-
16		190	182	-
17		222	221	-
18	Skew	-0.63667038	-0.35467488	

	F	G	H	I	J	K	L	M	N
14	Count +	4			Count -	12			
15	Count	16			Count	16			
16	CritBin	5			CritBin	11			

Figure: 5.44

1. In cell G14, enter `=COUNTIF(D2:D17,"+")`, and in cell G15, enter `=COUNTA(D2:D17)`.
2. In cell G16, you test whether a plus count of 4 (cell G14) out of 16 cases (cell G15) is significantly below the 5% mark. So you are testing against a null hypothesis of 50%: `=CRITBINOM(G15,0.5,5%)`.
3. In cell K16, do the opposite: Determine whether a minus count of 12 (cell K14) out of 16 cases (cell K15) is significantly above the 95% mark, compared with a null hypothesis of 50%. The answer is `=CRITBINOM(K15,0.5,0.95)`.

As you can see, both the plus and minus tests come up with the same result: The treatment had a significant impact.

* * *

Chapter 54

Significant Frequencies

Researchers often have to deal with frequencies instead of means and binomial proportions because they have their cases categorized in bins. The chi-squared distribution allows you to compare observed frequencies with expected frequencies. The chi distribution works with qualitative variables—with data based on categories rather than measurements. Until now, you have only dealt with quantitative variables.

Chi-tests are usually based on tables with a two-way structure. Each cell in a two-way table contains counts. However, the chi-test has one important condition: Each cell count must be at least 5. Could you use percentages instead? No, because percentages disregard sample size. For instance, "80 out of 100" is statistically better than "8 out of 10"—but in either case, the percentage would be 80%.

The basic idea behind a chi-test is that the observed frequencies have to be tested against the expected frequencies. In other words, you need to create a copy of the observed table and replace the frequencies with expectations. How do you do this?

Figure 5.45 shows a situation in which the chi-test is an appropriate choice. You test the effect of a certain pill on the recurrence of estrogen-fed tumors versus the effect of placebos, and you end up with four categories and their frequency counts. Because of the frequencies, the chi-test is called for. The table of observed frequencies (A3:D6) has total calculations in the end row and column. The table of expected frequencies (A11:D14) is an exact replica of the table above it, except for the observed frequencies (B12:C13). The observed frequencies have to be replaced by calculated, expected frequencies. Here's how you handle it:

1. In the first cell (B12), add a proportional value based on a null hypothesis of independence: `=Subtotal`$_{\texttt{Pillx}}$ `* Subtotal`$_{\texttt{Recur}}$ `/ Total`. If you do this with the proper relative/absolute settings, you can fill the table by using a single formula: `=$D4*B$6/$D$6`.
2. Apply the function `CHITEST` in cell G2: `=CHITEST(B4:C5,B12:C13)`. It quickly does all the tedious work for you. Like all other `TEST` functions, `CHITEST` returns the probability of the difference being random. Apparently, this combination of observed and expected frequencies is very unlikely (0.1%, not 0.1). In other words, the effect of the pill is highly significant.
3. In G3, enter `=IF(G2<F3,"significant","random")`. In G4, enter `=IF(G2<F4, "highly","not highly")`.

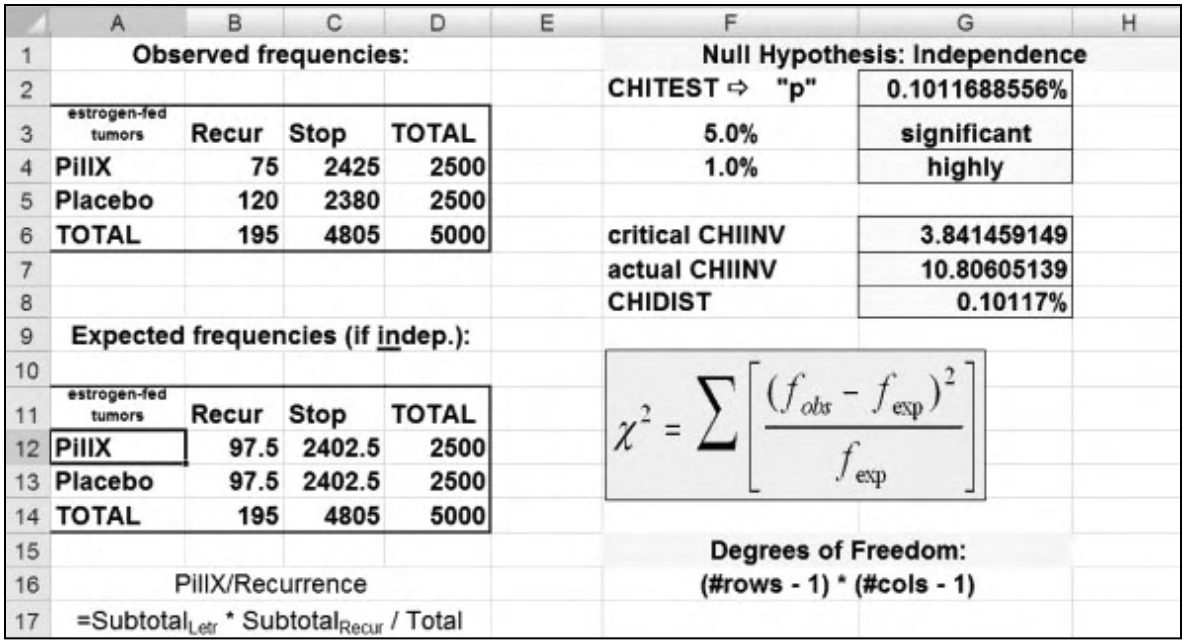


Figure: 5.45

> **Caution**: Unlike most other sampling distributions, the chi-distribution curve becomes steeper when the degrees of freedom decrease. The larger and finer the matrix system, the more degrees of freedom you have and therefore the slower the curve decline.

4. When you work with `CHIINV`, degrees of freedom are calculated this way: `(#rows - 1) * (#cols - 1)`. So calculate the critical chi-value in cell G6 with the formula `=CHIINV(signif,1)`. use The actual chi-value in G7: `=CHIINV(G2,1)`. The actual chi-value (10.81) is far beyond the critical chi-value (3.84) and is, therefore, highly significant. And `CHIDIST` would/should come up in G8 with the same probability as `CHITEST`: `=CHIDIST(G7,1)`.

You can also use the chi sampling distribution for cases in which you want to test law-like predictions for frequencies, as shown in Figure 5.46. Let's consider Mendel's law of independent segregation as an example. Two genes, each with two alleles (A and a plus B and b), are assumed to have independent segregation or no linkage—and this assumption acts here as the null hypothesis. Here's what you do:

1. Calculate the expected frequencies in B11:C12 according to Mendel's second law: `½ * ½ * total`.
2. Use CHITEST in G2: `=CHITEST(B4:C5,B11:C12)`.
3. Enter the verdict in G3: `=IF(G2<F3,"significant","random")`. The outcome is highly significant.
4. The rest speaks for itself.

	A	B	C	D	E	F	G	H
1	Observed frequencies:					Hypothesis of Independence		
2						CHITEST	0.82%	
3	AaBb x aabb	*Bb*	*bb*	TOTAL		5.0%	significant	
4	*Aa*	15	21	36		1.0%	highly	
5	*aa*	20	32	52				
6	TOTAL	35	53	88		critical CHIINV	3.841	
7						actual CHIINV	7.000	
8	Expected frequencies (no linkage):					CHIDIST	0.82%	
9								
10	AaBb x aabb	*Bb* 50%	*bb* 50%	TOTAL				
11	*Aa* 50%	22	22	44				
12	*aa* 50%	22	22	44				
13	TOTAL	44	44	88				

Figure: 5.46

Let's again consider Figure 5.45: Say that someone discovered that there is one more factor involved in the effect of the tested pill on the recurrence of estrogen-fed tumors. It has turned out that some of these women also received radiation and some didn't—which is a *confounding* factor. Confounding factors can have quite an impact on the verdict. Let's find out what the impact is with the help of Figure 5.47:

- Calculate the expected frequencies again in the second table.
- Set the degrees of freedom as follows: `(2-1) rows * (4-1) columns = 3`.

The (random) probability of this combination of observed and expected values went up from 0.1% to 1.3%. It is still a significant result, but it is not so impressive any more. The bottom line is that you need to always be on the lookout for confounding factors.

	A	B	C	D	E	F	G	H	I
1	Observed frequencies:							Null Hypothesis: Independence	
2								CHITEST ⇨ "p"	1.3311887%
3	estrogen-fed tumors	No radiation		Plus radiation				5.0%	significant
4		Recur	Stop	Recur	Stop	TOTAL		1.0%	not highly
5	PillX	70	400	5	2025	2500			
6	Placebo	111	368	5	2000	2484			
7	TOTAL	181	768	10	4025	4984			
8									
9									
10	Expected frequencies (if indep.):								
11									
12	estrogen-fed tumors	No radiation		Plus radiation					
13		Recur	Stop	Recur	Stop	TOTAL			
14	PillX	90.791	385.23	5.0161	2019	2500			
15	Placebo	90.209	382.77	4.9839	2006	2484			
16	TOTAL	181	768	10	4025	4984		Degrees of Freedom:	
17								(#rows - 1) * (#cols - 1)	

$$\chi^2 = \sum \left[\frac{(f_{obs} - f_{exp})^2}{f_{exp}} \right]$$

Figure: 5.47

* * *

Chapter 55

More on the Chi-Squared Test

The chi-squared sampling distribution is great for frequencies, but in that capacity, it may also be a good alternative for situations in which other types of distributions fail. This chapter examines cases in which chi-values can come to your aid.

