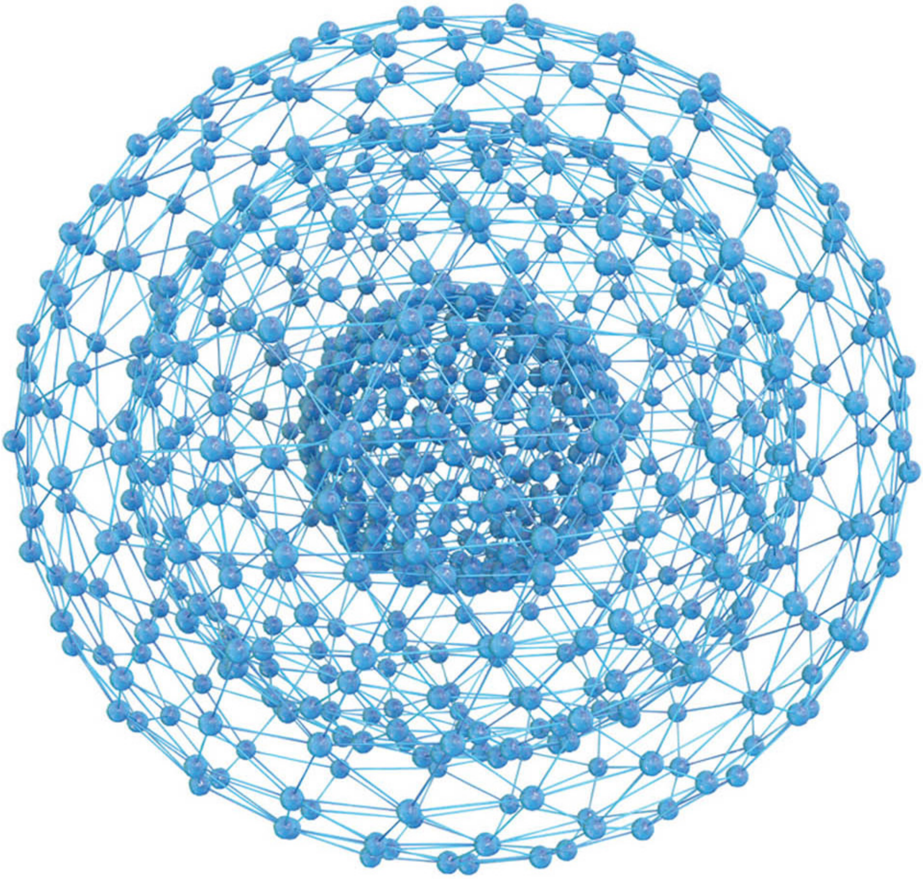
REGRESSION ANALYSIS

Microsoft® Excel®



QUE®

Conrad Carlberg

FREE SAMPLE CHAPTER

SHARE WITH OTHERS

Regression Analysis Microsoft® Excel®

Conrad Carlberg

QUE®

800 East 96th Street,
Indianapolis, Indiana 46240 USA

Contents at a Glance

Regression Analysis Microsoft® Excel®

ISBN-13: 978-0-7897-5655-8
ISBN-10: 0-7897-5655-2

Library of Congress Control Number: 2016930324

Printed in the United States of America

First Printing April 2016

Trademarks

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Que Publishing cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and Disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The information provided is on an “as is” basis. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

Special Sales

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearson.com

Editor-in-Chief
Greg Wiegand

Acquisitions Editor
Joan Murray
Trina MacDonald

Development Editor
Mark Renfrow

Managing Editor
Sandra Schroeder

Project Editor
Seth Kerney
Mandie Frank

Copy Editor
Paula Lowell

Indexer
WordWise Publishing Services

Proofreader
Debbie Williams

Technical Editor
Bill Pate

Editorial Assistant
Cindy Teeters

Cover Designer
Chuti Prasertsith

Compositor
codeMantra

Contents

About the Author

Conrad Carlberg (www.conradcarlberg.com) is a nationally recognized expert on quantitative analysis and on data analysis and management applications such as Microsoft Excel, SAS, and Oracle. He holds a Ph.D. in statistics from the University of Colorado and is a many-time recipient of Microsoft's Excel MVP designation.

Carlberg is a Southern California native. After college he moved to Colorado, where he worked for a succession of startups and attended graduate school. He spent two years in the Middle East, teaching computer science and dodging surly camels. After finishing graduate school, Carlberg worked at US West (a Baby Bell) in product management and at Motorola.

In 1995 he started a small consulting business that provides design and analysis services to companies that want to guide their business decisions by means of quantitative analysis—approaches that today we group under the term "analytics." He enjoys writing about those techniques and, in particular, how to carry them out using the world's most popular numeric analysis application, Microsoft Excel.

Acknowledgments

I thank Joan Murray for the skill and tact she employed in seeing this book through from a proposed table of contents to a bound volume (or, if your preferences in media run that way, an electronic gadget). Bill Pate provided a careful technical edit, not a straightforward task when the platform gets as idiosyncratic as Excel does from time to time. Paula Lowell's copy edit unmixed my metaphors, got my subjects to reach agreement with my verbs, and gently pointed out those occasions when I accidentally duplicated figure captions. Seth Kerney, Mandie Frank and Mark Renfrow kept the whole thing from careening out of control, and I'm the sole owner of any errors remaining in this book.

We Want to Hear from You!

As the reader of this book, *you* are our most important critic and commentator. We value your opinion and want to know what we're doing right, what we could do better, what areas you'd like to see us publish in, and any other words of wisdom you're willing to pass our way.

We welcome your comments. You can email or write to let us know what you did or didn't like about this book—as well as what we can do to make our books better.

Please note that we cannot help you with technical problems related to the topic of this book.

When you write, please be sure to include this book's title and author as well as your name and email address. We will carefully review your comments and share them with the author and editors who worked on the book.

Email: feedback@quepublishing.com

Mail: Que Publishing
ATTN: Reader Feedback
800 East 96th Street
Indianapolis, IN 46240 USA

Reader Services

Register your copy of Regression Analysis Microsoft Excel at quepublishing.com for convenient access to downloads, updates, and corrections as they become available. To start the registration process, go to quepublishing.com/register and log in or create an account*. Enter the product ISBN, 9780789756558, and click Submit. Once the process is complete, you will find any available bonus content under Registered Products.

*Be sure to check the box that you would like to hear from us in order to receive exclusive discounts on future editions of this product.

This page intentionally left blank

INTRODUCTION

Like a lot of people, I slogged through my first undergraduate classes in inferential statistics. I'm not talking here about the truly basic, everyday stats like averages, medians, and ranges. I'm talking about things you don't commonly run into outside the classroom, like randomized block designs and the analysis of variance.

I hated it. I didn't understand it. Assistant professors and textbooks inflicted formulas on us, formulas that made little sense. We were supposed to pump data through the formulas, but the results had mysterious names like "mean square within." All too often, the formulas appeared to bear no relationship to the concept they were supposed to quantify. Quite a bit later I came to understand that those formulas were "calculation formulas," meant to be quicker to apply, and less error-prone, than the more intuitively useful definitional formulas.

Eventually I came to understand why the analysis of variance, or ANOVA, is used to evaluate the differences between means—as counterintuitive as that sounded—but all those sums of squares between and sums of squares within and degrees of freedom just did not make sense. I knew that I had to calculate them to satisfy a requirement, and I knew how to do so, but I did not understand why.

Eventually I came across a book on regression analysis. Another student recommended it to me—it had clarified many of the issues that had confused him and that were still confusing me. The book, now long out of print, discussed the analysis of variance and covariance in terms of regression analysis. It resorted to computer analysis where that made sense. It stressed correlations and proportions of shared variance in its explanations. Although it also discussed sums of squares and mean squares, the book talked about them principally to help show the relationship between conventional Fisherian analysis and the regression approach.

The concepts began to clarify for me and I realized that they had been there all the time, but they were hidden behind the arcane calculations of ANOVA. Those calculations were used, and taught, because they were developed during the early twentieth century, when twenty-first century computing power wasn't merely hard to find, it just didn't exist. It was much easier to compute sums of squares (particularly using calculation formulas) than it was to calculate the staples of regression analysis, such as multiple correlations and squared semipartials. You didn't have to find the inverse of a matrix in traditional ANOVA, as was once necessary in regression. (Calculating by hand the inverse of any matrix larger than 3 × 3 is a maddening experience.)

Today, all those capabilities exist in Excel worksheets, and they make the concepts behind the analysis of variance much more straightforward. Furthermore, the Excel worksheet application makes things much easier than was hinted at in that book I read. The book was written long before Excel first emerged from its early shrink-wrap, and I shake my head that once upon a time it was necessary to pick individual pieces from the inverse of a matrix and fool around with them to get a result. Today, you can get the same result in an Excel worksheet just by combining fixed and relative addressing properly.

We still rely heavily on the analysis of variance and covariance in various fields of research, from medical and pharmaceutical studies to financial analysis and econometrics, from agricultural experiments to operations research. Understanding the concepts is important in those fields—and I maintain that understanding comes a lot easier from viewing the problems through the prism of regression than through that of conventional ANOVA.

More important, I think, is that understanding the concepts that you routinely use in regression makes it much easier to understand even more advanced methods such as logistic regression and factor analysis. Those techniques expand your horizons beyond the analysis of one-variable-at-a-time methods. They help you move into areas that involve latent, unobserved factors and multinomial dependent variables. The learning curve is much steeper in principal components analysis if you don't already have the concept of shared variance in your hip pocket.

And that's why I've written this book. I've had enough experience, first as a suit and then in my own consulting practice, with inferential statistics to know how powerful a tool it can be, if used correctly. I've also been using Excel to that end for more than 20 years. Some deride Excel as a numeric analysis application. I think they're wrong. On the other hand, Microsoft's history as Excel's publisher is, well, checkered. Not long ago a colleague forwarded to me an email in which his correspondent wondered, a little plaintively, whether it was "safe" to use Excel's statistical functions. At the time I was finishing this book up, and much of the book has to do with the use of Excel's LINEST() worksheet function. Here's what I wrote back:

> The question of whether it's "safe" to use Excel for statistical analysis is a messy one. Microsoft is at fault to some degree, and those who rant that it's dangerous to use Excel for statistical analysis share that fault.
>
> Since 1995, MS has done nothing to improve the Data Analysis add-in (aka the Analysis Toolpak) other than to convert it from the old V4 macro language to VBA.

That's a shame, because the add-in has plenty of problems that could easily be corrected. But the add-in is not Excel any more than the old Business Planner add-in is Excel. Nevertheless, I've seen plenty of papers published both privately and in refereed journals that rightly complain about the add-in's statistical tools, and then lay the blame on the actual application.

There were, through either 2003 or 2007—I can't now recall which—two principal problems with LINEST(). One had to do with the way that the regression and residual sums of squares were calculated when LINEST()'s third, *const* argument is set to FALSE. This was known as early as 1995, but MS didn't fix it until much later.

Another was that LINEST() solved what are termed the "normal equations" using matrix algebra—long the preferred method in statistical apps. But on rare occasions it's possible for multicollinearity (the presence of strong correlations among the predictor variables) to result in a matrix with a zero determinant. Such a matrix cannot be inverted, and that makes it impossible to return LINEST()'s usual results. In 2003 or 2007, MS fixed that by replacing it with something called *QR decomposition*.

But the multicollinearity problem caused LINEST() to return the #NUM! error value. No one could be led down an unsafe, dangerous path by that. And the problem with the third, *const* argument resulted in such things as a negative R-squared. Only someone utterly untutored in regression analysis could be misled by a negative R-squared. It cannot come about legitimately, so something must be wrong somewhere.

Finally, various bread-and-butter statistical functions in Excel have been improved to enhance their accuracy when they're pointed at really extreme values. This is useful—more accuracy is always better than less accuracy. But it's an instance of what Freud called the "narcissism of small differences." If I'm a biostatistician and I'm called upon to make a decision based on a difference between 10^–16 and 10^–17, I'm going to replicate the experiment. The difference is too small, both substantively and technically, to use as the basis for an important decision—regardless of whether I'm using SAS, R, or Excel.

Which brings me to the Chicken Little alarmists who scare people with lengthy screeds regarding this stuff. Badmouthing sound statistical applications has a long, dishonorable history. When I was still in school, other students who had sweated blood to learn an application named BMD said that it was a bad idea to use a different application. They said the competing app wasn't accurate, but their real motive was to prevent the erosion of their own hard-acquired expertise—more precisely, the perception of that expertise. (The new, competing application was SPSS.)

I spend some ink in the introduction to my book *Statistical Analysis Excel 2013*, and in its sixth chapter, on these and closely related matters. If it's unsafe to use Excel for statistical analysis, the danger lies in the use of an accurate tool by someone who hasn't a clue what he's doing, either with inferential statistics or with Excel.

That's my screed for 2016. I hope you enjoy this book as much as I enjoyed revisiting old friends.

This page intentionally left blank

Using Regression to Test Differences Between Group Means

7

IN THIS CHAPTER

This is a book about regression analysis. Nevertheless, I'm going to start this chapter by discussing different scales of measurement. When you use regression analysis, your *predicted* (or *outcome*, or *dependent*) variable is nearly always measured on an interval or ratio scale, one whose values are numeric quantities. Your *predictor* (or *independent*, or *regressor*) variables are also frequently measured on such numeric scales.

However, the predictor variables can also represent nominal or category scales. Because functions such as Excel's LINEST() do not respond directly to predictors with values such as STATIN and PLACEBO, or REPUBLICAN and DEMOCRAT, you need a system to convert those nominal values to numeric values that LINEST() can deal with.

The system you choose has major implications for the information you get back from the analysis. So I'll be taking a closer look at some of the underlying issues that inform your choice.

It will also be helpful to cover some terminology issues early on. This book's first six chapters have discussed the use of regression analysis to assess the relationships between variables measured on an interval or a ratio scale. There are a couple of reasons for that:

- Discussing interval variables only allows us to wait until now to introduce the slightly greater complexity of using regression to assess differences between groups.

- Most people who have heard of regression analysis at all have heard of it in connection with prediction and explanation: for example, predicting weight from known height. That sort of usage *tends* to imply interval or ratio variables as both the predicted variable and the predictor variables.

With this chapter we move into the use of regression analysis to analyze the influence of nominal variables (such as make of car or type of medical treatment) on interval variables (such as gas mileage or levels of indicators in blood tests). That sort of assessment tends to focus on the effects of belonging to different groups upon variables that quantify the outcome of group membership (gas mileage for different auto makes or cholesterol levels after different medical treatments).

We get back to the effects of interval variables in Chapter 8, "The Analysis of Covariance," but in this chapter I'll start referring to what earlier chapters called *predicted variables* as *outcome variables*, and what I have called *predictor variables* as *factors*. Lots of theorists and writers prefer terms other than *outcome variable*, because it implies a cause-and-effect relationship, and inferring that sort of situation is a job for your experimental design, not your statistical analysis. But as long as that's understood, I think we can get along with *outcome variable*—at least, it's less pretentious than some of its alternatives.

Dummy Coding

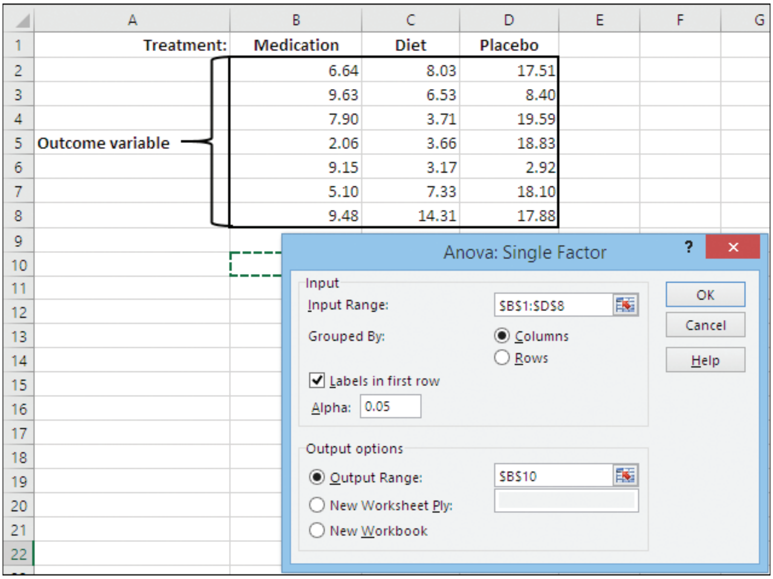
Perhaps the simplest approach to coding a nominal variable is termed *dummy coding*. I don't mean the word "simplest" to suggest that the approach is underpowered or simple-minded. For example, I prefer dummy coding in logistic regression, where it can clarify the interpretation of the coefficients used in that method.

Dummy coding can also be useful in standard linear regression when you want to compare one or more treatment groups with a comparison or *control* group.

An Example with Dummy Coding

Figures 7.1 and 7.2 show how the data from a small experiment could be set up for analysis by an application that returns a traditional analysis of variance, or *ANOVA*.

In ANOVA jargon, a variable whose values constitute the different conditions to which the subjects are exposed is called a *factor*. In this example, the factor is Treatment. The different values that the Treatment factor can take on are called *levels*. Here, the levels are the three treatments: Medication, Diet, and Placebo as a means of lowering amounts of an undesirable component in the blood.



	A	B	C	D	E	F	G
1	Treatment:	Medication	Diet	Placebo			
2		6.64	8.03	17.51			
3		9.63	6.53	8.40			
4		7.90	3.71	19.59			
5	Outcome variable	2.06	3.66	18.83			
6		9.15	3.17	2.92			
7		5.10	7.33	18.10			
8		9.48	14.31	17.88			

Anova: Single Factor

Input

Input Range: $B$1:$D$8

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

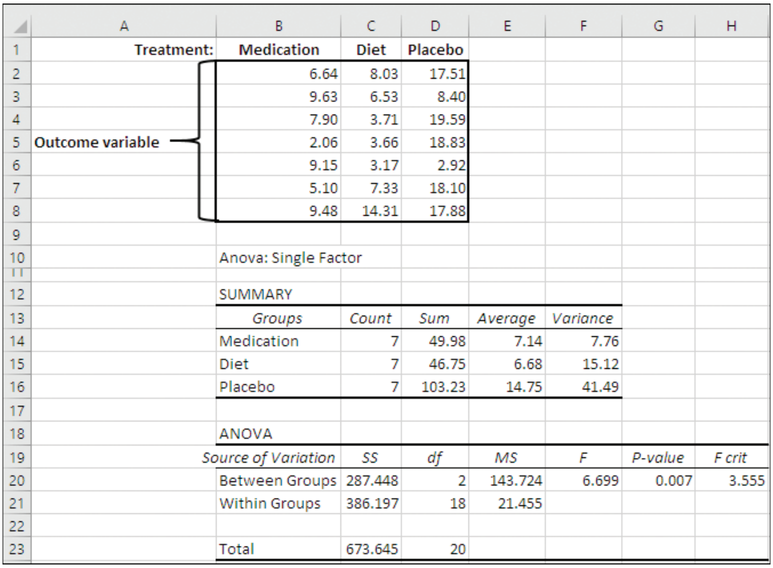
Output options

● Output Range: $B$10

○ New Worksheet Ply:

○ New Workbook

Figure 7.1
The Data Analysis tool requires that the factor levels occupy different columns or different rows.



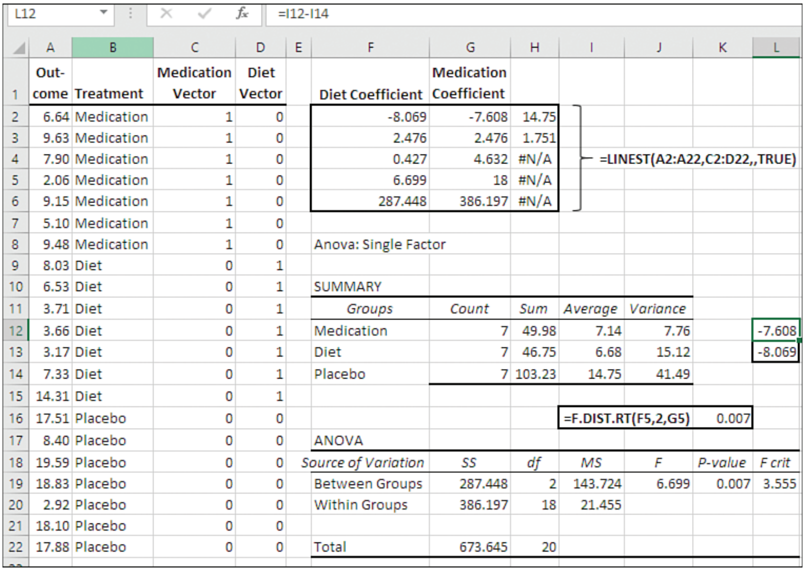
	A	B	C	D	E	F	G	H
1	Treatment:	Medication	Diet	Placebo				
2		6.64	8.03	17.51				
3		9.63	6.53	8.40				
4		7.90	3.71	19.59				
5	Outcome variable	2.06	3.66	18.83				
6		9.15	3.17	2.92				
7		5.10	7.33	18.10				
8		9.48	14.31	17.88				
9								
10		Anova: Single Factor						
11								
12		SUMMARY						
13		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
14		Medication	7	49.98	7.14	7.76		
15		Diet	7	46.75	6.68	15.12		
16		Placebo	7	103.23	14.75	41.49		
17								
18		ANOVA						
19		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
20		Between Groups	287.448	2	143.724	6.699	0.007	3.555
21		Within Groups	386.197	18	21.455			
22								
23		Total	673.645	20				

Figure 7.2
If you choose Labels in First Row in the dialog box, the output associates the summary statistics with the label.

