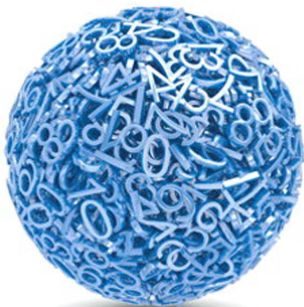


# STATISTICAL ANALYSIS:

Microsoft® Excel® 2013



que®

Conrad Carlberg

FREE SAMPLE CHAPTER

SHARE WITH OTHERS



# Statistical Analysis: Microsoft® Excel® 2013

Conrad Carlberg

**que**®

800 E. 96th Street  
Indianapolis, Indiana 46240

## Contents at a Glance

Introduction.....	xi
1 About Variables and Values.....	1
2 How Values Cluster Together.....	29
3 Variability: How Values Disperse.....	55
4 How Variables Move Jointly: Correlation.....	73
5 How Variables Classify Jointly: Contingency Tables.....	109
6 Telling the Truth with Statistics.....	149
7 Using Excel with the Normal Distribution.....	171
8 Testing Differences Between Means: The Basics.....	199
9 Testing Differences Between Means: Further Issues.....	227
10 Testing Differences Between Means: The Analysis of Variance.....	263
11 Analysis of Variance: Further Issues.....	293
12 Experimental Design and ANOVA.....	315
13 Statistical Power.....	331
14 Multiple Regression Analysis and Effect Coding: The Basics.....	355
15 Multiple Regression Analysis: Further Issues.....	385
16 Analysis of Covariance: The Basics.....	433
17 Analysis of Covariance: Further Issues.....	453
Index.....	473

# Statistical Analysis: Microsoft® Excel® 2013

Copyright © 2014 by Pearson Education

All rights reserved. No part of this book shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Nor is any liability assumed for damages resulting from the use of the information contained herein.

ISBN-13: 978-0-7897-5311-3

ISBN-10: 0-7897-5311-1

Library of Congress Control Number: 2013956944

Printed in the United States of America

First Printing: April 2014

## 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](mailto:corpsales@pearsoned.com) or (800) 382-3419.

For government sales inquiries, please contact [governmentsales@pearsoned.com](mailto:governmentsales@pearsoned.com).

For questions about sales outside the U.S., please contact [international@pearsoned.com](mailto:international@pearsoned.com).

## Editor-in-Chief

Greg Wiegand

## Acquisitions Editor

Loretta Yates

## Development Editor

Brandon Cackowski-Schnell

## Managing Editor

Kristy Hart

## Project Editor

Elaine Wiley

## Copy Editor

Keith Cline

## Indexer

Tim Wright

## Proofreader

Sara Schumacher

## Technical Editor

Michael Turner

## Editorial Assistant

Cindy Teeters

## Cover Designer

Matt Coleman

## Composer

Nonie Ratcliff

# Table of Contents

<b>Introduction</b> .....	<b>xi</b>
Using Excel for Statistical Analysis .....	xi
About You and About Excel.....	xii
Clearing Up the Terms .....	xii
Making Things Easier .....	xiii
The Wrong Box? .....	xiv
Wagging the Dog .....	xvi
What's in This Book .....	xvi
<b>1 About Variables and Values</b> .....	<b>1</b>
Variables and Values.....	1
Recording Data in Lists.....	2
Scales of Measurement.....	4
Category Scales.....	5
Numeric Scales .....	7
Telling an Interval Value from a Text Value .....	8
Charting Numeric Variables in Excel.....	10
Charting Two Variables .....	10
Understanding Frequency Distributions.....	12
Using Frequency Distributions .....	15
Building a Frequency Distribution from a Sample .....	18
Building Simulated Frequency Distributions .....	26
<b>2 How Values Cluster Together</b> .....	<b>29</b>
Calculating the Mean .....	30
Understanding Functions, Arguments, and Results.....	31
Understanding Formulas, Results, and Formats.....	34
Minimizing the Spread.....	36
Calculating the Median.....	41
Choosing to Use the Median .....	41
Calculating the Mode .....	42
Getting the Mode of Categories with a Formula .....	47
From Central Tendency to Variability .....	54
<b>3 Variability: How Values Disperse</b> .....	<b>55</b>
Measuring Variability with the Range .....	56
The Concept of a Standard Deviation .....	58
Arranging for a Standard.....	59
Thinking in Terms of Standard Deviations .....	60
Calculating the Standard Deviation and Variance.....	62
Squaring the Deviations.....	65
Population Parameters and Sample Statistics .....	66
Dividing by $N - 1$ .....	66

Bias in the Estimate . . . . .	68
Degrees of Freedom . . . . .	69
Excel's Variability Functions . . . . .	70
Standard Deviation Functions . . . . .	70
Variance Functions . . . . .	71
<b>4 How Variables Move Jointly: Correlation . . . . .</b>	<b>73</b>
Understanding Correlation . . . . .	73
The Correlation, Calculated . . . . .	75
Using the CORREL() Function . . . . .	81
Using the Analysis Tools . . . . .	84
Using the Correlation Tool . . . . .	86
Correlation Isn't Causation . . . . .	88
Using Correlation . . . . .	90
Removing the Effects of the Scale . . . . .	91
Using the Excel Function . . . . .	93
Getting the Predicted Values . . . . .	95
Getting the Regression Formula . . . . .	96
Using TREND() for Multiple Regression . . . . .	99
Combining the Predictors . . . . .	99
Understanding "Best Combination" . . . . .	100
Understanding Shared Variance . . . . .	104
A Technical Note: Matrix Algebra and Multiple Regression in Excel . . . . .	106
Moving on to Statistical Inference . . . . .	107
<b>5 How Variables Classify Jointly: Contingency Tables . . . . .</b>	<b>109</b>
Understanding One-Way Pivot Tables . . . . .	109
Running the Statistical Test . . . . .	112
Making Assumptions . . . . .	117
Random Selection . . . . .	118
Independent Selections . . . . .	119
The Binomial Distribution Formula . . . . .	120
Using the BINOM.INV() Function . . . . .	121
Understanding Two-Way Pivot Tables . . . . .	127
Probabilities and Independent Events . . . . .	130
Testing the Independence of Classifications . . . . .	131
The Yule Simpson effect . . . . .	137
Summarizing the Chi-Square Functions . . . . .	140
Using CHISQ.DIST() . . . . .	140
Using CHISQ.DIST.RT() and CHIDIST() . . . . .	141
Using CHISQ.INV() . . . . .	143
Using CHISQ.INV.RT() and CHIINV() . . . . .	143
Using CHISQ.TEST() and CHITEST() . . . . .	144
Using Mixed and Absolute References to Calculate Expected Frequencies . . . . .	145
Using the Pivot Table's Index Display . . . . .	146

<b>6</b>	<b>Telling the Truth with Statistics</b>	<b>149</b>
	A Context for Inferential Statistics	150
	Establishing Internal Validity	151
	Threats to Internal Validity	152
	Problems with Excel's Documentation	156
	The F-Test Two-Sample for Variances	157
	Why Run the Test?	158
	A Final Point	169
<b>7</b>	<b>Using Excel with the Normal Distribution</b>	<b>171</b>
	About the Normal Distribution	171
	Characteristics of the Normal Distribution	171
	The Unit Normal Distribution	176
	Excel Functions for the Normal Distribution	177
	The NORM.DIST() Function	177
	The NORM.INV() Function	180
	Confidence Intervals and the Normal Distribution	182
	The Meaning of a Confidence Interval	183
	Constructing a Confidence Interval	184
	Excel Worksheet Functions That Calculate Confidence Intervals	187
	Using CONFIDENCE.NORM() and CONFIDENCE()	188
	Using CONFIDENCE.T()	191
	Using the Data Analysis Add-In for Confidence Intervals	192
	Confidence Intervals and Hypothesis Testing	194
	The Central Limit Theorem	194
	Making Things Easier	196
	Making Things Better	198
<b>8</b>	<b>Testing Differences Between Means: The Basics</b>	<b>199</b>
	Testing Means: The Rationale	200
	Using a z-Test	201
	Using the Standard Error of the Mean	204
	Creating the Charts	208
	Using the t-Test Instead of the z-Test	216
	Defining the Decision Rule	218
	Understanding Statistical Power	222
<b>9</b>	<b>Testing Differences Between Means: Further Issues</b>	<b>227</b>
	Using Excel's T.DIST() and T.INV() Functions to Test Hypotheses	227
	Making Directional and Nondirectional Hypotheses	228
	Using Hypotheses to Guide Excel's t-Distribution Functions	229
	Completing the Picture with T.DIST()	237
	Using the T.TEST() Function	238
	Degrees of Freedom in Excel Functions	238
	Equal and Unequal Group Sizes	239
	The T.TEST() Syntax	242

Using the Data Analysis Add-in t-Tests .....	255
Group Variances in t-Tests.....	255
Visualizing Statistical Power.....	260
When to Avoid t-Tests .....	261
<b>10 Testing Differences Between Means: The Analysis of Variance.....</b>	<b>263</b>
Why Not t-Tests?.....	263
The Logic of ANOVA.....	265
Partitioning the Scores.....	265
Comparing Variances.....	268
The F Test.....	273
Using Excel's Worksheet Functions for the F Distribution .....	277
Using F.DIST() and F.DIST.RT() .....	277
Using F.INV() and FINV() .....	278
The F Distribution .....	279
Unequal Group Sizes .....	280
Multiple Comparison Procedures .....	282
The Scheffé Procedure .....	284
Planned Orthogonal Contrasts .....	289
<b>11 Analysis of Variance: Further Issues.....</b>	<b>293</b>
Factorial ANOVA.....	293
Other Rationales for Multiple Factors .....	294
Using the Two-Factor ANOVA Tool.....	297
The Meaning of Interaction .....	299
The Statistical Significance of an Interaction .....	300
Calculating the Interaction Effect.....	302
The Problem of Unequal Group Sizes .....	307
Repeated Measures: The Two Factor Without Replication Tool.....	309
Excel's Functions and Tools: Limitations and Solutions .....	310
Mixed Models .....	312
Power of the F Test.....	312
<b>12 Experimental Design and ANOVA.....</b>	<b>315</b>
Crossed Factors and Nested Factors .....	315
Depicting the Design Accurately.....	317
Nuisance Factors.....	317
Fixed Factors and Random Factors .....	318
The Data Analysis Add-In's ANOVA Tools .....	319
Data Layout.....	320
Calculating the F Ratios .....	322
Adapting the Data Analysis Tool for a Random Factor .....	322
Designing the F Test.....	323
The Mixed Model: Choosing the Denominator.....	325
Adapting the Data Analysis Tool for a Nested Factor.....	326

Data Layout for a Nested Design .....	327
Getting the Sums of Squares .....	328
Calculating the F Ratio for the Nesting Factor .....	329
<b>13 Statistical Power .....</b>	<b>331</b>
Controlling the Risk .....	331
Directional and Nondirectional Hypotheses .....	332
Changing the Sample Size .....	332
Visualizing Statistical Power .....	333
Quantifying Power .....	335
The Statistical Power of t-Tests .....	337
Nondirectional Hypotheses .....	338
Making a Directional Hypothesis .....	340
Increasing the Size of the Samples .....	341
The Dependent Groups t-Test .....	342
The Noncentrality Parameter in the F Distribution .....	344
Variance Estimates .....	344
The Noncentrality Parameter and the Probability Density Function .....	348
Calculating the Power of the F Test .....	350
Calculating the Cumulative Density Function .....	350
Using Power to Determine Sample Size .....	352
<b>14 Multiple Regression Analysis and Effect Coding: The Basics .....</b>	<b>355</b>
Multiple Regression and ANOVA .....	356
Using Effect Coding .....	358
Effect Coding: General Principles .....	358
Other Types of Coding .....	359
Multiple Regression and Proportions of Variance .....	360
Understanding the Segue from ANOVA to Regression .....	363
The Meaning of Effect Coding .....	365
Assigning Effect Codes in Excel .....	368
Using Excel's Regression Tool with Unequal Group Sizes .....	370
Effect Coding, Regression, and Factorial Designs in Excel .....	372
Exerting Statistical Control with Semipartial Correlations .....	374
Using a Squared Semipartial to Get the Correct Sum of Squares .....	376
Using Trend() to Replace Squared Semipartial Correlations .....	377
Working With the Residuals .....	379
Using Excel's Absolute and Relative Addressing to Extend the Semipartials .....	381
<b>15 Multiple Regression Analysis and Effect Coding: Further Issues .....</b>	<b>385</b>
Solving Unbalanced Factorial Designs Using Multiple Regression .....	385
Variables Are Uncorrelated in a Balanced Design .....	386
Variables Are Correlated in an Unbalanced Design .....	388
Order of Entry Is Irrelevant in the Balanced Design .....	388
Order Entry Is Important in the Unbalanced Design .....	391
About Fluctuating Proportions of Variance .....	393



Experimental Designs, Observational Studies, and Correlation .....	394
Using All the LINEST() Statistics .....	397
Using the Regression Coefficients .....	398
Using the Standard Errors .....	398
Dealing with the Intercept .....	399
Understanding LINEST()'s Third, Fourth, and Fifth Rows .....	400
Getting the Regression Coefficients .....	406
Getting the Sum of Squares Regression and Residual .....	410
Calculating the Regression Diagnostics .....	412
How LINEST() Handles Multicollinearity .....	416
Forcing a Zero Constant .....	421
The Excel 2007 Version .....	422
A Negative R <sup>2</sup> ? .....	425
Managing Unequal Group Sizes in a True Experiment .....	428
Managing Unequal Group Sizes in Observational Research .....	430
<b>16 Analysis of Covariance: The Basics .....</b>	<b>433</b>
The Purposes of ANCOVA .....	434
Greater Power .....	434
Bias Reduction .....	434
Using ANCOVA to Increase Statistical Power .....	435
ANOVA Finds No Significant Mean Difference .....	436
Adding a Covariate to the Analysis .....	437
Testing for a Common Regression Line .....	445
Removing Bias: A Different Outcome .....	447
<b>17 Analysis of Covariance: Further Issues .....</b>	<b>453</b>
Adjusting Means with LINEST() and Effect Coding .....	453
Effect Coding and Adjusted Group Means .....	458
Multiple Comparisons Following ANCOVA .....	461
Using the Scheffé Method .....	462
Using Planned Contrasts .....	466
The Analysis of Multiple Covariance .....	468
The Decision to Use Multiple Covariates .....	469
Two Covariates: An Example .....	470
<b>Index .....</b>	<b>473</b>

## About the Author

**Conrad Carlberg** started writing about Excel, and its use in quantitative analysis, before workbooks had worksheets. As a graduate student, he had the great good fortune to learn something about statistics from the wonderfully gifted Gene Glass. He remembers much of that and has learned more since. This is a book he has wanted to write for years, and he is grateful for the opportunity.

## Dedication

*For Toni, who has been putting up with this sort of thing for 17 years now, with all my love.*

## Acknowledgments

I'd like to thank Loretta Yates, who guided this book's overall progress, and who treats my self-imposed crises with an unexpected sort of pragmatic optimism. Michael Turner's technical edit was just right, and it was a delight to see how, at the stats lab anyway, the more things change...well, you know. Keith Cline kept the prose on track, despite my occasional howls of protest, with his copy edit. And in the end, Elaine Wiley somehow managed to get the whole thing put together. My thanks to each of you.

## 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](mailto:feedback@quepublishing.com)

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

## Reader Services

Visit our website and register this book at [quepublishing.com/register](http://quepublishing.com/register) for convenient access to any updates, downloads, or errata that might be available for this book.

# Introduction

There was no reason I shouldn't have already written a book about statistical analysis using Excel. But I didn't, although I knew I wanted to. Finally, I talked Pearson into letting me write it for them.

Be careful what you ask for. It's been a struggle, but at last I've got it out of my system, and I want to start by talking here about the reasons for some of the choices I made in writing this book.

## Using Excel for Statistical Analysis

The problem is that it's a huge amount of material to cover in a book that's supposed to be only 400 to 500 pages. The text used in the first statistics course I took was about 600 pages, and it was purely statistics, no Excel. In 2001, I co-authored a book about Excel (no statistics) that ran to 750 pages. To shoe-horn statistics *and* Excel into 400 pages or so takes some picking and choosing.

Furthermore, I did not want this book to be an expanded Help document, like one or two others I've seen. Instead, I take an approach that seemed to work well in an earlier book of mine, *Business Analysis with Excel*. The idea in both that book and this one is to identify a topic in statistical (or business) analysis; discuss the topic's rationale, its procedures, and associated issues; and only then get into how it's carried out in Excel.