Chapter 49 discusses the fact that some sampling distributions have specific conditions on their applicability. If the variable under investigation has unequal variances or is not normally distributed, you may not be able to use *z*, *t*, or *F*. Fortunately, the chi-distribution is less demanding—it requires only a minimum cell count of five.

Figure 5.48 demonstrates a situation in which you have measured cholesterol levels among several ethnic groups. Here's how you determine whether you can use *z* and *t* distributions:

1. Use `SKEW` in A13:D13. You find out that at least one subsample is extremely skewed.
2. Use `FTEST` in G15:I17. You find out that some probabilities are extremely low, which means unequal variances.

Based on these results, *z* and *t* values may not be reliable for this project. An additional problem is that you would need *z*/*t* values for A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D. Because each z/t value would have a 5% error chance, the collective error chance would be quite large.

	A	B	C	D	E	F	G	H	I
1	Cholesterol Measurements								
2	Afr.Amer.	Caucasians	Hispanics	Nat.Amer.					
3	179	160	171	156					
4	183	163	173	158					
5	186	165	174	159					
6	189	167	175	160					
7	192	169	176	164					
8	195	171	177						
9	198	173	178						
10	201								
11	204								
12									
13	-0.07602807	-0.18058846	-0.36727694	0.88488732					
14						Afr.Amer.	Caucasians	Hispanics	Nat.Amer.
15					Afr.Amer.		0.15392	0.00687	0.06109
16					Caucasian			0.14579	0.42424
17					Hispanics				0.61768
18					Nat.Amer.				

Figure: 5.48

Figure 5.49 may offer a way out of this dilemma. You cannot use TTEST here for three reasons: You would need six tests combined, most subgroups are skewed (A25:D25), and some subgroups vary significantly (M2:O4). The alternative is a chi-test, which has only one condition: Each cell must hold at least five elements. The problem is, however, that you may have to reshuffle the data: You need categories instead, and for each category, you need frequencies (observed vs. expected). Here's what you do:

1. Create one category for high cholesterol levels (that is, >180). In cell H7, use the formula `=COUNTIF(A2:A23,$G$7)`.
2. Determine the expected frequencies based on a null hypothesis by typing 34/4 in cell H9, or using the formula `=$L$7/COUNT($H$7:$K$7)`.
3. Use CHITEST. It tells you that the probability of finding these frequencies together is very low (2.3%). In other words, the alternative hypothesis kicks in: The racial difference of high cholesterol counts in this sample is significant. Needless to say, any changes in the category's borders (cell G7) would affect the outcome of the chi-test. Make sure, though, that each cell still holds at least five elements.
4. In the second table, expand the number of categories—and therefore, the number of degrees of freedom. In cells H11:H12, enter this array function: `=FREQUENCY(A2:A23,$G$11:$G$12)`.
5. Use cell H15 to find the expected count if the results were random: `=$L11*H$13/$L$13`.
6. Use three categories in the third table, calculate their frequencies, determine their expected frequencies (according to a Null hypothesis), and apply the chi-test again..

Creating a finer matrix with more cells usually means that randomness can play a larger role, so the probability of the difference may go up. It is not a good policy, however, to adjust the number of categories to your needs. You should set up your categories ahead of time so you don't manipulate them afterward in order to force a favorable verdict.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Afr.Am.	Cauc.	Hisp.	Nat.Am.		Ethnicity	X	SD	n			Afr.Am.	Cauc.	Hisp.	Nat.Am.
2	169	172	168	191		Afr.Amer.	190.64	14.30	22		Afr.Am.		0.34157	0.52175	0.16323
3	168	163	173	193		Caucasians	176.61	11.82	22		Cauc.			0.75452	0.02089
4	169	191	174	195		Hispanics	179.22	13.18	22		Hisp.				0.04401
5	193	167	165	146		Nat.Amer.	166.95	19.51	22		Nat.Am.				
6	196	169	176	161											
7	189	171	191	171			>180	16	8	5	5	34			
8	209	173	178	175										2.302%	
9	195	169	167	176				8.5	8.5	8.5	8.5	34			
10	207	168	173	173											
11	201	169	174	172			low	180	6	14	17	17	54		
12	204	167	199	199			high	250	16	8	5	5	34		
13	180	191	194	158					22	22	22	22	88		
14	207	192	197	140										0.142%	
15	186	201	210	165			low	180	13.5	13.5	13.5	13.5	54		
16	189	169	169	151			high	250	8.5	8.5	8.5	8.5	34		
17	199	171	176	142					22	22	22	22	88		
18	203	195	177	164											
19	188	181	165	168			low	170	5	9	7	12	33		
20	201	169	169	156			moderate	190	5	7	10	5	27		
21	204	171	176	135			high	210	12	6	5	5	28		
22	168	195	177	197					22	22	22	22	88		
23	169	181	165	145										10.222%	
24							low	170	8.25	8.25	8.25	8.25	33		
25	-0.541	0.85	1.25	0.1514			moderate	190	6.75	6.75	6.75	6.75	27		
26							high	210	7	7	7	7	28		
27									22	22	22	22	88		

Figure: 5.49

* * *

Chapter 56

Analysis of Variance

When two or more samples are normally distributed but differ in their variances, you have another option left: the analysis of variance (also called ANOVA). With this test, the difference among sample variances is used to estimate the population's variance, depending on the number of factors and the number of samples you have from the population under investigation. ANOVA compares samples and uses *F*-values to determine the ratio between the larger variance and the smaller variance.

Let's work with the *F*-test in Figure 5.50 before tackling the ANOVA tool. This sheet shows how two measuring methods have produced a set of differences in precision. Do these methods vary significantly as to their precision? Here's how you figure it out:

1. In cell B10, enter `=VAR(B3:B8)`.
2. In cell B11, enter `=COUNT(B3:B8)-1`.
3. In cell B13, divide the larger variance by the smaller variance: `=C10/B10`.

	A	B	C
1	Difference in precision for two measuring methods		
2		Results Method 1	Results Method 2
3		-0.3	0.6
4		0.1	-0.7
5		0.4	-0.2
6		0.2	0.4
7			0.8
8			-0.3
9			
10	variance	0.087	0.344
11	d.f.	3	5
12			
13	actual F-value	3.969	
14	critical F-value 2.5%	14.885	1-tailed (=5% 2-t)
15	verdict	random	
16			
17	use FTest instead	0.285849187	2-tailed
18	compare with FDist	0.142924594	1-tailed
19			

Figure: 5.50

4. In cell B14, calculate the critical *F*-value: `=FINV(signif/2,C11,B11)`. The function `FINV` is always one-tailed; a 2.5% one-tailed value is the same as a 5% two-tailed value. Because the actual *F*-value is less than the critical *F*-value, the difference in precision of these two measuring methods is random and not significant.
5. Use the `FTEST` function in cell B17: `=FTEST(B3:B8,C3:C8)`. This function is always two-tailed.
6. Compare the result from step 5 with `FDIST` in cell B18: `=FDIST(B13,C11,B11)`. Because FDIST is one-tailed, its result is half the result of `FTEST` (which is two-tailed).

Because `ANOVA` is an elaborate process, this chapter only discusses the Anova tool from the Analysis Toolpak. In Figure 5.51, could the three samples in columns A:C be analyzed with a *t*-test? Yes, they could, but that would require three tests (A-B, A-C, and B-C). Because each test has a 5% error chance, the total error chance would be 1-(1-0.05)^3 = 14%. An *F*-test, on the other hand, requires only one test: variance of the sample means / variance of all items.