Excel’s Data Analysis add-in includes a tool named *ANOVA: Single Factor*. To operate correctly, the data set must be arranged as in the range B2:C8 of Figure 7.2. (Or it may be turned 90 degrees, to have different factor levels in different rows and different subjects in different columns.) With the data laid out as shown in the figure, you can run the

ANOVA: Single Factor tool and in short order get back the results shown in the range A12:H23. The Data Analysis tool helpfully provides descriptive statistics as shown in B14:F16.

Figure 7.3 has an example of how you might use dummy coding to set up an analysis of the same data set by means of regression analysis via dummy coding.



L12 =I12-I14

	A	B	C	D	E	F	G	H	I	J	K	L
1	Out-come	Treatment	Medication Vector	Diet Vector		Diet Coefficient	Medication Coefficient					
2	6.64	Medication	1	0		-8.069	-7.608	14.75				
3	9.63	Medication	1	0		2.476	2.476	1.751				
4	7.90	Medication	1	0		0.427	4.632	#N/A	=LINEST(A2:A22,C2:D22,,TRUE)			
5	2.06	Medication	1	0		6.699	18	#N/A				
6	9.15	Medication	1	0		287.448	386.197	#N/A				
7	5.10	Medication	1	0								
8	9.48	Medication	1	0		Anova: Single Factor						
9	8.03	Diet	0	1								
10	6.53	Diet	0	1		SUMMARY						
11	3.71	Diet	0	1		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
12	3.66	Diet	0	1		Medication	7	49.98	7.14	7.76		-7.608
13	3.17	Diet	0	1		Diet	7	46.75	6.68	15.12		-8.069
14	7.33	Diet	0	1		Placebo	7	103.23	14.75	41.49		
15	14.31	Diet	0	1								
16	17.51	Placebo	0	0					=F.DIST.RT(F5,2,G5)		0.007	
17	8.40	Placebo	0	0		ANOVA						
18	19.59	Placebo	0	0		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
19	18.83	Placebo	0	0		Between Groups	287.448	2	143.724	6.699	0.007	3.555
20	2.92	Placebo	0	0		Within Groups	386.197	18	21.455			
21	18.10	Placebo	0	0								
22	17.88	Placebo	0	0		Total	673.645	20				

Figure 7.3 One minor reason to prefer the regression approach is that you use standard Excel layouts for the data.

When you use any sort of coding there are a couple of rules to follow. These are the rules that apply to dummy coding:

- You need to reserve as many columns for new data as the factor has levels, minus 1. Notice that this is the same as the number of degrees of freedom for the factor. With three levels, as in the present example, that's 3 – 1, or 2. It's useful to term these columns *vectors*.
- Each vector represents one level of the factor. In Figure 7.3, Vector 1 represents Medication, so every subject who receives the medication gets a 1 on Vector 1, and everyone else receives a 0 on that vector. Similarly, every subject receives a 0 on Vector 2 except those who are treated by Diet—they get a 1 on Vector 2.
- Subjects in one level, which is often a control group, receive a 0 on all vectors. In Figure 7.3, this is the case for those who take a placebo.

With the data laid out as shown in the range A2:D22 in Figure 7.3, array-enter this LINEST() function in a blank range five rows high and three columns wide, such as F2:H6 in the figure:

=LINEST(A2:A22,C2:D22,,TRUE)

Don't forget to array-enter the formula with the keyboard combination Ctrl+Shift+Enter. The arguments are as follows:

- The first argument, the range A2:A22, is the address that contains the outcome variable. (Because the description of this study suggests that it's a true, controlled experiment, it's not misleading to refer to the levels of a given component in the blood as an outcome variable, thus implying cause and effect.)
- The second argument, the range C2:D22, is the address that contains the vectors that indicate which level of the factor a subject belongs to. In other experimental contexts you might refer to these as *predictor variables*.
- The third argument is omitted, as indicated by the consecutive commas with nothing between them. If this argument is TRUE or omitted, Excel is instructed to calculate the regression equation's constant normally. If the argument is FALSE, Excel is instructed to set the constant to 0.0.
- The fourth argument, TRUE, instructs Excel to calculate and return the third through fifth rows of the results, which contain summary information, mostly about the reliability of the regression equation.

In Figure 7.3 I have repeated the results of the traditional ANOVA from Figure 7.2, to make it easier to compare the results of the two analyses. Note these points:

- The sum of squares regression and the sum of squares residual from the LINEST() results in cells F6 and G6 are identical to the sum of squares between groups and the sum of squares within groups returned by the Data Analysis add-in in cells G19 and G20.
- The degrees of freedom for the residual in cell G5 is the same as the degrees of freedom within groups in cell H20. Along with the sums of squares and knowledge of the number of factor levels, this enables you to calculate the mean square between and the mean square within if you want.
- The F-ratio returned in cell F5 by LINEST() is identical to the F-ratio reported by the Data Analysis add-in in cell J19.
- The constant (also termed the intercept) returned by LINEST() in cell H2 is identical to the mean of the group that's assigned codes of 0 throughout the vectors. In this case that's the Placebo group: Compare the value of the constant in cell H2 with the mean of the Placebo group in cell I14. (That the constant equals the group with codes of 0 throughout is true of dummy coding, not effect or orthogonal coding, discussed later in this chapter.)

The regression coefficients in cells F2 and G2, like the t-tests in Chapter 6, express the differences between group means. In the case of dummy coding, the difference is between the group assigned a code of 1 in a vector and the group assigned 0's throughout.

For example, the difference between the means of the group that took a medication and the group that was treated by placebo is 7.14 – 14.75 (see cells I12 and I14). That difference equals –7.608, and it's calculated in cell L12. That's the regression coefficient for Vector 1, returned by LINEST() in cell G2. Vector 1 identifies the Medication group with a 1.

Similarly, the difference between the mean of the group treated by diet and that treated by placebo is 6.68 – 14.75 (see cells I13 and I14). The difference equals –8.069, calculated in cell L13, which is also the regression coefficient for Vector 2.

NOTE

It's here that LINEST( )'s peculiarity in the order of the coefficients shows up again. Recall that if predictor variables A, B, and C appear in that left-to-right order on the worksheet, they appear in the left-to-right order C, B, and then A in the LINEST( ) results.

So in Figure 7.3, the vector that represents the Medication treatment is in column C, and the vector that represents Diet is to its right, in column D. However, LINEST( ) puts the regression coefficient for Medication in cell G2, and the regression coefficient for Diet to its *left*, in cell F2. The potential for confusion is clear and it's a good idea to label the columns in the LINEST( ) result to show which variable each coefficient refers to.

One bit of information that LINEST() does not provide you is statistical significance of the regression equation. In the context of ANOVA, where we're evaluating the differences between group means, that test of statistical significance asks whether *any* of the mean differences is large enough that the null hypothesis of no difference between the means in the population can be rejected. The F-ratio, in concert with the degrees of freedom for the regression and the residual, speaks to that question.

You can determine the probability of observing a given F-ratio if the null hypothesis is true by using Excel's F.DIST.RT() function. In this case, you use it in this way (it's also in cell K16):

```
=F.DIST.RT(F5,2,G5)
```

Notice that the value it returns, 0.007, is identical to that returned in cell K19 by the Data Analysis add-in's ANOVA: Single Factor tool. If there is no difference, measured by group means, in the populations of patients who receive the medication, or whose diet was controlled, or who took a placebo, then the chance of observing an F-ratio of 6.699 is 7 in 1,000. It's up to you whether that's rare enough to reject the null hypothesis. It would be for most people, but a sample of 21 is a very small sample, and that tends to inhibit the generalizability of the findings—that is, how confidently you can generalize your observed outcome from 21 patients to your entire target population.

Populating the Vectors Automatically

So: What does all this buy you? Is there enough advantage to running your ANOVA using regression in general and LINEST() in particular that it justifies any extra work involved?

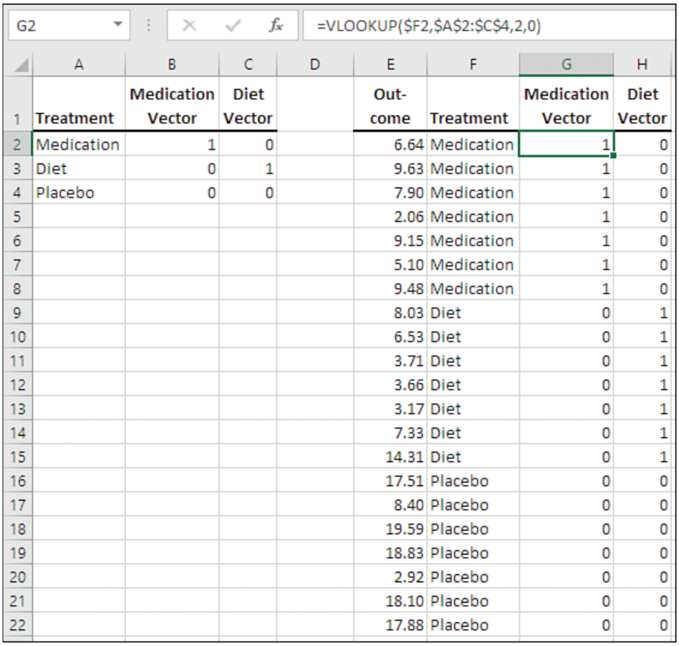
I think it does, and the decision isn't close. First, what are the steps needed to prepare for the Data Analysis tool, and what steps to prepare a regression analysis?

To run the Data Analysis ANOVA: Single Factor tool, you have to arrange your data as shown in the range B1:D8 in Figure 7.2. That's not a natural sort of arrangement of data in either a true database or in Excel. A list or table structure of the sort shown in A1:D22 of Figure 7.3 is much more typical, and as long as you provide columns C and D for the dummy 0/1 codes, it's ready for you to point LINEST() at.

To prepare for a regression analysis, you do need to supply the 0s and 1s in the proper rows and the proper columns. This is *not* a matter of manually entering 0s and 1s one by one. Nor is it a matter of copying and pasting values or using Ctrl+Enter on a multiple selection. I believe that the fastest, and most accurate, way of populating the coded vectors is by way of Excel's VLOOKUP() function. See Figure 7.4.

To prepare the ground, enter a key such as the one in the range A2:C4 in Figure 7.4. That key should have as many columns and as many rows as the factor has levels. In this case, the factor has three levels (Medication, Diet, and Placebo), so the key has three columns, and there's one row for each level. It's helpful but not strictly necessary to provide column headers, as is done in the range A1:C1 of Figure 7.4.

Figure 7.4
Practice in the use of the VLOOKUP( ) function can save you considerable time in the long run.



G2 =VLOOKUP($F2,$A$2:$C$4,2,0)

	A	B	C	D	E	F	G	H
1	Treatment	Medication Vector	Diet Vector		Out-come	Treatment	Medication Vector	Diet Vector
2	Medication	1	0		6.64	Medication	1	0
3	Diet	0	1		9.63	Medication	1	0
4	Placebo	0	0		7.90	Medication	1	0
5					2.06	Medication	1	0
6					9.15	Medication	1	0
7					5.10	Medication	1	0
8					9.48	Medication	1	0
9					8.03	Diet	0	1
10					6.53	Diet	0	1
11					3.71	Diet	0	1
12					3.66	Diet	0	1
13					3.17	Diet	0	1
14					7.33	Diet	0	1
15					14.31	Diet	0	1
16					17.51	Placebo	0	0
17					8.40	Placebo	0	0
18					19.59	Placebo	0	0
19					18.83	Placebo	0	0
20					2.92	Placebo	0	0
21					18.10	Placebo	0	0
22					17.88	Placebo	0	0

The first column—in this case, A2:A4—should contain the labels you use to identify the different levels of the factor. In this case those levels are shown for each subject in the range F2:F22.

You can save a little time by selecting the range cells in the key starting with its first row and its *second* column—so, B2:C4. Type 0, hold down the Ctrl key and press Enter. All the selected cells will now contain the value 0.

In the same row as a level's label, enter a 1 in the column that will represent that level. So, in Figure 7.4, cell B2 gets a 1 because column B represents the Medication level, and cell C3 gets a 1 because column C represents the Diet level. There will be no 1 to represent Placebo because we'll treat that as a control or comparison group, and so it gets a 0 in each column.

With the key established as in A2:C4 of Figure 7.4, select the first row of the first column where you want to establish your matrix of 0's and 1's. In Figure 7.4, that's cell G2. Enter this formula:

```
=VLOOKUP($F2,$A$2:$C$4,2,0)
```

Where:

- $F2 is the label that you want to represent with a 1 or a 0.
- $A$2:$C$4 contains the key (Excel terms this a *table lookup*).
- 2 identifies the column in the key that you want returned.
- 0 specifies that an exact match for the label is required, and that the labels in the first column of the key are not necessarily sorted.

I've supplied dollar signs where needed in the formula so that it can be copied to other columns and rows without disrupting the reference to the key's address, and to the column in which the level labels are found.

Now copy and paste cell G2 into H2 (or use the cell's selection handle to drag it one column right). In cell H2, edit the formula so that VLOOKUP()'s third argument has a 3 instead of a 2—this directs Excel to look in the key's third column for its value.

Finally, make a multiple selection of cells G2 and H2, and drag them down into G3:H22. This will populate columns G and H with the 0's and 1's that specify which factor level each record belongs to.

You can now obtain the full LINEST() analysis by selecting a range such as A6:C10, and array-entering this formula:

```
=LINEST(E2:E22,G2:H22,,TRUE)
```

By the way, you might find it more convenient to switch the contents of columns E and F in Figure 7.4. I placed the treatment labels in column F to ease the comparison of the labels with the dummy codes. If you swap columns E and F, you might find the LINEST() formula easier to handle. You'll also want to change the first VLOOKUP() formula from this:

```
=VLOOKUP($F2,$A$2:$C$4,2,0)
```

to this:

```
=VLOOKUP($E2,$A$2:$C$4,2,0)
```

The Dunnett Multiple Comparison Procedure

When you have completed a test of the differences between the means of three or more groups—whether by way of traditional ANOVA methods or a regression approach—you have learned the probability that *any* of the means in the population is different from any of the remaining means in the population. You have not learned *which* mean or means is different from others.

Statisticians studied this issue and wrote an intimidatingly comprehensive literature on the topic during the middle years of the twentieth century. The procedures they developed came to be known as *multiple comparisons*. Depending on how you count them, the list of different procedures runs to roughly ten. The procedures differ from one another in several ways, including the nature of the error involved (for example, per comparison or per experiment), the reference distribution (for example, F, t, or q), planned beforehand (a priori) or after the fact (post hoc), and on other dimensions.

If you choose to use dummy coding in a regression analysis, in preference to another coding method, it might well be because you want to compare all the groups but one to the remaining group. That approach is typical of an experiment in which you want to compare the results of two or more treatments to a control group. In the context of dummy coding, the control group is the one that receives 0's throughout the vectors that represent group membership. One result of dummy coding, as you've seen, is that the regression coefficient for a particular group has a value that is identical to the difference between the group's mean and the mean of the control group.

These procedures tend to be named for the statisticians who developed them, and one of them is called the Dunnett multiple comparison procedure. It makes a minor modification to the formula for the t-ratio. It also relies on modifications to the reference t-distribution. In exchange for those modifications, the Dunnett provides you with comparisons that have somewhat more statistical power than alternative procedures, given that you start your experiment intending to compare two or more treatments to a control.

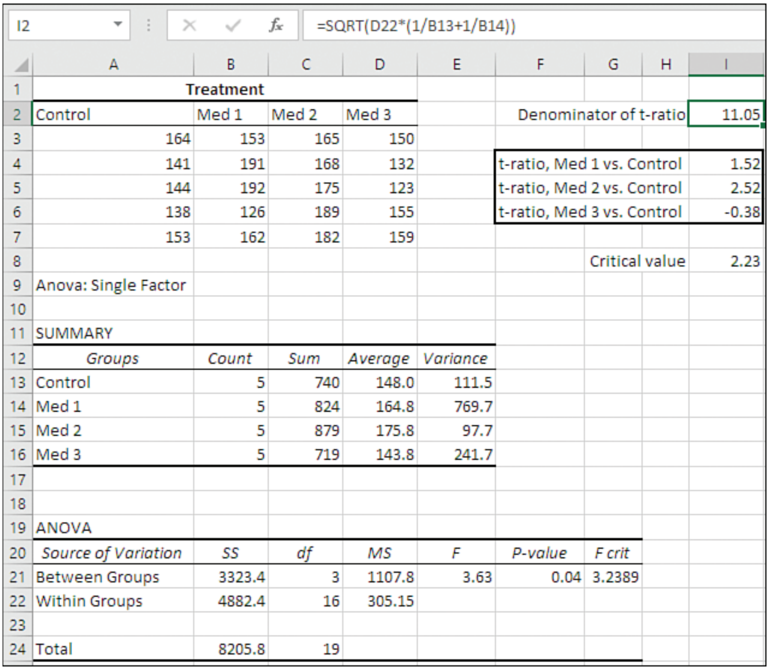
As you'll see, the calculation of the t-ratios is particularly easy when you have access to the LINEST() worksheet function. Access to the reference distribution for Dunnett's t-ratio is a little more complicated. Excel offers you direct access to, for example, the t-distribution and the F-distribution by way of its T.DIST(), T.INV(), F.DIST(), and F.INV() functions, and their derivatives due to the RT and 2T tags. But Excel does not have a DUNNETT() function that tells you the t-ratio that demarks the 95%, 99% or any other percent of the area beneath the distribution as it does for t and F.

NOTE

You cannot legitimately calculate a t-ratio using Dunnett's methods and then compare it to a standard t-distribution of the sort returned by T.DIST( ) and T.INV( ). Dunnett's t-distribution has a different shape than the "standard" t-distribution.

Although the values for Dunnett's t are not directly available in Excel, they are available on various online sites. It's easy enough to download and print the tables (they occupy two printed pages). A search using the keywords “Dunnett,” “multiple comparison,” and “tables” will locate more sites than you want, but many of them show the necessary tables. The tables also appear as an appendix in most intermediate-level, general statistics textbooks in print.

Let’s look at how you could conduct a Dunnett multiple comparison after running the Data Analysis ANOVA: Single Factor tool. See Figure 7.5.



I2 =SQRT(D22*(1/B13+1/B14))

	A	B	C	D	E	F	G	H	I
1		Treatment							
2	Control	Med 1	Med 2	Med 3		Denominator of t-ratio			11.05
3	164	153	165	150					
4	141	191	168	132		t-ratio, Med 1 vs. Control			1.52
5	144	192	175	123		t-ratio, Med 2 vs. Control			2.52
6	138	126	189	155		t-ratio, Med 3 vs. Control			-0.38
7	153	162	182	159					
8							Critical value		2.23
9	Anova: Single Factor								
10									
11	SUMMARY								
12	*Groups*	*Count*	*Sum*	*Average*	*Variance*				
13	Control	5	740	148.0	111.5				
14	Med 1	5	824	164.8	769.7				
15	Med 2	5	879	175.8	97.7				
16	Med 3	5	719	143.8	241.7				
17									
18									
19	ANOVA								
20	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*		
21	Between Groups	3323.4	3	1107.8	3.63	0.04	3.2389		
22	Within Groups	4882.4	16	305.15					
23									
24	Total	8205.8	19						

Figure 7.5
When the group sizes are equal, as here, all the comparisons’ t-ratios have the same denominator.

The data is laid out for the Data Analysis tool in A2:D7. The ANOVA: Single Factor tool in the Data Analysis add-in returns the results shown in the range A11:G24. You’ll need the group means, the Mean Square Error from the ANOVA table, and the group counts.

The first step is to calculate the denominator of the t-ratios. With equal group sizes, the same denominator is used for each t-ratio. The formula for the denominator is:

$$\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where MSE is the mean square error from the ANOVA table, shown in Figure 7.5 in cell D22. (*Mean square error* is simply another term for *mean square within* or *mean square residual*.)

When, as here, the group sizes are equal, you can also use this arithmetically equivalent formula:

$$\sqrt{2MSE/n}$$

The denominator for the t-ratios in this example is given in cell I2. It uses this formula:

=SQRT(D22∗(1/B13+1/B14))

where cell D22 contains the mean square error and cells B13 and B14 contain group counts. With all groups of the same size, it doesn't matter which group counts you use in the formula. As suggested earlier, with equal group sizes the Excel formula could also be:

=SQRT((2∗D22)/B13)

The next step is to find the difference between the mean of each treatment group and the mean of the control group, and divide those differences by the denominator. The result is one or more t-ratios. For example, here's the formula in cell I4 of Figure 7.5.