You shouldn't expect to find discussions of, say, the Weibull function or the lognormal distribution here. They have their uses, and Excel provides them as statistical functions, but my picking and choosing forced me to ignore them—at my peril, probably—and to use the space saved for material on more bread-and-butter topics such as statistical regression.

Using Excel for Statistical Analysis ..... xi

What's in This Book ..... xvi



## About You and About Excel

How much background in statistics do you need to get value from this book? My intention is that you need none. The book starts out with a discussion of different ways to measure things—by categories, such as models of cars, by ranks, such as first place through tenth, by numbers, such as degrees Fahrenheit—and how Excel handles those methods of measurement in its worksheets and its charts.

This book moves on to basic statistics, such as averages and ranges, and only then to intermediate statistical methods such as t-tests, multiple regression, and the analysis of covariance. The material assumes knowledge of nothing more complex than how to calculate an average. You do not need to have taken courses in statistics to use this book.

As to Excel itself, it matters little whether you're using Excel 97, Excel 2013, or any version in between. Very little statistical functionality changed between Excel 97 and Excel 2003. The few changes that did occur had to do primarily with how functions behaved when the user stress-tested them using extreme values or in very unlikely situations.

The Ribbon showed up in Excel 2007 and is still with us in Excel 2013. But nearly all statistical analysis in Excel takes place in worksheet functions—very little is menu driven—and there was almost no change to the function list, function names, or their arguments between Excel 97 and Excel 2007. The Ribbon does introduce a few differences, such as how to get a trendline into a chart. This book discusses the differences in the steps you take using the traditional menu structure and the steps you take using the Ribbon.

In Excel 2010, several apparently new statistical functions appeared, but the differences were more apparent than real. For example, through Excel 2007, the two functions that calculate standard deviations are `STDEV()` and `STDEVP()`. If you are working with a sample of values, you should use `STDEV()`, but if you happen to be working with a full population, you should use `STDEVP()`. Of course, the *P* stands for *population*.

Both `STDEV()` and `STDEVP()` remain in Excel 2010 and 2013, but they are termed *compatibility functions*. It appears that they may be phased out in some future release. Excel 2010 added what it calls *consistency functions*, two of which are `STDEV.S()` and `STDEV.P()`. Note that a period has been added in each function's name. The period is followed by a letter that, for consistency, indicates whether the function should be used with a sample of values or a population of values.

Other consistency functions were added to Excel 2010, and the functions they are intended to replace are still supported in Excel 2013. There are a few substantive differences between the compatibility version and the consistency version of some functions, and this book discusses those differences and how best to use each version.

## Clearing Up the Terms

Terminology poses another problem, both in Excel and in the field of statistics (and, it turns out, in the areas where the two overlap). For example, it's normal to use the word *alpha* in a statistical context to mean the probability that you will decide that there's a true difference

between the means of two groups when there really isn't. But Excel extends *alpha* to usages that are related but much less standard, such as the probability of getting some number of heads from flipping a fair coin. It's not wrong to do so. It's just unusual, and therefore it's an unnecessary hurdle to understanding the concepts.

The vocabulary of statistics itself is full of names that mean very different things in slightly different contexts. The word *beta*, for example, can mean the probability of deciding that a true difference does *not* exist, when it does. It can also mean a coefficient in a regression equation (for which Excel's documentation unfortunately uses the letter *m*), and it's also the name of a distribution that is a close relative of the binomial distribution. None of that is due to Excel. It's due to having more concepts than there are letters in the Greek alphabet.

You can see the potential for confusion. It gets worse when you hook Excel's terminology up with that of statistics. For example, in Excel the word *cell* means a rectangle on a worksheet, the intersection of a row and a column. In statistics, particularly the analysis of variance, *cell* usually means a group in a factorial design: If an experiment tests the joint effects of sex and a new medication, one cell might consist of men who receive a placebo, and another might consist of women who receive the medication being assessed. Unfortunately, you can't depend on seeing "cell" where you might expect it: *within cell error* is called *residual error* in the context of regression analysis.

So this book presents you with some terms you might otherwise find redundant: I use *design cell* for analysis contexts and *worksheet cell* when referring to the software context where there's any possibility of confusion about which I mean.

For consistency, though, I try always to use *alpha* rather than *Type I error* or *statistical significance*. In general, I use just one term for a given concept throughout. I intend to complain about it when the possibility of confusion exists: when *mean square* doesn't mean *mean square*, you ought to know about it.

## Making Things Easier

If you're just starting to study statistical analysis, your timing's much better than mine was. You have avoided some of the obstacles to understanding statistics that once—as recently as the 1980s—stood in the way. I'll mention those obstacles once or twice more in this book, partly to vent my spleen but also to stress how much better Excel has made things.

Suppose that 25 years ago you were calculating something as basic as the standard deviation of twenty numbers. You had no access to a computer. Or, if there was one around, it was a mainframe or a mini, and whoever owned it had more important uses for it than to support a Psychology 101 assignment.

So you trudged down to the Psych building's basement, where there was a room filled with gray metal desks with adding machines on them. Some of the adding machines might even have been plugged into a source of electricity. You entered your twenty numbers very carefully because the adding machines did not come with Undo buttons or Ctrl+Z. The

electricity-enabled machines were in demand because they had a memory function that allowed you to enter a number, square it, and add the result to what was already in the memory.

It could take half an hour to calculate the standard deviation of twenty numbers. It was all incredibly tedious and it distracted you from the main point, which was the concept of a standard deviation and the reason you wanted to quantify it.

Of course, 25 years ago our teachers were telling us how lucky we were to have adding machines instead of having to use paper, pencil, and a box of erasers.

Things are different in 2013, and truth be told, they have been changing since the mid 1980s when applications such as Lotus 1-2-3 and Microsoft Excel started to find their way onto personal computers' floppy disks. Now, all you have to do is enter the numbers into a worksheet—or maybe not even that, if you downloaded them from a server somewhere. Then, type `=STDEV.S()` (and drag across the cells with the numbers before you press Enter. It takes half a minute at most, not half an hour at least.

Several statistics have relatively simple *definitional* formulas. The definitional formula tends to be straightforward and therefore gives you actual insight into what the statistic means. But those same definitional formulas often turn out to be difficult to manage in practice if you're using paper and pencil, or even an adding machine or hand calculator. Rounding errors occur and compound one another.

So statisticians developed *computational* formulas. These are mathematically equivalent to the definitional formulas, but are much better suited to manual calculations. Although it's nice to have computational formulas that ease the arithmetic, those formulas make you take your eye off the ball. You're so involved with accumulating the sum of the squared values that you forget that your purpose is to understand how values vary around their average.

That's one primary reason that an application such as Excel, or an application specifically and solely designed for statistical analysis, is so helpful. It takes the drudgery of the arithmetic off your hands and frees you to think about what the numbers actually mean.

Statistics is conceptual. It's not just arithmetic. And it shouldn't be taught as though it is.

## The Wrong Box?

But should you even be using Excel to do statistical calculations? After all, people have been moaning about inadequacies in Excel's statistical functions for twenty years. The Excel forum on CompuServe had plenty of complaints about this issue, as did the Usenet newsgroups. As I write this introduction, I can switch from Word to Firefox and see that some people are still complaining on Wikipedia talk pages, and others contribute angry screeds to publications such as *Computational Statistics & Data Analysis*, which I believe are there as a reminder to us all of the importance of taking our prescription medication.

I have sometimes found myself as upset about problems with Excel's statistical functions as anyone. And it's true that Excel has had, and in some cases continues to have, problems with the algorithms it uses to manage certain functions such as the inverse of the F distribution.

But most of the complaints that are voiced fall into one of two categories: those that are based on misunderstandings about either Excel or statistical analysis, and those that are based on complaints that Excel isn't accurate enough.

If you read this book, you'll be able to avoid those kinds of misunderstandings. As to inaccuracies in Excel results, let's look a little more closely at that. The complaints are typically along these lines:

I enter into an Excel worksheet two different formulas that should return the same result. Simple algebraic rearrangement of the equations proves that. But then I find that Excel calculates two different results.

Well, for the data the user supplied, the results differ at the fifteenth decimal place, so Excel's results disagree with one another by approximately five in 111 trillion.

Or this:

I tried to get the inverse of the F distribution using the formula `FINV(0.025,4198986,1025419)`, but I got an unexpected result. Is there a bug in `FINV`?

No. Once upon a time, `FINV` returned the `#NUM!` error value for those arguments, but no longer. However, that's not the point. With so many degrees of freedom (over four million and one million, respectively), the person who asked the question was effectively dealing with populations, not samples. To use that sort of inferential technique with so many degrees of freedom is a striking instance of "unclear on the concept."

Would it be better if Excel's math were more accurate—or at least more internally consistent? Sure. But even the finger-waggers admit that Excel's statistical functions are acceptable at least, as the following comment shows.

They can rarely be relied on for more than four figures, and then only for  $0.001 < p < 0.999$ , plenty good for routine hypothesis testing.

Now look. Chapter 6, "Telling the Truth with Statistics," goes into this issue further, but the point deserves a better soapbox, closer to the start of the book. Regardless of the accuracy of a statement such as "They can rarely be relied on for more than four figures," it's pointless to make it. It's irrelevant whether a finding is "statistically significant" at the 0.001 level instead of the 0.005 level, and to worry about whether Excel can successfully distinguish between the two findings is to miss the context.

There are many possible explanations for a research outcome other than the one you're seeking: a real and replicable treatment effect. Random chance is only one of these. It's one that gets a lot of attention because we attach the word *significance* to our tests to rule out chance, but it's not more important than other possible explanations you should be concerned about when you design your study. It's the design of your study, and how well you implement it, that allows you to rule out alternative explanations such as selection bias and disproportionate dropout rates. Those explanations—bias and dropout rates—are just two



examples of possible explanations for an apparent treatment effect: explanations that might make a treatment look like it had an effect when it actually didn't.

Even the strongest design doesn't enable you to rule out a chance outcome. But if the design of your study is sound, and you obtained what looks like a meaningful result, you'll want to control chance's role as an alternative explanation of the result. So, you certainly want to run your data through the appropriate statistical test, which *does* help you control the effect of chance.

If you get a result that doesn't clearly rule out chance—or rule it in—you're much better off to run the experiment again than to take a position based on a borderline outcome. At the very least, it's a better use of your time and resources than to worry in print about whether Excel's F tests are accurate to the fifth decimal place.

## Wagging the Dog

And ask yourself this: Once you reach the point of planning the statistical test, are you going to reject your findings if they might come about by chance five times in 1,000? Is that too loose a criterion? What about just one time in 1,000? How many angels are on that pinhead anyway?

If you're concerned that Excel won't return the correct distinction between one and five chances in 1,000 that the result of your study is due to chance, you allow what's really an irrelevancy to dictate how, and using what calibrations, you're going to conduct your statistical analysis. It's pointless to worry about whether a test is accurate to one point in a thousand or two in a thousand. Your decision rules for risking a chance finding should be based on more substantive grounds.

Chapter 9, "Testing Differences Between Means: Further Issues," goes into the matter in greater detail, but a quick summary of the issue is that you should let the risk of making the wrong decision be guided by the costs of a bad decision and the benefits of a good one—not by which criterion appears to be the more selective.

## What's in This Book

You'll find that there are two broad types of statistics. I'm not talking about that scurrilous line about lies, damned lies and statistics—both its source and its applicability are disputed. I'm talking about *descriptive* statistics and *inferential* statistics.

No matter if you've never studied statistics before this, you're already familiar with concepts such as averages and ranges. These are descriptive statistics. They describe identified groups: The average age of the members is 42 years; the range of the weights is 105 pounds; the median price of the houses is \$270,000. A variety of other sorts of descriptive statistics exists, such as standard deviations, correlations, and skewness. The first five chapters of this book take a fairly close look at descriptive statistics, and you might find that they have some aspects that you haven't considered before.

Descriptive statistics provides you with insight into the characteristics of a restricted set of beings or objects. They can be interesting and useful, and they have some properties that aren't at all well known. But you don't get a better understanding of the world from descriptive statistics. For that, it helps to have a handle on inferential statistics. That sort of analysis is based on descriptive statistics, but you are asking and perhaps answering broader questions. Questions such as this:

The average systolic blood pressure in this group of patients is 135. How large a margin of error must I report so that if I took another 99 samples, 95 of the 100 would capture the true population mean within margins calculated similarly?

Inferential statistics enables you to make inferences about a population based on samples from that population. As such, inferential statistics broadens the horizons considerably.

Therefore, I have prepared two new chapters on inferential statistics for this 2013 edition of *Statistical Analysis: Microsoft Excel*. Chapter 12, "Experimental Design and ANOVA," explores the effects of fixed versus random factors on the nature of your F tests. It also examines crossed and nested factors in factorial designs, and how a factor's status in a factorial design affects the mean square you should use in the F ratio's denominator.

I have also expanded coverage of the topic of statistical power, and this edition devotes an entire chapter to it. Chapter 13, "Statistical Power," discusses how to use Excel's worksheet functions to generate F distributions with different noncentrality parameters. (Excel's native F() functions all assume a noncentrality parameter of zero.) You can use this capability to calculate the power of an F test without resorting to 80-year-old charts.

But you have to take on some assumptions about your samples, and about the populations that your samples represent, to make the sort of generalization that inferential statistics makes available to you. From Chapter 6 through the end of this book, you'll find discussions of the issues involved, along with examples of how those issues work out in practice. And, by the way, how you work them out using Microsoft Excel.

*This page intentionally left blank*

# About Variables and Values

# 1

## Variables and Values

It must seem odd to start a book about statistical analysis using Excel with a discussion of ordinary, everyday notions such as variables and values. But variables and values, along with scales of measurement (covered in the next section), are at the heart of how you represent data in Excel. And how you choose to represent data in Excel has implications for how you run the numbers.

With your data laid out properly, you can easily and efficiently combine records into groups, pull groups of records apart to examine them more closely, and create charts that give you insight into what the raw numbers are really doing. When you put the statistics into tables and charts, you begin to understand what the numbers have to say.

When you lay out your data without considering how you will use the data later, it becomes much more difficult to do any sort of analysis. Excel is generally very flexible about how and where you put the data you're interested in, but when it comes to preparing a formal analysis, you want to follow some guidelines. In fact, some of Excel's features don't work at all if your data doesn't conform to what Excel expects. To illustrate one useful arrangement, you won't go wrong if you put different variables in different columns and different records in different rows.

A *variable* is an attribute or property that describes a person or a thing. Age is a variable that describes you. It describes all humans, all living organisms, all objects—anything that exists for some period of time. Surname is a variable, and so are Weight in Pounds and Brand of Car. Database jargon often

## IN THIS CHAPTER

Variables and Values .....	1
Scales of Measurement .....	4
Charting Numeric Variables in Excel .....	10
Understanding Frequency Distributions.....	12



refers to variables as *fields*, and some Excel tools use that terminology, but in statistics you generally use the term *variable*.

Variables have *values*. The number 20 is a value of the variable Age, the name Smith is a value of the variable Surname, 130 is a value of the variable Weight in Pounds, and Ford is a value of the variable Brand of Car. Values vary from person to person and from object to object—hence the term *variable*.

## Recording Data in Lists

When you run a statistical analysis, your purpose is generally to summarize a group of numeric values that belong to the same variable. For example, you might have obtained and recorded the weight in pounds for 20 people, as shown in Figure 1.1.

**Figure 1.1**  
This layout is ideal for analyzing data in Excel.

	A	B	C	D
1	Weight in pounds			
2	129			
3	187			
4	212			
5	215			
6	150			
7	170			
8	159			
9	225			
10	167			
11	184			
12	162			
13	116			
14	156			
15	218			
16	141			
17	147			
18	114			
19	124			
20	172			
21	169			

The way the data is arranged in Figure 1.1 is what Excel calls a *list*—a variable that occupies a column, records that each occupy a different row, and values in the cells where the records' rows intersect the variable's column. (The *record* is the individual being, object, location—whatever—that the list brings together with other, similar records. If the list in Figure 1.1 is made up of students in a classroom, each student constitutes a record.)

A list always has a *header*, usually the name of the variable, at the top of the column. In Figure 1.1, the header is the label Weight in Pounds in cell A1.

NOTE

A *list* is an informal arrangement of headers and values on a worksheet. It's not a formal structure that has a name and properties, such as a chart or a pivot table. Excel 2007 through 2013 offer a formal structure called a *table* that acts much like a list, but has some bells and whistles that a list doesn't have. This book has more to say about tables in subsequent chapters.

There are some interesting questions that you can answer with a single-column list such as the one in Figure 1.1. You could select all the values and look at the status bar at the bottom of the Excel window to see summary information such as the average, the sum, and the count of the selected values. Those are just the quickest and simplest statistical analyses you might do with this basic single-column list.