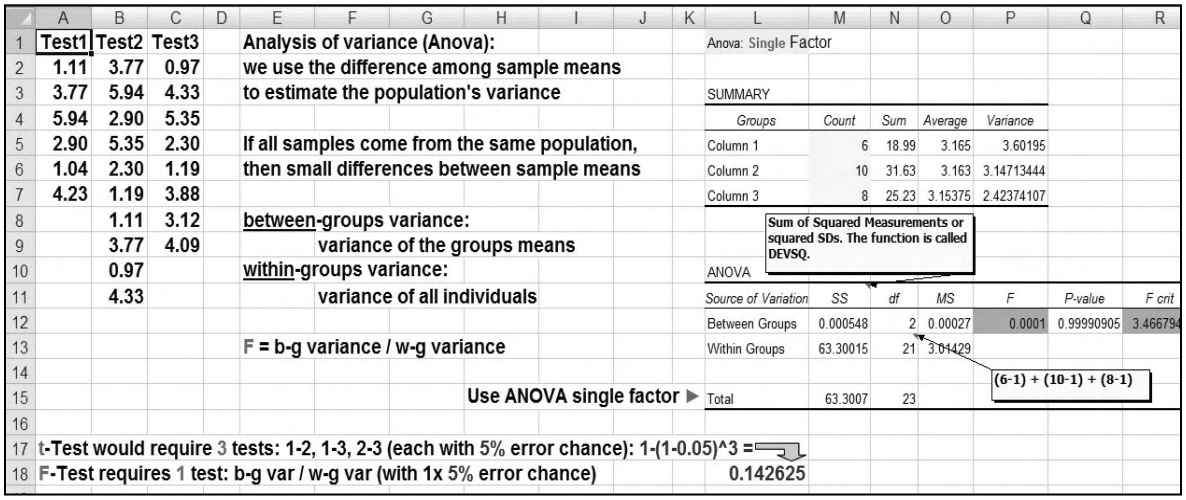


Figure: 5.51

As mentioned earlier, it is quite a procedure to calculate all the steps needed for a between-group variance and a within-group variance—not to mention the final analysis. You can instead use the super-speedy Anova tool from the Analysis Toolpak instead:

1. Open the Anova Single Factor tool.
2. Set the following settings: Input Range A1:C11, Grouped by Columns, with Labels in First Row, Alpha 0.05, and Output Range L1.
3. Anova displays all the basic information first, such as counts, sums, and variances (L4:P7). Then it calculates the between-groups and within-groups variance (L11:R15).

Note: When the samples are not of equal size, the sum of squares (SS) should be used instead of variances.

4. Based on the SS (which can be found with `DEVSQ`) and the degrees of freedom, determine that the *F*-value can be calculated (cell P12). Because the actual *F*-value is well within the range of the critical *F*-value (cell R12) and thus has a high probability of occurring by mere chance (cell Q12), the samples do not differ significantly.

In Figure 5.52, the number of colonies on Petri dishes has been counted for two sets of conditions. This is called a *two-factor analysis with replication*. It is a two-factor analysis because it deals with the factor nutrient level (rows) and the factor pH (columns). It is considered to be a test *with replication* because there are several readings per combination of factors (there are actually 10). Using the Analysis Toolpak is your best bet. You need to choose a two-factor analysis with replication and then set Input Range to A1:D21, Rows per Sample to 10, Alpha to 0.05, and Output Range to F1. This is what happens:

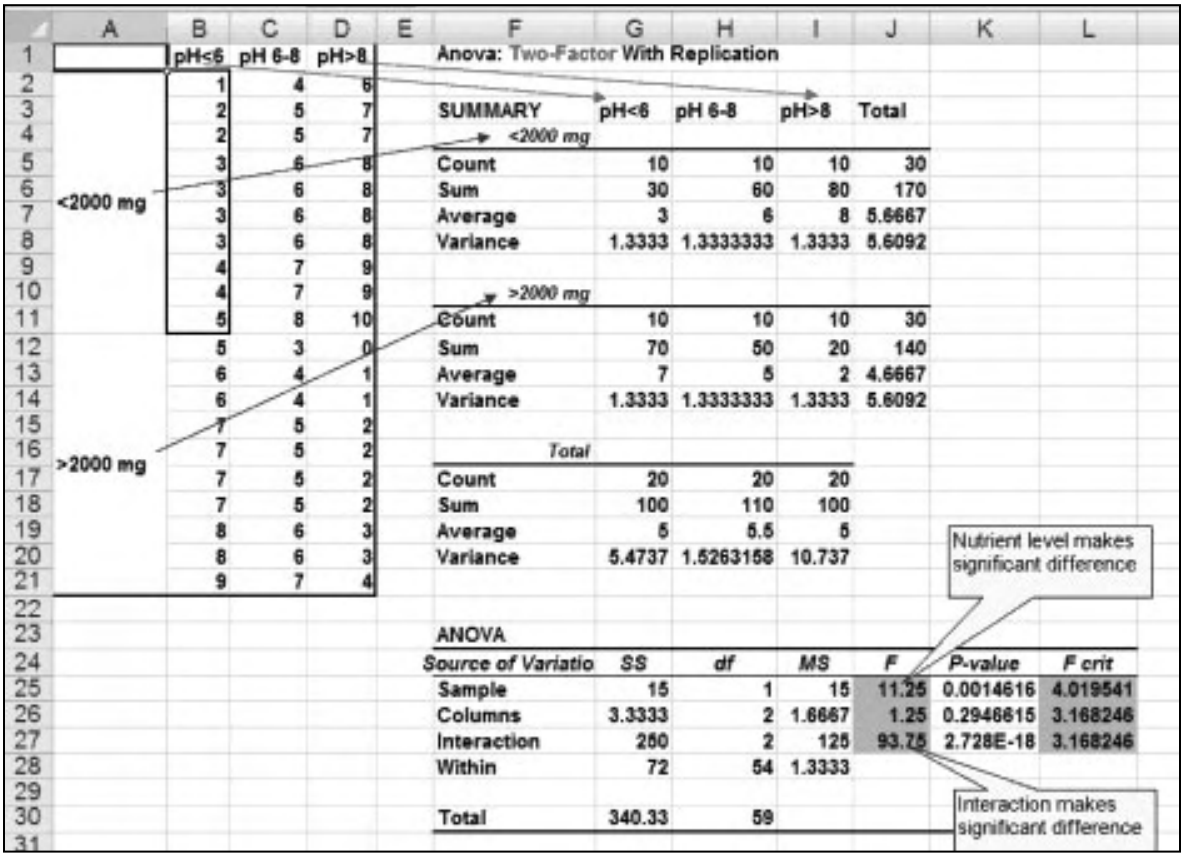


Figure: 5.52

First there is a summary table with the categories of each factor. At the bottom is a table that includes *F*-values:

- The first row represents the impact of the first factor, nutrient level. You find here the between-rows variance (B-R SS).
- The second row stands for the effect of the second factor, pH. It displays the between-columns variance (B-C SS)
- The third row shows their interaction, which is explain later in this chapter.
- The fourth row is for the within-group variance (W-G SS).
- The between-group variance (B-G SS) is missing.

What is the interaction? The B-G SS is usually larger than the combination of B-R SS and B-C SS. The remaining part is due to the combined effects of rows and columns—which is the interaction of both variables. That's what the third row is about.

The conclusion is three-fold:

- The differences for nutrient level are significant (11.25>4).
- The differences for pH are not significant (1.25<3.2).
- The interaction between both factors is significant (93.75>3.2).

However, you should go for the conclusion that corresponds with your preexisting alternative hypothesis because each conclusion comes with a 5% error chance:

- If you are testing the effect of nutrient levels, you did find one!
- If you are interested as to whether there is an optimum combination of level and pH, the answer is yes!
- But don't draw both conclusions at the same time because that would result in a 9.75% error!

Figure 5.53 revisits an example from Chapter 55. Do you remember this situation, where the `TTEST` would fail? Earlier, you used `CHITEST` to do the job, but it was not a very good solution. Would the Anova tool be a good alternative? When you apply the Anova tool, you notice that there is definitely not a great match here with the chi-test. Why not? Because Anova has its own requirements:

- Every subgroup must be independent (which is okay in this case).
- Every subgroup must have a normal distribution (which is not okay in this case).
- The variances of these distributions must be equal (which is not okay in this case).

So the chi-test may still be your best bet for a case like the one shown in Figure 5.53. You need to make sure, though, to set up your categories ahead of time so you don't manipulate

them afterward in order to force a favorable verdict.

In addition to what you've learned in this book, there are many more issues you should know in statistics. This book is not a handbook on statistics, so you need to gather more detailed statistical knowledge somewhere else. Even if you don't opt to learn more about statistics, Excel may still be an excellent tool for applying your statistical knowledge in all your scientific work.