=(D14–D13)/I2

The formula divides the difference between the mean of the Med 1 group (D14) and the mean of the Control group (D13) by the denominator of the t-ratio (I2). The formulas in cells I5 and I6 follow that pattern:

I5: =(D15–D13)/I2

I6: =(D16–D13)/I2

At this point you look up the value in the Dunnett tables that corresponds to three criteria:

- The Degrees of Freedom Within from the ANOVA table (in Figure 7.5, the value 16 in cell C22)
- The total number of groups, including the control group
- The value of alpha that you selected before seeing the data from your experiment

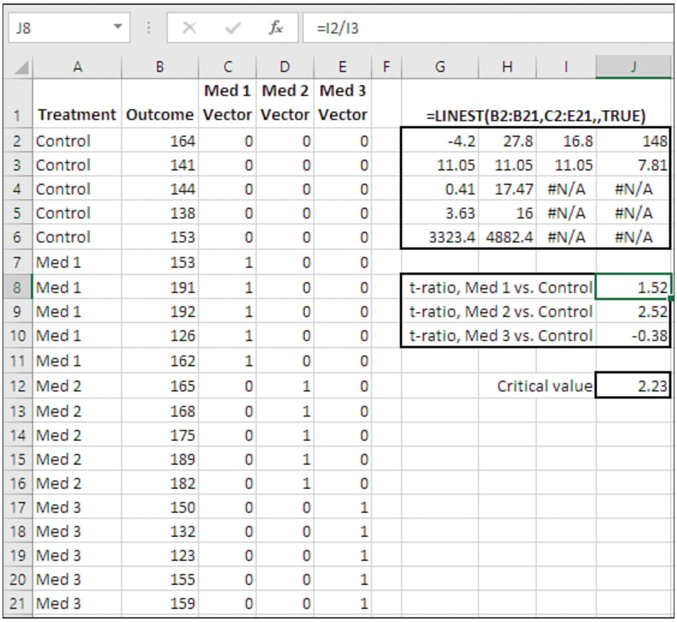
Most printed tables give you a choice of 0.05 and 0.01. That's restrictive, of course, and it delights me that Excel offers exact probabilities for any probability level you might present it for various distributions including the chi-square, the binomial, the t and the F.

For the present data set, the printed Dunnett tables give a critical value of 2.23 for four groups and 16 Degrees of Freedom Within, at the 0.05 alpha level. They give 3.05 at the 0.01 alpha level.

Because the t-ratio in cell I5, which contrasts Med 2 with Control, is the only one to exceed the critical value of 2.23, you could reject the null hypothesis of no difference in the population means for those two groups at the .05 confidence level. You could not reject it at the 0.01 confidence level because the t-ratio does not exceed the 0.01 level of 3.05.

Compare all that with the results shown in Figure 7.6.

Start by noticing that cells G2, H2, and I2, which contain the regression coefficients for Med 3 Vector, Med 2 Vector, and Med 1 Vector (in that order), express the differences between the means of the treatment groups and the control group.



J8 =I2/I3

	A	B	C	D	E	F	G	H	I	J
1	Treatment	Outcome	Med 1 Vector	Med 2 Vector	Med 3 Vector		=LINEST(B2:B21,C2:E21,,TRUE)			
2	Control	164	0	0	0		-4.2	27.8	16.8	148
3	Control	141	0	0	0		11.05	11.05	11.05	7.81
4	Control	144	0	0	0		0.41	17.47	#N/A	#N/A
5	Control	138	0	0	0		3.63	16	#N/A	#N/A
6	Control	153	0	0	0		3323.4	4882.4	#N/A	#N/A
7	Med 1	153	1	0	0					
8	Med 1	191	1	0	0		t-ratio, Med 1 vs. Control			1.52
9	Med 1	192	1	0	0		t-ratio, Med 2 vs. Control			2.52
10	Med 1	126	1	0	0		t-ratio, Med 3 vs. Control			-0.38
11	Med 1	162	1	0	0					
12	Med 2	165	0	1	0				Critical value	2.23
13	Med 2	168	0	1	0					
14	Med 2	175	0	1	0					
15	Med 2	189	0	1	0					
16	Med 2	182	0	1	0					
17	Med 3	150	0	0	1					
18	Med 3	132	0	0	1					
19	Med 3	123	0	0	1					
20	Med 3	155	0	0	1					
21	Med 3	159	0	0	1					

Figure 7.6
Much less calculation is needed when you start by analyzing the data with regression.

For example, the regression coefficient in cell H2 (27.8) is the difference between the Med 2 mean (175.8, in cell D15 of Figure 7.5) and the Control mean (148.0, in cell D13 of Figure 7.5). So right off the bat you're relieved of the need to calculate those differences. (You can, however, find the mean of the control group in the regression equation's constant, in cell J2.)

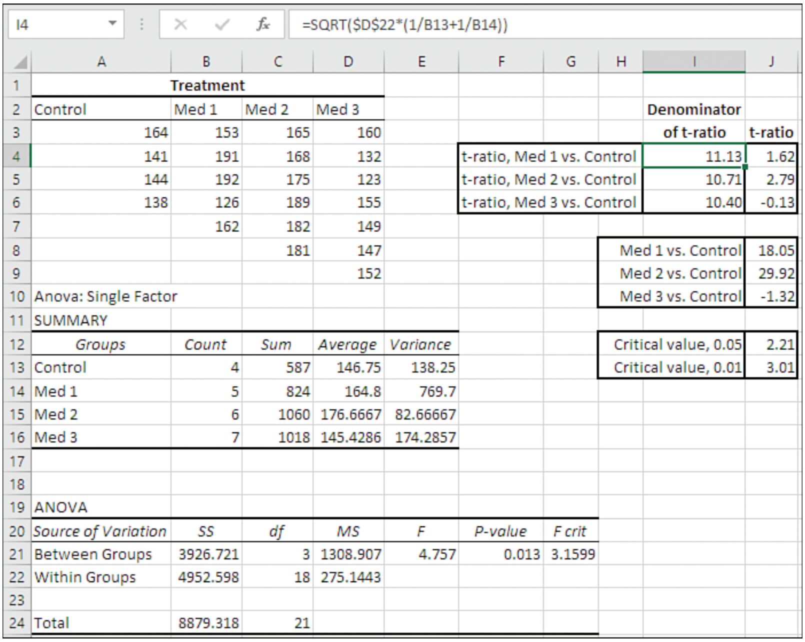
Now notice the standard errors of the regression coefficients, in cells G3, H3, and I3. They are all equal to 11.05, and in any equal-cell-size situation with dummy coding, the standard errors will all have the same value. That value is also the one calculated from the mean square error and the group sizes in Figure 7.5 (see that figure, cell I2).

So, all you need to do if you start the data analysis with LINEST() is to divide the regression coefficients by their standard errors to get the t-ratios that correspond to the Dunnett procedure. Notice that the t-ratios in cells J8, J9, and J10 are identical to those calculated in Figure 7.5, cells I4, I5, and I6.

Now let's have a look at a slightly more complicated situation, one in which you have different numbers of subjects in your groups. See Figure 7.7.

In Figure 7.7, the basic calculations are the same, but instead of using just one denominator as was done in Figure 7.5 (because the groups all had the same number of subjects), we need three denominators because the group sizes are different. The three denominators appear in the range I4:I6, and use the version of the formula given earlier:

$$\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$



I4 =SQRT($D$22*(1/B13+1/B14))

	A	B	C	D	E	F	G	H	I	J
1		Treatment								
2	Control	Med 1	Med 2	Med 3					Denominator	
3	164	153	165	160					of t-ratio	t-ratio
4	141	191	168	132		t-ratio, Med 1 vs. Control			11.13	1.62
5	144	192	175	123		t-ratio, Med 2 vs. Control			10.71	2.79
6	138	126	189	155		t-ratio, Med 3 vs. Control			10.40	-0.13
7		162	182	149						
8			181	147					Med 1 vs. Control	18.05
9				152					Med 2 vs. Control	29.92
10	Anova: Single Factor								Med 3 vs. Control	-1.32
11	SUMMARY									
12	*Groups*	*Count*	*Sum*	*Average*	*Variance*				Critical value, 0.05	2.21
13	Control	4	587	146.75	138.25				Critical value, 0.01	3.01
14	Med 1	5	824	164.8	769.7					
15	Med 2	6	1060	176.6667	82.66667					
16	Med 3	7	1018	145.4286	174.2857					
17										
18										
19	ANOVA									
20	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*			
21	Between Groups	3926.721	3	1308.907	4.757	0.013	3.1599			
22	Within Groups	4952.598	18	275.1443						
23										
24	Total	8879.318	21							

Figure 7.7
Using traditional ANOVA on a data set with unequal group sizes, you need to calculate a different denominator for each t-ratio.

So, the formulas to return the t-ratio's denominator are:

I4: =SQRT($D$22*(1/B13+1/B14))

I5: =SQRT($D$22*(1/B13+1/B15))

I6: =SQRT($D$22*(1/B13+1/B16))

Notice that the only difference between the formulas is that they alter a reference from B14 to B15 to B16, as the number of observations in the Med 1, Med 2, and Med 3 groups increases from 5 to 6 to 7. The formulas make use of the group counts returned by the Data Analysis add-in to pick up the number of observations in each treatment group.

The differences between the treatment group means and the control group mean are shown in the range J8:J10. They are the numerators for the t-ratios, which appear in the range J4:J6. Each t-ratio is the result of dividing the difference between two group means by the associated denominator, as follows:

J4: =J8/I4

J5: =J9/I5

J6: =J10/I6

In sum, when your group sizes are unequal, traditional methods have you calculate different denominators for each of the t-ratios that contrast all group means but one (here, Med 1, Med 2, and Med 3) with another mean (here, Control). Then for each pair of means, calculate the mean difference and divide by the denominator for that pair.

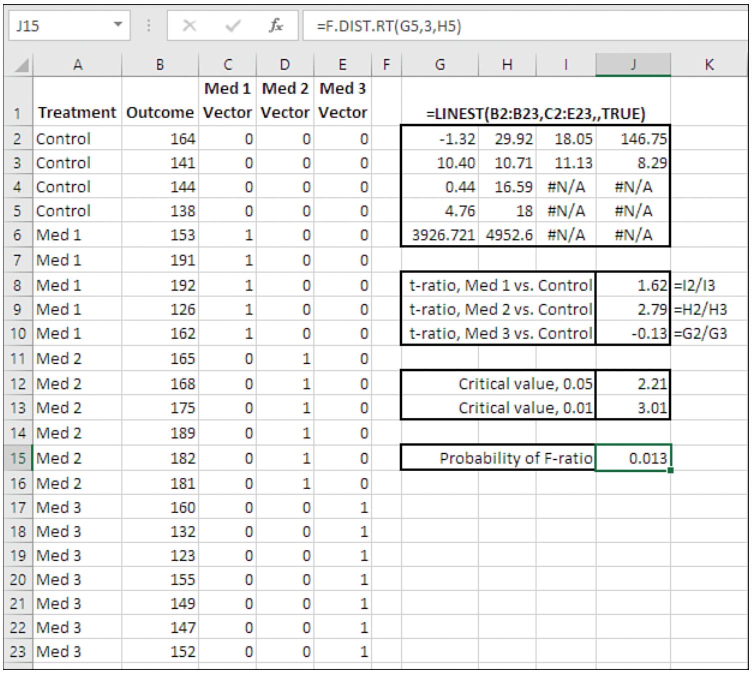
You'll also want to compare the values of the t-ratios with the values in Dunnett's tables. In this case you would want to locate the values associated with 18 within-groups degrees of freedom (from cell C22 in the ANOVA table) and 4 groups. The intersection of those values in the table is 2.21 for an alpha level of 0.05 and 3.01 for an alpha level of 0.01 (see cells J12 and J13 in Figure 7.7). Therefore, only the difference between Med 2 and Control, with a t-ratio of 2.79, is beyond the cutoff for 5% of the Dunnett t distribution, and it does not exceed the cutoff for 1% of the distribution. You can reject the null for Med 2 versus Control at the 5% level of confidence but not at the 1% level. You cannot reject the null hypothesis for the other two contrasts at even the 5% level of confidence.

Notice that the F-ratio in the ANOVA table, 4.757 in cell E21, will appear in a central F distribution with 3 and 18 degrees of freedom only 1.3% of the time. (A central F distribution in the context of an ANOVA is one in which the estimate of the population variance due to the differences among group means is equal to the estimate of the population variance due to the average within-group variance.) So the ANOVA informs you that an F-ratio of 4.757 with 3 and 18 degrees of freedom is unlikely to occur by chance if the population means equal one another.

That likelihood, 1.3%, echoes the result of the contrast of the Med 2 group with the Control group. The t-ratio for that contrast, 2.79, exceeds the critical value for 5% of the Dunnett distribution but not the critical value for 1% of the distribution. So the objective of the multiple comparison procedure, to pinpoint the difference in means that the ANOVA's F-ratio tells you must exist, has been met.

Things go a lot more smoothly if you use LINEST() instead. See Figure 7.8.

Figure 7.8
LINEST( ) calculates the mean differences and t-ratio denominators for you.



J15 =F.DIST.RT(G5,3,H5)

	A	B	C	D	E	F	G	H	I	J	K
1	Treatment	Outcome	Med 1 Vector	Med 2 Vector	Med 3 Vector		=LINEST(B2:B23,C2:E23,,TRUE)				
2	Control	164	0	0	0		-1.32	29.92	18.05	146.75	
3	Control	141	0	0	0		10.40	10.71	11.13	8.29	
4	Control	144	0	0	0		0.44	16.59	#N/A	#N/A	
5	Control	138	0	0	0		4.76	18	#N/A	#N/A	
6	Med 1	153	1	0	0		3926.721	4952.6	#N/A	#N/A	
7	Med 1	191	1	0	0						
8	Med 1	192	1	0	0		t-ratio, Med 1 vs. Control			1.62	=I2/I3
9	Med 1	126	1	0	0		t-ratio, Med 2 vs. Control			2.79	=H2/H3
10	Med 1	162	1	0	0		t-ratio, Med 3 vs. Control			-0.13	=G2/G3
11	Med 2	165	0	1	0						
12	Med 2	168	0	1	0		Critical value, 0.05			2.21	
13	Med 2	175	0	1	0		Critical value, 0.01			3.01	
14	Med 2	189	0	1	0						
15	Med 2	182	0	1	0		Probability of F-ratio			0.013	
16	Med 2	181	0	1	0						
17	Med 3	160	0	0	1						
18	Med 3	132	0	0	1						
19	Med 3	123	0	0	1						
20	Med 3	155	0	0	1						
21	Med 3	149	0	0	1						
22	Med 3	147	0	0	1						
23	Med 3	152	0	0	1						

Figure 7.8 has the same underlying data as Figure 7.7: Four groups with a different number of subjects in each. (The group membership vectors were created just as shown in Figure 7.6, using VLOOKUP(), but to save space in the figure the formulas were converted to values and the key deleted.)

This LINEST() formula is array-entered in the range G2:J6:

```
=LINEST(B2:B23,C2:E23,,TRUE)
```

Compare the regression coefficients in cells G2, H2, and I2 with the mean differences shown in Figure 7.7 (J8:J10). Once again, just as in Figures 7.5 and 7.6, the regression coefficients are exactly equal to the mean differences between the groups that have 1's in the vectors and the group that has 0's throughout. So there's no need to calculate the mean differences explicitly.

The standard errors of the regression coefficients in Figure 7.8 also equal the denominators of the t-ratios in Figure 7.7 (in the range I4:I6). LINEST() automatically takes the differences in the group sizes into account. All there's left to do is divide the regression coefficients by their standards errors, as is done in the range J8:J10. The formulas in those cells are given as text in K8:K10. But don't forget, when you label each t-ratio with verbiage that states which two means are involved, that LINEST() returns the coefficients and their standard errors backwards: Med 3 versus Control in G2:G3, Med 2 versus Control in cell H2:H3, and Med 1 versus Control in I2:I3.

Figure 7.8 repeats in J12 and J13 the critical Dunnett values for 18 within-group degrees of freedom (picked up from cell H5 in the LINEST() results) and 4 groups at the 0.05 and 0.01 cutoffs. The outcome is, of course, the same: Your choice of whether to use regression or traditional ANOVA makes no difference to the outcome of the multiple comparison procedure.

Finally, as mentioned earlier, the LINEST() function does not return the probability of the F-ratio associated with the R^2 for the full regression. That figure is returned in cell J15 by this formula:

```
=F.DIST.RT(G5,3,H5)
```

Where G5 contains the F-ratio and H5 contains the within-group (or "residual") degrees of freedom. You have to supply the second argument (here, 3) yourself: It's the number of groups minus 1 (notice that it equals the number of vectors in LINEST()'s second argument, C2:E23) also known as the degrees of freedom between in an ANOVA or degrees of freedom regression in the context of LINEST().

Effect Coding

Another type of coding, called *effect coding*, contrasts each group mean following an ANOVA with the grand mean of all the observations.

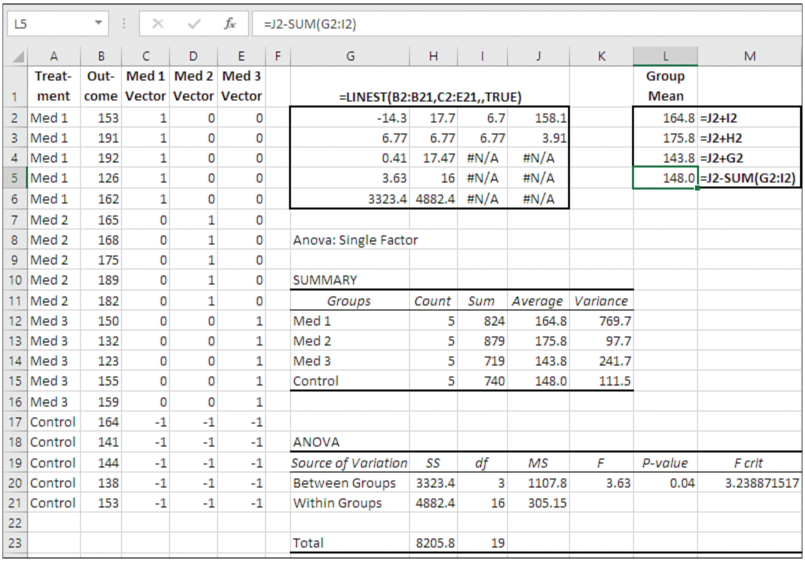
NOTE

More precisely, effect coding contrasts each group mean with the mean of all the group means. When each group has the same number of observations, the grand mean of all the observations is equal to the mean of the group means. With unequal group sizes, the two are not equivalent. In either case, though, effect coding contrasts each group mean with the mean of the group means.

This aspect of effect coding—contrasting group means with the grand mean rather than with a specified group, as with dummy coding—is due to the use of –1 instead of 0 as the code for the group that gets the same code throughout the coded vectors. Because the contrasts are with the grand mean, each contrast represents the effect of being in a particular group.

Coding with –1 Instead of 0

Let's take a look at an example before getting into the particulars of effect coding. See Figure 7.9.



L5 =J2-SUM(G2:I2)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Treat-ment	Out-come	Med 1 Vector	Med 2 Vector	Med 3 Vector		=LINEST(B2:B21,C2:E21,,TRUE)					Group Mean	
2	Med 1	153	1	0	0		-14.3	17.7	6.7	158.1		164.8	=J2+I2
3	Med 1	191	1	0	0		6.77	6.77	6.77	3.91		175.8	=J2+H2
4	Med 1	192	1	0	0		0.41	17.47	#N/A	#N/A		143.8	=J2+G2
5	Med 1	126	1	0	0		3.63	16	#N/A	#N/A		148.0	=J2-SUM(G2:I2)
6	Med 1	162	1	0	0		3323.4	4882.4	#N/A	#N/A			
7	Med 2	165	0	1	0								
8	Med 2	168	0	1	0		Anova: Single Factor						
9	Med 2	175	0	1	0								
10	Med 2	189	0	1	0		SUMMARY						
11	Med 2	182	0	1	0		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
12	Med 3	150	0	0	1		Med 1	5	824	164.8	769.7		
13	Med 3	132	0	0	1		Med 2	5	879	175.8	97.7		
14	Med 3	123	0	0	1		Med 3	5	719	143.8	241.7		
15	Med 3	155	0	0	1		Control	5	740	148.0	111.5		
16	Med 3	159	0	0	1								
17	Control	164	-1	-1	-1								
18	Control	141	-1	-1	-1		ANOVA						
19	Control	144	-1	-1	-1		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
20	Control	138	-1	-1	-1		Between Groups	3323.4	3	1107.8	3.63	0.04	3.238871517
21	Control	153	-1	-1	-1		Within Groups	4882.4	16	305.15			
22													
23							Total	8205.8	19				

Figure 7.9
The coefficients in LINEST( ) equal each group's distance from the grand mean.