TIP

You can turn the display of indicators such as simple statistics on and off. Right-click the status bar and select or deselect the items you want to show or hide. However, you won't see a statistic unless the current selection contains at least two values. The status bar of Figure 1.1 shows the average, count, and sum of the selected values. (The worksheet tabs have been suppressed to unclutter the figure.)

Again, this book has much more to say about the richer analyses of a single variable that are available in Excel. But first, suppose that you add a second variable, Sex, to the list in Figure 1.1.

You might get something like the two-column list in Figure 1.2. All the values for a particular record—here, a particular person—are found in the same row. So, in Figure 1.2, the person whose weight is 129 pounds is female (row 2), the person who weighs 187 pounds is male (row 3), and so on.

Using the list structure, you can easily do the simple analyses that appear in Figure 1.3, where you see a *pivot table* and a *pivot chart*. These are powerful tools and well suited to statistical analysis, but they're also very easy to use.

All that's needed for the pivot chart and pivot table in Figure 1.3 is the simple, informal, unglamorous list in Figure 1.2. But that list, and the fact that it keeps related values of weight and sex together in records, makes it possible to do the analyses shown in Figure 1.3. With the list in Figure 1.2, you're just a few clicks away from analyzing and charting average weight by sex.

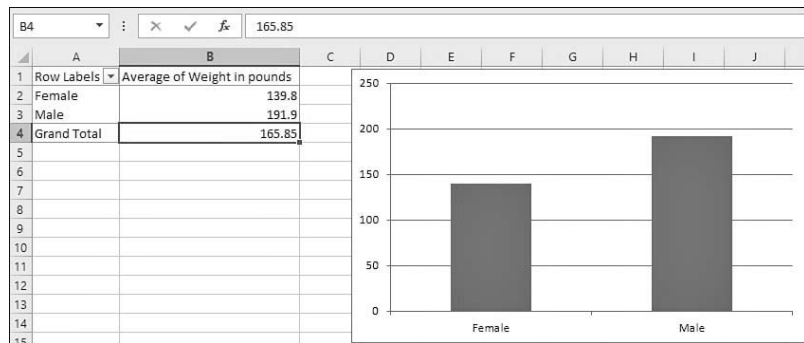
NOTE

In Excel 2013, it's eleven clicks if you do it all yourself; you save a click if you start with the Recommended Pivot Tables button on the Ribbon's Insert tab. And if you select the full list or even just a subset of the records in the list (say, cells A4:B4) the Quick Analysis tool gets you a weight-by-sex pivot table in only three clicks.

**Figure 1.2**  
The list structure helps you keep related values together.

	A	B	C	D	E
1	Weight in pounds	Sex			
2	129	Female			
3	187	Male			
4	212	Male			
5	215	Male			
6	150	Female			
7	170	Male			
8	159	Female			
9	225	Male			
10	167	Male			
11	184	Male			
12	162	Female			
13	116	Female			
14	156	Female			
15	218	Male			
16	141	Female			
17	147	Female			
18	114	Female			
19	124	Female			
20	172	Male			
21	169	Male			

**Figure 1.3**  
The pivot table and pivot chart summarize the individual records shown in Figure 1.2.



Note that you cannot create a standard Excel column chart directly from the data as displayed in Figure 1.2. You first need to get the average weight of men and women, then associate those averages with the appropriate labels, and finally create the chart. A pivot chart is much quicker, more convenient, and more powerful.

## Scales of Measurement

There's a difference in how weight and sex are measured and reported in Figure 1.2 that is fundamental to all statistical analysis—and to how you bring Excel's tools to bear on the numbers. The difference concerns scales of measurement.

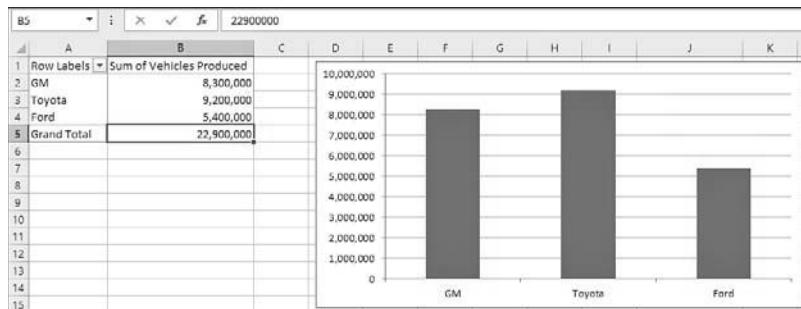
## Category Scales

In Figures 1.2 and 1.3, the variable Sex is measured using a *category* scale, often called a *nominal* scale. Different values in a category variable merely represent different groups, and there's nothing intrinsic to the categories that does anything but identify them. If you throw out the psychological and cultural connotations that we pile onto labels, there's nothing about Male and Female that would lead you to put one on the left and the other on the right in Figure 1.3's pivot chart, the way you'd put June to the left of July.

Another example: Suppose that you want to chart the annual sales of Ford, General Motors, and Toyota cars. There is no order that's necessarily implied by the names themselves: They're just categories. This is reflected in the way that Excel might chart that data (see Figure 1.4).

**Figure 1.4**

Excel's Column charts always show categories on the horizontal axis and numeric values on the vertical axis.



Notice these two aspects of the car manufacturer categories in Figure 1.4:

- Adjacent categories are equidistant from one another. No additional information is supplied by the distance of GM from Toyota, or Toyota from Ford.
- The chart conveys no information through the order in which the manufacturers appear on the horizontal axis. There's no implication that GM has less "car-ness" than Toyota, or Toyota less than Ford. You could arrange them in alphabetical order if you wanted, or in order of number of vehicles produced, but there's nothing intrinsic to the scale of manufacturers' names that suggests any rank order.

NOTE

This is one of many quirks of terminology in Excel. The name Ford is of course a value, but Excel prefers to call it a category and to reserve the term value for numeric values only.

In contrast, the vertical axis in the chart shown in Figure 1.4 is what Excel terms a *value* axis. It represents numeric values.

Notice in Figure 1.4 that a position on the vertical, value axis conveys real quantitative information: the more vehicles produced, the taller the column. The vertical and the

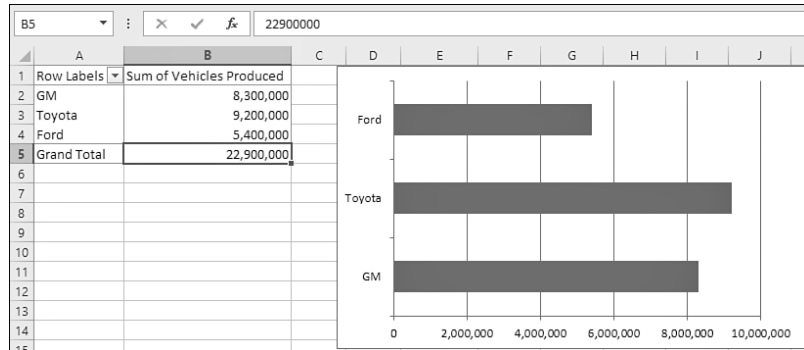


horizontal axes in Excel's Column charts differ in several ways, but the most crucial is that the vertical axis represents numeric quantities, while the horizontal axis simply indicates the existence of categories.

In general, Excel charts put the names of groups, categories, products, or any other designation on a category axis and the numeric value of each category on the value axis. But the category axis isn't always the horizontal axis (see Figure 1.5).

**Figure 1.5**

In contrast to Column charts, Excel's Bar charts always show categories on the vertical axis and numeric values on the horizontal axis.



The Bar chart provides precisely the same information as does the Column chart. It just rotates this information by 90 degrees, putting the categories on the vertical axis and the numeric values on the horizontal axis.

I'm not belaboring the issue of measurement scales just to make a point about Excel charts. When you do statistical analysis, you choose a technique based in large part on the sort of question you're asking. In turn, the way you ask your question depends in part on the scale of measurement you use for the variable you're interested in.

For example, if you're trying to investigate life expectancy in men and women, it's pretty basic to ask questions such as, "What is the average life span of males? of females?" You're examining two variables: sex and age. One of them is a category variable, and the other is a numeric variable. (As you'll see in later chapters, if you are generalizing from a sample of men and women to a population, the fact that you're working with a category variable and a numeric variable might steer you toward what's called a *t-test*.)

In Figures 1.3 through 1.5, you see that numeric summaries—average and sum—are compared across different groups. That sort of comparison forms one of the major types of statistical analysis. If you design your samples properly, you can then ask and answer questions such as these:

- Are men and women paid differently for comparable work? Compare the average salaries of men and women who hold similar jobs.
- Is a new medication more effective than a placebo at treating a particular disease? Compare, say, average blood pressure for those taking an alpha blocker with that of those taking a sugar pill.

- Do Republicans and Democrats have different attitudes toward a given political issue? Ask a random sample of people their party affiliation, and then ask them to rate a given issue or candidate on a numeric scale.

Notice that each of these questions can be answered by comparing a *numeric* variable across different *categories* of interest.

## Numeric Scales

Although there is only one type of category scale, there are three types of numeric scales: ordinal, interval, and ratio. You can use the value axis of any Excel chart to represent any type of numeric scale, and you often find yourself analyzing one numeric variable, regardless of type, in terms of another variable. Briefly, the numeric scale types are as follows:

- Ordinal scales are often rankings, and tell you who finished first, second, third, and so on. These rankings tell you who came out ahead, but not how far ahead, and often you don't care about that. Suppose that in a qualifying race Jane ran 100 meters in 10.54 seconds, Mary in 10.83 seconds, and Ellen in 10.84 seconds. Because it's a preliminary heat, you might care only about their order of finish, and not about how fast each woman ran. Therefore, you might convert the time measurements to order of finish (1, 2 and 3), and then discard the timings themselves. Ordinal scales are sometimes used in a branch of statistics called *nonparametrics* but are used infrequently in the parametric analyses discussed in this book.
- Interval scales indicate differences in measures such as temperature and elapsed time. If the high temperature Fahrenheit on July 1 is 100 degrees, 101 degrees on July 2, and 102 degrees on July 3, you know that each day is one degree hotter than the previous day. So, an interval scale conveys more information than an ordinal scale. You know, from the order of finish on an ordinal scale, that in the qualifying race Jane ran faster than Mary and Mary ran faster than Ellen, but the rankings by themselves don't tell you how much faster. It takes elapsed time, an interval scale, to tell you that.
- Ratio scales are similar to interval scales, but they have a true zero point, one at which there is a complete absence of some quantity. The Celsius temperature scale has a zero point, but it doesn't indicate a complete absence of heat, just that water freezes there. Therefore, 10 degrees Celsius is not twice as warm as 5 degrees Celsius, so Celsius is not a ratio scale. Degrees kelvin does have a true zero point, one at which there is no molecular motion and therefore no heat. Kelvin is a ratio scale, and 100 degrees kelvin is twice as warm as 50 degrees kelvin. Other familiar ratio scales are height and weight.

It's worth noting that converting between interval (or ratio) and ordinal measurement is a one-way process. If you know how many seconds it takes three people to run 100 meters, you have measures on a ratio scale that you can convert to an ordinal scale—gold, silver, and bronze medals. You can't go the other way, though: If you know who won each medal, you're still in the dark as to whether the bronze medal was won with a time of 10 seconds or 10 minutes.

## Telling an Interval Value from a Text Value

Excel has an astonishingly broad scope, and not only in statistical analysis. As much skill as has been built in to it, though, it can't quite read your mind. It doesn't know, for example, whether the 1, 2, and 3 you just entered into a worksheet's cells represent the number of teaspoons of olive oil you use in three different recipes or 1st, 2nd, and 3rd place in a political primary. In the first case, you meant to indicate liquid measures on an interval scale. In the second case, you meant to enter the first three places in an ordinal scale. But they both look alike to Excel.

**NOTE**

This is a case in which you must rely on your own knowledge of numeric scales because Excel can't tell whether you intend a number as a value on an ordinal or an interval scale. Ordinal and interval scales have different characteristics—for one thing, ordinal scales do not follow a normal distribution, a “bell curve.” An ordinal variable has one instance of the value 1, one instance of 2, one instance of 3, and so on, so its distribution is flat instead of curved. Excel can't tell the difference between an ordinal and an interval variable, though, so you have to take control if you're to avoid using a statistical technique that's wrong for a given scale of measurement.

Text is a different matter. You might use the letters A, B and C to name three different groups, and in that case you're using text values on a nominal, category scale. You can also use numbers: 1, 2 and 3 to represent the same three groups. But if you use a number as a nominal value, it's a good idea to store it in the worksheet as a text value. For example, one way to store the number 2 as a text value in a worksheet cell is to precede it with an apostrophe: '2. (You'll see the apostrophe in the formula box but not in the cell.)

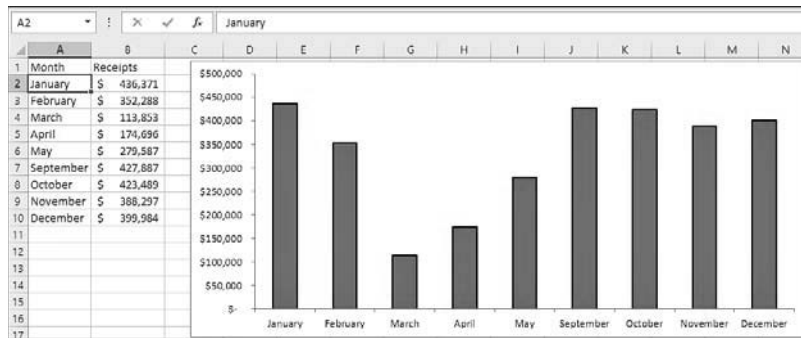
On a chart, Excel has some complicated decision rules that it uses to determine whether a number is only a number. (Excel 2013 has some additional tools to help you participate in the decision-making process, as you'll see later in this chapter). Some of those rules concern the type of chart you request. For example, if you request a Line chart, Excel treats numbers on the horizontal axis as though they were nominal, text values. But if instead you request an XY chart using the same data, Excel treats the numbers on the horizontal axis as values on an interval scale. You'll see more about this in the next section.

So, as disquieting as it may sound, a number in Excel may be treated as a number in one context and not in another. Excel's rules are pretty reasonable, though, and if you give them a little thought when you see their results, you'll find that they make good sense.

If Excel's rules don't do the job for you in a particular instance, you can provide an assist. Figure 1.6 shows an example.

**Figure 1.6**

You don't have data for all the months in the year.



Suppose that you run a business that operates only when public schools are in session, and you collect revenues during all months except June, July and August. Figure 1.6 shows that Excel interprets dates as categories—but only if they are entered as text, as they are in the figure. Notice these two aspects of the worksheet and chart in Figure 1.6:

- The dates are entered in the worksheet cells A2:A10 as text values. One way to tell is to look in the formula box, just to the right of the  $f_x$  symbol, where you see the text value January.
- Because they are text values, Excel has no way of knowing that you mean them to represent dates, and so it treats them as simple categories—just like it does for GM, Ford, and Toyota. Excel charts the dates-as-text accordingly, with equal distances between them: May is as far from April as it is from September.

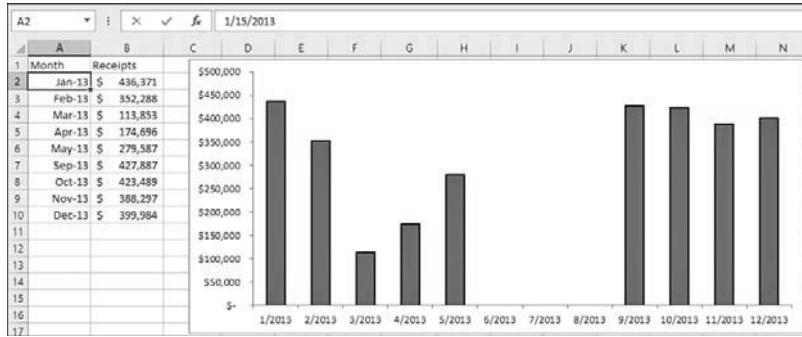
Compare Figure 1.6 with Figure 1.7, where the dates are real numeric values, not simply text:

- You can see in the formula box that it's an actual date, not just the name of a month, in cell A2, and the same is true for the values in cells A3:A10.
- The Excel chart automatically responds to the type of values you have supplied in the worksheet. The program recognizes that the numbers entered represent monthly intervals and, although there is no data for June through August, the chart leaves places for where the data would appear if it were available. Because the horizontal axis now represents a numeric scale, not simple categories, it faithfully reflects the fact that in the calendar, May is four times as far from September as it is from April.