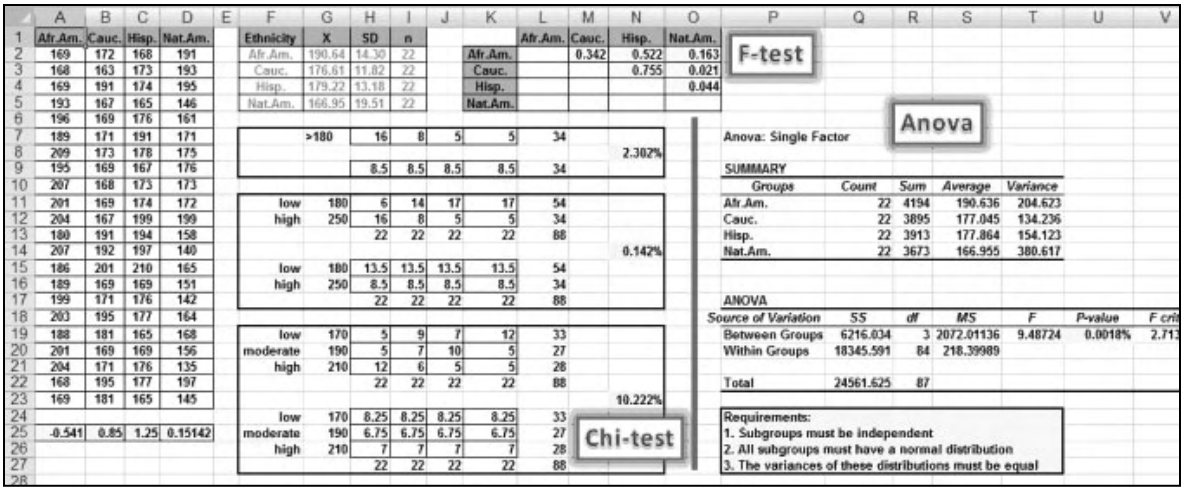


Figure: 5.53

* * *

Chapter 57

Power Curves

We certainly need statistics when we want to make a conclusion based on a sample, because "results may vary." Since the sample is not the entire population, we always take a chance of making an error. The null hypothesis claims that an observed difference is not real (but a random result), whereas the alternative hypothesis claims that the difference is actually real (and not fake). Consequently, we can make two types of errors:

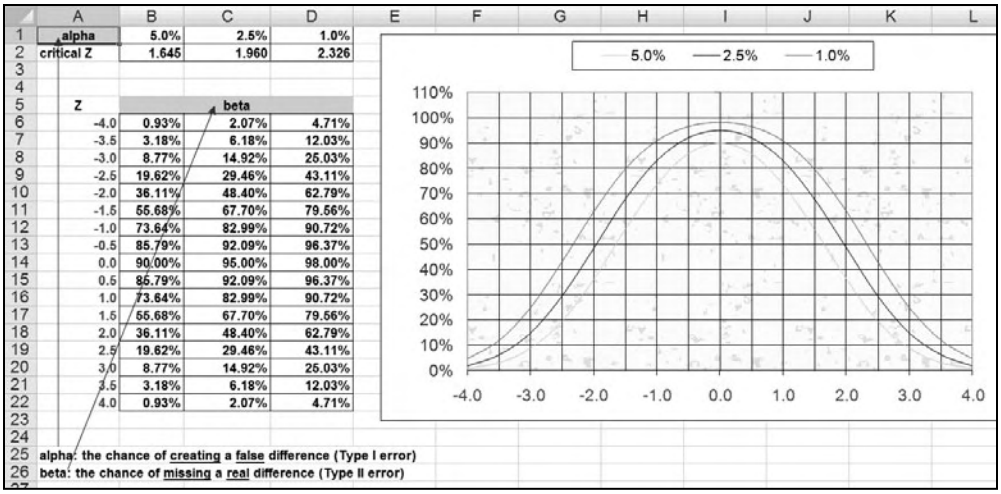
- Alpha (α) is the chance of making a Type I error by creating a false difference. At an alpha level of 1%, for instance, we would go for the alternative hypothesis by claiming a real difference, but there is still a 1% chance of creating a fake difference.
- Beta (β) is the chance of making a Type II error by missing a real difference. This happens when we go for the null hypothesis by claiming there is no difference, but we still run the risk of missing a very real difference.

How can we reduce our chances of making these errors? To reduce the chance of a Type I error, we could reduce the value of alpha – but by doing so, we would also increase the chance of making a Type II error. To reduce the chance of making a Type II error, we could take larger samples – but by doing so, we would also increase the chance of a Type I error. So we are stuck in the middle: Choose a large sample size and a small α level (<0.05).

So let us take a different approach. Instead of looking at the chances of missing a real difference, we could look at the chances of detecting a real difference. Instead of looking at the weakness of a hypothesis test, we could look at the power of a hypothesis test. The power of a hypothesis test is one minus the probability of making a Type II error: $1 - \beta$.

And that is where power curves come into the picture. Power curves give us an idea of how much of a difference we can detect with the sample size we have and given the magnitude of the difference we are trying to detect. As the sample size increases and/or the magnitude of the difference increases, the power gets closer and closer to 1.

Figure 5.54 shows the chances of making a Type II error – which means missing a real difference. It does so for 3 different α levels (5%, 2.5%, and 1%) – which represent the chances of creating a false difference. It is obvious that the β-curve goes down when the observed difference gets larger and larger.



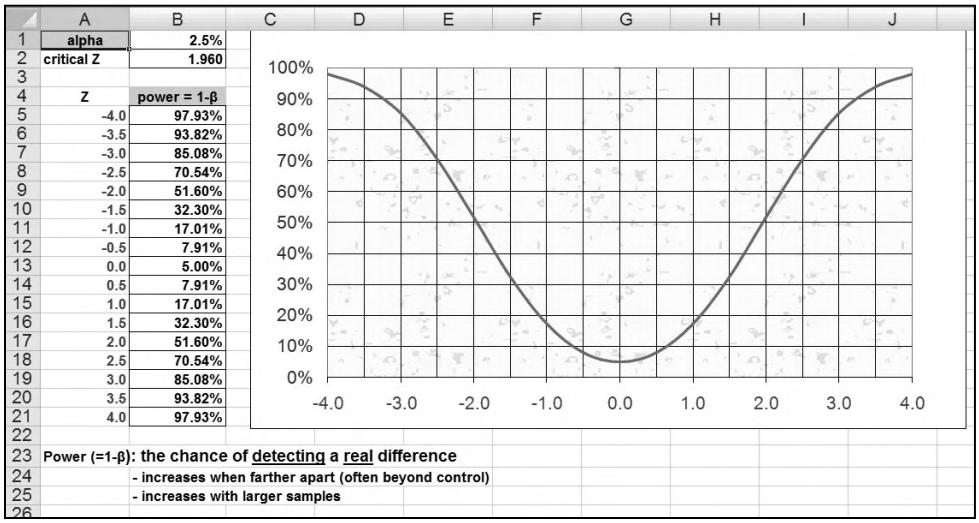
	A	B	C	D
1	alpha	5.0%	2.5%	1.0%
2	critical Z	1.645	1.960	2.326
3				
4				
5	Z	beta		
6	-4.0	0.93%	2.07%	4.71%
7	-3.5	3.18%	6.18%	12.03%
8	-3.0	8.77%	14.92%	25.03%
9	-2.5	19.62%	29.46%	43.11%
10	-2.0	36.11%	48.40%	62.79%
11	-1.5	55.68%	67.70%	79.56%
12	-1.0	73.64%	82.99%	90.72%
13	-0.5	85.79%	92.09%	96.37%
14	0.0	90.00%	95.00%	98.00%
15	0.5	85.79%	92.09%	96.37%
16	1.0	73.64%	82.99%	90.72%
17	1.5	55.68%	67.70%	79.56%
18	2.0	36.11%	48.40%	62.79%
19	2.5	19.62%	29.46%	43.11%
20	3.0	8.77%	14.92%	25.03%
21	3.5	3.18%	6.18%	12.03%
22	4.0	0.93%	2.07%	4.71%
23				
24				
25	alpha: the chance of creating a false difference (Type I error)			
26	beta: the chance of missing a real difference (Type II error)			

Figure: 5.54

- Cell B2 finds the critical *Z*-value with the following formula: `=ABS(NORMSINV(B1))`.
- Cell B6 is a bit more complicated: `=(NORMSDIST($A6+B$2)-NORMSDIST($A6-B$2))`. What we need here is the surface under the curve between the probability of ($Z + Z_{crit}$) and the probability of ($Z - Z_{crit}$).

Figure 5.55 shows the power curve for an alpha level of 2.5%: The power (1-β) increases when the magnitude of the difference increases. Obviously, the curve never reaches 1 or 0, because there is never 100% certainty in statistics.

- Cell B2 calculates the critical *Z*-value for an alpha level of 2.5%: `=ABS(NORMSINV(B1))`.
- Cell B5 calculates the power (1-β): `=1-(NORMSDIST($A5+B$2)-NORMSDIST($A5-B$2))`.



	A	B
1	alpha	2.5%
2	critical Z	1.960
3		
4	Z	power = 1-β
5	-4.0	97.93%
6	-3.5	93.82%
7	-3.0	85.08%
8	-2.5	70.54%
9	-2.0	51.60%
10	-1.5	32.30%
11	-1.0	17.01%
12	-0.5	7.91%
13	0.0	5.00%
14	0.5	7.91%
15	1.0	17.01%
16	1.5	32.30%
17	2.0	51.60%
18	2.5	70.54%
19	3.0	85.08%
20	3.5	93.82%
21	4.0	97.93%
22		
23	Power (=1-β): the chance of detecting a real difference	
24	- increases when farther apart (often beyond control)	
25	- increases with larger samples	
26		

Figure: 5.55

In Figure 5.56, we are doing something similar, but this time for different sample sizes (5, 25, and 100), and given a specific mean (10) and standard deviation (0.5). The power curves become steeper when the sample size gets larger – in other words, the probability of detecting a real difference decreases when the sample size decreases. On the other hand, these power curves also tell us how much of a difference we would be able to detect for a given sample size.