Figure 7.9 has the same data set as Figure 7.6, except that the Control group has the value –1 throughout the three coded vectors, instead of 0 as in dummy coding. Some of the LINEST() results are therefore different than in Figure 7.6. The regression coefficients in Figure 7.9 differ from those in Figure 7.6, as do their standard errors. All the remaining values are the same: R^2, the standard error of estimate, the F-ratio, the residual degrees of freedom, and the regression and residual sums of squares—all those remain the same, just as they do with the third method of coding that this chapter considers, planned orthogonal contrasts.

In dummy coding, the constant returned by LINEST() is the mean of the group that's assigned 0's throughout the coded vectors—usually a control group. In effect coding, the constant is the grand mean. The constant is easy to find. It's the value in the first row (along with the regression coefficients) and in the rightmost column of the LINEST() results.

Because the constant equals the grand mean, it's easy to calculate the group means from the constant and the regression coefficients. Each coefficient, as I mentioned at the start of this section, represents the difference between the associated group's mean and the grand mean. So, to calculate the group means, add the constant to the regression coefficients. That's been done in Figure 7.9, in the range L2:L4. The formulas used in that range are given as text in M2:M4.

Notice that the three formulas add the constant (the grand mean) to a regression coefficient (a measure of the effect of being in that group, the distance of the group mean above or below the grand mean). The fourth formula in L5 is specific to the group assigned codes of –1, and it subtracts the other coefficients from the grand mean to calculate the mean of that group.

Also notice that the results of the formulas in L2:L5 equal the group means reported in the range J12:J15 by the Data Analysis add-in's ANOVA: Single Factor tool. It's also worth verifying that the F-ratio, the residual degrees of freedom, and the regression and residual sums of squares equal those reported by that tool in the range H20:K21.

Relationship to the General Linear Model

The general linear model is a useful way of conceptualizing the components of a value on an outcome variable. Its name makes it sound a lot more forbidding than it really is. Here's the general linear model in its simplest form:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

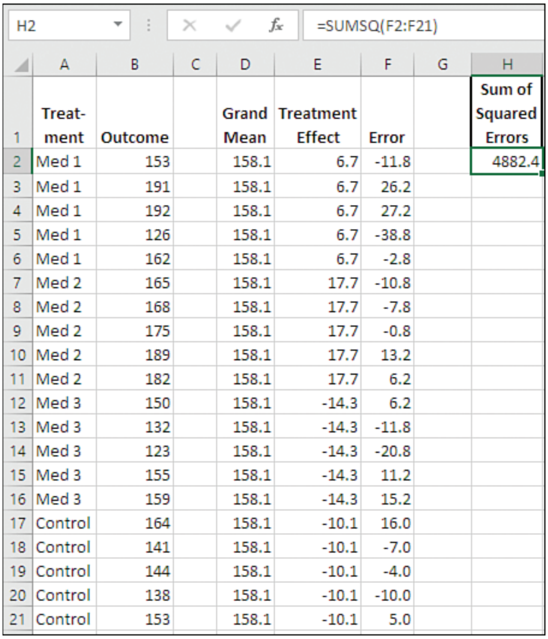
The formula uses Greek instead of Roman letters to emphasize that it's referring to the population from which observations are sampled, but it's equally useful to consider that it refers to a sample taken from that population:

$$Y_{ij} = \overline{Y} + a_j + e_{ij}$$

The idea is that each observation Y_{ij} can be considered as the sum of three components:

- The grand mean, μ
- The effect of treatment j, α_j
- The quantity ε_{ij} that represents the deviation of an individual score Y_{ij} from the combination of the grand mean and the j^th^ treatment's effect.

Here it is in the context of a worksheet (see Figure 7.10):



H2 =SUMSQ(F2:F21)

	A	B	C	D	E	F	G	H
1	Treat-ment	Outcome		Grand Mean	Treatment Effect	Error		Sum of Squared Errors
2	Med 1	153		158.1	6.7	-11.8		4882.4
3	Med 1	191		158.1	6.7	26.2		
4	Med 1	192		158.1	6.7	27.2		
5	Med 1	126		158.1	6.7	-38.8		
6	Med 1	162		158.1	6.7	-2.8		
7	Med 2	165		158.1	17.7	-10.8		
8	Med 2	168		158.1	17.7	-7.8		
9	Med 2	175		158.1	17.7	-0.8		
10	Med 2	189		158.1	17.7	13.2		
11	Med 2	182		158.1	17.7	6.2		
12	Med 3	150		158.1	-14.3	6.2		
13	Med 3	132		158.1	-14.3	-11.8		
14	Med 3	123		158.1	-14.3	-20.8		
15	Med 3	155		158.1	-14.3	11.2		
16	Med 3	159		158.1	-14.3	15.2		
17	Control	164		158.1	-10.1	16.0		
18	Control	141		158.1	-10.1	-7.0		
19	Control	144		158.1	-10.1	-4.0		
20	Control	138		158.1	-10.1	-10.0		
21	Control	153		158.1	-10.1	5.0		

Figure 7.10
Observations broken down in terms of the components of the general linear model.

In Figure 7.10, each of the 20 observations in Figure 7.9 have been broken down into the three components of the general linear model: the grand mean in the range D2:D21, the effect of each treatment group in E2:E21, and the so-called "error" involved with each observation.

NOTE

The term *error* is used for some not especially good historical reasons, and it's made its way into other terms such as *mean square error* and even the symbol ∈ .There's nothing erroneous about these values. *Residuals* is a perfectly descriptive term that isn't misleading, but statistical jargon tends to prefer *error*.

If you didn't expect that one or more treatments would have an effect on the subjects receiving that treatment, then your best estimate of the value of a particular observation would be the grand mean (in this case, that's 158.1).

But suppose you expected that the effect of a treatment would be to raise the observed values for the subjects receiving that treatment above, or lower them below, the grand mean. In that case your best estimate of a given observation would be the grand mean plus the effect, whether positive or negative, associated with that treatment. In the case of, say, the observation in row 5 of Figure 7.10, your expectation would be 158.1 + 6.7, or 164.8. If you give the matter a little thought, you'll see why that figure, 164.8, must be the mean outcome score for the Med 1 group.

Although the mean of its group is your best expectation for any one of its members, most—typically all—of the members of a group will have a score on the outcome variable different from the mean of the group. Those quantities (differences, deviations, residuals, errors, or whatever you prefer to call them) are shown in the range F2:F21 as the result of subtracting the grand mean and the group's treatment effect from the actual observation. For example, the value in cell F2 is returned by this formula:

```
=B2-D2-E2
```

The purpose of a regression equation is to minimize the sum of the squares of those errors. When that's done, the minimized result is called the Residual Sum of Squares in the context of regression, and the Sum of Squares Within in the context of ANOVA.

Note the sum of the squared errors in cell H2. It's returned by this formula:

```
=SUMSQ(F2:F21)
```

The SUMSQ() function squares the values in its argument and totals them. That's the same value as you'll find in Figure 7.9, cells H21, the Sum of Squares Within Groups from the ANOVA, and H6, the residual sum of squares from LINEST(). As Figure 7.10 shows, the sum of squares is based on the mean deviations from the grand mean, and on the individual deviations from the group means.

It's very simple to move from dummy coding to effect coding. Rather than assigning codes of 0 throughout the coding vectors to one particular group, you assign codes of –1 to one of the groups—not necessarily a control group—throughout the vectors. If you do that in the key range used by VLOOKUP(), you need to make that replacement in only as many key range cells as you have vectors. You can see in Figure 7.9 that this has been done in the range C17:E21, which contains –1's rather than 0's. I assigned the –1's to the control group not because it's necessarily desirable to do so, but to make comparisons with the dummy coding used in Figure 7.6.

Regression analysis with a single factor and effect coding handles unequal group sizes accurately, and so does traditional ANOVA. Figure 7.11 shows an analysis of a data set with unequal group sizes.

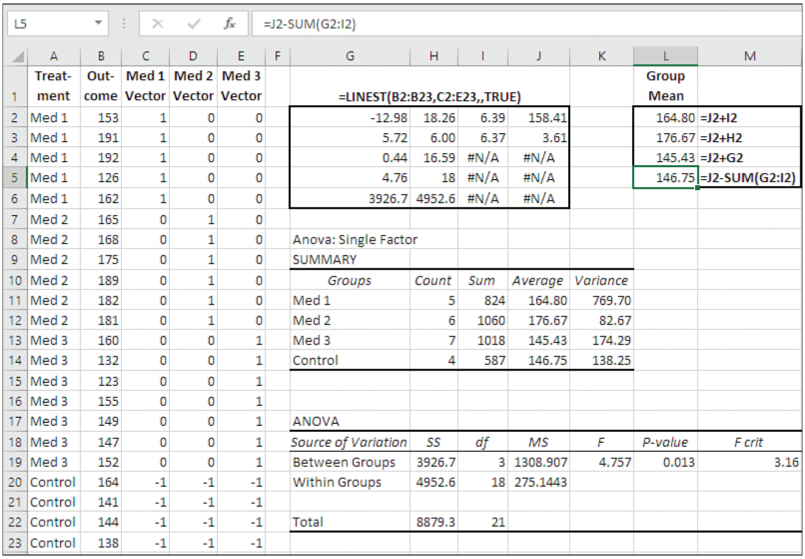


Figure 7.11 The ANOVA: Single Factor tool returns the group counts in the range H11:H14.

L5 =J2-SUM(G2:I2)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Treat-ment	Out-come	Med 1 Vector	Med 2 Vector	Med 3 Vector		=LINEST(B2:B23,C2:E23,,TRUE)					Group Mean	
2	Med 1	153	1	0	0		-12.98	18.26	6.39	158.41		164.80	=J2+I2
3	Med 1	191	1	0	0		5.72	6.00	6.37	3.61		176.67	=J2+H2
4	Med 1	192	1	0	0		0.44	16.59	#N/A	#N/A		145.43	=J2+G2
5	Med 1	126	1	0	0		4.76	18	#N/A	#N/A		146.75	=J2-SUM(G2:I2)
6	Med 1	162	1	0	0		3926.7	4952.6	#N/A	#N/A			
7	Med 2	165	0	1	0								
8	Med 2	168	0	1	0		Anova: Single Factor						
9	Med 2	175	0	1	0		SUMMARY						
10	Med 2	189	0	1	0		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
11	Med 2	182	0	1	0		Med 1	5	824	164.80	769.70		
12	Med 2	181	0	1	0		Med 2	6	1060	176.67	82.67		
13	Med 3	160	0	0	1		Med 3	7	1018	145.43	174.29		
14	Med 3	132	0	0	1		Control	4	587	146.75	138.25		
15	Med 3	123	0	0	1								
16	Med 3	155	0	0	1								
17	Med 3	149	0	0	1		ANOVA						
18	Med 3	147	0	0	1		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
19	Med 3	152	0	0	1		Between Groups	3926.7	3	1308.907	4.757	0.013	3.16
20	Control	164	-1	-1	-1		Within Groups	4952.6	18	275.1443			
21	Control	141	-1	-1	-1								
22	Control	144	-1	-1	-1		Total	8879.3	21				
23	Control	138	-1	-1	-1								

Notice that the regression equation returns as the regression coefficients the effect of being in each of the treatment groups. In the range L2:L5, the grand mean (which is the constant in the regression equation) is added to each of the regression coefficients to return the actual mean for each group. Compare the results with the means returned by the Data Analysis add-in in the range J11:J14.

Notice that the grand mean is the average of the group means, 158.41, rather than the mean of the individual observations, 158.6. This situation is typical of designs in which the groups have different numbers of observations.

Both the traditional ANOVA approach and the regression approach manage the situation of unequal group sizes effectively. But if you have groups with very discrepant numbers of observations *and* very discrepant variances, you'll want to keep in mind the discussion from Chapter 6 regarding their combined effects on probability estimates: If your larger groups also have the larger variances, your apparent tests will tend to be conservative. If the larger groups have the smaller variances, your apparent tests will tend to be liberal.

Multiple Comparisons with Effect Coding

Dummy coding largely defines the comparisons of interest to you. The fact that you choose dummy coding as the method of populating the vectors in the data matrix implies that you want to compare one particular group mean, usually that of a control group, with the other group means in the data set. The Dunnett method of multiple comparisons is often the method of choice when you've used dummy coding.

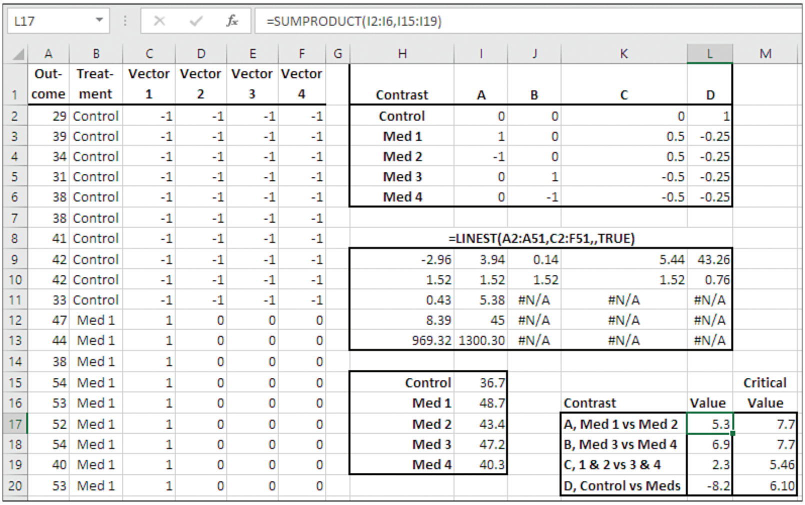
A more flexible method of multiple comparisons is called the Scheffé method. It is a post hoc method, meaning that you can use it after you've seen the results of the overall analysis and that you need not plan ahead of time what comparisons you'll make. The Scheffé method also enables you to make complex contrasts, such as the mean of two groups versus the mean of three other groups.

There's a price to that flexibility, and it's in the statistical power of the Scheffé method. The Scheffé will fail to declare comparisons as statistically significant that other methods would. That's a problem and it's a good reason to consider other methods such as planned orthogonal contrast (discussed later in this chapter).

To use the Scheffé method, you need to set up a matrix that defines the contrasts you want to make. See Figure 7.12.

Consider Contrast A, in the range I2:I6 of Figure 7.12. Cell I3 contains a 1 and cell I4 contains a –1; the remaining cells in I2:I6 contain 0's. The values in the matrix are termed *contrast coefficients*. You multiply each contrast coefficient by the mean of the group it belongs to. Therefore, I2:I6 defines a contrast in which the mean of Med 2 (coefficient of –1) is subtracted from the mean of Med 1 (coefficient of 1), and the remaining group means do not enter the contrast.

Similarly, Contrast B, in J2:J6, also contains a 1 and a –1, but this time it's the difference between Med 3 and Med 4 that's to be tested.



L17 =SUMPRODUCT(I2:I6,I15:I19)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Outcome	Treatment	Vector 1	Vector 2	Vector 3	Vector 4		Contrast	A	B	C	D	
2	29	Control	-1	-1	-1	-1		Control	0	0	0	1	
3	39	Control	-1	-1	-1	-1		Med 1	1	0	0.5	-0.25	
4	34	Control	-1	-1	-1	-1		Med 2	-1	0	0.5	-0.25	
5	31	Control	-1	-1	-1	-1		Med 3	0	1	-0.5	-0.25	
6	38	Control	-1	-1	-1	-1		Med 4	0	-1	-0.5	-0.25	
7	38	Control	-1	-1	-1	-1							
8	41	Control	-1	-1	-1	-1		=LINEST(A2:A51,C2:F51,,TRUE)					
9	42	Control	-1	-1	-1	-1		-2.96	3.94	0.14	5.44	43.26	
10	42	Control	-1	-1	-1	-1		1.52	1.52	1.52	1.52	0.76	
11	33	Control	-1	-1	-1	-1		0.43	5.38	#N/A	#N/A	#N/A	
12	47	Med 1	1	0	0	0		8.39	45	#N/A	#N/A	#N/A	
13	44	Med 1	1	0	0	0		969.32	1300.30	#N/A	#N/A	#N/A	
14	38	Med 1	1	0	0	0							
15	54	Med 1	1	0	0	0		Control	36.7				Critical
16	53	Med 1	1	0	0	0		Med 1	48.7		Contrast	Value	Value
17	52	Med 1	1	0	0	0		Med 2	43.4		A, Med 1 vs Med 2	5.3	7.7
18	54	Med 1	1	0	0	0		Med 3	47.2		B, Med 3 vs Med 4	6.9	7.7
19	40	Med 1	1	0	0	0		Med 4	40.3		C, 1 & 2 vs 3 & 4	2.3	5.46
20	53	Med 1	1	0	0	0					D, Control vs Meds	-8.2	6.10

Figure 7.12
The matrix of contrasts defines how much weight each group mean is given.

More complex contrasts are possible, of course. Contrast C compares the *average* of Med 1 and Med 2 with the average of Med 3 and Med 4, and Contrast D compares Control with the average of the four medication groups.

A regression analysis of the effect-coded data in A2:F51 appears in H9:L13. An F-test of the full regression (which is equivalent to a test of the deviation of the R^2 value in H11 from 0.0) could be managed with this formula:

```
=F.DIST.RT(H12,4,I12)
```

It returns 0.00004, and something like 4 in 100,000 replications of this experiment would return an F-ratio of 8.39 or greater if there were no differences between the population means. So you move on to a multiple comparisons procedure to try to pinpoint the differences that bring about so large an F-ratio.

You can pick up the group means by combining the constant returned by LINEST() in cell L9 (which, with effect coding, is the mean of the group means) with the individual regression coefficients. For example, the mean of the Med 1 group is returned in cell I16 with this formula:

```
=K9+$L$9
```

The formula for the mean of the group assigned –1's throughout the coding matrix is just a little more complicated. It is the grand mean minus the sum of the remaining regression coefficients. So, the formula for the control group in cell I15 is:

```
=$L$9-SUM(H9:K9)
```

With the five group means established in the range I15:I19, you can apply the contrasts you defined in the range I2:L6 by multiplying each group mean by the associated contrast coefficient for that contrast. Excel's SUMPRODUCT() function is convenient for that: It multiplies the corresponding elements in two arrays and returns the sum of the products. Therefore, this formula in cell L17:

```
=SUMPRODUCT(I2:I6,I15:I19)
```

has this effect:

```
=I2*I15 + I3*I16 + I4*I17 + I5*I18 + I6*I19
```

which results in the value 5.3. The formula in cell L18 moves one column to the right in the contrast matrix:

```
=SUMPRODUCT(J2:J6,I15:I19)
```

and so on through the fourth contrast.

The final step in the Scheffé method is to determine a critical value that the contrast values in L17:L20 must exceed to be regarded as statistically significant. Here's the formula, which looks a little forbidding in Excel syntax:

```
=SQRT((5–1)*F.INV(0.95,4,I12))*SQRT(($I$13/$I$12)
  *(J2^2/10+J3^2/10+J4^2/10+J5^2/10+J6^2/10))
```

Here it is using more conventional notation:

$$\sqrt{(k - 1)F_{df1,df2}}\sqrt{MSR\sum C_j^2/n_j}$$

where:

- k is the number of groups.
- $F_{df1,df2}$ is the value of the F distribution at the alpha level you select, such as 0.05 or 0.01. In the Excel version of the formula just given, I chose the .05 level, using 0.95 as the argument to the F.INV() function because it returns the F-ratio that has, in this case, 0.95 of the distribution to its left. I could have used, instead, F.INV.RT(0.05,4,I12) to return the same value.
- MSR is the mean square residual from the LINEST() results, obtained by dividing the residual sum of squares by the degrees of freedom for the residual.
- C is the contrast coefficient. Each contrast coefficient is squared and divided by n_j, the number of observations in the group. The results of the divisions are summed.

The critical value varies across the contrasts that have different coefficients. To complete the process, compare the value of each contrast with its critical value. If the absolute value of the contrast exceeds the critical value, then the contrast is considered significant at the level you chose for the F value in the formula for the critical value.

In Figure 7.12, the critical values are shown in the range M17:M20. Only one contrast has an absolute value that exceeds its associated critical value: Contrast D, which contrasts the mean of the Control group with the average of the means of the four remaining groups.

I mentioned at the start of this section that the Scheffé method is at once the least statistically powerful and the most flexible of the multiple comparison methods. You might want to compare the results reported here with the results of planned orthogonal contrasts, discussed in the next section. Planned orthogonal contrasts are at once the most statistically powerful and the least flexible of the multiple comparison methods. When we get to the multiple comparisons in the next section, you'll see that the same data set returns very different outcomes.