NOTE

A date value in Excel is just a numeric value: the number of days that have elapsed between the date in question and January 1, 1900. Excel assumes that when you enter a value such as 1/1/14, three numbers separated by two slashes, you intend it as a date. Excel treats it as a number but applies a date format such as mm/yy or mm/dd/yyyy to that number. You can demonstrate this for yourself by entering a legitimate date (not something such as 34/56/78) in a worksheet cell and then setting the cell's number format to Number with zero decimal places.

**Figure 1.7**  
The horizontal axis accounts for the missing months.



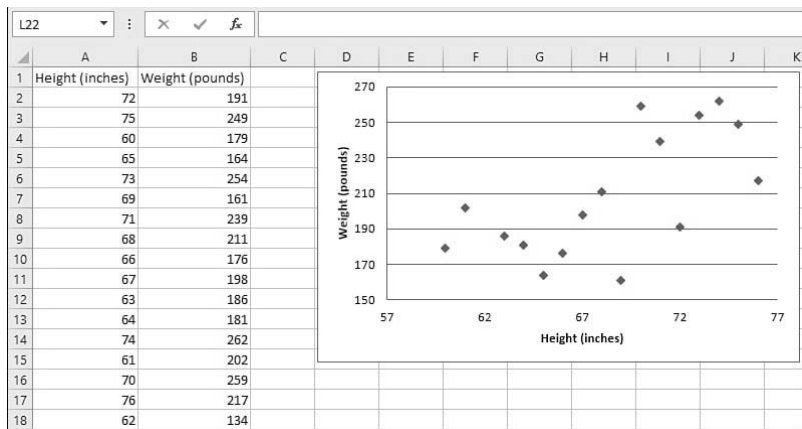
## Charting Numeric Variables in Excel

Several chart types in Excel lend themselves beautifully to the visual representation of numeric variables. This book relies heavily on charts of that type because most of us find statistical concepts that are difficult to grasp in the abstract are much clearer when they're illustrated in charts.

### Charting Two Variables

Earlier this chapter briefly discussed two chart types that use a category variable on one axis and a numeric variable on the other: Column charts and Bar charts. There are other, similar types of charts, such as Line charts, that are useful for analyzing a numeric variable in terms of different categories—especially time categories such as months, quarters, and years. However, one particular type of Excel chart, called an *XY (Scatter)* chart, shows the relationship between exactly two numeric variables. Figure 1.8 provides an example.

**Figure 1.8**  
In an XY (Scatter) chart, both the horizontal and vertical axes are value axes.



NOTE

Since the 1990s at least, Excel has called this sort of chart an XY (Scatter) chart. In its 2007 version, Excel started referring to it as an XY chart in some places, as a Scatter chart in others, and as an XY (Scatter) chart in still others. For the most part, this book opts for the brevity of XY chart, and when you see that term you can be confident it's the same as an XY (Scatter) chart.

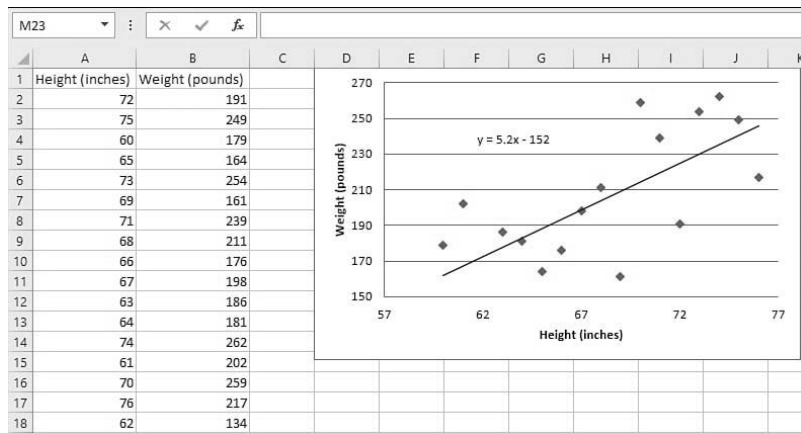
The markers in an XY chart show where a particular person or object falls on each of two numeric variables. The overall pattern of the markers can tell you quite a bit about the relationship between the variables, as expressed in each record's measurement. Chapter 4, "How Variables Move Jointly: Correlation," goes into considerable detail about this sort of relationship.

In Figure 1.8, for example, you can see the relationship between a person's height and weight: Generally, the greater the height, the greater the weight. The relationship between the two variables differs fundamentally from those discussed earlier in this chapter, where the emphasis is placed on the sum or average of a numeric variable, such as number of vehicles, according to the category of a nominal variable, such as make of car.

However, when you are interested in the way that two numeric variables are related, you are asking a different sort of question, and you use a different sort of statistical analysis. How are height and weight related, and how strong is the relationship? Does the amount of time spent on a cell phone correspond in some way to the likelihood of contracting cancer? Do people who spend more years in school eventually make more money? (And if so, does that relationship hold all the way from elementary school to post-graduate degrees?) This is another major class of empirical research and statistical analysis: the investigation of how different variables change together—or, in statistical jargon, how they *covary*.

Excel's XY charts can tell you a considerable amount about how two numeric variables are related. Figure 1.9 adds a trendline to the XY chart in Figure 1.8.

**Figure 1.9**  
A trendline graphs a numeric relationship, which is almost never an accurate way to depict reality.



The diagonal line you see in Figure 1.9 is a *trendline*. It is an idealized representation of the relationship between men's height and weight, at least as determined from the sample of 17 men whose measures are charted in the figure. The trendline is based on this formula:

$$\text{Weight} = 5.2 * \text{Height} - 152$$

Excel calculates the formula based on what's called the *least squares* criterion. You'll see much more about this in Chapter 4.

Suppose that you picked several—say, 20—different values for height in inches, plugged them into that formula, and then used the formula to calculate the resulting weight. If you now created an Excel XY chart that shows those values of height and weight, you would get a chart that shows a straight line similar to the trendline you see in Figure 1.9.

That's because arithmetic is nice and clean and doesn't involve errors. The formula applies arithmetic which results in a set of predicted weights that, plotted against height on a chart, describe a straight line. Reality, though, is seldom free from errors. Some people weigh more than a formula thinks they should, given their height. Other people weigh less. (Statistical analysis terms these discrepancies *errors* or *deviations*.) The result is that if you chart the measures you get from actual people instead of from a mechanical formula, you're going to get a set of data that looks like the somewhat scattered markers in Figures 1.8 and 1.9.

Reality is messy, and the statistician's approach to cleaning it up is to seek to identify regular patterns lurking behind the real-world measures. If those real-world measures don't precisely fit the pattern that has been identified, there are several explanations, including these (and they're not mutually exclusive):

- People and things just don't always conform to ideal mathematical patterns. Deal with it.
- There may be some problem with the way the measures were taken. Get better yardsticks.
- Some other, unexamined variable may cause the deviations from the underlying pattern. Come up with some more theory, and then carry out more research.

## Understanding Frequency Distributions

In addition to charts that show two variables—such as numbers broken down by categories in a Column chart, or the relationship between two numeric variables in an XY chart—there is another sort of Excel chart that deals with one variable only. It's the visual representation of a *frequency distribution*, a concept that's absolutely fundamental to intermediate and advanced statistical methods.

A frequency distribution is intended to show how many instances there are of each value of a variable. For example:

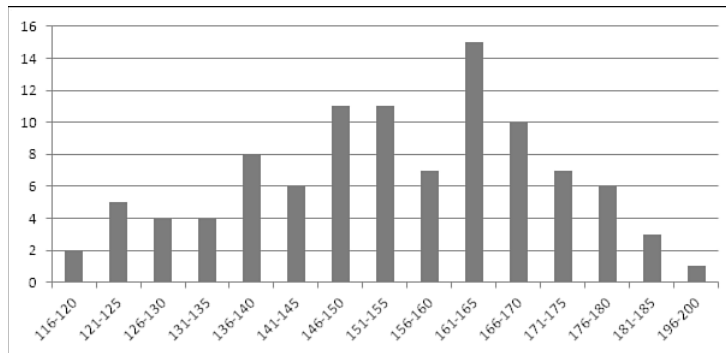
- The number of people who weigh 100 pounds, 101 pounds, 102 pounds, and so on
- The number of cars that get 18 miles per gallon (mpg), 19 mpg, 20 mpg, and so on
- The number of houses that cost between \$200,001 and \$205,000, between \$205,001 and \$210,000, and so on

Because we usually round measurements to some convenient level of precision, a frequency distribution tends to group individual measurements into classes. Using the examples just given, two people who weigh 100.2 and 100.4 pounds might each be classed as 100 pounds; two cars that get 18.8 and 19.2 mpg might be grouped together at 19 mpg; and any number of houses that cost between \$220,001 and \$225,000 would be treated as in the same price level.

As it's usually shown, the chart of a frequency distribution puts the variable's values on its horizontal axis and the count of instances on the vertical axis. Figure 1.10 shows a typical frequency distribution.

**Figure 1.10**

Typically, most records cluster toward the center of a frequency distribution.

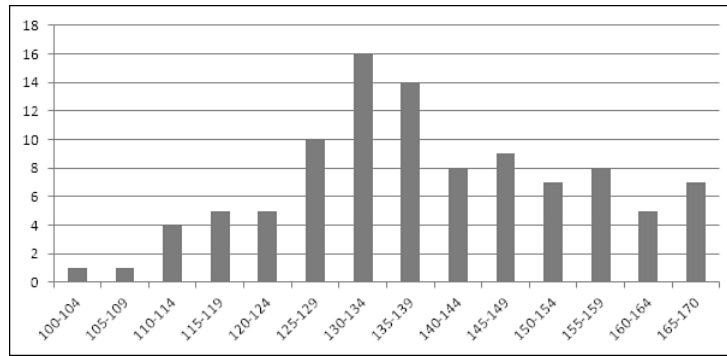


You can tell quite a bit about a variable by looking at a chart of its frequency distribution. For example, Figure 1.10 shows the weights of a sample of 100 people. Most of them are between 140 and 180 pounds. In this sample, there are about as many people who weigh a lot (say, over 175 pounds) as there are whose weight is relatively low (say, up to 130). The range of weights—that is, the difference between the lightest and the heaviest weights—is about 85 pounds, from 116 to 200.

There are lots of ways that a different sample of people might provide different weights than those shown in Figure 1.10. For example, Figure 1.11 shows a sample of 100 vegans. (Notice that the distribution of their weights is shifted down the scale somewhat from the sample of the general population shown in Figure 1.10.)



**Figure 1.11**  
Compared to Figure 1.10, the location of the frequency distribution has shifted to the left.



The frequency distributions in Figures 1.10 and 1.11 are relatively symmetric. Their general shapes are not far from the idealized normal “bell” curve, which depicts the distribution of many variables that describe living beings. This book has much more to say in later chapters about the normal curve, partly because it describes so many variables of interest, but also because Excel has so many ways of dealing with the normal curve.

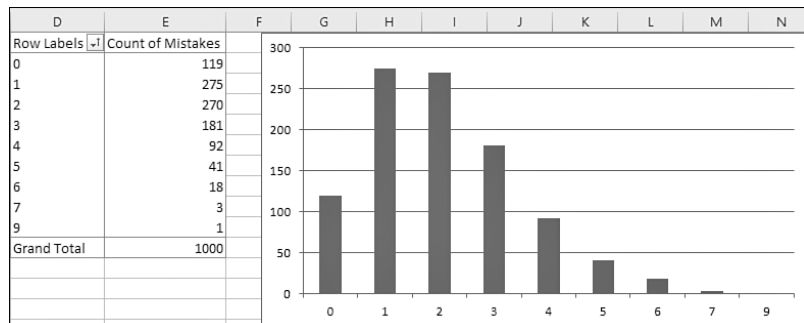
Still, many variables follow a different sort of frequency distribution. Some are skewed right (see Figure 1.12) and others left (see Figure 1.13).

Figure 1.12 shows counts of the number of mistakes on individual federal tax forms. It’s normal to make a few mistakes (say, one or two), and it’s abnormal to make several (say, five or more). This distribution is positively skewed.

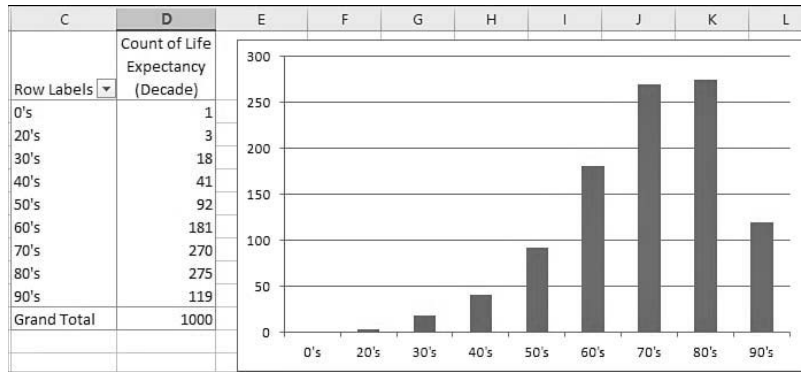
Another variable, home prices, tends to be positively skewed, because although there’s a real lower limit (a house cannot cost less than \$0) there is no theoretical upper limit to the price of a house. House prices therefore tend to bunch up between \$100,000 and \$300,000, with fewer between \$300,000 and \$400,000, and fewer still as you go up the scale.

A quality control engineer might sample 100 ceramic tiles from a production run of 10,000 and count the number of defects on each tile. Most would have zero, one, or two defects, several would have three or four, and a very few would have five or six. This is another positively skewed distribution—quite a common situation in manufacturing process control.

**Figure 1.12**  
A frequency distribution that stretches out to the right is called positively skewed.



**Figure 1.13**  
Negatively skewed distributions are not as common as positively skewed distributions.



Because true lower limits are more common than true upper limits, you tend to encounter more positively skewed frequency distributions than negatively skewed. But negative skews certainly occur. Figure 1.13 might represent personal longevity: Relatively few people die in their twenties, thirties and forties, compared to the numbers who die in their fifties through their eighties.

### Using Frequency Distributions

It's helpful to use frequency distributions in statistical analysis for two broad reasons. One concerns visualizing how a variable is distributed across people or objects. The other concerns how to make inferences about a population of people or objects on the basis of a sample.

Those two reasons help define the two general branches of statistics: *descriptive* statistics and *inferential* statistics. Along with descriptive statistics such as averages, ranges of values, and percentages or counts, the chart of a frequency distribution puts you in a stronger position to understand a set of people or things because it helps you visualize how a variable behaves across its range of possible values.

In the area of inferential statistics, frequency distributions based on samples help you determine the type of analysis you should use to make inferences about the population. As you'll see in later chapters, frequency distributions also help you visualize the results of certain choices that you must make—choices such as the probability of coming to the wrong conclusion.

### Visualizing the Distribution: Descriptive Statistics

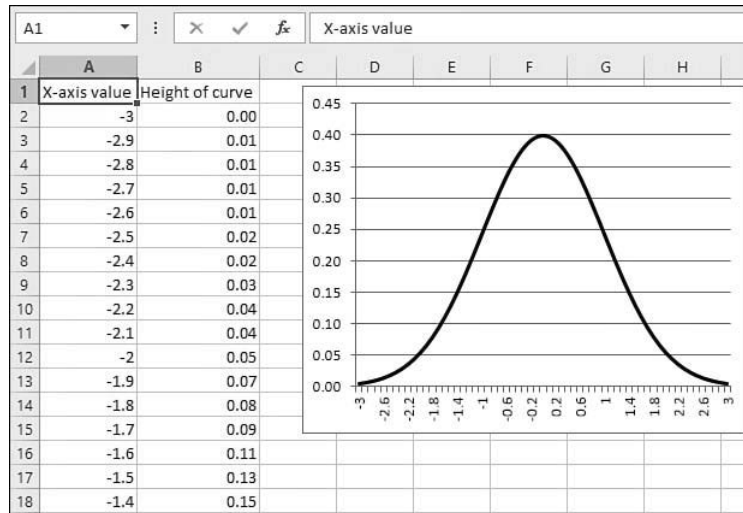
It's usually much easier to understand a variable—how it behaves in different groups, how it may change over time, and even just what it looks like—when you see it in a chart. For example, here's the formula that defines the normal distribution:

$$u = (1 / ((2\pi)^{-5}) \sigma) e ^ { - (X - \mu)^2 / 2 \sigma^2}$$

And Figure 1.14 shows the normal distribution in chart form.

1

**Figure 1.14**  
The familiar normal curve is just a frequency distribution.



The formula itself is indispensable, but it doesn't convey understanding. In contrast, the chart informs you that the frequency distribution of the normal curve is symmetric and that most of the records cluster around the center of the horizontal axis.