- Cell C3 has this formula: `=$A3/SQRT(C2)`.
- Cell C6 uses the formula shown on the insert in the right lower corner: `=NORMSDIST(-1.96+(ABS($B$12-$B6)*SQRT(C$2))/$A$3)`.

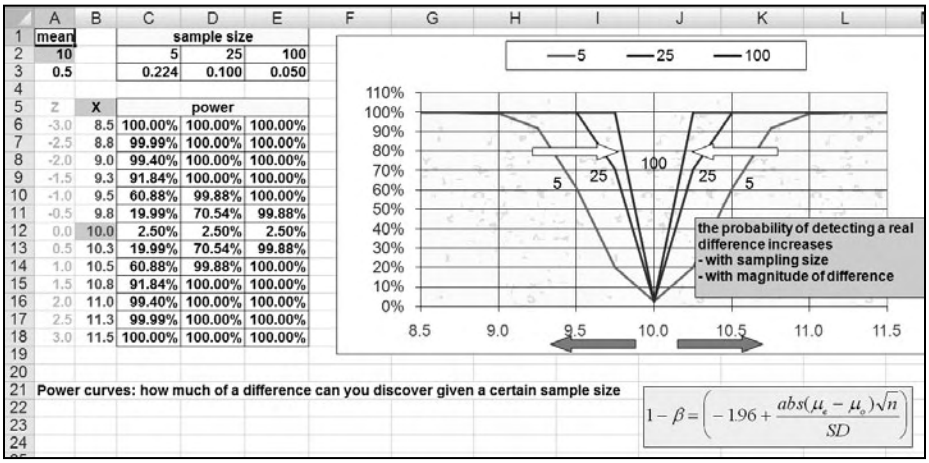


Figure: 5.56

Figure 5.57 shows the same phenomenon, but this time by using `TDIST` instead of `NORMSDIST`. The message is the same again: The power curves become steeper when the samples get larger. So it is harder to detect a real difference in smaller samples.

- Cell B1: `=TINV($A$1,B3-1)`.
- Cell B4: `=1-(TDIST($A4-B$1,B$3-1,2)-TDIST($A4+B$1,B$3-1,2))`. Be aware that the graph hides the calculations created in the table.

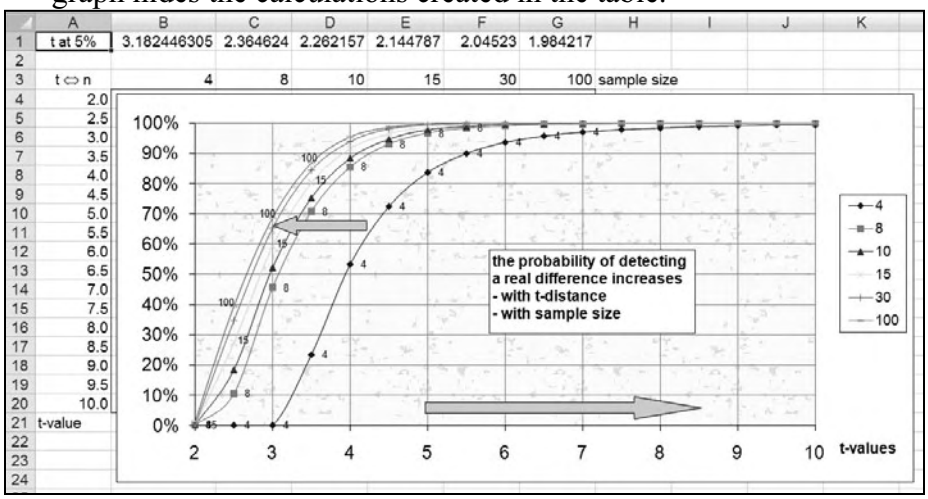


Figure: 5.57

Figure 5.58 demonstrates again, by way of a summary, that the probability of detecting a real difference (versus a fake difference) depends on two factors: the sample size and the magnitude of the difference (margin). Therefore, we could determine the minimum sample size needed in order to detect a certain difference (in proportion to the mean) given a certain standard deviation (proportional to the mean). Finer differences and relatively larger standard deviations require larger samples. You probably didn't need statistics to find this out. But now we are better equipped to quantify these requirements.

- Cell D6 uses the formula shown in the upper insert: `=$C6^2*($C$5/D$5)^2`.
- Cell C25 uses the formula shown in the lower insert: `=((C24*C21)/C23)^2`.

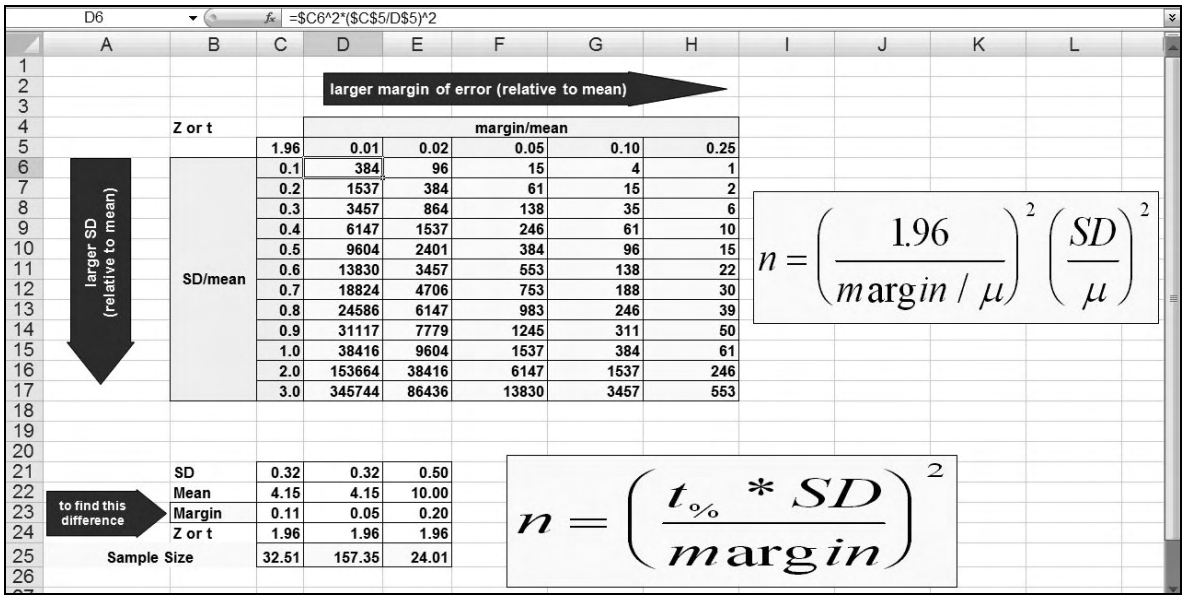
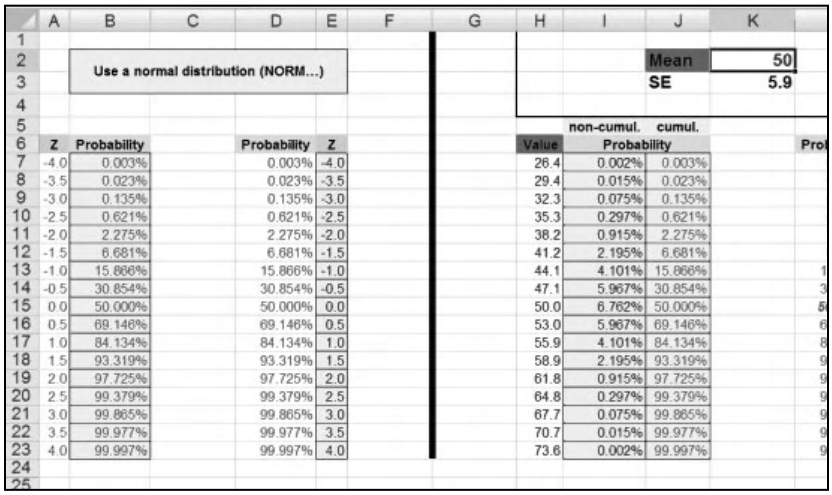


Figure: 5.58

* * *

Excercises - Part 5

You can download all the files used in this book from `www.genesispc.com/Science2007.htm`, where you can find each file in its original version (to work on) and in its finished version (to check your solutions).