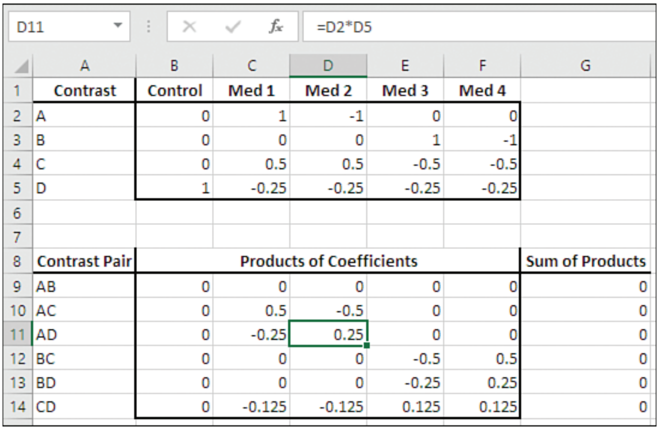
Orthogonal Coding

A third useful type of coding, besides dummy coding and effect coding, is *orthogonal coding*. You can use orthogonal coding in both planned and post hoc situations. I'll be discussing planned orthogonal coding (also termed planned orthogonal *contrasts*) here, because this approach is most useful when you already know something about how your variables work, and therefore are in a position to specify in advance which comparisons you will want to make.

Establishing the Contrasts

Orthogonal coding (I'll explain the term *orthogonal* shortly) depends on a matrix of values that define the contrasts that you want to make. Suppose that you plan an experiment with five groups: say, four treatments and a control. To define the contrasts that interest you, you set up a matrix such as the one shown in Figure 7.13.

Figure 7.13
The sums of products in G9:G14 satisfy the condition of orthogonality.



D11 =D2*D5

	A	B	C	D	E	F	G
1	Contrast	Control	Med 1	Med 2	Med 3	Med 4	
2	A	0	1	-1	0	0	
3	B	0	0	0	1	-1	
4	C	0	0.5	0.5	-0.5	-0.5	
5	D	1	-0.25	-0.25	-0.25	-0.25	
6							
7							
8	Contrast Pair			Products of Coefficients			Sum of Products
9	AB	0	0	0	0	0	0
10	AC	0	0.5	-0.5	0	0	0
11	AD	0	-0.25	0.25	0	0	0
12	BC	0	0	0	-0.5	0.5	0
13	BD	0	0	0	-0.25	0.25	0
14	CD	0	-0.125	-0.125	0.125	0.125	0

In orthogonal coding, just defining the contrasts isn't enough. Verifying that the contrasts are orthogonal to one another is also necessary. One fairly tedious way to verify that is also shown in Figure 7.13. The range B9:F14 contains the products of corresponding coefficients for each pair of contrasts defined in B2:F5. So row 10 tests Contrasts A and C, and the coefficients in row 2 and row 4 are multiplied to get the products in row 10. For example, the formula in cell C10 is:

=C2*C4

In row 11, testing Contrast A with Contrast D, cell D11 contains this formula:

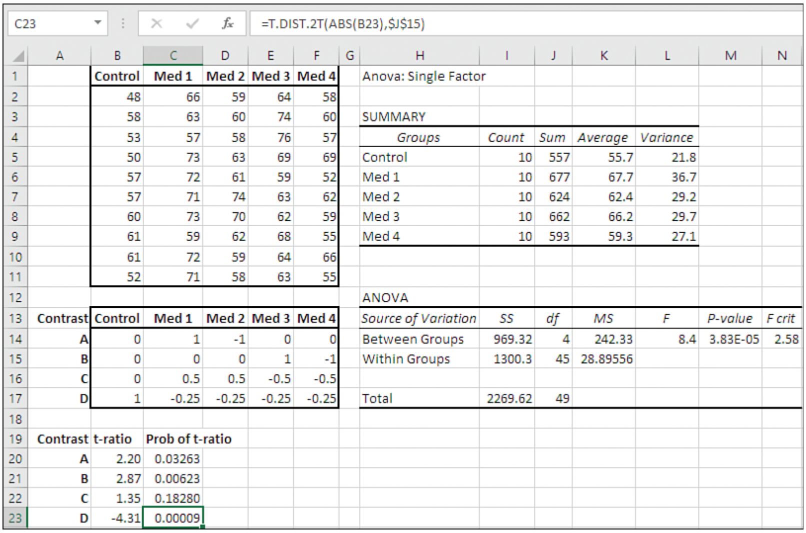
=D2*D5

Finally, total up the cells in each row of the matrix of coefficient products. If the total is 0, those two contrasts are orthogonal to one another. This is done in the range G9:G14. All the totals in that range are 0, so each of the contrasts defined in B2:F5 are orthogonal to one another.

Planned Orthogonal Contrasts Via ANOVA

Figure 7.14 shows how the contrast coefficients are used in the context of an ANOVA. I'm inflicting this on you to give you a greater appreciation of how much easier the regression approach makes all this.

Figure 7.14
The calculation of the t-ratios involves the group means and counts, the mean square within and the contrast coefficients.



C23 =T.DIST.2T(ABS(B23),$J$15)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Control	Med 1	Med 2	Med 3	Med 4		Anova: Single Factor						
2		48	66	59	64	58								
3		58	63	60	74	60		SUMMARY						
4		53	57	58	76	57		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5		50	73	63	69	69		Control	10	557	55.7	21.8		
6		57	72	61	59	52		Med 1	10	677	67.7	36.7		
7		57	71	74	63	62		Med 2	10	624	62.4	29.2		
8		60	73	70	62	59		Med 3	10	662	66.2	29.7		
9		61	59	62	68	55		Med 4	10	593	59.3	27.1		
10		61	72	59	64	66								
11		52	71	58	63	55								
12								ANOVA						
13	Contrast	Control	Med 1	Med 2	Med 3	Med 4		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
14	A	0	1	-1	0	0		Between Groups	969.32	4	242.33	8.4	3.83E-05	2.58
15	B	0	0	0	1	-1		Within Groups	1300.3	45	28.89556			
16	C	0	0.5	0.5	-0.5	-0.5								
17	D	1	-0.25	-0.25	-0.25	-0.25		Total	2269.62	49				
18														
19	Contrast	t-ratio	Prob of t-ratio											
20	A	2.20	0.03263											
21	B	2.87	0.00623											
22	C	1.35	0.18280											
23	D	-4.31	0.00009											

Figure 7.14 shows a new data set, laid out for analysis by the ANOVA: Single Factor tool. That tool has been run on the data, and the results are shown in H1:N17. The matrix of contrast coefficients, which has already been tested for orthogonality, is in the range B14:F17. Each of these is needed to compute the t-ratios that test the significance of the difference established in each contrast.

The formulas to calculate the t-ratios are complex. Here's the formula for the first contrast, Contrast A, which tests the difference between the mean of the Med 1 group and the Med 2 group:

```
=SUMPRODUCT(B14:F14,TRANSPOSE($K$5:$K$9))/
 SQRT($K$15*SUM(B14:F14^2/TRANSPOSE($I$5:$I$9)))
```

The formula must be array-entered using Ctrl+Shift+Enter. Here it is in general form, using summation notation:

$$t = \sum C_j \overline{X}_j / \sqrt{MSE \sum C_j^2 / n_j}$$

where:

- C_j is the contrast coefficient for the jth mean.
- $\overline{X}_j$ is the jth sample mean.
- MSE is the mean square error from the ANOVA table. If you don't want to start by running an ANOVA, just take the average of the sample group variances. In this case, MSE is picked up from cell K15, calculated and reported by the Data Analysis tool.
- n_j is the number of observations in the jth sample.

The prior two formulas, in Excel and summation syntax, are a trifle more complicated than they need be. They allow for unequal sample sizes. As you'll see in the next section, unequal sample sizes generally—not always—result in nonorthogonal contrasts. If you have equal sample sizes, the formulas can treat the sample sizes as a constant and simplify as a result.

Returning to Figure 7.14, notice the t-ratios and associated probability levels in the range B20:C23. Each of the t-ratios is calculated using the Excel array formula just given, adjusted to pick up the contrast coefficients for different contrasts.

The probabilities are returned by the T.DIST.2T() function, the non-directional version of the t-test. The probability informs you how much of the area under the t-distribution with 45 degrees of freedom is to the left of, in the case of Contrast A, –2.20 and to the right of +2.20. If you had specified alpha as 0.01 prior to seeing the data, you could reject the null hypothesis of no population difference for Contrast B and Contrast D. The probabilities of the associated t-ratios occurring by chance in a central t distribution are lower than your alpha level. The probabilities for Contrast A and Contrast C are higher than alpha and you must retain the associated null hypotheses.

Planned Orthogonal Contrasts Using LINEST( )

As far as I'm concerned, there's a lot of work—and opportunity to make mistakes—involved with planned orthogonal contrasts in the context of the traditional ANOVA. Figure 7.15 shows how much easier things are using regression, and in an Excel worksheet that means LINEST().

Using regression, you still need to come up with the orthogonal contrasts and their coefficients. But they're the same ones needed for the ANOVA approach. Figure 7.15 repeats them, transposed from Figure 7.14, in the range I1:M6.

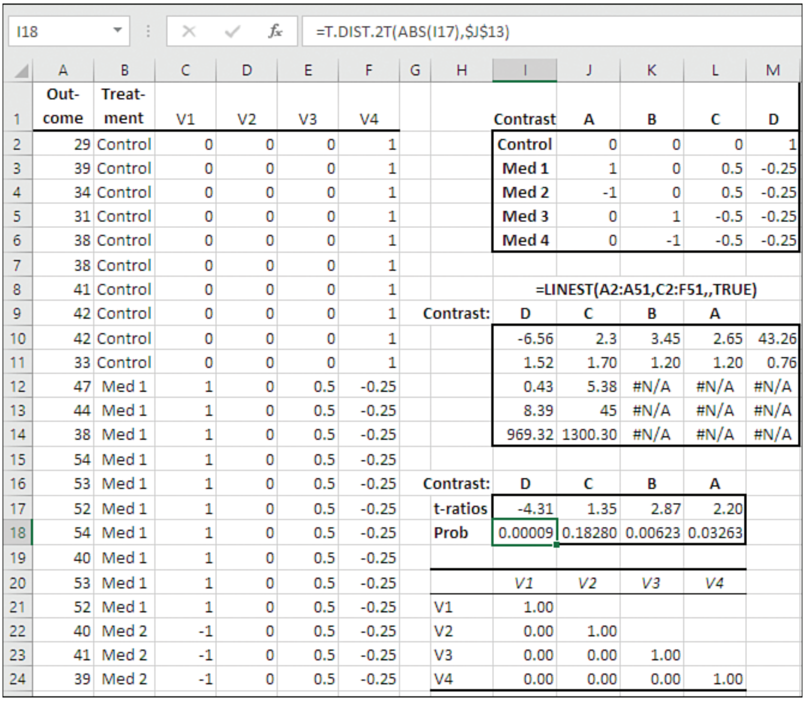


Figure 7.15
With orthogonal coding, the regression coefficients and their standard errors do most of the work for you.

I18 =T.DIST.2T(ABS(I17),$J$13)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Outcome	Treatment	V1	V2	V3	V4			Contrast	A	B	C	D
2	29	Control	0	0	0	1			Control	0	0	0	1
3	39	Control	0	0	0	1			Med 1	1	0	0.5	-0.25
4	34	Control	0	0	0	1			Med 2	-1	0	0.5	-0.25
5	31	Control	0	0	0	1			Med 3	0	1	-0.5	-0.25
6	38	Control	0	0	0	1			Med 4	0	-1	-0.5	-0.25
7	38	Control	0	0	0	1							
8	41	Control	0	0	0	1			=LINEST(A2:A51,C2:F51,,TRUE)				
9	42	Control	0	0	0	1		Contrast:	D	C	B	A	
10	42	Control	0	0	0	1			-6.56	2.3	3.45	2.65	43.26
11	33	Control	0	0	0	1			1.52	1.70	1.20	1.20	0.76
12	47	Med 1	1	0	0.5	-0.25			0.43	5.38	#N/A	#N/A	#N/A
13	44	Med 1	1	0	0.5	-0.25			8.39	45	#N/A	#N/A	#N/A
14	38	Med 1	1	0	0.5	-0.25			969.32	1300.30	#N/A	#N/A	#N/A
15	54	Med 1	1	0	0.5	-0.25							
16	53	Med 1	1	0	0.5	-0.25		Contrast:	D	C	B	A	
17	52	Med 1	1	0	0.5	-0.25		t-ratios	-4.31	1.35	2.87	2.20	
18	54	Med 1	1	0	0.5	-0.25		Prob	0.00009	0.18280	0.00623	0.03263	
19	40	Med 1	1	0	0.5	-0.25							
20	53	Med 1	1	0	0.5	-0.25			*V1*	*V2*	*V3*	*V4*	
21	52	Med 1	1	0	0.5	-0.25		V1	1.00				
22	40	Med 2	-1	0	0.5	-0.25		V2	0.00	1.00			
23	41	Med 2	-1	0	0.5	-0.25		V3	0.00	0.00	1.00		
24	39	Med 2	-1	0	0.5	-0.25		V4	0.00	0.00	0.00	1.00	

The difference with orthogonal coding and regression, as distinct from the traditional ANOVA approach shown in Figure 7.14, is that you use the coefficients to populate the vectors, just as you do with dummy coding (1's and 0's) and effect coding (1's, 0's, and –1's). Each vector represents a contrast and the values in the vector are the contrast's coefficients, each associated with a different group.

So, in Figure 7.15, Vector 1 in Column C has 0's for the Control group, 1's for Med 1, –1's for Med 2, and—although you can't see them in the figure—0's for Med 3 and Med 4. Those are the values called for in Contrast A, in the range J2:J6. Similar comments apply to vectors 2 through 4. The vectors make the contrast coefficients a formal part of the analysis.

The regression approach also allows for a different slant on the notion of orthogonality. Notice the matrix of values in the range I21:L24. It's a correlation matrix showing the correlations between each pair of vectors in columns C through F. Notice that each vector has a 0.0 correlation with each of the other vectors. They are independent of one another. That's another way of saying that if you plotted them, their axes would be at right angles to one another (*orthogonal* means *right angled*).

NOTE

As an experiment, I suggest that you try adding at least one case to at least one of the groups in columns A through F of the worksheet for Figure 7.15—it's in the workbook for this chapter, which you can download from quepublishing.com/title/9780789756558. For example, to add a case to the control group, insert cells in A12:F12 and put 0's in C12:E12 and a 1 in F12. Then rebuild the correlation matrix starting in cell I21, either entering the CORREL( ) functions yourself or running the Data Analysis add-in's Correlation tool on the data in columns C through F. Notice that the correlations involving vectors where you have changed the group count no longer equal 0.0. They're no longer orthogonal.

This effect has implications for designs with two or more factors and unequal group frequencies. A distinction is made between situations in which the treatments might be causally related to the unequal frequencies—differential experimental mortality by treatment—and inequality in group counts due to causes unrelated to the treatments.

Planned orthogonal contrasts have the greatest amount of statistical power of any of the multiple comparison methods. That means that planned orthogonal contrasts are more likely to identify true population differences than the alternatives (such as Dunnett and Scheffé). However, they require that you be able to specify your hypotheses in the form of contrasts before the experiment, and that you are able to obtain equal group sizes. If you add even one observation to any of the groups, the correlations among the vectors will no longer be 0.0, you'll have lost the orthogonality, and you'll need to resort to (probably) planned nonorthogonal contrasts, which, other things equal, are less powerful.

It's easy to set up the vectors using the general VLOOKUP() approach described earlier in this chapter. For example, this formula is used to populate Vector 1:

```
=VLOOKUP($B2,$I$2:$M$6,2,0)
```

It's entered in cell C2 and can be copied and pasted into columns D through F (you'll need to adjust the third argument from 2 to 3, 4 and 5). Then make a multiple selection of C2:F2 and drag down through the end of the Outcome values.

With the vectors established, array-enter this LINEST() formula into a five-row by five-column range:

```
=LINEST(A2:A51,C2:F51,,TRUE)
```

You now have the regression coefficients and their standard errors. The t-ratios—the same ones that show up in the range B20:B23 of Figure 7.14—are calculated by dividing a regression coefficient by its standard error. So the t-ratio in cell L17 of Figure 7.15 is returned by this formula:

```
=L10/L11
```

The coefficients and standard errors come back from LINEST() in reverse of the order that you would like, so the t-ratios are in reverse order, too. However, if you compare them to the t-ratios in Figure 7.14, you'll find that their values are precisely the same.

You calculate the probabilities associated with the t-ratios just as in Figure 7.14, using the T.DIST() function that's appropriate to the sort of research hypothesis (directional or nondirectional) that you would specify at the outset.

Factorial Analysis

One of the reasons that the development of the analysis of variance represents such a major step forward in the science of data analysis is that it provides the ability to study the simultaneous effects of two or more factors on the outcome variable. Prior to the groundwork that Fisher did with ANOVA, researchers were limited to studying one variable at a time, usually just two levels of that factor at a time.

This situation meant that researchers could not investigate the *joint* effect of two or more factors. For example, it may be that men have a different attitude toward a politician when they are over 50 years of age than they do earlier in their lives. Furthermore, it may be that women's attitude toward that politician do not change as a function of their age. If we had to study the effects of sex and age separately, we wouldn't be able to determine that a joint effect—termed an *interaction* in statistical jargon—exists.

But we can accommodate more than just one factor in an ANOVA—or, of course, in a regression analysis. When you simultaneously analyze how two or more factors are related to an outcome variable, you're said to be using *factorial analysis*.

And when you can study and analyze the effects of more than just one variable at a time, you get more bang for your buck. The costs of running an experiment are often just trivially greater when you study additional variables than when you study only one.

It also happens that adding one or more factors to a single factor ANOVA can increase its statistical power. In a single-factor ANOVA, variation in the outcome variable that can't be attributed to the factor gets tossed into the mean square residual. It can happen that such variation might be associated with another factor (or, as you'll see in the next chapter, a covariate). Then that variation could be removed from the mean square error—which, when decreased, increases the value of F-ratios in the analysis, thus increasing the tests' statistical power.

Excel's Data Analysis add-in includes a tool that accommodates two factors at once, but it has drawbacks. In addition to a problem I've noted before, that the results do not come back as formulas but as static values, the ANOVA: Two-Factor with Replication tool requires that you arrange your data in a highly idiosyncratic fashion, and it cannot accommodate unequal group sizes, nor can it accommodate more than two factors. Covariates are out.

If you use regression instead, you don't have to live with those limits. To give you a basis for comparison, let's look at the results of the ANOVA: Two-Factor with Replication tool.

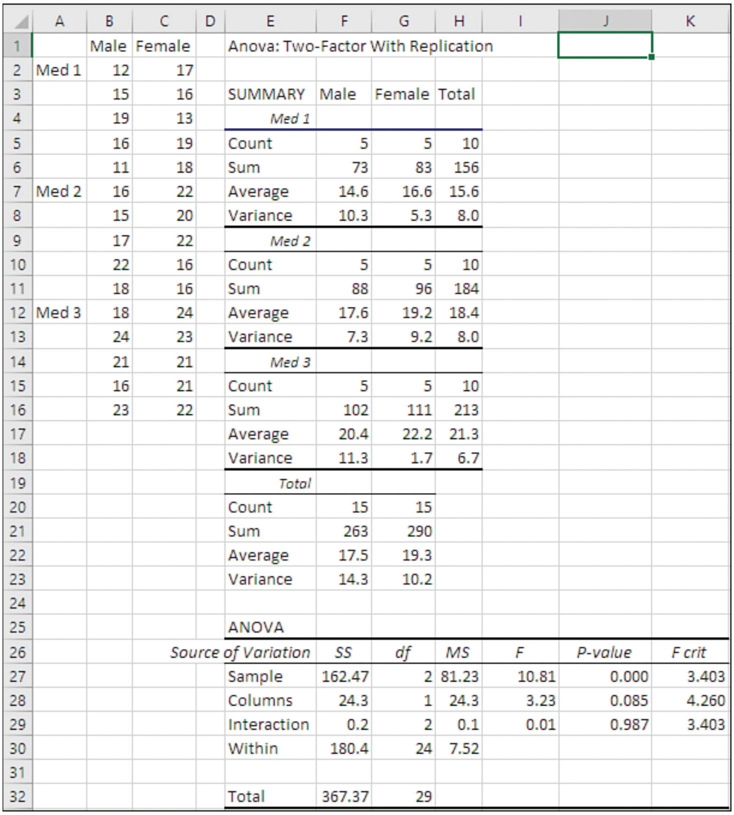
NOTE

Excel's Data Analysis add-in also offers a tool called ANOVA: Two-Factor Without Replication. It is a fairly old-fashioned approach to analyzing what's often termed a *repeated measures design*. This is the last you'll hear of it in this book.

The ANOVA tool used in Figure 7.16 is helpful in that it returns the average and variance of the outcome variable, as well as the count, for each group in the design. My own preference would be to use a pivot table to report these descriptive statistics, because that's a live analysis and the table returned by the ANOVA tool is, again, static values. With a pivot table I can add, delete, or edit observations and have the pivot table update itself. With static values I have to run the ANOVA tool over again.

The ANOVA table at the end of the results shows a couple of features that don't appear in the Data Analysis add-in's Single Factor version. Notice that rows 27 and 28 show a Sample and a Column source of variation. The Column source of variation refers to sex: Values for males are in column B and values for females are in column C. The Sample data source refers to whatever variable has values that occupy different rows. In Figure 7.16, values for Med 1 are in rows 2 through 6, Med 2 in rows 7 through 11, and Med 3 in rows 12 through 16.