NOTE

The formula was developed by a seventeenth-century French mathematician named Abraham De Moivre. Excel simplifies it to this:

`=NORMDIST(1,0,1,FALSE)`

In Excel 2010 and 2013, it's this:

`=NORM.S.DIST(1,FALSE)`

Those are *major* simplifications.

Again, personal longevity tends to bulge in the higher levels of its range (and therefore skews left as in Figure 1.13). Home prices tend to bulge in the lower levels of their range (and therefore skew right). The height of human beings creates a bulge in the center of the range, and is therefore symmetric and *not* skewed.

Some statistical analyses assume that the data comes from a normal distribution, and in some statistical analyses that assumption is an important one. This book does not explore the topic in great detail because it comes up infrequently. Be aware, though, that if you want to analyze a skewed distribution there are ways to normalize it and therefore comply with

the requirements of the analysis. Very generally, you can use Excel's `SQRT()` and `LOG()` functions to help normalize a negatively skewed distribution, and an exponentiation operator (for example, `=A2^2` to square the value in A2) to help normalize a positively skewed distribution.

NOTE

Finding just the right transformation for a particular data set can be a matter of trial and error, however, and the Excel Solver add-in can help in conjunction with Excel's `SKEW()` function. See Chapter 2, "How Values Cluster Together," for information on Solver, and Chapter 7, "Using Excel with the Normal Distribution," for information on `SKEW()`. The basic idea is to use `SKEW()` to calculate the skewness of your transformed data and to have Solver find the exponent that brings the result of `SKEW()` closest to zero.

## Visualizing the Population: Inferential Statistics

The other general rationale for examining frequency distributions has to do with making an inference about a population, using the information you get from a sample as a basis. This is the field of inferential statistics. In later chapters of this book, you will see how to use Excel's tools—in particular, its functions and its charts—to infer a population's characteristics from a sample's frequency distribution.

A familiar example is the political survey. When a pollster announces that 53% of those who were asked preferred Smith, he is reporting a descriptive statistic. Fifty-three percent of the sample preferred Smith, and no inference is needed.

But when another pollster reports that the margin of error around that 53% statistic is plus or minus 3%, she is reporting an inferential statistic. She is extrapolating from the sample to the larger population and inferring, with some specified degree of confidence, that between 50% and 56% of all voters prefer Smith.

The size of the reported margin of error, six percentage points, depends heavily on how confident the pollster wants to be. In general, the greater degree of confidence you want in your extrapolation, the greater the margin of error that you allow. If you're on an archery range and you want to be virtually certain of hitting your target, you make the target as large as necessary.

Similarly, if the pollster wants to be 99.9% confident of her projection into the population, the margin might be so great as to be useless—say, plus or minus 20%. And although it's not headline material to report that somewhere between 33% and 73% of the voters prefer Smith, the pollster can be confident that the projection is accurate.

But the size of the margin of error also depends on certain aspects of the frequency distribution in the sample of the variable. In this particular (and relatively straightforward) case, the accuracy of the projection from the sample to the population depends in part on the level of confidence desired (as just briefly discussed), in part on the size of the sample, and in part on the percent favoring Smith in the sample. The latter two issues, sample size and

percent in favor, are both aspects of the frequency distribution you determine by examining the sample's responses.

Of course, it's not just political polling that depends on sample frequency distributions to make inferences about populations. Here are some other typical questions posed by empirical researchers:

- What percent of the nation's existing houses were resold last quarter?
- What is the incidence of cardiovascular disease today among diabetics who took the drug Avandia before questions about its side effects arose in 2007? Is that incidence reliably different from the incidence of cardiovascular disease among those who never took the drug?
- A sample of 100 cars from a particular manufacturer, made during 2013, had average highway gas mileage of 26.5 mpg. How likely is it that the average highway mpg, for all that manufacturer's cars made during that year, is greater than 26.0 mpg?
- Your company manufactures custom glassware. Your contract with a customer calls for no more than 2% defective items in a production lot. You sample 100 units from your latest production run and find 5 that are defective. What is the likelihood that the entire production run of 1,000 units has a maximum of 20 that are defective?

In each of these four cases, the specific statistical procedures to use—and therefore the specific Excel tools—would be different. But the basic approach would be the same: Using the characteristics of a frequency distribution from a sample, compare the sample to a population whose frequency distribution is either known or founded in good theoretical work. Use the numeric functions in Excel to estimate how likely it is that your sample accurately represents the population you're interested in.

## Building a Frequency Distribution from a Sample

Conceptually, it's easy to build a frequency distribution. Take a sample of people or things and measure each member of the sample on the variable that interests you. Your next step depends on how much sophistication you want to bring to the project.

### Tallying a Sample

One straightforward approach continues by dividing the relevant range of the variable into manageable groups. For example, suppose that you obtained the weight in pounds of each of 100 people. You might decide that it's reasonable and feasible to assign each person to a weight class that is ten pounds wide: 75 to 84, 85 to 94, 95 to 104, and so on. Then, on a sheet of graph paper, make a tally in the appropriate column for each person, as suggested in Figure 1.15.

The approach shown in Figure 1.15 uses a *grouped* frequency distribution, and tallying by hand into groups was the only practical option as recently as the 1980s, before personal computers came into truly widespread use. But using an Excel function named `FREQUENCY()`, you can get the benefits of grouping individual observations without the tedium of manually assigning individual records to groups.

**Figure 1.15**  
This approach helps clarify the process, but there are quicker and easier ways.

	A	B	C	D	E	F	G
1							
2				✓			
3				✓			
4				✓			
5			✓	✓			
6			✓	✓	✓		
7			✓	✓	✓		
8			✓	✓	✓		
9			✓	✓	✓		
10			✓	✓	✓		
11			✓	✓	✓		
12			✓	✓	✓		
13			✓	✓	✓		
14			✓	✓	✓		
15			✓	✓	✓		
16		✓	✓	✓	✓		
17		✓	✓	✓	✓		
18		✓	✓	✓	✓		
19		✓	✓	✓	✓		
20		✓	✓	✓	✓	✓	
21		✓	✓	✓	✓	✓	
22		✓	✓	✓	✓	✓	
23	✓	✓	✓	✓	✓	✓	
24	✓	✓	✓	✓	✓	✓	✓
25	✓	✓	✓	✓	✓	✓	✓
26	✓	✓	✓	✓	✓	✓	✓
27	✓	✓	✓	✓	✓	✓	✓
28	75 to 84	85 to 94	95 to 104	105 to 114	115 to 124	125 to 134	135 to 144

### Grouping with FREQUENCY()

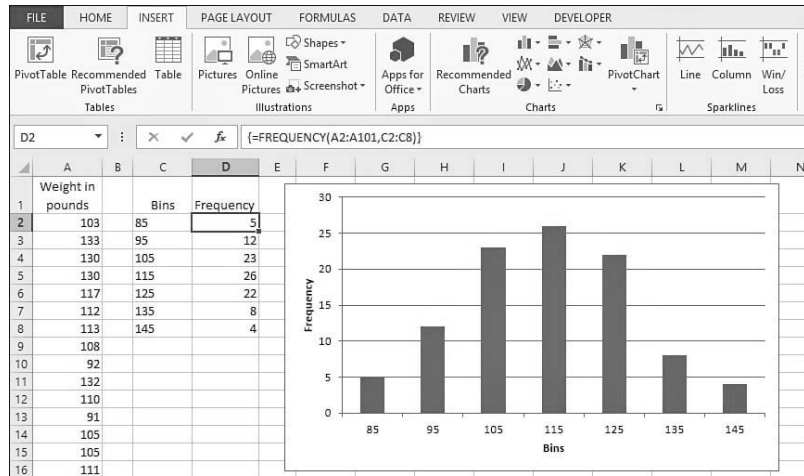
If you assemble a frequency distribution as just described, you have to count up all the records that belong to each of the groups that you define. Excel has a function, `FREQUENCY()`, that will do the heavy lifting for you. All you have to do is decide on the boundaries for the groups and then point the `FREQUENCY()` function at those boundaries and at the raw data.

Figure 1.16 shows one way to lay out the data.

In Figure 1.16, the weight of each person in your sample is recorded in column A. The numbers in cells C2:C8 define the upper boundaries of what this section has called *groups*, and what Excel calls *bins*. Up to 85 pounds defines one bin; from 86 to 95 defines another; from 96 to 105 defines another, and so on.

**NOTE** There's no special need to use the column headers shown in Figure 1.16, cells A1, C1, and D1. In fact, if you're creating a standard Excel chart as described here, there's no great need to supply column headers at all. If you don't include the headers, Excel names the data Series1 and Series2. If you use the pivot chart instead of a standard chart, though, you will need to supply a column header for the data shown in column A in Figure 1.16.

**Figure 1.16**  
The groups are defined by the numbers in cells C2:C8.



The count of records within each bin appears in D2:D8. You don't count them yourself—you call on Excel to do that for you, and you do that by means of a special kind of Excel formula, called an *array formula*. You'll read more about array formulas in Chapter 2, as well as in later chapters, but for now here are the steps needed to get the bin counts shown in Figure 1.16:

1. Select the range of cells that the results will occupy. In this case, that's the range of cells D2:D8.
2. Type, but don't yet enter, the following formula:  
**=FREQUENCY(A2:A101,C2:C8)**  
which tells Excel to count the number of records in A2:A101 that are in each bin defined by the numeric boundaries in C2:C8.
3. After you have typed the formula, hold down the Ctrl and Shift keys simultaneously and press Enter. Then release all three keys. This keyboard sequence notifies Excel that you want it to interpret the formula as an array formula.

NOTE

When Excel interprets a formula as an array formula, it places curly brackets around the formula in the formula box.

TIP

You can use the same range for the Data argument and the Bins argument in the FREQUENCY() function: for example, =FREQUENCY(A1:A101,A1:A101). Don't forget to enter it as an array formula. This is a convenient way to get Excel to treat every recorded value as its own bin, and you get the count for every unique value in the range A1:A101.

The results appear very much like those in cells D2:D8 of Figure 1.16, of course depending on the actual values in A2:A101 and the bins defined in C2:C8. You now have the frequency distribution but you still should create the chart.

Compared to earlier versions, Excel 2013 makes it quicker and easier to create certain basic charts such as the one shown in Figure 1.16. Assuming the data layout used in that figure, here are the steps you might use in Excel 2013 to create the chart:

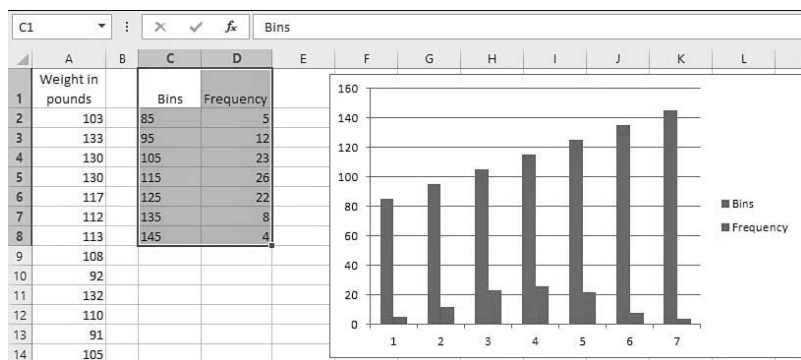
1. Select the data you want to chart—that is, the range C1:D8. (If the relevant data range is surrounded by empty cells or worksheet boundaries, all you need to select is a single cell in the range you want to chart.)
2. Click the Insert tab, and then click the Recommended Charts button in the Charts group.
3. Click the Clustered Column chart example in the Insert Chart window, and then click OK.

You can get other variations on chart types in Excel 2013 by clicking, for example, the Insert Column Chart button (in the Charts group on the Insert tab). Click More Chart Types at the bottom of the drop-down to see various ways of structuring Bar charts, Line charts, and so on given the layout of your underlying data.

Things weren't as simple in earlier versions of Excel. For example, here are the steps in Excel 2010, again assuming the data is located as in Figure 1.16:

1. Select the data you want to chart—that is, the range C1:D8.
2. Click the Insert tab, and then click the Insert Column Chart button in the Charts group.
3. Choose the Clustered Column chart type from the 2-D charts. A new chart appears, as shown in Figure 1.17. Because columns C and D on the worksheet both contain numeric values, Excel initially thinks that there are two data series to chart: one named Bins and one named Frequency.

**Figure 1.17**  
Values from both columns are charted as data series at first because they're all numeric.

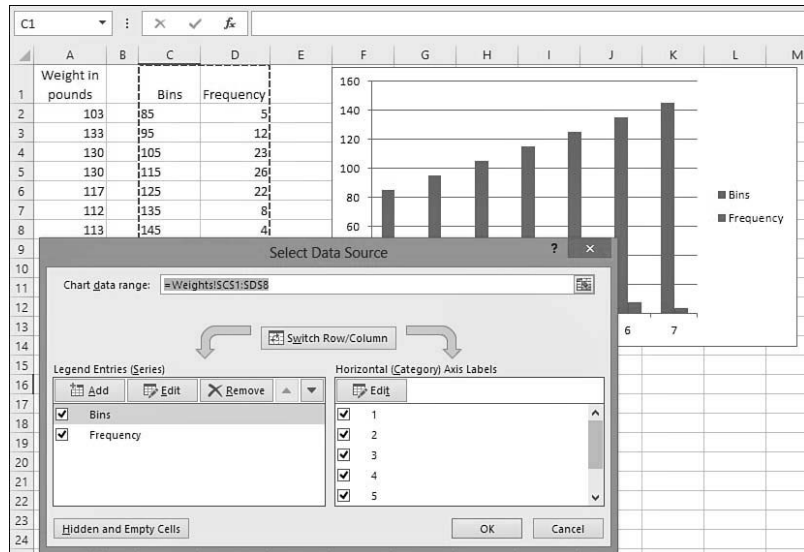




- Fix the chart by clicking Select Data in the Design tab that appears when a chart is active. The dialog box shown in Figure 1.18 appears.

**Figure 1.18**

You can also use the Select Data dialog box to add another data series to the chart.



- Click the Edit button under Horizontal (Category) Axis Labels. A new Axis Labels dialog box appears; drag through cells C2:C8 to establish that range as the basis for the horizontal axis. Click OK.
- Click the Bins label in the left list box shown in Figure 1.18. Click the Remove button to delete it as a charted series. Click OK to return to the chart.
- Remove the chart title and series legend, if you want, by clicking each and pressing Delete.

At this point, you will have a normal Excel chart that looks much like the one shown in Figure 1.16.

## USING NUMERIC VALUES AS CATEGORIES

The differences between how Excel 2010 and Excel 2013 handle charts present a good illustration of the problems created by the use of numeric values as categories. The “Charting Two Variables” section earlier in this chapter alludes to the ambiguity involved when you want Excel to treat numeric values as categories.

In the example shown in Figure 1.16, you present two numeric variables—Bins and Frequency—to Excel’s charting facility. Because both variables are numeric (and their values are stored as numbers rather than text), there are various ways that Excel can treat them in charts:

- Treat each *column*—the Bins variable and the Frequency variable—as data series to be charted. This is the approach you might take if you wanted to chart both Income and Expenses over time: you would have Excel treat each variable as a data series, and the different rows in the underlying data range would represent different time periods. You get this chart if you choose Clustered Chart in the Insert Column Chart drop-down.
- Treat each *row* in the underlying data range as a data series. Then, the columns are treated as different categories on the column chart's horizontal axis. You get this result if you click More Column Charts at the bottom of the Insert Column Chart drop-down—it's the third example chart in the Insert Chart window.
- Treat one of the variables—Bins or Frequency—as a category variable for use on the horizontal axis. This is the column chart you see in Figure 1.16 and is the first of the recommended charts.

Excel 2013, at least in the area of charting, recognizes the possibility that you will want to use numeric values as nominal categories. It lets you express an opinion without forcing you to take all the extra steps required by Excel 2010. Still, if you're to participate effectively, you need to recognize the differences between, say, interval and nominal variables. You also need to recognize the ambiguities that crop up when you want to use a number as a category.

## Grouping with Pivot Tables

Another approach to constructing the frequency distribution is to use a pivot table. A related tool, the pivot chart, is based on the analysis that the pivot table provides. I prefer this method to using an array formula that employs `FREQUENCY()`. With a pivot table, once the initial groundwork is done, I can use the same pivot table to do analyses that go beyond the basic frequency distribution. But if all I want is a quick group count, `FREQUENCY()` is usually the faster way.

Again, there's more on pivot tables and pivot charts in Chapter 2 and later chapters, but this section shows you how to use them to establish the frequency distribution.