	A	B	C	D	E	F	G	H	I	J	K
1											
2		Use a normal distribution (NORM...)								Mean	50
3										SE	5.9
4											
5									non-cumul.	cumul.	
6		Z	Probability		Probability	Z		Value	Probability		
7		-4.0	0.003%		0.003%	-4.0		26.4	0.002%	0.003%	
8		-3.5	0.023%		0.023%	-3.5		29.4	0.015%	0.023%	
9		-3.0	0.135%		0.135%	-3.0		32.3	0.075%	0.135%	
10		-2.5	0.621%		0.621%	-2.5		35.3	0.297%	0.621%	
11		-2.0	2.275%		2.275%	-2.0		38.2	0.915%	2.275%	
12		-1.5	6.681%		6.681%	-1.5		41.2	2.195%	6.681%	
13		-1.0	15.866%		15.866%	-1.0		44.1	4.101%	15.866%	
14		-0.5	30.854%		30.854%	-0.5		47.1	5.967%	30.854%	
15		0.0	50.000%		50.000%	0.0		50.0	6.762%	50.000%	
16		0.5	69.146%		69.146%	0.5		53.0	5.967%	69.146%	
17		1.0	84.134%		84.134%	1.0		55.9	4.101%	84.134%	
18		1.5	93.319%		93.319%	1.5		58.9	2.195%	93.319%	
19		2.0	97.725%		97.725%	2.0		61.8	0.915%	97.725%	
20		2.5	99.379%		99.379%	2.5		64.8	0.297%	99.379%	
21		3.0	99.865%		99.865%	3.0		67.7	0.075%	99.865%	
22		3.5	99.977%		99.977%	3.5		70.7	0.015%	99.977%	
23		4.0	99.997%		99.997%	4.0		73.6	0.002%	99.997%	
24											
25											

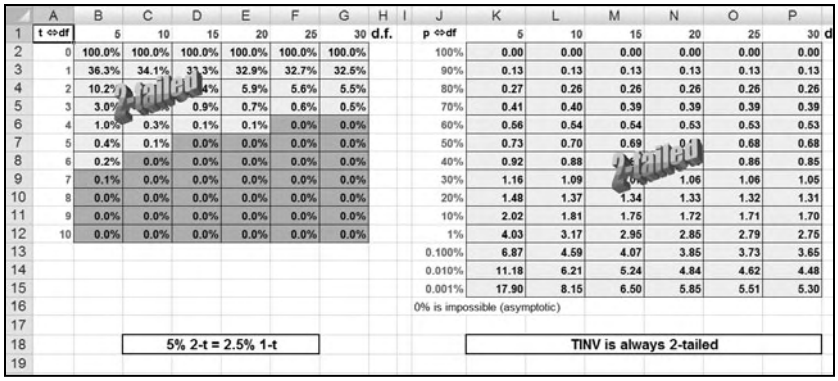
Figure: Ex-1

Exercise 1

1. Types of Distributions
   1.1. Use the proper normal distribution functions in columns B and E.
   1.2. Use the proper normal distribution functions in columns I, J, and M, based on a specific mean (cell K2) and a specific SE (cell K3).

Exercise 2

2. Types of Distributions
   2.1. Use the proper t-distribution function in the table on the left to find the probability of certain *t*-values (column A) with certain degrees of freedom (row 1).
   2.2. Use the proper *t*-distribution function in the table on the right to find the *t*-values for certain probabilities (column J) with certain degrees of freedom (row 1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	t ⇔df	5	10	15	20	25	30	d.f.		p ⇔df	5	10	15	20	25	30 d
2	0	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%			100%	0.00	0.00	0.00	0.00	0.00	0.00
3	1	36.3%	34.1%	33.3%	32.9%	32.7%	32.5%			90%	0.13	0.13	0.13	0.13	0.13	0.13
4	2	10.2%	7.7%	6.4%	5.9%	5.6%	5.5%			80%	0.27	0.26	0.26	0.26	0.26	0.26
5	3	3.0%	1.4%	0.9%	0.7%	0.6%	0.5%			70%	0.41	0.40	0.39	0.39	0.39	0.39
6	4	1.0%	0.3%	0.1%	0.1%	0.0%	0.0%			60%	0.56	0.54	0.54	0.53	0.53	0.53
7	5	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%			50%	0.73	0.70	0.69	0.68	0.68	0.68
8	6	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%			40%	0.92	0.88	0.87	0.86	0.86	0.85
9	7	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%			30%	1.16	1.09	1.08	1.06	1.06	1.05
10	8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%			20%	1.48	1.37	1.34	1.33	1.32	1.31
11	9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%			10%	2.02	1.81	1.75	1.72	1.71	1.70
12	10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%			1%	4.03	3.17	2.95	2.85	2.79	2.75
13										0.100%	6.87	4.59	4.07	3.85	3.73	3.65
14										0.010%	11.18	6.21	5.24	4.84	4.62	4.48
15										0.001%	17.90	8.15	6.50	5.85	5.51	5.30
16										0% is impossible (asymptotic)						
17																
18		5% 2-t = 2.5% 1-t								TINV is always 2-tailed						
19																

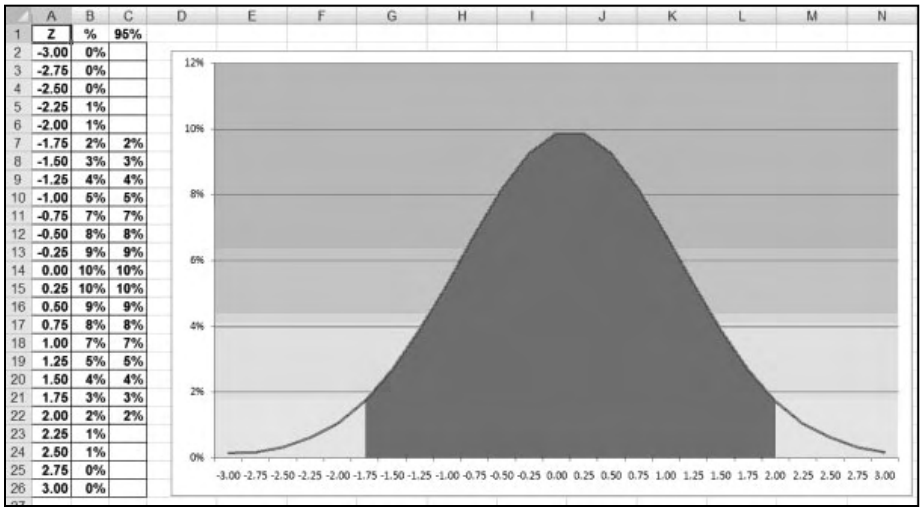
Figure: Ex-2

Exercise 3

3. Simulating Distributions

   3.1. Retrieve the values in column B by using the proper function.

   3.2. Create a formula for the values in column C.



Z	%	95%
-3.00	0%	
-2.75	0%	
-2.50	0%	
-2.25	1%	
-2.00	1%	
-1.75	2%	2%
-1.50	3%	3%
-1.25	4%	4%
-1.00	5%	5%
-0.75	7%	7%
-0.50	8%	8%
-0.25	9%	9%
0.00	10%	10%
0.25	10%	10%
0.50	9%	9%
0.75	8%	8%
1.00	7%	7%
1.25	5%	5%
1.50	4%	4%
1.75	3%	3%
2.00	2%	2%
2.25	1%	
2.50	1%	
2.75	0%	
3.00	0%	

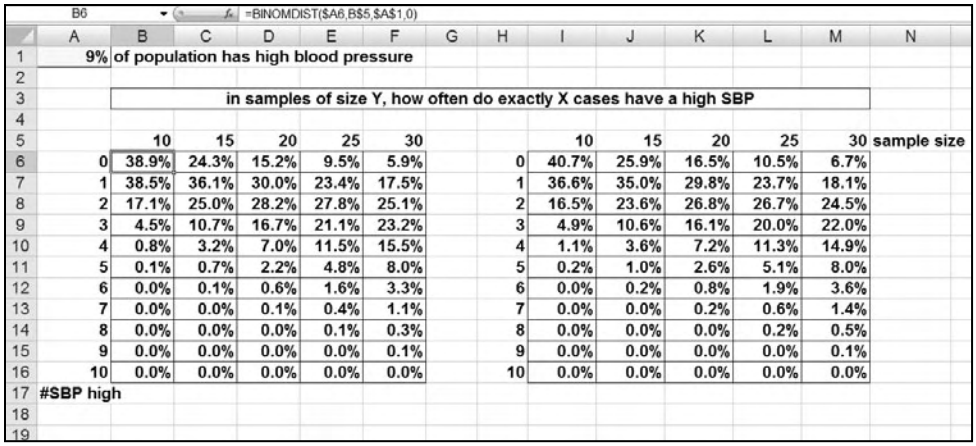
Figure: Ex-3

Exercise 4

4. Sampling Techniques

   4.1. In the table on the left, use a binomial function to calculate how often you find exactly X cases (column A) of high blood pressure in samples of size y (row 5), given the fact that 9% of the population is known to have high blood pressure.

   4.2. In the table on the right, use a Poisson function to calculate how often you find exactly x cases (column H) of high blood pressure in samples of size y (row 5), given the fact that 9% of the population is known to have high blood pressure.