Figure 7.16
The ANOVA: Two-Factor with Replication tool will not run if different groups have different numbers of observations.



	A	B	C	D	E	F	G	H	I	J	K
1		Male	Female		Anova: Two-Factor With Replication						
2	Med 1	12	17								
3		15	16		SUMMARY	Male	Female	Total			
4		19	13		*Med 1*						
5		16	19		Count	5	5	10			
6		11	18		Sum	73	83	156			
7	Med 2	16	22		Average	14.6	16.6	15.6			
8		15	20		Variance	10.3	5.3	8.0			
9		17	22		*Med 2*						
10		22	16		Count	5	5	10			
11		18	16		Sum	88	96	184			
12	Med 3	18	24		Average	17.6	19.2	18.4			
13		24	23		Variance	7.3	9.2	8.0			
14		21	21		*Med 3*						
15		16	21		Count	5	5	10			
16		23	22		Sum	102	111	213			
17					Average	20.4	22.2	21.3			
18					Variance	11.3	1.7	6.7			
19					*Total*						
20					Count	15	15				
21					Sum	263	290				
22					Average	17.5	19.3				
23					Variance	14.3	10.2				
24											
25					ANOVA						
26					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
27					Sample	162.47	2	81.23	10.81	0.000	3.403
28					Columns	24.3	1	24.3	3.23	0.085	4.260
29					Interaction	0.2	2	0.1	0.01	0.987	3.403
30					Within	180.4	24	7.52			
31											
32					Total	367.37	29				

You'll want to draw your own conclusions regarding the convenience of the data layout (required, by the way, by the Data Analysis tool) and regarding the labeling of the factors in the ANOVA table.

The main point is that both factors, Sex (labeled *Columns* in cell E28) and Medication (labeled *Sample* in cell E27), exist as sources of variation in the ANOVA table. Males' averages differ from females' averages, and that constitutes a source of variation. The three kinds of medication also differ from one another's averages—another source of variation.

There is also a third source labeled *Interaction*, which refers to the joint effect of the Sex and Medication variables. At the interaction level, groups are considered to constitute combinations of levels of the main factors: For example, Males who get Med 2 constitute a group, as do Females who get Med 1. Differences due to the combined main effects—not just Male compared to Female, or Med 1 compared to Med 3—are collectively referred to as the *interaction* between, here, Sex and Treatment.

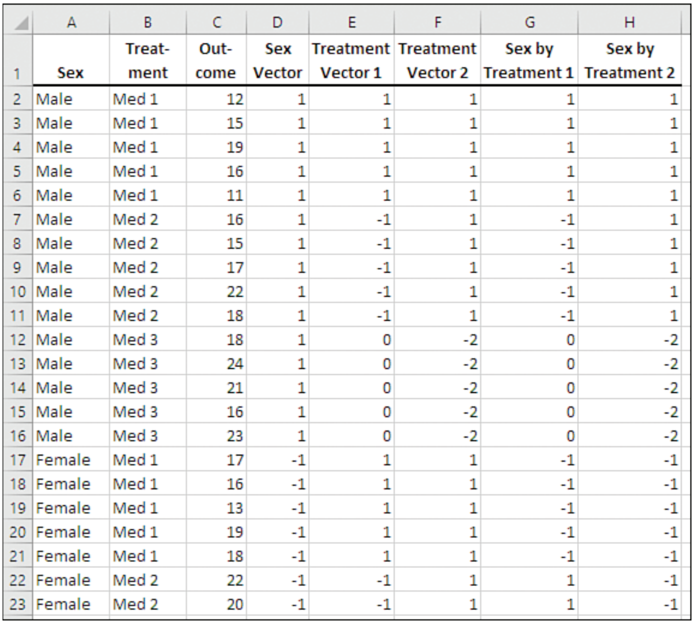
The ANOVA shown in Figure 7.16 evaluates the effect of Sex as not significant at the .05 level (see cell J28, which reports the probability of an F-ratio of 3.23 with 1 and 24 degrees of freedom as 8.5% when there is no difference in the populations). Similarly, there is no significant difference due to the interaction of Sex with Treatment. Differences between the means of the six design cells (two sexes times three treatments) are not great enough to reject the null hypothesis of no differences among the six groups. Only the Treatment main effect is statistically significant. If, from the outset, you intended to use planned orthogonal contrasts to test the differences between specific means, you could do so now and enjoy the statistical power available to you. In the absence of such planning, you could use the Scheffé procedure, hoping that you wouldn't lose too much statistical power as a penalty for having failed to plan your contrasts.

Factorial Analysis with Orthogonal Coding

However, there's no reason that you couldn't use orthogonal coefficients in the vectors. You wouldn't do so on a post hoc basis to increase statistical power, because that requires you to choose your comparisons before seeing the results. However, with equal group sizes you could still use orthogonal codes in the vectors to make some of the computations more convenient. Figure 7.17 shows the data from Figure 7.16 laid out as a list, with vectors that represent the Sex and the Treatment variables.

The data set in Figure 7.17 has one vector in column D to represent the Sex variable. Because that factor has only two levels, one vector is sufficient to represent it. The data set also has two vectors in columns E and F to represent the Treatment factor. That factor has three levels, so two vectors are needed. Finally, two vectors representing the interaction between Sex and Treatment occupy columns G and H.

The interaction vectors are easily populated by multiplying the main effect vectors. The vector in column G is the result of multiplying the Sex vector by the Treatment 1 vector. The vector in column H results from the product of the Sex vector and the Treatment 2 vector.



	A	B	C	D	E	F	G	H
1	Sex	Treat-ment	Out-come	Sex Vector	Treatment Vector 1	Treatment Vector 2	Sex by Treatment 1	Sex by Treatment 2
2	Male	Med 1	12	1	1	1	1	1
3	Male	Med 1	15	1	1	1	1	1
4	Male	Med 1	19	1	1	1	1	1
5	Male	Med 1	16	1	1	1	1	1
6	Male	Med 1	11	1	1	1	1	1
7	Male	Med 2	16	1	-1	1	-1	1
8	Male	Med 2	15	1	-1	1	-1	1
9	Male	Med 2	17	1	-1	1	-1	1
10	Male	Med 2	22	1	-1	1	-1	1
11	Male	Med 2	18	1	-1	1	-1	1
12	Male	Med 3	18	1	0	-2	0	-2
13	Male	Med 3	24	1	0	-2	0	-2
14	Male	Med 3	21	1	0	-2	0	-2
15	Male	Med 3	16	1	0	-2	0	-2
16	Male	Med 3	23	1	0	-2	0	-2
17	Female	Med 1	17	-1	1	1	-1	-1
18	Female	Med 1	16	-1	1	1	-1	-1
19	Female	Med 1	13	-1	1	1	-1	-1
20	Female	Med 1	19	-1	1	1	-1	-1
21	Female	Med 1	18	-1	1	1	-1	-1
22	Female	Med 2	22	-1	-1	1	1	-1
23	Female	Med 2	20	-1	-1	1	1	-1

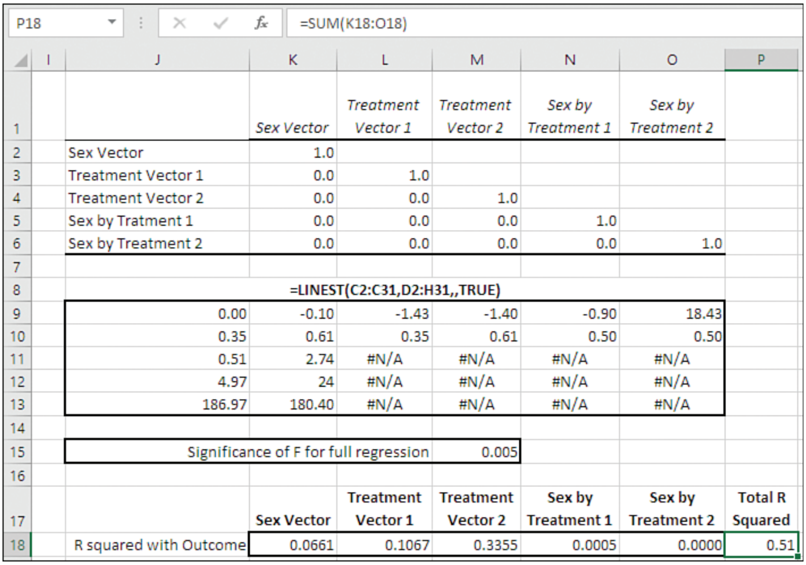
Figure 7.17
The two-factor problem from Figure 7.16 laid out for regression analysis.

The choice of codes in the Sex and Treatment vectors is made so that all the vectors will be mutually orthogonal. That's a different reason from the one used in Figure 7.15, where the idea is to specify contrasts that are of particular theoretical interest—the means of particular groups, and the combinations of group means, that you hope will inform you about the way that independent variables work together and with the dependent variable to bring about the observed outcomes.

But in Figure 7.17, the codes are chosen simply to make the vectors mutually orthogonal because it makes the subsequent analysis easier. The most straightforward way to do this is as follows.

1. Supply the first vector with codes that will contrast the first level of the factor with the second level, and ignore other levels. In Figure 7.17, the first factor has only two levels, so it requires only one vector, and the two levels exhaust the factor's information. Therefore, give one level of Sex a code of 1 and the other level a code of –1.
2. Do the same for the first level of the second factor. In this case the second factor is Treatment, which has three levels and therefore two vectors. The first level, Med 1, gets a 1 in the first vector and the second level, Med 2, gets a –1. All other levels, in this case Med 3, get 0's. This conforms to what was done with the Sex vector in Step 1.
3. In the second (and subsequent) vectors for a given factor, enter codes that contrast the first two levels with the third level (or the first three with the fourth, or the first four with the fifth, and so on). That's done in the second Treatment variable by assigning the code 1 to both Med 1 and Med 2, and –2 to Med 3. This contrasts the first two levels from the third. If there were other levels shown in this vector they would be assigned 0's.

The interaction vectors are obtained by multiplication of the main effect vectors, as described in the preceding steps. Now, Figure 7.18 shows the analysis of this data set.



P18 =SUM(K18:O18)

	I	J	K	L	M	N	O	P
1			Sex Vector	Treatment Vector 1	Treatment Vector 2	Sex by Treatment 1	Sex by Treatment 2	
2		Sex Vector	1.0					
3		Treatment Vector 1	0.0	1.0				
4		Treatment Vector 2	0.0	0.0	1.0			
5		Sex by Tratment 1	0.0	0.0	0.0	1.0		
6		Sex by Treatment 2	0.0	0.0	0.0	0.0	1.0	
7								
8			=LINEST(C2:C31,D2:H31,,TRUE)					
9		0.00	-0.10	-1.43	-1.40	-0.90	18.43	
10		0.35	0.61	0.35	0.61	0.50	0.50	
11		0.51	2.74	#N/A	#N/A	#N/A	#N/A	
12		4.97	24	#N/A	#N/A	#N/A	#N/A	
13		186.97	180.40	#N/A	#N/A	#N/A	#N/A	
14								
15		Significance of F for full regression			0.005			
16								
17			Sex Vector	Treatment Vector 1	Treatment Vector 2	Sex by Treatment 1	Sex by Treatment 2	Total R Squared
18		R squared with Outcome	0.0661	0.1067	0.3355	0.0005	0.0000	0.51

Figure 7.18
The orthogonal vectors all correlate 0.0 with one another.

In Figure 7.18, the correlation matrix in the range K2:O6 shows that the correlations between each pair of vectors is 0.0.

TIP

The Correlation tool in the Data Analysis add-in is a convenient way to create a matrix such as the one in K2:06.

The fact that all the correlations between the vectors are 0.0 means that the vectors share no variance. Because they share no variance, it's impossible for the relationships of two vectors with the outcome variable to overlap. Any variance shared by the outcome variable and, say, the first Treatment vector is unique to that Treatment vector. *When all the vectors are mutually orthogonal, there is no ambiguity about where to assign variance shared with the outcome variable.*

In the range J9:O13 of Figure 7.18 you'll find the results returned by LINEST() for the data shown in Figure 7.17. Not that it matters for present purposes, but the statistical significance of the overall regression is shown in cell M15, again using the F.DIST.RT() function. More pertinent is that the R^2 for the regression, 0.51, is found in cell J11.

The range K18:O18 contains the R^2 values for each coded vector with the outcome variable. Excel provides a function, RSQ(), that returns the square of the correlation between two variables. So the formula in cell K18 is:

```
=RSQ($C$2:$C$31,D2:D31)
```

Cell P18 shows the sum of the five R^2 values. That sum, 0.51, is identical to the R^2 for the full regression equation that's returned by LINEST() in cell J11. We have now partitioned the R^2 for the full equation into five constituents.

Figure 7.19 ties the results of the regression analysis back to the two-factor ANOVA in Figure 7.16.

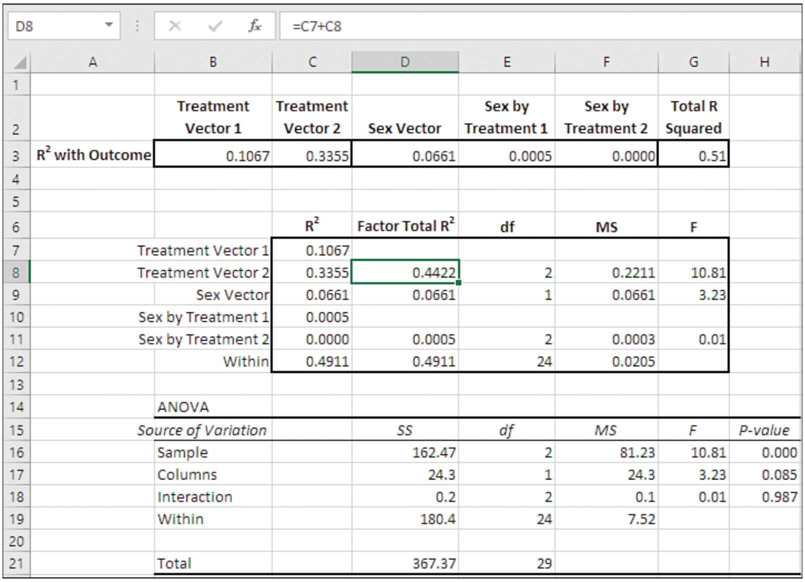


Figure 7.19
Compare the result of using sums of squares with the using proportions of variance.

D8 =C7+C8

	A	B	C	D	E	F	G	H
1								
2		Treatment Vector 1	Treatment Vector 2	Sex Vector	Sex by Treatment 1	Sex by Treatment 2	Total R Squared	
3	R^2 with Outcome	0.1067	0.3355	0.0661	0.0005	0.0000	0.51	
4								
5								
6			R^2	Factor Total R^2	df	MS	F	
7		Treatment Vector 1	0.1067					
8		Treatment Vector 2	0.3355	0.4422	2	0.2211	10.81	
9		Sex Vector	0.0661	0.0661	1	0.0661	3.23	
10		Sex by Treatment 1	0.0005					
11		Sex by Treatment 2	0.0000	0.0005	2	0.0003	0.01	
12		Within	0.4911	0.4911	24	0.0205		
13								
14		ANOVA						
15		*Source of Variation*		*SS*	*df*	*MS*	*F*	*P-value*
16		Sample		162.47	2	81.23	10.81	0.000
17		Columns		24.3	1	24.3	3.23	0.085
18		Interaction		0.2	2	0.1	0.01	0.987
19		Within		180.4	24	7.52		
20								
21		Total		367.37	29			

In Figure 7.19 I have brought forward the R^2 values for the coded vectors, and the total regression R^2, from the range K18:P18 in Figure 7.18. Recall that these R^2 values are actually measures of the proportion of total variance in the outcome variable associated with each vector. So the total amount of variance explained by the coded vectors is 51%, and therefore 49% of the total variance remains unexplained. That 49% of the variance is represented by the mean square residual (or mean square within, or mean square error) component—the divisor for the F-ratios.

The three main points to take from Figure 7.19 are discussed next.

Unique Proportions of Variance

The individual R^2 values for each vector can simply be summed to get the R^2 for the total regression equation. The simple sum is accurate because the vectors are mutually orthogonal. Each vector accounts for a unique proportion of the variance in the outcome. Therefore there is no double-counting of the variance, as there would be if the vectors were correlated. The total of the individual R^2 values equals the total R^2 for the full regression.

Proportions of Variance Equivalent to Sums of Squares

Compare the F-ratios from the analysis in the range C7:G12, derived from proportions of variance, with the F-ratios from the ANOVA in B15:H21, derived from the sums of squares. The F-ratios are identical, and the conclusions that you would draw from each analysis: that the sole significant difference is due to the treatments, and no difference emerges as a function of either sex or the interaction of sex with treatment.

The proportions of variance bear the same relationships to one another as do the sums of squares. That's not surprising. It's virtually by definition, because each proportion of variance is simply the sum of squares for that component divided by the total sum of squares—a constant. All the proportions of variance, including that associated with mean square residual, total to 1.0, so if you multiply an individual proportion of variance such as 0.4422 in cell D8, by the total sum of squares (367.37 in cell D21), you wind up with the sum of squares for that component (162.47 in cell D16). Generally, proportions of variance speak for themselves, while sums of squares don't. If you say that 44.22% of the variance in the outcome variable is due to the treatments, I immediately know how important the treatments are. If you say that the sum of squares due to treatments is 162.47, I suggest that you're not communicating with me.

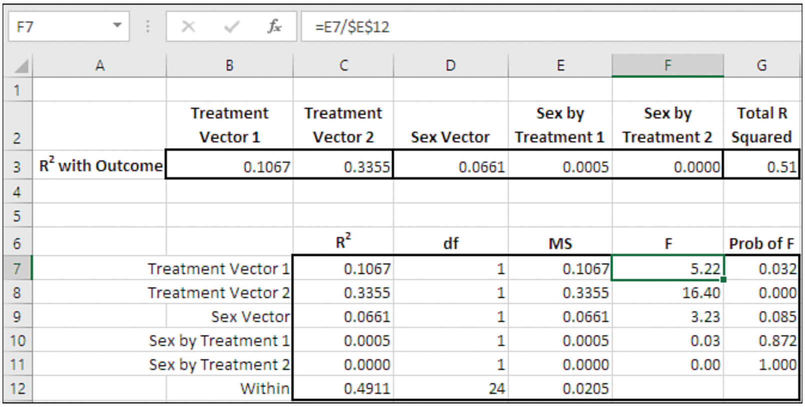
Summing Component Effects

The traditional ANOVA shown in B15:H21 of Figure 7.19 does not provide inferential information for each comparison. The sums of squares, the mean squares and the F-ratios are for the full factor. Traditional methods cannot distinguish, for example, the effect of Med 1 versus Med 2 from the effect of Med 2 versus Med 3. That's what multiple comparisons are for.

However, we can get an R^2 for each vector in the regression analysis. For example, the R^2 values in cells C7 and C8 are 0.1067 and 0.3355. Those proportions of variance are attributable to whatever comparison is implied by the codes in their respective vectors. As coded in Figure 7.17, Treatment Vector 1 compares Med 1 with Med 2, and Treatment Vector 2 compares the average of Med 1 and Med 2 with Med 3. Notice that if you add the two proportions of variance together and multiply by the total sum of squares, 367.37, you get the sum of squares associated with the Treatment factor in cell D16, returned by the traditional ANOVA.

The individual vectors can be tested in the same way as the collective vectors (one for each main effect and one for the interaction). Figure 7.20 demonstrates that more fine-grained analysis.

The probabilities in the range G7:G11 of Figure 7.20 indicate the likelihoods of obtaining an F-ratio as large as those in F7:F11 if the differences in means that are defined by the vectors' codes were all 0.0 in the population. So, if you had specified an alpha of 0.05, you could reject the null hypothesis for the Treatment 1 vector (Med 1 versus Med 2) and the Treatment 2 vector (the average of Med 1 and Med 2 versus Med 3). But if you had selected an alpha of 0.01, you could reject the null hypothesis for only the comparison in Treatment Vector 2.