Building the pivot table (and the pivot chart) requires you to specify bins, just as the use of `FREQUENCY()` does, but that happens a little further on.

NOTE

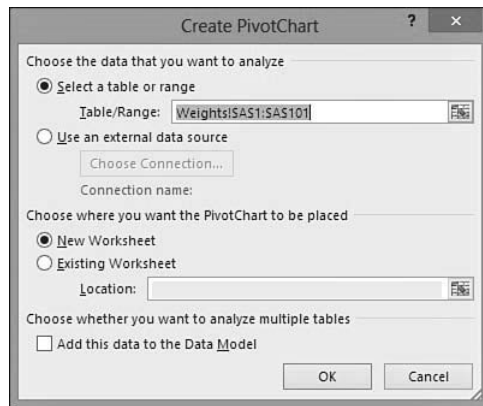
A reminder: When you use the `FREQUENCY()` method described in the prior section, a header at the top of the column of raw data can be helpful but is not required. When you use the pivot table method discussed in this section, the header is required.

Begin with your sample data in A1:A101 of Figure 1.16, just as before. Select any one of the cells in that range and then follow these steps:

1. Click the Insert tab. Click the PivotChart button in the Charts group. (Prior to Excel 2013, click the PivotTable drop-down in the Tables group and choose PivotChart from the drop-down list.) When you choose a pivot chart, you automatically get a pivot table along with it. The dialog box in Figure 1.19 appears.

**Figure 1.19**

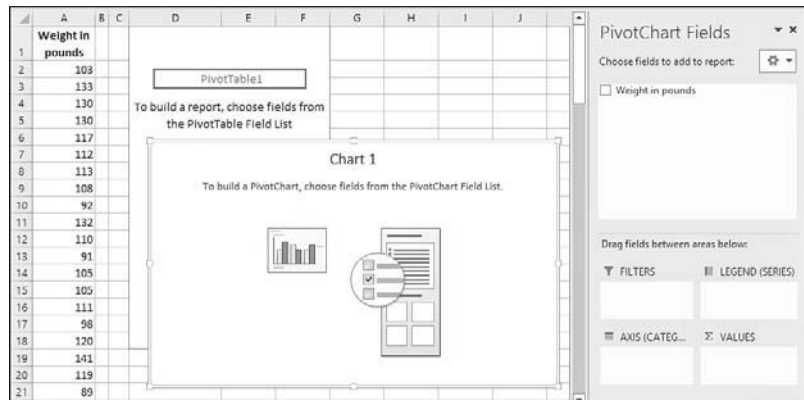
If you begin by selecting a single cell in the range containing your input data, Excel automatically proposes the range of adjacent cells that contain data.



2. Click the Existing Worksheet option button. Click in the Location range edit box. Then, to avoid overwriting valuable data, click some blank cell in the worksheet that has other empty cells to its right and below it.
3. Click OK. The worksheet now appears as shown in Figure 1.20.

**Figure 1.20**

With one field only, you normally use it for both Axis Fields (Categories) and Summary Values.

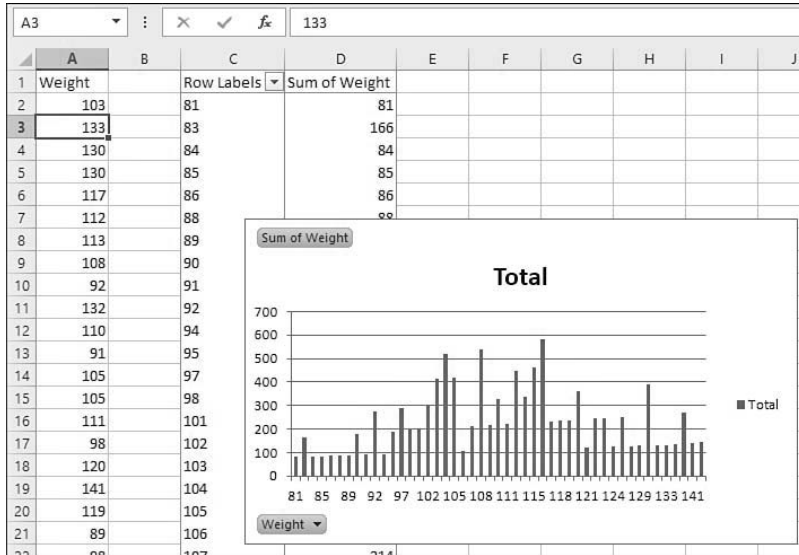


4. Click the Weight In Pounds field in the PivotTable Fields list and drag it into the Axis (Categories) area.
5. Click the Weight In Pounds field again and drag it into the  $\Sigma$  Values area. Despite the uppercase Greek sigma, which is a summation symbol, the  $\Sigma$  Values in a pivot table can show averages, counts, standard deviations, and a variety of statistics other than the sum. However, Sum is the default statistic for a field that contains numeric values only.

- The pivot table and pivot chart are both populated as shown in Figure 1.21. Right-click any cell that contains a row label, such as C2. Choose Group from the shortcut menu. The Grouping dialog box shown in Figure 1.22 appears.

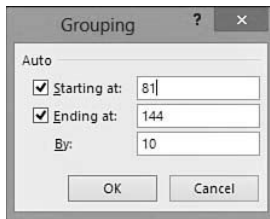
**Figure 1.21**

The Weight field contains numeric values only, so the pivot table defaults to Sum as the summary statistic.



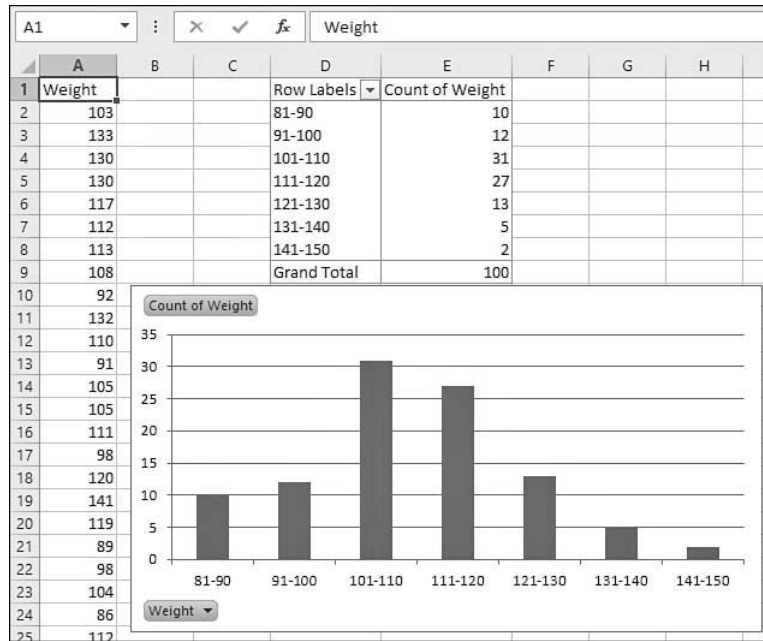
**Figure 1.22**

This step establishes the groups that the FREQUENCY() function refers to as bins.



- In the Grouping dialog box, set the Starting At value to **81** and enter **10** in the By box. Click OK.
- Right-click a cell in the pivot table under the header Sum of Weight. Choose Value Field Settings from the shortcut menu. Select Count in the Summarize Value Field By list box, and then click OK.
- The pivot table and chart reconfigure themselves to appear as in Figure 1.23. To remove the field buttons in the upper- and lower-left corners of the pivot chart, select the chart, click the Analyze tab, click Field Buttons, and select Hide All.

**Figure 1.23**  
This sample's frequency distribution has a slight right skew but is reasonably close to a normal curve.



## Building Simulated Frequency Distributions

It can be helpful to see how a frequency distribution assumes a particular shape as the number of underlying records increases. *Statistical Analysis: Excel 2013* has a variety of worksheets and workbooks for you to download from this book's website ([www.quepublishing.com/title/9780789753113](http://www.quepublishing.com/title/9780789753113)). The workbook for Chapter 1 has a worksheet named Figure 1.24 that samples records at random from a population of values that follows a normal distribution. The following figure, as well as the worksheet on which it's based, shows how a frequency distribution comes closer and closer to the population distribution as the number of sampled records increases.

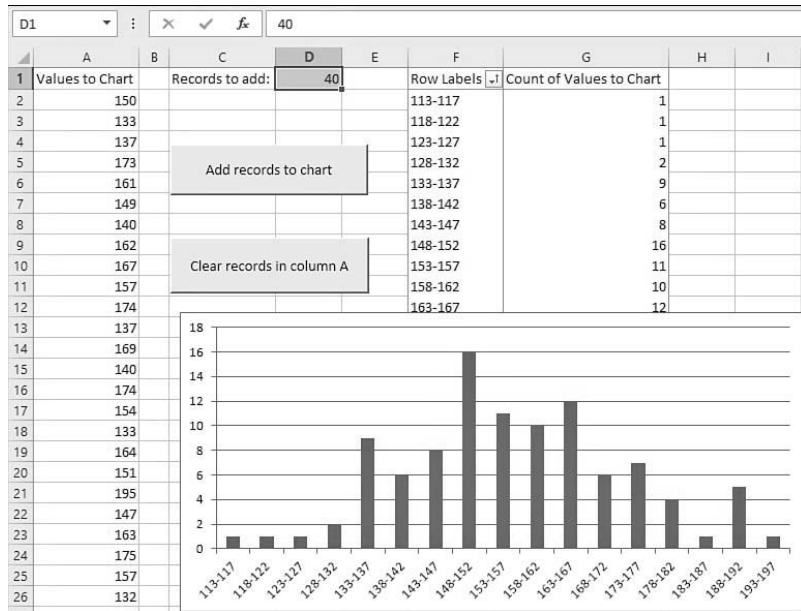
Begin by clicking the button labeled Clear Records in column A. All the numbers will be deleted from column A, leaving only the header value in cell A1. (The pivot table and pivot chart will remain as they were: It's a characteristic of pivot tables and pivot charts that they do not respond immediately to changes in their underlying data sources.)

Decide how many records you'd like to add, and then enter that number in cell D1. You can always change it to another number.

Click the button labeled Add Records to Chart. When you do so, several events take place, all driven by Visual Basic procedures that are stored in the workbook:

**Figure 1.24**

This frequency distribution is based on a population of records that follow a normal distribution.



- A sample is taken from the underlying normal distribution. The sample has as many records as specified in cell D1. (The underlying, normally distributed population is stored in a separate, hidden worksheet named Random Normal Values; you can display the worksheet by right-clicking a worksheet tab and selecting Unhide from the shortcut menu.)
- The sample of records is added to column A. If there were no records in column A, the new sample is written starting in cell A2. If there were already, say, 100 records in column A, the new sample would begin in cell A102.
- The pivot table and pivot chart are updated (or, in Excel terms, *refreshed*). As you click the Add Records to Chart button repeatedly, more and more records are used in the chart. The greater the number of records, the more nearly the chart comes to resemble the underlying normal distribution.

In effect, this is what happens in an experiment when you increase the sample size. Larger samples resemble more closely the population from which you draw them than do smaller samples. That greater resemblance isn't limited to the shape of the distribution: It includes the average value and measures of how the values vary around the average. Other things being equal, you would prefer a larger sample to a smaller one because it's likely to represent the population more closely.

But this effect creates a cost-benefit problem. It is usually the case that the larger the sample, the more accurate the experimental findings—and the more expensive the experiment. Many issues are involved here (and this book discusses them), but at some point the incremental accuracy of adding, say, ten more experimental subjects no longer justifies the incremental expense of adding them. One of the bits of advice that statistical analysis provides is to tell you when you're reaching the point when the returns begin to diminish.

With the material in this chapter—scales of measurement, the nature of axes on Excel charts, and frequency distributions—in hand, Chapter 2 moves on to the beginnings of practical statistical analysis, the measurement of central tendency.

*This page intentionally left blank*



# Index

## A

---

**a priori ordering approach, 396**

**absolute references**

calculating expected frequencies,  
145-146

semipartial correlations, 381-384

**adapting ANOVA Data Analysis tool**

for nested factors, 326-327

for random factors, 322-323

**adjusted group means, 458-461**

**adjusting means with LINES() and effect  
coding, 453-458**

**alpha, 126-127**

**alternative hypotheses, 113**

directionality of, 332

**The Analysis of Covariance and  
Alternatives (Wiley, 2011), 445**

**Analysis ToolPak. See Data Analysis  
add-in**

**ANCOVA (analysis of covariance), 433**

adjusted group means, 458-461

common regression line, testing for,  
445-447

increasing statistical power, 435-444

multiple comparison procedures

planned contrasts, 466-468

Scheffé procedure, 462-466

multiple covariates, 469-471

purposes of

bias reduction, 434-435

greater power, 434

removing bias, 447-452

**ANOVA (analysis of variance)**

- comparing variances
  - sums of squares within groups, 269-270
- comparing with multiple regression, 355-356
- Data Analysis add-in tools, 306-307
- experimental design
  - accurate design depiction, 317
  - data layout, 320-322
  - mixed models, 318
  - nested designs, 327-328
  - nuisance factors, 317-318
- F distribution, 279-280
- F test, 273-276
  - alpha, calculating, 276
  - designing, 323-325
  - statistical power, 350-354
- factorial ANOVA, 293-299
  - crossed factors, 315-316
  - fixed factors, 312
  - interaction, 294, 299-305
  - nested factors, 294, 315-316
  - random factors, 318-319
- factorial designs, 293
- F.DIST() function, 277
- F.DIST.RT() function, 277
- FINV() function, 278-279
- F.INV() function, 278-279
- means, adjusting with LINEST() and effect coding, 453-458
- multiple comparison procedures, 282-291
  - orthogonal contrasts, 289-290
  - planned contrasts, 289
  - Scheffé procedure, 284-289
- noncentral F distribution, 313, 344-350
  - PDF, 348-350
  - variance estimates, 344-347
- partitioning the scores, 265-268

- proportional cell frequencies, 309
- replication, 310
- single-factor ANOVA, unequal group sizes, 280-282
- sum of squares
  - between groups, 266-267, 270-273
  - within groups, 267-268
- versus t-tests, 263-265
- Two-Factor ANOVA tool (Data Analysis add-in), 297-299
- unequal group sizes, 305-310
- variance estimates, 363-364

**ANOVA: Single Factor tool (Data Analysis add-in), 319****ANOVA: Two-Factor With Replication tool (Data Analysis add-in), 320-321**

- adapting for nested factors, 326-327
- adapting for random factors, 322-323

**ANOVA: Two-Factor Without Replication tool (Data Analysis add-in), 309-310, 320**

- limitations of, 310-313

**arguments, 32-34**

- for BINOM.DIST() function, 115
- for BINOM.INV() function, 122
- Tails argument, 243-244
- for TREND() function, 94-95
- for T.TEST() function
  - Type argument, 248

**array formulas, 50-51**

- data arrays, 33
- values, counting, 48-49

**arrays, identifying for T.TEST() function, 242-243****assigning**

- effect codes in Excel, 368-370
- nominal value to numbers, 8-9

**assumptions**

- independent selections, 119-120
- random selection, 118-119

**average, 29-30****AVERAGE() function, 30-31**

## B

- balanced factorial designs, 386-387**
  - comparing with unbalanced designs, 386-393
- Bar charts, 6**
- bell curves, 14**
- best combination, 100-104**
- beta**
  - and statistical power, 224
- bias reduction function, 434-435**
- BINOM.DIST() function, 113-115**
  - arguments, 115
  - interpreting results of, 116
  - setting decision rules, 116-117
- binomial distributions, 112-117**
  - BINOM.DIST() function, 113-115
  - BINOM.INV() function, 121-127
  - complexity of, 123-125
  - formula, 120-121
  - hypothesis testing, 125-126
  - normal approximation to the binomial, 198
- BINOM.INV() function, 121-127**
  - alpha, 126-127
  - arguments, 122
- building**
  - frequency distributions, 18-26
    - grouping with FREQUENCY(), 19-23
    - grouping with pivot tables, 22-26
    - tallying the sample, 18
  - pivot charts, 45-46
  - simulated frequency distributions, 26-28