B6 =BINOMDIST($A6,B$5,$A$1,0)

9%	of population has high blood pressure												
	in samples of size Y, how often do exactly X cases have a high SBP												
	10	15	20	25	30		10	15	20	25	30	sample size
0	38.9%	24.3%	15.2%	9.5%	5.9%	0	40.7%	25.9%	16.5%	10.5%	6.7%	
1	38.5%	36.1%	30.0%	23.4%	17.5%	1	36.6%	35.0%	29.8%	23.7%	18.1%	
2	17.1%	25.0%	28.2%	27.8%	25.1%	2	16.5%	23.6%	26.8%	26.7%	24.5%	
3	4.5%	10.7%	16.7%	21.1%	23.2%	3	4.9%	10.6%	16.1%	20.0%	22.0%	
4	0.8%	3.2%	7.0%	11.5%	15.5%	4	1.1%	3.6%	7.2%	11.3%	14.9%	
5	0.1%	0.7%	2.2%	4.8%	8.0%	5	0.2%	1.0%	2.6%	5.1%	8.0%	
6	0.0%	0.1%	0.6%	1.6%	3.3%	6	0.0%	0.2%	0.8%	1.9%	3.6%	
7	0.0%	0.0%	0.1%	0.4%	1.1%	7	0.0%	0.0%	0.2%	0.6%	1.4%	
8	0.0%	0.0%	0.0%	0.1%	0.3%	8	0.0%	0.0%	0.0%	0.2%	0.5%	
9	0.0%	0.0%	0.0%	0.0%	0.1%	9	0.0%	0.0%	0.0%	0.0%	0.1%	
10	0.0%	0.0%	0.0%	0.0%	0.0%	10	0.0%	0.0%	0.0%	0.0%	0.0%	
#SBP high												

Figure: Ex-4

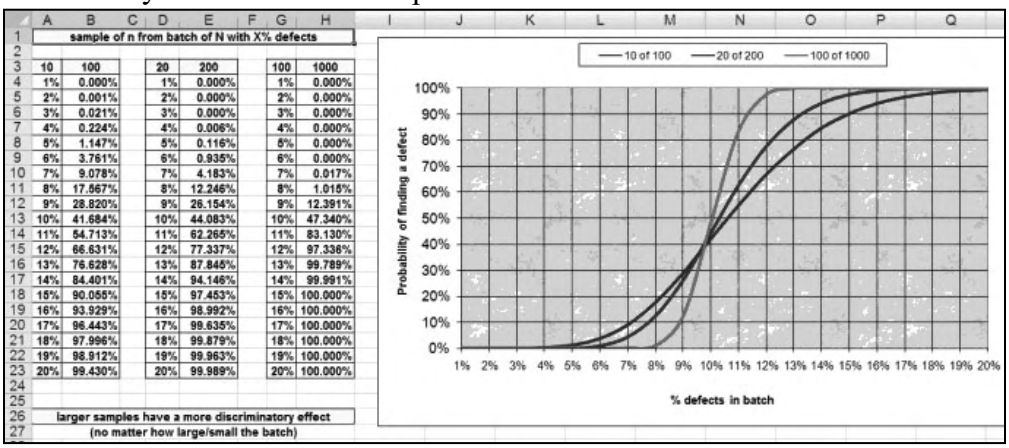
Exercise 5

5. Sampling Techniques

5.1. Calculate in column B what the probability is of finding x% defects (column A) in a sample of 10 (cell A3) from a batch of 100 (cell B3).

5.2. Calculate in column E what the probability is of finding x% defects in a sample of 20 from a batch of 200.

5.3. Do the previous step also in column H, for a sample of 100 from a batch of 1,000 – and you will see that sample size is all that matters.



sample of n from batch of N with X% defects						
10	100		20	200	100	1000
1%	0.000%		1%	0.000%	1%	0.000%
2%	0.001%		2%	0.000%	2%	0.000%
3%	0.021%		3%	0.000%	3%	0.000%
4%	0.224%		4%	0.006%	4%	0.000%
5%	1.147%		5%	0.116%	5%	0.000%
6%	3.761%		6%	0.935%	6%	0.000%
7%	9.078%		7%	4.183%	7%	0.017%
8%	17.567%		8%	12.246%	8%	1.015%
9%	28.820%		9%	26.154%	9%	12.391%
10%	41.684%		10%	44.083%	10%	47.340%
11%	54.713%		11%	62.265%	11%	83.130%
12%	66.631%		12%	77.337%	12%	97.336%
13%	76.628%		13%	87.845%	13%	99.789%
14%	84.401%		14%	94.146%	14%	99.991%
15%	90.055%		15%	97.453%	15%	100.000%
16%	93.929%		16%	98.992%	16%	100.000%
17%	96.443%		17%	99.635%	17%	100.000%
18%	97.996%		18%	99.879%	18%	100.000%
19%	98.912%		19%	99.963%	19%	100.000%
20%	99.430%		20%	99.989%	20%	100.000%

larger samples have a more discriminatory effect
(no matter how large/small the batch)

Figure: Ex-5

Exercise 6

6. Some Conditions

6.1. Calculate the SE of the sampling distribution in G5 and J5.

6.2. What is the probability of being -1.96 SE units away from 56?

6.3. Calculate the *z* or *t* value that has a 5% chance of occurring.

6.4. Find the *z* or t value that covers 95% of the cases.

6.5. Calculate how far away 50 is from 56 in SE units. (Don't use a function to do this.)

6.6. Which cells change when you change the mean in cells G1 and J1 from 56 to 54?

					Mean	56		Mean	56
					SD	11		SD	11
					Size	20		Size	20
What is the SE of the sampling distribution?						2.460			2.460
The probability of being -1.96 SE-Units away from 56 is:						2.50%			3.24%
Which Z or t value has a 5% chance of occurring?						-1.645			1.729
Which Z or t value covers 95% of the cases?						1.645			1.729
How many SE-units is 50 away from 56? (don't use a function)						2.439			2.439
Which cells change when you change the mean on top from 56 to 54?						none			none
Make both sample sizes 49									
Make both sample sizes 20									

Figure: Ex-6

Exercise 7

7. Estimating Means

7.1. Calculate the 95% confidence intervals (or margins of error) for row 9 by using *z* and SE.

7.2. Calculate the 95% confidence intervals (or margins of error) for row 10 by using the function `CONFIDENCE`.

7.3. Calculate the 95% confidence intervals (or margins of error) for row 15 by using *t* and SE.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1										confidence limit				confidenc intervals
2														
3		SAMPLES				Confidence						Mean		
4	Feature	Mean	SD	Count		Level	2-tails	1-tail	Z or t	StErr	Margin	Min	Max	
5														
6														
7		size over 30							Z	use *Normal* Distribution				
8														
9	pH	6.80	0.04	50		95%	5%	2.5%	-1.96	0.006	0.011	6.789	6.811	
10							Confidence uses 5%+SD:				0.011	6.789	6.811	
11														
12														
13		size under 30							t	use t-table: *t*-Distribution				
14														
15	pH	6.80	0.04	15		95%	5%	2.5%	2.14	0.010	0.022	6.778	6.822	
16														

Figure: Ex-7

	A	B	C	D	E	F	G
1	Hb A1C	Glucose					
2	5.5	110					
3	5.5	118					
4	5.8	115				Hb A1C	Glucose
5	6.1	120			Mean	7.513	138.733
6	6.5	125			SD	1.61328	17.7179
7	6.8	146			Count	15	15
8	7.1	135			5%?	0.05	0.05
9	7.4	140			t	2.14479	2.14479
10	7.7	145			SE	0.41655	4.57474
11	8.0	145			Margin	0.8934	9.81184
12	8.0	150			Min	6.620	128.921
13	8.3	147			Max	8.407	148.545
14	9.0	155					
15	10.0	160					
16	11.0	170					