F7 =E7/$E$12

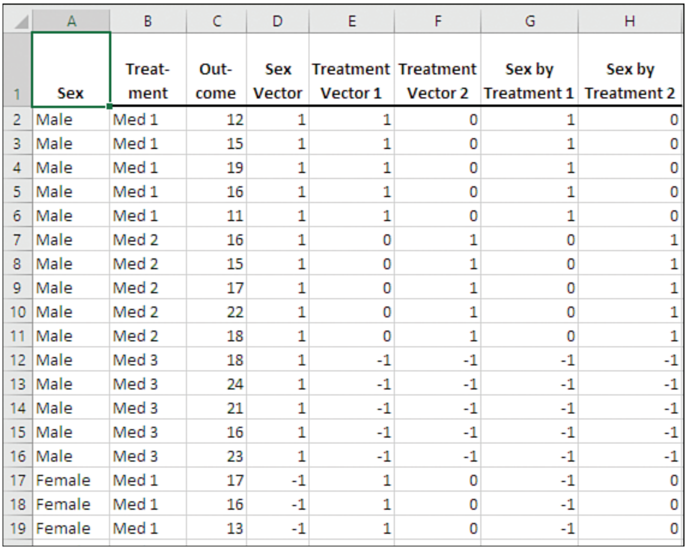
	A	B	C	D	E	F	G
1							
2		Treatment Vector 1	Treatment Vector 2	Sex Vector	Sex by Treatment 1	Sex by Treatment 2	Total R Squared
3	R^2 with Outcome	0.1067	0.3355	0.0661	0.0005	0.0000	0.51
4							
5							
6			R^2	df	MS	F	Prob of F
7		Treatment Vector 1	0.1067	1	0.1067	5.22	0.032
8		Treatment Vector 2	0.3355	1	0.3355	16.40	0.000
9		Sex Vector	0.0661	1	0.0661	3.23	0.085
10		Sex by Treatment 1	0.0005	1	0.0005	0.03	0.872
11		Sex by Treatment 2	0.0000	1	0.0000	0.00	1.000
12		Within	0.4911	24	0.0205		

Figure 7.20
The question of what is being tested by a vector's F-ratio depends on how you have coded the vector.

Factorial Analysis with Effect Coding

Chapter 3, in the section titled "Partial and Semipartial Correlations," discussed how the effect of a third variable can be statistically removed from the correlation between two other variables. The third variable's effect can be removed from both of the other two variables (partial correlation) or from just one of the other two (semipartial correlation). We'll make use of semipartial correlations—actually, the squares of the semipartial correlations—in this section. The technique also finds broad applicability in situations that involve unequal numbers of observations per group.

Figure 7.21 shows how the orthogonal coding from Figure 7.17, used in Figures 7.18 through 7.20, has been changed to effect coding.



	A	B	C	D	E	F	G	H
1	Sex	Treat-ment	Out-come	Sex Vector	Treatment Vector 1	Treatment Vector 2	Sex by Treatment 1	Sex by Treatment 2
2	Male	Med 1	12	1	1	0	1	0
3	Male	Med 1	15	1	1	0	1	0
4	Male	Med 1	19	1	1	0	1	0
5	Male	Med 1	16	1	1	0	1	0
6	Male	Med 1	11	1	1	0	1	0
7	Male	Med 2	16	1	0	1	0	1
8	Male	Med 2	15	1	0	1	0	1
9	Male	Med 2	17	1	0	1	0	1
10	Male	Med 2	22	1	0	1	0	1
11	Male	Med 2	18	1	0	1	0	1
12	Male	Med 3	18	1	-1	-1	-1	-1
13	Male	Med 3	24	1	-1	-1	-1	-1
14	Male	Med 3	21	1	-1	-1	-1	-1
15	Male	Med 3	16	1	-1	-1	-1	-1
16	Male	Med 3	23	1	-1	-1	-1	-1
17	Female	Med 1	17	-1	1	0	-1	0
18	Female	Med 1	16	-1	1	0	-1	0
19	Female	Med 1	13	-1	1	0	-1	0

Figure 7.21
As you'll see, effect coding results in vectors that are not wholly orthogonal.

In the Sex vector, Males are assigned 1's and Females are assigned –1's. With a two-level factor such as Sex, orthogonal coding is identical to effect coding.

The first Treatment vector assigns 1's to Med 1, 0's to Med 2, and –1's to Med 3. So Treatment Vector 1 contrasts Med 1 with Med 3. Treatment Vector 2 assigns 0's to Med 1, 1's to Med 2 and (again) –1's to Med 3, resulting in a contrast of Med 2 with Med 3. Thus, although they both provide tests of the Treatment variable, the two Treatment vectors define different contrasts than are defined by the orthogonal coding used in Figure 7.17.

Figure 7.22 displays the results of effect coding on the outcome variable, which has the same values as in Figure 7.17.

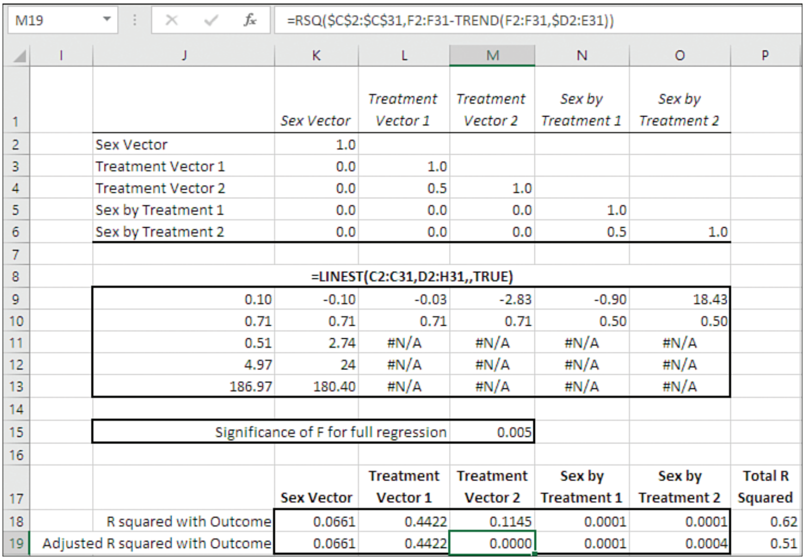


Figure 7.22
Vectors that represent different levels of a given factor are correlated if you use effect coding.

M19 =RSQ($C$2:$C$31,F2:F31-TREND(F2:F31,$D2:E31))

	I	J	K	L	M	N	O	P
1			*Sex Vector*	*Treatment Vector 1*	*Treatment Vector 2*	*Sex by Treatment 1*	*Sex by Treatment 2*	
2		Sex Vector	1.0					
3		Treatment Vector 1	0.0	1.0				
4		Treatment Vector 2	0.0	0.5	1.0			
5		Sex by Treatment 1	0.0	0.0	0.0	1.0		
6		Sex by Treatment 2	0.0	0.0	0.0	0.5	1.0	
7								
8		=LINEST(C2:C31,D2:H31,,TRUE)						
9		0.10	-0.10	-0.03	-2.83	-0.90	18.43	
10		0.71	0.71	0.71	0.71	0.50	0.50	
11		0.51	2.74	#N/A	#N/A	#N/A	#N/A	
12		4.97	24	#N/A	#N/A	#N/A	#N/A	
13		186.97	180.40	#N/A	#N/A	#N/A	#N/A	
14								
15		Significance of F for full regression			0.005			
16								
17			Sex Vector	Treatment Vector 1	Treatment Vector 2	Sex by Treatment 1	Sex by Treatment 2	Total R Squared
18		R squared with Outcome	0.0661	0.4422	0.1145	0.0001	0.0001	0.62
19	Adjusted R squared with Outcome		0.0661	0.4422	0.0000	0.0001	0.0004	0.51

Notice that not all the off-diagonal entries in the correlation matrix, in the range K2:O6, are 0.0. Treatment Vector 1 has a 0.50 correlation with Treatment Vector 2, and the two vectors that represent the Sex by Treatment interaction also correlate at 0.50. This is typical of effect coding, although it becomes evident in the correlation matrix only when a main effect has at least three levels (as does Treatment in this example).

The result is that the vectors are not all mutually orthogonal, and therefore we cannot simply add up each variable's R^2 to get the R^2 for the full regression equation, as is done in Figures 7.18 through 7.20. Furthermore, because the R^2 of the vectors do not represent unique proportions of variance, we can't simply use those R^2 values to test the statistical significance of each vector.

Instead, it's necessary to use squared semipartial correlations to adjust the R^2 values so that they *are* orthogonal, representing unique proportions of the variance of the outcome variable.

In Figure 7.22, the LINEST() analysis in the range J9:O13 returns the same values as the LINEST() analysis with orthogonal coding in Figure 7.18, *except* for the regression coefficients, the constant, and their standard errors. In other words, the differences between orthogonal and effect coding make no difference to the equation's R^2, its standard error of estimate, the F-ratio, the degrees of freedom for the residual, or the regression and residual sums of squares. This is not limited to effect and orthogonal coding. Regardless of the method you apply—dummy coding, for example—it makes no difference to the statistics that pertain to the equation generally. The differences in coding methods show up when you start to look at variable-to-variable quantities, such as a vector's regression coefficient or its simple R^2 with the outcome variable.

Notice the table of R^2 values in rows 18 and 19 of Figure 7.22. The R^2 values in row 18 are raw, unadjusted proportions of variance. They do not represent unique proportions shared with the outcome variable. As evidence of that, the totals of the R^2 values in rows 18 and 19 are shown in cells P18 and P19. The value in cell P18, the total of the unadjusted R^2 values in row 18, is 0.62, well in excess of the R^2 for the full regression reported by LINEST() in cell J11.

Most of the R^2 values in row 19, by contrast, are actually squared semipartial correlations. Two that should catch your eye are those in L19 and M19. Because the two vectors for the Treatment variable are correlated, the proportions of variance attributed to them in row 18 via the unadjusted R^2 values double-count some of the variance shared with the outcome variable. It's that double-counting that inflates the total of the R^2 values to 0.62 from its legitimate value of 0.51.

What we want to do is remove the effect of the vectors to the left of the second Treatment vector from the second Treatment vector itself. You can see how that's done most easily by starting with the R^2 in cell K19, and then following the trail of bread crumbs through cells L19 and M19, as follows:

Cell K19: The formula is as follows:

```
=RSQ($C$2:$C$31,D2:D31).
```

The vector in column D, Sex, is the leftmost variable in LINEST()'s X-value arguments. (Space limits prevent the display of column D in Figure 7.22, but it's visible in Figure 7.21 and, of course, in the downloaded workbook for Chapter 7.) No variables precede the vector in column D and so there's nothing to partial out of the Sex vector: We just accept the raw R^2.

Cell L19: The formula is:

```
=RSQ($C$2:$C$31,E2:E31-TREND(E2:E31,$D2:D31))
```

The fragment TREND(E2:E31,$D2:D31) predicts the values in E2:E31 (the first Treatment vector) from the values in the Sex vector (D2:D31). You can see from the correlation matrix at the top of Figure 7.22 that the correlation between the Sex vector and the first Treatment vector is 0.0. In that case, the regression of Treatment 1 on Sex predicts the mean of the Treatment 1 vector. The vector has equal numbers of 1's, 0's and –1's, so its mean is 0.0. In short, the R^2 formula subtracts 0.0 from the codes in E2:E31 and we wind up with the same result in L19 as we do in L18.

Cell M19: The formula is:

```
=RSQ($C$2:$C$31,F2:F31-TREND(F2:F31,$D2:E31))
```

Here's where the squared semipartial kicks in. This fragment:

```
TREND(F2:F31,$D2:E31)
```

predicts the values for the second Treatment vector, F2:F31, based on its relationship to the vectors in column D *and* column E, via the TREND() function. When those predicted values are subtracted from the actual codes in column F, via this fragment:

```
F2:F31-TREND(F2:F31,$D2:E31)
```

you're left with residuals: the values for the second Treatment vector in F2:F31 that have their relationship with the Sex and the first Treatment vector removed. With those effects gone, what's left of the codes in F2:F31 is unique and unshared with either Sex or Treatment Vector 1. As it happens, the result of removing the effects of the Sex and the Treatment 1 vectors eliminates the original relationship between the Treatment 2 vector and the outcome variable, leaving both a correlation and an R^2 of 0.0. The double-counting of the shared variance is also eliminated and, when the adjusting formulas are extended through O19, the sum of the proportions of variance in K19:O19 equals the R^2 for the full equation in cell J11.

The structure of the formulas that calculate the squared semipartial correlations deserves a little attention because it can save you time and headaches. Here again is the formula used in cell L19:

```
=RSQ($C$2:$C$31,E2:E31-TREND(E2:E31,$D2:D31))
```

The address $C$2:$C$31 contains the outcome variable. It is a fixed reference (although it might as well be treated as a mixed reference, $C2:$C31, because we won't intentionally paste it outside row 19). As we copy and paste it to the right of column L, we want to continue to point the RSQ function at C2:C31, and anchoring the reference to column C accomplishes that.

The other point to note in the RSQ() formula is the mixed reference $D2:D31. Here's what the formula changes to as you copy and paste it, or drag and drop it, from L19 one column right into M19:

```
=RSQ($C$2:$C$31,F2:F31-TREND(F2:F31,$D2:E31))
```

Notice first that the references to E2:E31 in L19 have changed to F2:F31, in response to copying the formula one column right. We're now looking at the squared semipartial correlation between the outcome variable in column C and the second Treatment vector in column F.

But the TREND() fragment shows that we're adjusting the codes in F2:F31 for their relationship to the codes in columns D *and* E. By dragging the formula on column to the right:

- $C$2:$C$31 remains unchanged. That's where the outcome variable is located.
- E2:E31 changes to F2:F31. That's the address of the second Treatment vector.
- $D2:D31 changes to $D2:E31. That's the address of the preceding predictor variables. We want to remove from the second Treatment vector the variance it shares with the preceding predictors: the Sex vector and the first Treatment vector.

By the time the formula reaches O19:

- $C$2:$C$31 remains unchanged.
- E2:E31 changes to H2:H31.
- $D2:D31 changes to $D2:G31.

The techniques I've outlined in this section become even more important in Chapter 8, where we take up the analysis of covariance (ANCOVA). In ANCOVA you use variables that are measured on an interval or ratio scale as though they were factors measured on a nominal scale. The idea is not just to make successive R^2 values unique, as discussed in the present section, but to equate different groups of subjects as though they entered the experiment on a common footing—in effect, giving random assignment an assist.

Statistical Power, Type I and Type II Errors

In previous chapters I have mentioned a topic termed *statistical power* from time to time. Because it is a major reason to carry out factorial analyses as discussed in this chapter, and to carry out the analysis of covariance as discussed in Chapter 8, it's important to develop a more thorough understanding of what statistical power is and how to quantify it.

On a purely conceptual level, statistical power refers to a statistical test's ability to identify the difference between two or more group means as genuine, when in fact the difference *is* genuine at the population level. You might think of statistical power as the sensitivity of a test to the difference between groups.

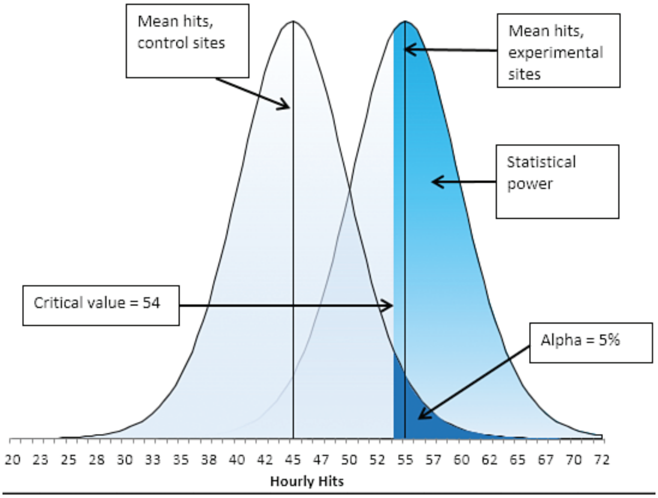
Suppose you're responsible for bringing a collection of websites to the attention of consumers who are shopping online. Your goal is to increase the number of hits that your websites experience; any resulting revenue and profit are up to the people who choose which products to market and how much to charge for them.

You arrange with the owner of a popular site for web searches to display links to 16 of your sites, randomly selected from among those that your company controls. The other randomly selected 16 of your sites will, for a month, get no special promotion.

Your intent is to compare the average number of hourly hits for the sites whose links get prominent display with the average number of hourly hits for the remaining sites. You decide to make a directional hypothesis at the 0.05 alpha level: Only if the specially promoted sites have a higher average number of hits, and only if the difference between the two groups of sites is so large that it could come about by chance only once in 20 replications of this trial, will you reject the hypothesis that the added promotion makes no difference to the hourly average number of hits.

Your data come in a month later and you find that your control group—the sites that received no special promotion—have an average of 45 hits each hour, and the specially promoted sites have an average hourly hit rate of 55. The standard error of the mean is 5. Figure 7.23 displays the situation graphically.

Figure 7.23
Both power and alpha can be thought of as probabilities and depicted as areas under a curve.



Assume that two populations exist: The first consists of websites like yours that get no special promotion. The second consists of websites that are promoted via links on another popular site, but that are otherwise equivalent to the first population. If you repeated your month-long study hundreds or perhaps thousands of times, you might get two distributions that look like the two curves in Figure 7.23.

The curve on the left represents the population of websites that get no special promotion. Over the course of a month, some of those sites—a very few—get as few as 25 hits per hour, and an equally small number get 62 hits per hour. The great majority of those sites average 45 hits per hour: the mode, mean and median of the curve on the left.

The curve on the right represents the specially promoted websites. They tend to get about 10 hits more per hour than the sites represented by the curve on the left. Their overall average is 55 hits per hour.

Now, most of this information is hidden from you. You don't have access to information about the full populations, just the results of the two samples you took—but that's enough.

Suppose that at the end of the month the two populations have the same mean, as would be the case if the extra promotion had no effect on the average hourly hits.

In that case, the difference in the average hit rate returned by your 16 experimental sites would have been due to nothing more than sampling error. That average of 55 hourly hits is among the averages in the right-hand tail of the curve on the left: the portion of the curve designated as *alpha*, shown in the chart in Figure 7.23 in a darker shade than the rest of the curve on the left.

Calculating Statistical Power

The boundary between alpha and the rest of the curve on the left is the critical value established by alpha. When you adopted 5% as your alpha level, with a directional hypothesis, you committed to the 5% of the right-hand tail of the curve. The critical value cuts off that 5%, and you can find that critical value using Excel's T.INV() function:

=T.INV(0.95,30)

That is, what is the value in the t distribution with 30 degrees of freedom that separates the lowest 95% of the values in the distribution from the top 5%? The result is 1.7. If you go up from the mean of the distribution by 1.7 standard errors, you account for the lowest 95% of the distribution. In this case the standard error is 5 (you learned that when you got the data on mean hourly hits), and 5 times 1.7 is 8.5. Add that to the mean of the curve on the left, and you get a critical value of 53.5.

In sum: The value of alpha is entirely under your control—it's *your* decision rule. You have made a directional hypothesis and you have set alpha to 0.05. Therefore, you have decided to reject the null hypothesis of no difference between the groups at the population level if, and only if, the experimental group's sample mean turns out to be at least 1.7 standard errors above the control group's mean.

Sometimes, the experimental group's mean will come from that right-hand tail of the left curve's distribution, just because of sampling error. Because the experimental group's mean, in that case, is at least 1.7 standard errors above the control group's mean, you'll reject the null hypothesis even though both populations have the same mean. That's Type I error, the probability of incorrectly rejecting a true null hypothesis.

Now suppose that in reality the populations are distributed as shown in Figure 7.23. If the sample experimental group has a mean at least 1.7 standard errors above the critical value of 54—which is 1.7 standard errors above the control group mean—then you'll *correctly* reject the null hypothesis of no difference at the population level.

Focus on the right curve in Figure 7.23. The area to the right of the critical value in that curve is the statistical power of your t-test. It is the probability that the experimental group mean comes from the curve on the right, in a reality where the two groups are distributed as shown at the population level.

Quantifying that probability is easy enough. Just take the difference between the critical value and the experimental group mean and divide by the standard error of 5:

=(54 – 55)/5

To get –0.2. That's a t-value. Evaluate it using the T.DIST() function:

=T.DIST(–0.2,15,TRUE)

using 15 as the degrees of freedom, because at this point we're working solely with the experimental group of 16 websites. The result is 0.422. That is, 42.2% of the area beneath the curve that represents the experimental group lies below the critical value of 54. Therefore 57.8% of the area under the curve lies to the right of the critical value, and the statistical power of the t-test is 57.8%. See Figure 7.24.

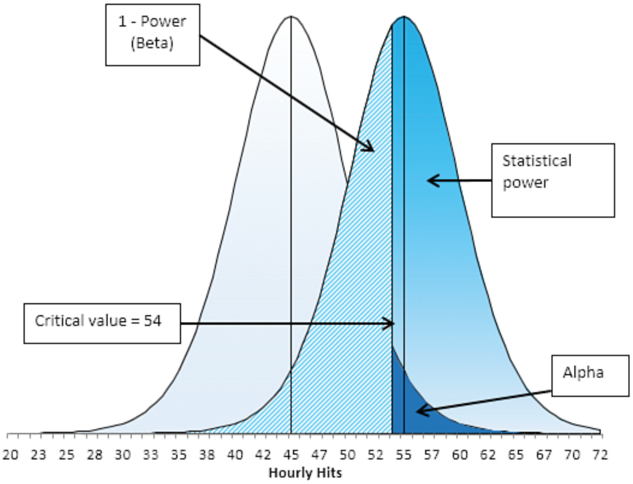


Figure 7.24
Type I error and alpha have counterparts in Type II error and beta.

In Figure 7.24 you can see the area that corresponds to statistical power in the curve on the right, to the right of the critical value. The remaining area under that curve is usually termed *beta*. It is alpha's counterpart.