## C

- calculating**
  - binomial probability, 120-121
  - CDF, 350-352
  - correlation, 75-81
    - CORREL() function, 75-76, 81-84
    - covariance, 77-79
  - exact probability, 196-198
  - expected frequencies, 145-146
  - F ratios, 322-323, 329
  - interaction effect, 302-305
  - mean, 30-40
    - minimizing the spread, 36
  - median, 41-42
  - mode, 42-54
    - with worksheet formula, 47-48
  - probability in t-tests, 254
  - regression, 96-99
  - standard deviation, 62-63
  - standard error of the mean, 202-204
  - t-statistic, 254
  - variance, 62-63
    - bias, 68-70
    - degrees of freedom, 68
    - dividing  $N - 1$ , 66-68
- Campbell, Donald, 151**
- capitalizing on chance, 117**
- category scales, 5-7**
  - numeric values as categories, 23
- causal relationships, 88-90**
- CDF (cumulative density function)**
  - calculating, 350-352
- Central Limit Theorem, 30, 194-198**
  - exact probability, calculating, 196-198
  - normal approximation to the binomial, 198

**central tendency, 30. See also variability**

- mean, minimizing the spread, 36
- median, calculating, 41-42
- mode, calculating, 42-54

**chance, as threat to internal validity, 154-155****characteristics of normal distribution, 171-176**

- kurtosis, 174-176
- skewness, 172-174

**charts**

- Bar charts, 6
- frequency distributions, 12-28
- pivot charts, 3-4
  - building, 45-46
- XY charts, 10-12
  - correlation analysis, 84

**CHIDIST() function, 141-142****CHIINV() function, 143-144****CHISQ.DIST() function, 135-137, 140-141****CHISQ.DIST.RT() function, 141-142****CHISQ.INV() function, 135-137, 143****CHISQ.INV.RT() function, 143-144****CHISQ.TEST() function, 132-135, 144-145****chi-square distribution**

- CHIDIST() function, 141-142
- CHIINV() function, 143-144
- CHISQ.DIST() function, 135-137, 140-141
- CHISQ.DIST.RT() function, 141-142
- CHISQ.INV() function, 135-137, 143
- CHISQ.INV.RT() function, 143-144
- CHISQ.TEST(), 144-145
- CHISQ.TEST() function, 132-135
- CHITEST() function, 144-145

**CHITEST() function, 144-145****coding**

- dummy coding, 360
- effect coding, 358-359, 365-367
  - assigning effect codes in Excel, 368-370

factorial designs, 372-377

means, adjusting, 453-458

orthogonal coding, 360

**coefficient of determination, 105****common regression line, testing for, 445-447****comparing**

- ANOVA and multiple regression, 355-356
- balanced and unbalanced factorial designs, 386-393
- BINOM.DIST() and BINOM.INV(), 126
- correlation and causal relationships, 88-90
- critical values, 221
- FDIST() and F.DIST() functions, 277
- means between two groups, 199-200
  - z-scores, 201-204
- variances
  - based on sum of squares between groups, 270-273
  - based on sum of squares within groups, 269-270

**compatibility functions, xii****complexity of binomial analysis, 123-125****computational formulas, xiv****CONFIDENCE() function, 188-191****confidence intervals, 183-194**

- CONFIDENCE() function, 188-191
- CONFIDENCE.NORM() function, 188-191
- CONFIDENCE.T() function, 191-192
- constructing, 184-187
- hypothesis testing, 194

**CONFIDENCE.NORM() function, 188-191****CONFIDENCE.T() function, 191-192****consistency in naming functions, 72-71****constructing confidence intervals, 184-187**

**contingency tables, 129**  
 chi-square distribution, CHISQ.  
 TEST() function, 132-135  
 Index display (pivot tables), 146-147  
 probabilities, 130-131  
 Simpson's paradox, 139  
 Yule Simpson effect, 137-139

**controlling risk of Type II errors, 331-337**

**converting between interval and ordinal measurement, 7**

**CORREL() function, 75-76, 81-84**

**correlation, 73-91**  
 analyzing with XY charts, 84  
 calculating, 75-81  
     CORREL() function, 75-76  
 versus causal relationships, 88-90  
 correlation coefficient, 74-75  
 covariance, 77-80  
 definitional formula, 80-81  
 imperfect correlations, 80  
 negative correlation, 73-74  
 nonlinear, 83  
 and observational studies, 394-397  
 positive correlation, 73-74  
 regression  
     calculating, 96-99  
     intercept, 97-98  
     multiple regression, 99-100  
     shared variance, 104-105  
     slope, 97  
 semipartial correlations, 374-375  
     absolute references, 381-384  
     sum of squares, obtaining, 376-377  
 TREND() function, 93-96

**correlation coefficient, 74-75**  
 calculating, 82-83  
 regression, 91-93

**Correlation tool (Data Analysis add-in), 84-88**  
 Output Range issue, 88

**counting values with array formulas, 48-49**

**covariance, 79-80**  
 calculating, 77-79  
 multiple covariance analysis, 469-471

**creating**  
 one-way pivot tables, 109-112  
 two-way pivot tables, 128

**CRITBINOM() function, 127**

**critical values**  
 comparing, 221  
 finding  
     for t-tests, 220-221  
     for z-tests, 220

**crossed factors, 294, 315-316**

---

## D

### Data Analysis add-in

ANOVA: Two Factor Without Replication tool, 309-310  
 ANOVA: Single Factor tool, 319  
 Correlation tool, 84-88  
 Descriptive Statistics tool, 192-193  
 Equal Variances t-Test tool, 256-258  
 F-Test Two-Sample for Variances tool, 157-170  
     directional hypotheses, 169  
     nondirectional hypotheses, 166-168  
 problems with Excel's documentation, 156-157  
 t-Tests tool, 255-262  
     group variances, 255-256  
 Two-Factor ANOVA tool, 297-299  
 Unequal Variances t-Test tool, 258-260

### data arrays, 33

### date values in Excel, 9

### De Moivre, Abraham, 16

**decision rules**

- defining for t-test, 218-219
- nondirectional, 246
- setting for BINOM.DIST() function, 116-117

**defining worksheet functions, 31-32****definitional formulas, xiv**

- correlation, 80-81
- standard deviation, 64
- variance, 63-64

**degrees of freedom, 68-70**

- F distribution, 279-280
- specifying in Excel functions, 238-239

**dependent group t-tests, 239-240, 252-253**

- statistical power, 342-344

**descriptive statistics, xvi-xvii**

- frequency distributions, 15-17

**Descriptive Statistics tool (Data Analysis add-in), 192-193****designing an F test, 323-325****DEVSQ() function, 268****directional hypotheses, 228-229**

- F-Test Two-Sample for Variances tool, 169
- T.DIST() function, 237-238
- T.INV() function, 229-237
- t-tests, 340-341

**distribution of sample means, charting for statistical tests, 212****distributions, PDF, 348-350****documentation (Excel), problems with, 156-157****dummy coding, 360**


---

## E

**effect coding, 365-367, 385**

- adjusted group means, 458-461
- factorial designs, 372-377
- means, adjusting, 453-458

**Equal Variances t-Test tool (Data Analysis add-in), 256-258****error rates**

- manipulating, 224-226
- standard error of the mean, 206-208
- Type I error, 331

**establishing internal validity, 151-152****evaluating formulas, 36****evaluating planned orthogonal contrasts, 290-291****exact probability, calculating, 196-198****Excel, xii**

- Bar charts, 6
- compatibility functions, xii
- Data Analysis add-in
  - ANOVA: Two Factor Without Replication tool, 309-310
  - Correlation tool, 84-88
  - Descriptive Statistics tool, 192-193
  - F-Test Two-Sample for Variances tool, 157-170
  - t-Tests tool, 255-262
  - Unequal Variances t-Test tool, 258-260

## date values, 9

## documentation, problems with, 156-157

## effect coding, 368-370

## formulas, 31, 34-35

- evaluating, 51-53

## inaccuracies in, xv-xvi

## lists, 2-3

## matrix algebra, 106-107

## Ribbon, xii

## Solver, 37

- installing, 37-38

- setting up worksheets for, 38-40

## terminology, xii-xiii

## treatment effect, xv-xvi

## value axis, 5

## XY charts, 10-12

**expected counts, 130-131**  
**expected frequencies, calculating, 145-146**  
**experimental design, 394-397**  
 accurate design depiction, 317  
 crossed factors, 315-316  
 data layout, 320-322  
 F ratios, calculating, 322-323  
 F test, designing, 323-325  
 mixed models, 318  
 nested designs, 327-328  
 nested factors, 315-316  
 nuisance factors, 317-318  
 unequal group sizes, managing, 428-429

**experimental mortality, 154**  
**exponential smoothing, 156**

---

## F

---

**F distribution, 279-280**  
**F ratios, 344**  
 calculating, 322-323, 329  
 mixed model, selecting denominator, 325-326

**F tests, 273-276, 312-313**  
 alpha, calculating, 276  
 CDF, calculating, 350-352  
 designing, 323-325  
 F ratios, 344  
 multiple comparison procedures, 282-291  
 noncentral F distribution, 313, 344-350  
 PDF, 348-350  
 variance estimates, 344-347  
 power, calculating, 350-354  
 reasons for running, 158-159

**factorial ANOVA, 293-299**  
 crossed factors, 294, 315-316  
 fixed factors, 312

interaction, 294, 299-305  
 interaction effect, calculating, 302-305  
 statistical significance of, 300-302  
 nested factors, 294, 315-316  
 noncentral F distribution, 313  
 random factors, 318-319  
 rationales for multiple factors, 294-296  
 Two-Factor ANOVA tool (Data Analysis add-in), 297-299

**factorial designs, 293**  
 comparing balanced and unbalanced designs, 386-393  
 effect coding, 372-377  
 unbalanced designs, solving with multiple regression, 385-394

**factors, 293**  
 crossed factors, 294  
 nested factors, 294  
 random factors  
 adapting ANOVA Data Analysis tool for, 322-323

**Fay, Leo, 445**  
**F.DIST() function, 165, 277**  
**F.DIST.RT() function, 165-166, 277**  
**F.INV() function, 165-166, 278-279**  
**FINV() function, 278-279**  
**Fisher, R.A., 117**  
**fixed factors, 312**  
**fluctuating proportions of variance, 393-394**  
**forcing zero constant, 421-422**  
**formatting**  
 formulas, 35

**formulas, 31, 34-35**  
 arguments, 32-34  
 array formulas, 30, 50-51  
 counting values with, 48-49  
 binomial distributions, 120-121  
 computational formulas, xiv  
 definitional formulas, xiv

- evaluating, 36
- formatting, 35
- recalculating, 53-54
- returning the result
  - visible results, 35
- visible formulas, 35
- frequency distributions, 12-28**
  - binomial distributions, 112-117
    - BINOM.DIST() function, 113-115
    - complexity, 123-125
    - hypothesis testing, 125-126
  - building from a sample, 18-26
    - grouping with FREQUENCY(), 19-23
    - grouping with pivot tables, 22-26
    - tallying the sample, 18
  - chi-square distribution
    - CHISQ.DIST() function, 135-137
    - CHISQ.INV() function, 135-137
    - CHISQ.TEST() function, 132-135
  - in descriptive statistics, 15-17
  - in inferential statistics, 17-18
  - normal distribution
    - Central Limit Theorem, 194-198
    - characteristics of, 171-176
    - unit normal distribution, 176-177
  - range, 56-58
  - reasons for using, 15
  - simulated frequency distributions
    - building, 26-28
  - standard deviation, 64-65
- FREQUENCY() function, 19-23, 43**
- F-Test Two-Sample for Variances tool, 157-170**
  - directional hypotheses, 169
  - nondirectional hypotheses, 166-168
  - numeric example, 159-161
- functions**
  - arguments, 32-34
  - AVERAGE(), 30-31
  - BINOM.DIST(), 113-115
    - arguments, 115
    - interpreting results of, 116
    - setting decision rules, 116-117
  - BINOM.INV(), 121-127
    - alpha, 126-127
    - arguments, 122
  - CHIDIST(), 141-142
  - CHIINV(), 143-144
  - CHISQ.DIST(), 135-137, 140-141
  - CHISQ.DIST.RT(), 141-142
  - CHISQ.INV(), 135-137, 143
  - CHISQ.INV.RT(), 143-144
  - CHISQ.TEST(), 132-135, 144-145
  - CHITEST(), 144-145
  - CONFIDENCE(), 188-191
  - CONFIDENCE.NORM(), 188-191
  - CONFIDENCE.T(), 191-192
  - consistency in naming, 72-71
  - CORREL(), 75-76, 81-84
  - CRITBINOM(), 127
  - degrees of freedom, specifying, 238-239
  - DEVSQ(), 268
  - F.DIST(), 165, 277
  - F.DIST.RT(), 165-166, 277
  - F.INV(), 165-166
  - FINV(), 278-279
  - F.INV(), 278-279
  - FREQUENCY(), 19-23
  - INTERCEPT(), 97
  - KURT(), 176
  - LINEST(), 100-103, 397-428
    - calculation of results, 404-406
    - Excel 2007 version, 422-425
    - intercept, 399-400
    - means, adjusting, 453-458
    - multicollinearity, handling, 416-421
    - negative R2, 425-428



QR decomposition, 417-419  
 regression coefficients, 398  
 regression diagnostics, calculating, 412-416  
 standard errors, 398-399  
 statistics, 401-404  
 sum of squares regression, 410-412  
 MATCH(), 48  
 MEDIAN(), 41-42  
 MINVERSE(), 107  
 MMULT(), 107  
 MODE(), 43-45  
 NORM.DIST(), 177-180  
 NORMDIST(), 210  
 NORM.INV(), 180-181  
 NORM.S.DIST(), 181-182  
 NORM.S.INV(), 182  
 PEARSON(), 76  
 regression coefficients, obtaining, 406-410  
 returning the result, 34  
 SKEW(), 17  
 SLOPE(), 97  
 STDEV(), 62-63, 70  
 STDEVA(), 70  
 STDEV.P(), 70  
 STDEV.P(), 71  
 STDEVPA(), 70  
 T.DIST(), 237-238  
 T.INV(), hypothesis testing, 229-237  
 TREND(), 93-96, 99-100  
   arguments, 94-95  
   replacing squared semipartial correlations, 377-384  
 T.TEST(), 254-255  
   arrays, identifying, 242-243  
   results, interpreting, 244-245  
   syntax, 242  
   Type argument, 248

T.TEST() function  
   Tails argument, 243-244  
 VAR(), 63, 71  
 VARA(), 71  
 VARP(), 68, 71  
 VAR.S(), 68  
 VLOOKUP(), 368-370  
 worksheet functions  
   defining, 31-32

---

## G

**Galton, Francis, 90**  
**gambler's fallacy, 130**  
**General Linear Model, 365**  
**grand mean, 366**  
**group variances, in t-tests, 255-256**

---

## H

**history, as threat to internal validity, 152-153**  
**horizontal axis, charting for statistical tests, 210**  
**How to Lie with Statistics, 149**  
**Huff, Darrell, 149**  
**Huitema, B.E., 445**  
**hypothesis testing, 227-238**  
   in binomial analysis, 125-126  
   confidence intervals, 194  
   directional hypotheses, 228-229  
     T.DIST() function, 237-238  
     T.INV() function, 229-237  
   inferential statistics, 150-151  
   nondirectional hypotheses, 228-229  
     t-tests, 338-340

I

---

**identifying arrays for T.TEST() function, 242-243**

**imperfect correlations, 80**

**inaccuracies in Excel, xv-xvi**

**increasing**

- sample size of t-tests, 341-342
- statistical power with ANCOVA, 435-444

**independent observations in t-tests, 249-250**

**independent selections, 119-120**

**Index display (pivot tables), 146-147**

**individual observations, effect coding, 365-367**

**inferential statistics, xvii, 150-155**

- frequency distributions, 17-18
- hypothesis testing, 150-151
- internal validity, establishing, 151-152
- validity, internal validity, 151-155

**installing Solver, 37-38**

**instrumentation, as threat to internal validity, 153**

**interaction, 294, 299-305**

- statistical significance of, 300-302

**intercept, 97-98**

- in LINEST() function, 399-400

**INTERCEPT() function, 97**

**internal validity**

- establishing, 151-152
- threats to
  - chance, 154-155
  - history, 152-153
  - instrumentation, 153
  - maturation, 153
  - mortality, 154
  - regression, 153-154
  - selection, 152
  - testing, 153