Exercise 8

8. Estimating Means

8.1. Calculate in column F the 95% confidence intervals (or margins of error) for HbA1C readings.

8.2. Calculate in column G the 95% confidence intervals (or margins of error) for glucose readings.

8.2. Use a name for the alpha error in cells F8 and G8.

Figure: Ex-8

Exercise 9

9. Estimating Proportions

9.1. Calculate in column D the SE for certain proportions (column A), given a certain sample size (cell B1).

9.2. Calculate in column E the *t*-value, given a certain confidence level (cell G1).

9.3. Calculate the minimum and maximum proportions, given a certain confidence level (cell G1).

9.4. Move the controls and watch the results.

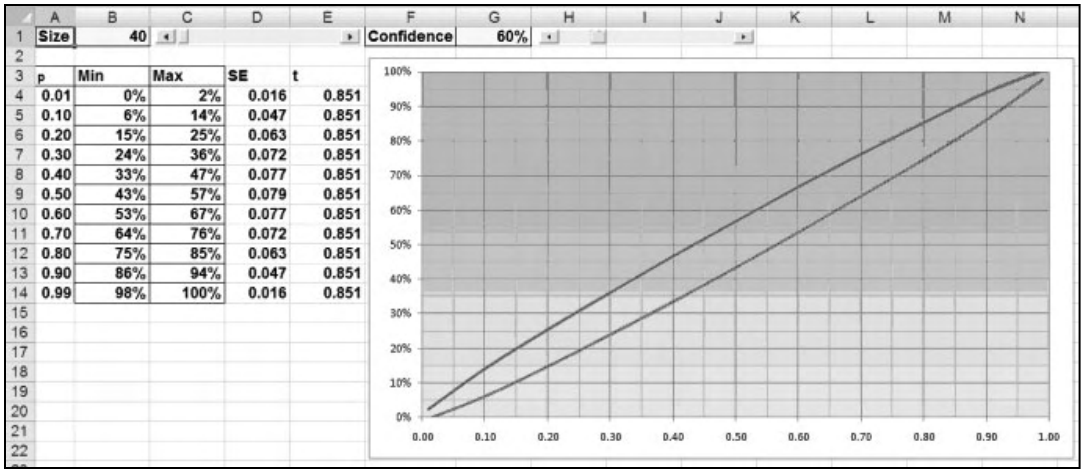


Figure: Ex-9

Exercise 10

10. Estimating Proportions

10.1. Do step 9.3. again, but this time use the function `CRITBINOM`.

10.2. Explain why the results of step 9.3. and 10.1 are slightly different.

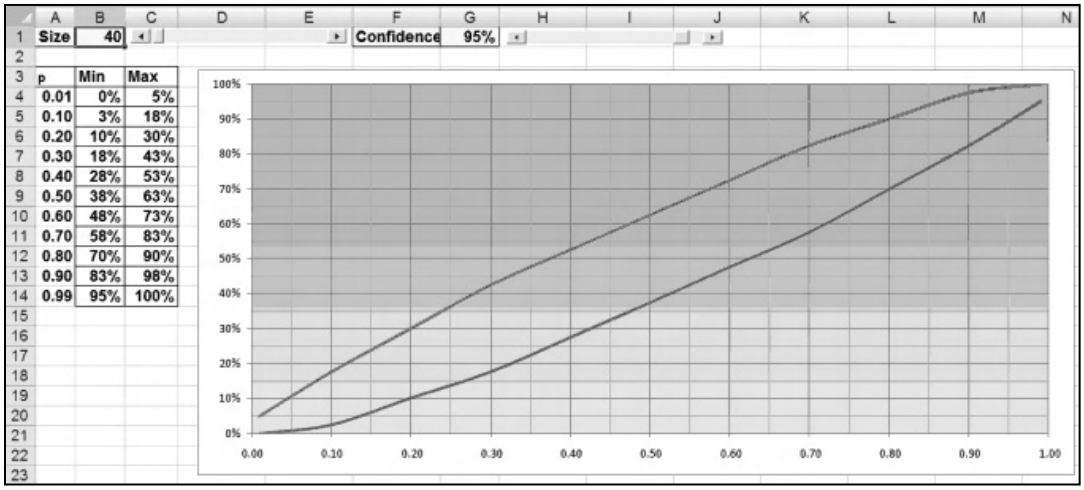


Figure: Ex-10

Exercise 11

11. Significant Means

11.1. Calculate the differences before and after treatment for each strain in column D.

11.2. Calculate the statistics in column G.

G13 =TINV(G12,G5-1)

	A	B	C	D	E	F	G	H
1		Treated	Non-treated	Diff.		Null Hypothesis: Difference is random		
2	Strain1	1.11	0.97	0.14				
3	Strain2	3.77	4.33	-0.56		mean of diff.	0.16	
4	Strain3	5.94	5.35	0.59		stdev of diff.	0.45	
5	Strain4	2.90	2.30	0.60		n	6	
6	Strain5	1.04	1.19	-0.15		level of probability	5%	
7	Strain6	4.23	3.88	0.35		stderror	0.19	
8						actual t-value	0.87	
9						critical t-value	2.57	
10						verdict on Null Hypothesis	Random	
11								
12						t-Test paired probability	42.274%	TTEST
13						actual t-value	0.87271	
14								

Figure: Ex-11

Exercise 12

12. Significant Frequencies

12.1. Calculate the expected frequencies in B12:C14.

12.2. Calculate the statistics in column G.

	A	B	C	D	E	F	G
1	Observed frequencies:					Null Hypothesis: independence	
2						CHITEST	0.646%
3		Cancer	No cancer	Total		5.0%	significant
4	>5 cig	32	12	44		1.0%	highly
5	1-5 cig	15	22	37			
6	no cig	6	9	15		critical CHIINV	5.99146455
7	Total	53	43	96		actual CHIINV	10.082974
8						CHIDIST	0.646%
9	Expected frequencies (if indep.):						
10							
11		Cancer	No cancer	Total			
12	>5 cig	24.292	19.708333	44			
13	1-5 cig	20.427	16.572917	37			
14	no cig	8.2813	6.71875	15			
15	Total	53	43	96			
16							
17		>5 cig./Cancer					
18		=Subtotal$_{>5\ cig.}$ * Subtotal$_{Cancer}$ / Total					

Figure: Ex-12

Exercise 13

13. Analysis of Variance

13.1. Use the Anova tool from the Analysis Toolpak for the table on the left.

13.2. What is your conclusion?

	A	B	C	D
1		50 ng/mL	25 ng/mL	10 ng/mL
2	Frozen	47.3	23.0	8.9
3		48.8	24.0	9.2
4		51.1	24.6	8.8
5		50.7	25.2	10.0
6		56.1	26.7	9.4
7		51.9	22.6	9.8
8		54.1	22.2	9.5
9	Fresh	50.1	21.9	10.0
10		49.9	23.2	9.1
11		49.8	23.6	9.1
12		49.6	25.7	11.1
13		53.7	27.7	9.8
14		49.2	22.5	10.2
15		50.4	22.1	9.5

SUMMARY	50 ng/mL	25 ng/mL	10 ng/mL	Total
Frozen				
Count	7	7	7	21
Sum	360	168.333333	65.6333333	593.966667
Average	51.4285714	24.047619	9.37619048	28.284127
Variance	8.96238095	2.52587302	0.19730159	322.398847
Fresh				
Count	7	7	7	21
Sum	352.733333	166.7	68.7666667	588.2
Average	50.3904762	23.8142857	9.82380952	28.0095238
Variance	2.2784127	4.56809524	0.49619048	299.432349
Total				
Count	14	14	14	
Sum	712.733333	335.033333	134.4	
Average	50.9095238	23.9309524	9.6	
Variance	5.47819292	3.28879731	0.37401709	

Figure: Ex-13

Exercise 14

14. Analysis of Variance

14.1. Use the Anova tool from the Analysis Toolpak for the table on the left.

14.2. What is your conclusion?

	A	B	C
1	Diastolic	Normal	Overweight
2	Non-diab.	75	85
3		80	80
4		83	90
5		85	95
6		65	88
7	Diabetic	85	90
8		90	95
9		95	100
10		90	105
11		86	110

SUMMARY	Normal	Overweight	Total
Non-diab.			
Count	5	5	10
Sum	388	438	826
Average	77.6	87.6	82.6
Variance	63.8	31.3	70.04444444
Diabetic			
Count	5	5	10
Sum	446	500	946
Average	89.2	100	94.6
Variance	15.7	62.5	67.15555556
Total			
Count	10	10	
Sum	834	938	
Average	83.4	93.8	
Variance	72.71111111	84.4	

Figure: Ex-13

Exercise 15

15. Analysis of Variance

15.1. Use the Anova tool from the Analysis Toolpak for the table on the left.

15.2. What is your conclusion?

	A	B	C	D	E	F
1	HbA1C after 3 months					
2	Insulin	Diet	Exercise			
3	5.8	6.0	6.2		SUMMARY	
4	6.0	6.2	6.0		*Groups*	*Count*
5	6.0	6.3	6.5		Insulin	5
6	6.2	6.5	7.0		Diet	5
7	6.3	6.2	6.8		Exercise	5
8						
9						
10					ANOVA	
11					*Source of Variation*	*SS*
12					Between Groups	0.48933333
13					Within Groups	0.964
14						
15					Total	1.45333333

* * *

Index

Shelving Category: Spreadsheets
Reader Level: Intermediate

LEARN EXCEL 2007 FROM A SCIENTIST

Excel 2007 offers significant new features over previous versions of Excel. From automatic formula replication to formula autocomplete, these features can have you working faster than ever before. Dr. Verschuuren responds with a second edition covering all of these new features, plus expanded content throughout. *- Bill Jelen, host of MrExcel.com*

Gerry Verschuuren

About the Author:

Dr. Gerard M. Verschuuren is a Microsoft Certified Professional specialized in VB, VBA, and VB.NET. He is the author of many textbooks and has more than 20 years of experience in teaching at colleges and corporations. He holds master's degrees in Biology (Human Genetics) and in Philosophy, plus a Doctorate in the Philosophy of Science from Universities in Europe. This book covers the same concepts taught during Dr. Verschuuren's Excel Seminars. He has taught this course to major pharmaceutical and other scientific companies with rave reviews! If you can't schedule Gerry at your office, then the next best thing is to buy this book.

Excel 2007 training written for scientists by a scientist. Are you tired of trying to learn Excel 2007 with examples from accounting? This book covers relevant examples from the world of science and engineering. Learn the pitfalls of Excel 2007 for scientists and how to work around them.

This book is published by:
Holy Macro! Books
13386 Judy Ave NW
Uniontown OH 44685
$29.95 USA | $ 29.95 CAN

ISBN 978-1-932802-35-1
52995 >
9 781932 802351

WOWIO® is proud to have sponsored this ebook for you. If you would like to know more about us, visit us at…

www.wowio.com