If you incorrectly reject a true null hypothesis (for example, by deciding that two population means differ when in fact they don't), that's a Type I error and it has a probability of alpha. You decide the value of alpha, and your decision is typically based on the cost of making a Type I error, in the context of the benefits of correctly rejecting a false null hypothesis.

If you incorrectly reject a true alternative hypothesis (for example, by deciding that two population means are identical when in fact they differ), that's a Type II error and it has a probability of beta. The value of beta is not directly in your control. However, you can influence it, along with the statistical power of your test, as discussed in the next section.

Increasing Statistical Power

One excellent time to perform a power analysis is right after concluding a pilot study. At that point you often have the basic numbers on hand to calculate the power of a planned full study, and you're still in a position to make changes to the experimental design if the power study warrants. While a comparison of costs and benefits does not always argue for an increase in statistical power, it can warn you against pointless use of costly resources.

For example, if you can't get the estimated statistical power above 50%, you might decide that the study just isn't feasible—your odds of getting a reliable treatment effect are too low. Or it might turn out that increasing the sample size by 50% will result in an increase of only 5% in statistical power, so you're not getting enough bang for your buck.

You have available several methods of increasing statistical power. Some are purely theoretical, and have little chance of helping in real-world conditions. Others can make good sense.

One way is to reduce the size of the denominator of the test statistic. That denominator is typically a measure of the variability in the individual measures: a t-test, for example, might use either the standard error of the mean or the standard error of the difference between two means as the denominator of the t-ratio. An F-test uses the mean square residual (depending on the context, also known as mean square within or mean square error) as the denominator of the F-ratio.

When the denominator of a ratio decreases, the ratio itself increases. Other things equal, a larger t-ratio is more likely to be significant in a statistical sense than is a smaller t-ratio. One way to decrease the standard error or the mean square residual is to increase the sample size. Recall that the standard error of the mean divides the standard deviation by the square root of the sample size, and the mean square residual is the result of dividing the residual sum of squares by the residual degrees of freedom. In either case, increasing the sample size decreases the size of the t-ratio's or the F-ratio's denominator, which in turn increases the t-ratio or the F-ratio—improving the statistical power.

Another method of decreasing the size of the denominator is directly pertinent to factorial analysis, discussed in this chapter, and the analysis of covariance, discussed in Chapter 8. Both techniques add one or more predictors to the analysis: predictors that might have a substantial effect on the outcome variable. In that case, some of the variability in the individual measures can be attributed to the added factor or covariate and in that way kept out of the ratio's denominator.

So, adding a factor or covariate to the analysis might result in moving some of the variation out of the t-test's or the F-test's denominator and into the regression sum of squares (or the sum of squares between), thus increasing the size of the ratio and therefore its statistical power. Furthermore, and perhaps more importantly, adding the factor or the covariate could better illuminate the outcome of the study, particularly if two or more of the factors turn out to be involved in significant interactions.

You should also bear in mind three other ways to increase statistical power (neither of them directly related to the topics discussed in this chapter or in Chapter 8). One is to increase the treatment effect—the numerator of the t-ratio or the F-ratio, rather than its denominator. If you can increase the size of the treatment without also increasing the individual variation, your statistical test will be more powerful.

Consider making directional hypotheses ("one-tailed tests") instead of nondirectional hypotheses ("two-tailed tests"). One-tailed tests put all of alpha into one tail of the distribution. That moves the critical value toward the distribution's mean value. The closer the critical value is to the mean, the more likely you are to obtain an experimental result that exceeds the critical value—again, increasing the statistical power.

A related technique is to relax alpha. Notice in Figure 7.24 that if you increase (or *relax*) alpha from 0.05 to, say, 0.10, one result takes place in the distribution the right: the area representing statistical power increases as the critical value moves toward the mean of the curve on the left. By increasing the likelihood of making a Type I error, you reduce the likelihood of making a Type II error.

Coping with Unequal Cell Sizes

In Chapter 6 we looked at the combined effects of unequal group sizes and unequal variances on the nominal probability of a given t-ratio with a given degrees of freedom. You saw that the results can differ from the expected probabilities depending on whether the larger or the smaller group has the larger variance. You saw how Welch's correction can help compensate for unequal cell sizes.

Things are more complicated with more than just two groups (as in a t-test), particularly in factorial designs with two or more factors. Then, there are several—rather than just two—groups to compare as to both group size and variance.

In a design with at least two factors, and therefore at least four cells, several options exist, based primarily on the models comparison approach that is discussed at some length in Chapter 5. Unfortunately, these approaches do not have names that are generally accepted. Of the two discussed in this section, one is sometimes termed the *regression* approach and sometimes the *experimental design* approach; the other is sometimes termed the *sequential* approach and sometimes the *a priori ordering* approach. There are other terms in use. I use *experimental design* and *sequential* here.

The additional difficulty imposed by factorial designs when cell sizes are unequal concerns correlations between the vectors that define group membership: the 1's, 0's and −1's used in dummy, effect and orthogonal coding. Recall from earlier chapters that with equal cell sizes, the correlations between the vectors are largely 0.0. That feature means the sums of squares (and equivalently the variance) of the outcome variable can be assigned unambiguously to one vector or another.

But when the vectors are correlated, the unambiguous assignment of variability to a given vector becomes ambiguous. The vectors share variance with the outcome variable, of course. If they didn't there would be little point to retaining them in the analysis. But with unequal cell frequencies, the vectors share variance not only with the outcome variable but with one another, and in that case it's not possible to tell whether, say, 2% of the outcome variance belongs to Factor A, to Factor B, or to some sort of shared assignment such as 1.5% and 0.5%.

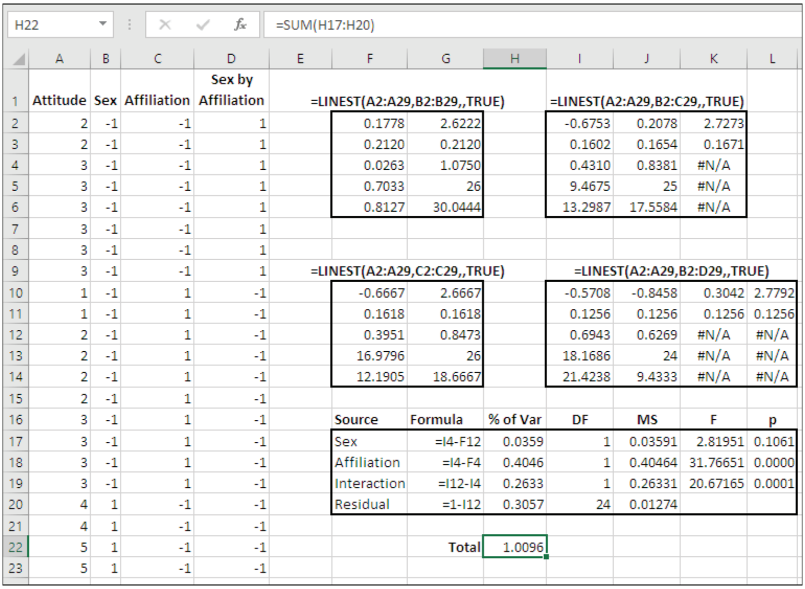
Despite the fact that I've cast this problem in terms of the vectors used in multiple regression analysis, the traditional ANOVA approaches are subject to the problem too.

But there's no single, generally applicable answer to the problem in the traditional framework either. The reliance there is typically on proportional (if unequal) cell frequencies and unweighted means analysis. Most current statistical packages use one of the approaches discussed here, or on one of their near relatives. (Not that it's representative of applications such as SAS or R, but Excel's Data Analysis add-in does not support designs with unequal cell frequencies in its 2-Factor ANOVA tool.)

Using the Regression Approach

This approach is also termed the *unique* approach because it treats each vector as though it were the last to enter the regression equation. If, say, a factor named Treatment is the last to enter the equation, all the other sources of regression variation are already in the equation and any variance shared by Treatment with the other vectors has already been assigned to them. Therefore, any remaining variance attributable to Treatment belongs to Treatment alone. It's unique to the Treatment vector.

Figure 7.25 shows an example of how this works.



H22

=SUM(H17:H20)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Attitude	Sex	Affiliation	Sex by Affiliation		=LINEST(A2:A29,B2:B29,,TRUE)			=LINEST(A2:A29,B2:C29,,TRUE)			
2	2	-1	-1	1		0.1778	2.6222		-0.6753	0.2078	2.7273	
3	2	-1	-1	1		0.2120	0.2120		0.1602	0.1654	0.1671	
4	3	-1	-1	1		0.0263	1.0750		0.4310	0.8381	#N/A	
5	3	-1	-1	1		0.7033	26		9.4675	25	#N/A	
6	3	-1	-1	1		0.8127	30.0444		13.2987	17.5584	#N/A	
7	3	-1	-1	1								
8	3	-1	-1	1								
9	3	-1	-1	1		=LINEST(A2:A29,C2:C29,,TRUE)			=LINEST(A2:A29,B2:D29,,TRUE)			
10	1	-1	1	-1		-0.6667	2.6667		-0.5708	-0.8458	0.3042	2.7792
11	1	-1	1	-1		0.1618	0.1618		0.1256	0.1256	0.1256	0.1256
12	2	-1	1	-1		0.3951	0.8473		0.6943	0.6269	#N/A	#N/A
13	2	-1	1	-1		16.9796	26		18.1686	24	#N/A	#N/A
14	2	-1	1	-1		12.1905	18.6667		21.4238	9.4333	#N/A	#N/A
15	2	-1	1	-1								
16	3	-1	1	-1		Source	Formula	% of Var	DF	MS	F	p
17	3	-1	1	-1		Sex	=I4-F12	0.0359	1	0.03591	2.81951	0.1061
18	3	-1	1	-1		Affiliation	=I4-F4	0.4046	1	0.40464	31.76651	0.0000
19	3	-1	1	-1		Interaction	=I12-I4	0.2633	1	0.26331	20.67165	0.0001
20	4	1	-1	-1		Residual	=1-I12	0.3057	24	0.01274		
21	4	1	-1	-1								
22	5	1	-1	-1			Total	1.0096				
23	5	1	-1	-1								

Figure 7.25
This design has four cells with different numbers of observations.

The idea is to use the models comparison approach to isolate the variance explained uniquely by each variable. Generally, we want to assess the factors one by one, before moving on to their interactions. So the process with this design is to subtract the variance explained by each main effect from the variance explained by all the main effects.

For example, in Figure 7.25, the range F10:G14 returns LINEST() results for Attitude, the outcome variable, regressed onto Affiliation, one of the two factors. Affiliation explains 39.51% of the variability in Attitude when Affiliation is the only vector entered: see cell F12.

Similarly, the range I2:K6 returns LINEST() results for the regression of Attitude on Sex *and* Affiliation. Cell I4 shows that together, Sex and Affiliation account for 43.1% of the variance in Attitude.

Therefore, with this data set, we can conclude that 43.10% – 39.51%, or 3.59%, of the variability in the outcome measure is specifically and uniquely attributable to the Sex factor: the proportion attributable to the two main effects less that attributable to Affiliation. This finding, 3.59%, differs from the result obtained from the LINEST() analysis in F2:G6, where cell F4 tells us that Sex accounts for 2.63% of the variance in the outcome measure. The difference between 3.59% and 2.63% is due to the fact that the unequal cell frequencies induce correlations between the vectors, introducing ambiguity into how the variance is allocated to the vectors.

NOTE

It's worth noting that if the cell frequencies were equal, the proportions of variance attributable to each factor would be the same whether a factor was isolated by means of the models comparison approach or by including only the one factor in LINEST( ). In terms of the example in Figure 7.25, if each of the four design cells had the same number of observations, the same proportion of variance would appear in both cells F4 and H17. The coded vectors would be orthogonal to one another, and it would make no difference whether a factor were the first or last to enter the regression equation: It would always account for the same proportion of variance.

The proportion of variance due to Affiliation is calculated in the same fashion. The proportion returned by LINEST() for the outcome measure regressed onto Sex, in cell F4, is subtracted from (once again) the proportion for both main effects, in cell I4. The result of that subtraction is 40.46%, a bit more than the 39.51% returned by the single factor LINEST() in cell F12.

After the two main effects, Sex and Affiliation, are assessed individually by subtracting their proportions of variance from the proportion accounted for by both, the analysis moves on to the interaction of the main effects. That's managed by subtracting the proportion of variance for the main effects, returned by LINEST() in cell I4, from the total proportion explained by the main effects and the interaction, returned by LINEST() in cell I12.

We can test the statistical significance of each main effect and the interaction in an ANOVA table, substituting proportions of total variance for sums of squares. That's done in the range F17:L20. Just divide the proportion of variance associated with each source of variation by its degrees of freedom to get a stand in for the mean square. Divide the mean square for each main effect and for the interaction by the mean square residual to obtain the F-ratio for each factor and for the interaction. The F-ratios are tested as usual with Excel's F.DIST.RT() function.

This analysis can be duplicated for sums of squares instead of proportions of variance, simply by multiplying each proportion by the sum of the squared deviations of the outcome variable from the grand mean.

Notice, by the way, in cell H22 that the proportions of variance do not total to precisely 100.00%, although the total is quite close. With unequal cell frequencies, even when managed by this unique variance approach, the total of the proportions of variance is not necessarily equal to exactly 100%. If you were working with sums of squares rather than proportions of variance, the factors' sums of squares do not necessarily add up precisely to the total sum of squares. Again, this is due to the adjustment of each factor and interaction vector for its correlation with the other vectors.

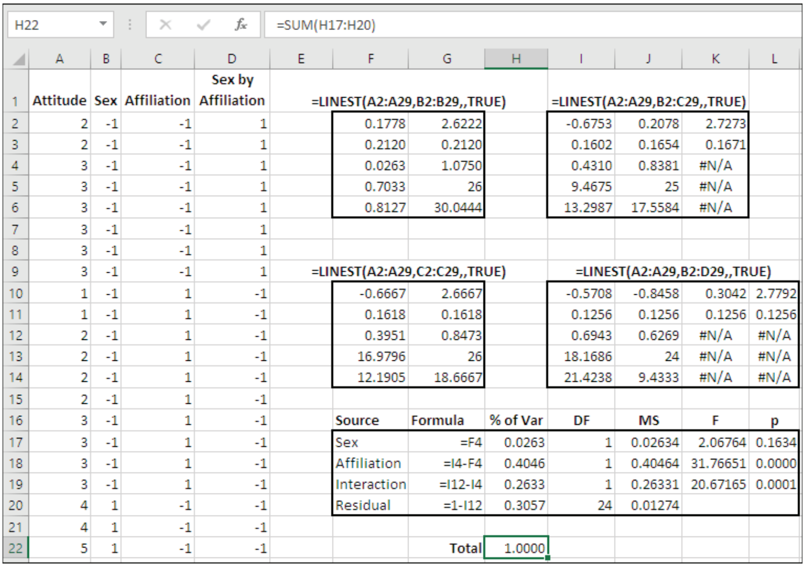
The sequential approach to dealing with unequal cell frequencies and correlated vectors, discussed in the next section, usually leads to a somewhat different outcome.

Sequential Variance Assignment

Bear in mind that the technique discussed in this section is just one of several methods for dealing with unequal cell frequencies in factorial designs. The unique assignment technique, described in the preceding section, is another such method. It differs from the sequential method in that it adjusts each factor's contribution to the regression sum of squares for that of the other factor or factors. In the sequential method, factors that are entered earlier are *not* adjusted for factors entered later.

Figure 7.26 shows how the sequential method works.

Figure 7.26
Consider using the sequential approach when one factor might have a causal effect on another.



H22 =SUM(H17:H20)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Attitude	Sex	Affiliation	Sex by Affiliation		=LINEST(A2:A29,B2:B29,,TRUE)			=LINEST(A2:A29,B2:C29,,TRUE)			
2	2	-1	-1	1		0.1778	2.6222		-0.6753	0.2078	2.7273	
3	2	-1	-1	1		0.2120	0.2120		0.1602	0.1654	0.1671	
4	3	-1	-1	1		0.0263	1.0750		0.4310	0.8381	#N/A	
5	3	-1	-1	1		0.7033	26		9.4675	25	#N/A	
6	3	-1	-1	1		0.8127	30.0444		13.2987	17.5584	#N/A	
7	3	-1	-1	1								
8	3	-1	-1	1								
9	3	-1	-1	1		=LINEST(A2:A29,C2:C29,,TRUE)			=LINEST(A2:A29,B2:D29,,TRUE)			
10	1	-1	1	-1		-0.6667	2.6667		-0.5708	-0.8458	0.3042	2.7792
11	1	-1	1	-1		0.1618	0.1618		0.1256	0.1256	0.1256	0.1256
12	2	-1	1	-1		0.3951	0.8473		0.6943	0.6269	#N/A	#N/A
13	2	-1	1	-1		16.9796	26		18.1686	24	#N/A	#N/A
14	2	-1	1	-1		12.1905	18.6667		21.4238	9.4333	#N/A	#N/A
15	2	-1	1	-1								
16	3	-1	1	-1		Source	Formula	% of Var	DF	MS	F	p
17	3	-1	1	-1		Sex	=F4	0.0263	1	0.02634	2.06764	0.1634
18	3	-1	1	-1		Affiliation	=I4-F4	0.4046	1	0.40464	31.76651	0.0000
19	3	-1	1	-1		Interaction	=I12-I4	0.2633	1	0.26331	20.67165	0.0001
20	4	1	-1	-1		Residual	=1-I12	0.3057	24	0.01274		
21	4	1	-1	-1								
22	5	1	-1	-1			Total	1.0000				

Figure 7.26 includes only two changes from Figure 7.25, but they can turn out to be important. In the sequential analysis shown in Figure 7.26, the variability associated with the Sex variable is unadjusted for its correlation with the Affiliation variable, whereas that adjustment occurs in Figure 7.25. Compare cell H17 in the two figures.

Here's the rationale for the difference. In this data set, the subjects are categorized according to their sex and their party affiliation. It is known that, nationally, women show a moderate preference for registering as Democrats rather than as Republicans, and the reverse is true for men. Therefore, any random sample of registered voters will have unequal cell frequencies (unless the researcher takes steps to ensure equal group sizes, a dubious practice at best when the variables are not directly under experimental control). And with those unequal cell frequencies come the correlations between the coded vectors that we're trying to deal with.

In this and similar cases, however, there's a good argument for assigning all the variance shared by Sex and Affiliation to the Sex variable. The reasoning is that a person's sex might influence his or her political preference (mediated, no doubt, by social and cultural variables that are sensitive to a person's sex). The reverse idea, that a person's sex is influenced by his or her choice of political party, is absurd.

Therefore, it's arguable that variance shared by Sex and Affiliation is due to Sex and not to Affiliation. In turn, that argues for allowing the Sex variable to retain all the variance that it can claim in a single factor analysis, and not to adjust the variance attributed to Sex according to its correlation with Affiliation.

In that case you can use the entire proportion of variance attributable to Sex in a single factor analysis as its proportion in the full analysis. That's what has been done in Figure 7.26, where the formula in cell H17 is:

```
=F4
```

instead of this:

```
=I4-F12
```

in cell H17 of Figure 7.25. In that figure, the variance attributable to the Sex factor is adjusted by subtracting the variance attributable to the Affiliation factor (cell F12) from the variance attributable to both main effects (cell I4). But in Figure 7.26, the variance attributable to Sex in a single-factor analysis in cell F4 is used in cell H17, unadjusted for variance it shares with Affiliation. Again, this is because in the researcher's judgment any variance shared by the two factors belongs to the Sex variable as, to some degree, causing variability in the Affiliation factor.

The adjustment, or the lack thereof, makes no practical difference in this case: The variance attributable to Sex is so small that it will not approach statistical significance whether it is adjusted or not. But with a different data set, the decision to adjust one variable for another as in the unique variance approach, or to retain the first factor's full variance, as in the sequential approach, could easily make a meaningful difference.

Notice, by the way, in Figure 7.26 that the coded vectors have in effect been made orthogonal to one another. The total of the proportions of variance in the range H17:H20 now comes to 1.000, as shown in cell H22. That demonstrates that the overlap in variability has been removed by the decision *not* to adjust Sex's variance for its correlation with Affiliation.

Why not follow the sequential approach in all cases? Because you need a good, sound reason to treat one factor as causal with respect to other factors. In this case, the fact of causality is underscored by the patterns in the full population, and the logic of the situation argues for the directionality of the cause: that Sex causes Affiliation rather than the other way around.

Nevertheless, this is a case in which the subjects assign themselves to groups by being of one sex or the other and by deciding which political party to belong to. If the researcher were to selectively discard subjects in order to achieve equal group sizes, without regard to causality, he would be artificially imposing orthogonality on the variables. So doing alters the reality of the situation, and you therefore need to be able to show that causality exists and what its direction is.

This consideration does not tend to arise in true experimental designs, where the researcher is in a position to randomly assign subjects to treatments and conditions. Broccoli plants are not in a position to decide whether they prefer organic or inorganic fertilizer, or whether they flourish in sun or in shade.

Let's move on to Chapter 8 now, and look further into the effect of adding a covariate to your design.

This page intentionally left blank

Index