**interpreting**

- BINOM.DIST() results, 116
- T.TEST() function results, 244-245

**interval scales, 7**

J-K

---

**Johnson, Palmer, 445**

**The Johnson-Neyman Technique, Its Theory and Application (Biometrika, December 1950), 445**

**KURT() function, 176**

**kurtosis**

- in normal distribution, 174-176
- quantifying, 176

L

---

**leptokurtic curves, 175**

**limitations**

- of ANOVA: Two Factor Without Replication tool, 310-313

**LINEST() function, 99-103, 397-428**

- calculation of results, 404-406
- Excel 2007 version, 422-425
- intercept, 399-400
- means, adjusting, 453-458
- multicollinearity, handling, 416-421
- negative R<sup>2</sup>, 425-428
- QR decomposition, 417-419
- regression coefficients, 398
- regression coefficients, obtaining, 406-410
- regression diagnostics, calculating, 412-416
- standard errors, 398-399
- statistics, 401-404
- sum of squares regression, 410-412
- zero constant, forcing, 421-422

lists, 2-3

locating Solver, 37-38

---

## M

---

**managing unequal group sizes**

in observational research, 430-432

in true experiments, 428-429

**manipulating error rates, 224-226**

**MATCH() function, 48**

**matrix algebra, 106-107**

**maturation, as threat to internal validity, 153**

**means**

adjusted group means, 458-461

calculating, 30-40

comparing between two groups, 199-200

z-scores, 201-204

deviation, 65

grand mean, 366

spread, minimizing, 36

standard error, 202-208

error rates, 206-208

statistical power, 222-224

beta, 224

testing, 200-201

**measuring**

standard deviation

variance, 60-61

z-scores, 60

variability, 56-58

**median, 29**

calculating, 41-42

**MEDIAN() function, 41-42**

**mesokurtic curves, 175**

**Microsoft Excel. See Excel**

**minimizing the spread, 36**

**MINVERSE() function, 107**

**mixed models, 318**

selecting denominator, 325-326

**mixed references, calculating expected frequencies, 145-146**

**MMULT() function, 107**

**mode, 30**

calculating, 42-54

with worksheet formula, 47-48

values, counting with array formulas, 48-49

**MODE() function, 43-45**

**mortality as threat to internal validity, 154**

**multicollinearity in LINEST() function, 416-421**

**multiple comparison procedures, 282-291**

orthogonal contrasts, 289-290

planned contrasts, 289, 466-468

Scheffé procedure, 284-289, 462-466

**multiple covariance analysis, 469-471**

**multiple factors, rationale for, 294-296**

**multiple regression, 355-356**

best combination, 100-104

coefficient of determination, 105

combining predictors, 99-100

comparing with ANOVA, 355-356

effect coding, 358-359

factorial designs, 372-377

predictor variables, 105-106

proportions of variance, 360-363

shared variance, 104-105

solving unbalanced factorial designs, 385-394

TREND() function, 99-100, 379-381

variance estimates, 364-365

---

## N

---

**negative correlation, 73-74**

**negative R<sup>2</sup>, 425-428**

**negatively skewed distributions, 14-15**

**nested designs, 327-328**

**nested factors, 294, 315-316**  
 adapting ANOVA Data Analysis tool for, 326-327

**nominal scales, 5-7**

**noncentral F distribution, 313, 344-350**  
 PDF, 348-350  
 variance estimates, 344-347

**nondirectional decision rules, 246**

**nondirectional hypotheses, 228-229**  
 F-Test Two-Sample for Variances tool, 166-168  
 t-tests, 338-340

**nondirectional tests, 246-248**

**nonlinear correlation, 83**

**normal approximation to the binomial, 198**

**normal distribution**  
 Central Limit Theorem, 194-198  
 exact probability, calculating, 196-198  
 normal approximation to the binomial, 198  
 characteristics of, 171-176  
 kurtosis, 174-176  
 skewness, 172-174  
 confidence intervals, 183-194  
 constructing, 184-187  
 hypothesis testing, 194  
 NORM.DIST() function, 177-180  
 NORM.INV() function, 180-181  
 NORM.S.DIST() function, 181-182  
 NORM.S.INV() function, 182  
 in t-tests, 249  
 unit normal distribution, 176-177

**NORM.DIST() function, 177-180**

**NORMDIST() function, 210**

**NORM.INV() function, 180-181**

**NORM.S.DIST() function, 181-182**

**NORM.S.INV() function, 182**

**nuisance factors, in experimental design, 317-318**

**null hypotheses, 113**  
 rejecting, 222  
 statistical power, 222-224  
 beta, 224  
 error rate, manipulating, 224-226

**numbers, as nominal value, 8-9**

**numeric example of F-Test tool, 165-161**

**numeric scales, 7**  
 interval scales, 7  
 ratio scales, 7

**numeric variables, XY charts, 10-12**

---

## O

**observational studies, 394-397**  
 a priori ordering approach, 396  
 unequal group sizes, managing, 430-432

**observations, effect coding, 365-367**

**observed counts, 130-131**

**obtaining**  
 regression coefficients with LINEST(), 406-410  
 sum of squares with semipartial correlation, 376-377

**one-tailed tests, 246**

**one-way pivot tables, creating, 109-112**

**ordinal scales, 7**

**orthogonal coding, 360**

**orthogonal contrasts, 289-290**

**Output Range issue (Correlation tool), 88**

---

## P

**parameters, 66**  
 confidence intervals, 183-194  
 constructing, 184-187

**partitioning scores (ANOVA), 265-268**

**PDF (probability density function), 348-350**

**Pearson, Karl, 76, 91**  
**PEARSON() function, 76**  
**percentages, displaying pivot table counts as, 111**  
**pivot charts, 3-4**  
 building, 45-46  
**pivot tables, 3-4, 22-26**  
 Index display, 146-147  
 one-way pivot tables, creating, 109-112  
 two-way pivot tables, 127-137  
 creating, 128  
 expected counts, 130-131  
 observed counts, 130-131  
**planned contrasts, 289**  
 ANCOVA, 466-468  
**planned orthogonal contrasts, evaluating, 290-291**  
**platykurtic curves, 175**  
**population frequency, 201**  
**population parameters, 66**  
**population values, charting for statistical tests, 210**  
**positive correlation, 73-74**  
**positively skewed distributions, 14-15**  
**power, 332**  
 determining sample size, 352-354  
 directionality of alternative hypotheses, 332  
 of F tests, 350-354  
 increasing with ANCOVA, 435-444  
 quantifying, 335-337  
 sample size, 332  
 of t-tests, 337-344  
 nondirectional hypotheses, 338-340  
 visualizing, 333-335  
**prediction, regression, 90-91**  
 TREND() function, 93-96  
**probability, 130-131**  
 calculating, 120-121  
 exact probability, calculating, 196-198

gambler's fallacy, 130  
 observed versus expected counts, 130-131  
 Simpson's paradox, 139  
 of Type II errors, controlling risk, 331-337  
**problems with Excel's documentation, 156-157**  
**proportional cell frequencies, 309**  
**proportions of variance, 360-363, 393-394**  
**purpose of ANCOVA**  
 bias reduction, 434-435  
 greater power, 434

---

## Q

**QR decomposition, 417-419**  
**quantifying**  
 kurtosis, 176  
 power, 335-337  
 statistical power, 223

---

## R

**random factors, 318-319**  
 adapting ANOVA Data Analysis tool for, 322-323  
**random selection, 118-119**  
**range, measuring variability, 56-58**  
**ratio scales, 7**  
**rationales for multiple factors, 294-296**  
**reasons for using frequency distributions, 15**  
**Recalculate key, 53-54**  
**regression, 90-93**  
 calculating, 96-99  
 common regression line, testing for, 445-447  
 intercept, 97-98

- multiple regression, 355-356
  - best combination, 100-104
  - coefficient of determination, 105
  - combining predictors, 99-100
  - comparing with ANOVA, 355-356
  - effect coding, 358-359
  - factorial designs, 372-377
  - predictor variables, 105-106
  - proportions of variance, 360-363
  - solving unbalanced factorial designs, 385-394
  - TREND() function, 99-100
- slope, 97
- as threat to internal validity, 153-154
- TREND() function, 93-96
- unequal group sizes, 370-372
- variance estimates, 364-365
- regression coefficients, obtaining from LINEST() function, 406-410**
- regression lines, 78**
- rejecting null hypotheses, 222**
- relative addressing, 381-384**
- removing bias using ANCOVA, 447-452**
- replacing squared semipartial correlations with TREND() function, 377-384**
- replication, 310**
- residuals, 379-381**
- results**
  - of BINOM.DIST(), interpreting, 116
  - of LINEST(), calculation, 404-406
  - of T.TEST() function, interpreting, 244-245
  - visible results, 35
- returning the result, 34**
- Ribbon (Excel), xii**
- risk of Type II errors, controlling, 331-337**

---

## S

- sample size, calculating with power, 352-354**
- scales of measurement, 4-9**
  - Bar charts, 6
  - category scales, 5-7
  - numeric scales, 7
    - interval scales, 7
    - ordinal scales, 7
    - ratio scales, 7
  - ordinal scales, 7
- scatter charts, 10-12**
- Scheffé procedure, 284-289**
  - ANCOVA, 462-466
- selection, as threat to internal validity, 152**
- semipartial correlations, 374-375**
  - absolute references, 381-384
  - squared semipartial correlations, replacing with TREND() function, 377-384
  - sum of squares, obtaining, 376-377
- setting up worksheets for Solver, 38-40**
- shared variance, 104-105**
- sigma, 66**
- Simpson's paradox, 139**
- simulated frequency distributions, building, 26-28**
- single-column lists, 2-3**
- single-factor ANOVA, unequal group sizes, 280-282**
- SKEW() function, 17**
- skewed distribution, 41**
- skewness, in normal distribution, 172-174**
- SLOPE() function, 97**
- Solver, 37**
  - installing, 37-38
  - setting up worksheets for, 38-40

- specifying degrees of freedom in Excel functions, 238-239**
- spread, minimizing, 36**
- squared semipartial correlations, replacing with TREND() function, 377-384**
- standard deviation, 58-62**
  - calculating, 62-63
    - bias, 68-70
  - charting for statistical tests, 211-212
  - definitional formula, 64
  - degrees of freedom, 69-70
  - mean deviation, 65
  - measuring, 56-58
  - population parameters, 66
  - squaring the deviations, 65
  - variance, 60-61
  - z-scores, 60
- standard error of the mean, 202-208**
  - error rates, 206-208
  - in t-tests, 253
- Stanley, Julian, 151**
- statistical inference**
  - assumptions
    - independent selections, 119-120
    - random selection, 118-119
  - binomial probability, calculating, 120-121
- statistical power, 222-224**
  - beta, 224
  - directionality of alternative hypotheses, 332
  - error rate, manipulating, 224-226
  - of F tests, 350-354
  - increasing with ANCOVA, 435-444
  - quantifying, 223, 335-337
  - risk, controlling, 331-337
  - sample size, 332
- of t-tests, 337-344
  - dependent group t-tests, 342-344
  - directional hypotheses, 340-341
  - nondirectional hypotheses, 338-340
  - visualizing, 260-261, 333-335
- statistical process control, 56**
- statistical significance of interaction, 302-305**
- statistical tests**
  - charting the data
    - creating the charts, 212-216
    - distribution of sample means, 212
    - horizontal axis, 210
    - mean of the sample, 212-213
    - population values, 210
    - standard deviation, 211-212
    - z-scores, 209-210
  - t-test, 216-226
  - t-tests
    - critical value, finding, 220-221
    - decision rule, defining, 218-219
    - z-tests, finding critical value, 220
- statistics, xvi-xvii**
  - descriptive statistics, frequency distributions, 15-17
  - inferential statistics, xvii, 150-155
    - frequency distributions, 17-18
    - hypothesis testing, 150-151
    - validity, 151
- STDEV() function, 62-63, 70**
- STDEVA() function, 70**
- STDEV.P() function, 70**
- STDEV.P() function, 71**
- STDEVPA() function, 70**
- STDEV.S() function, 71**
- studentized range statistic, 283**

sum of squares

**sum of squares**

- between groups, 266-267
- within groups, 267-270
- obtaining with semipartial correlation, 376-377

**syntax**

- TREND() function, 94-95
- T.TEST() function, 242

---

## T

---

**Tails argument, 243-244**

**T.DIST() function, 237-238**

**t-distributions, 175**

**terminology, xii-xiii**

**testing**

- for common regression line, 445-447
- F test, 273-276
- hypotheses, 227-238
  - directional hypotheses, 228-229
  - nondirectional hypotheses, 228-229
- T.DIST() function, 237-238
- T.INV() function, 229-237
- means, 200-201
- nondirectional tests, 246-248
- as threat to internal validity, 153
- t-tests
  - versus ANOVA, 263-265
  - correlation, 253-254
  - dependent group t-tests, 239-240, 252-253
  - Equal Variances t-Test tool (Data Analysis add-in), 256-258
  - group variability, 253
  - increasing sample size, 341-342
  - independent observations, 249-250
  - normal distributions, 249
  - probability, calculating, 254
  - standard error, calculating for dependent groups, 250-252

- standard error of the mean, 253
- statistical power, 337-344
- t-statistic, calculating, 254
- unequal group variances, 240-241
- Unequal Variances t-Test tool (Data Analysis add-in), 258-260
- when to avoid, 261-262

**threats to internal validity**

- chance, 154-155
- history, 152-153
- instrumentation, 153
- maturation, 153
- mortality, 154
- regression, 153-154
- selection, 152
- testing, 153

**T.INV() function, hypothesis testing, 229-237**

**treatment effect, xv-xvi**

**TREND() function, 93-96, 99-100**

- arguments, 94-95
- replacing squared semipartial correlations, 377-384

**t-statistic, calculating, 254**

**t-test, 216-226**

**T.TEST() function, 254-255**

- arrays, identifying, 242-243
- results, interpreting, 244-245
- syntax, 242
- Tails argument, 243-244
- Type argument, 248

**t-tests, 199**

- versus ANOVA, 263-265
- correlation, 253-254
- critical value, finding, 220-221
- decision rule, defining, 218-219
- dependent group t-tests, 239-240, 252-253
- Equal Variances t-Test tool, 256-258
- group variability, 253
- group variances, 255-256



increasing sample size, 341-342  
 independent observations, 249-250  
 normal distributions, 249  
 probability, calculating, 254  
 standard error, calculating for  
   dependent groups, 250-252  
 standard error of the mean, 253  
 statistical power, 337-344  
   dependent group t-tests, 342-344  
   directional hypotheses, 340-341  
   nondirectional hypotheses, 338-340  
 t-statistic, calculating, 254  
 T.TEST() function, Tails argument,  
   243-244  
 unequal group variances, 240-241  
 Unequal Variances t-Test tool (Data  
   Analysis add-in), 258-260  
 when to avoid, 261-262  
**t-Tests tool (Data Analysis add-in), 255-262**  
**two-column lists, 3-4**  
**Two-Factor ANOVA tool (Data Analysis  
 add-in), 297-299**  
**two-tailed tests, 246**  
**two-way pivot tables, 127-137**  
   creating, 128  
   expected counts, 130-131  
   observed counts, 130-131  
**Type argument, 248**  
**Type I errors, 331**  
**Type II errors, controlling risk of, 331-337**

---

## U

**unbalanced factorial designs**  
   comparing with balanced designs,  
     386-393  
   solving with multiple regression,  
     385-394

**unequal group sizes**  
   managing  
     in observational research, 430-432  
     in true experiments, 428-429  
   regression analysis, 370-372  
**unequal group variances, 240-241**  
   in single-factor ANOVA, 280-282  
**Unequal Variances t-Test tool (Data  
 Analysis add-in), 258-260**  
**unit normal distribution, 176-177**

---

## V

**validity, 151**  
   internal validity, threats to, 152-155  
**value axis, 5**  
**values, 1-4**  
   arguments, 32-34  
   numeric values as categories, 23  
**VAR() function, 63, 71**  
**VARA() function, 71**  
**variability, 55**  
   group variability in t-tests, 253  
   measuring with range, 56-58  
**variables, 1-4**  
   in balanced factorial designs, 386-387  
   frequency distributions, 12-28  
   numeric variables, XY charts, 10-12  
**variance, 60-61**  
   ANOVA  
     comparing variances, 268-273  
     F distribution, 279-280  
     factorial ANOVA, 293-299  
     factorial designs, 293  
     F.DIST() function, 277  
     F.DIST.RT() function, 277  
     FINV() function, 278-279  
     F.INV() function, 278-279

- partitioning the scores, 265-268
- proportional cell frequencies, 309
- replication, 310
- sum of squares between groups, 266-267, 270-273
- sum of squares within groups, 267-270
- versus t-tests, 263-265
- unequal group sizes, 280-282, 305-310
- calculating, 62-63
  - bias, 68-70
  - dividing  $N - 1$ , 66-68
- definitional formula, 63-64
- degrees of freedom, 68-70
- estimates in noncentral F distribution, 344-347
- fluctuating proportions of variance, 393-394
- F-Test Two-Sample for Variances tool, 157-170
- as parameter, 66
- proportions of variance, 360-363
- shared variance, 104-105
- unequal group variances, 240-241
- VARP() function, 68, 71**
- VAR.S() function, 68**
- visible formulas, 35**
- visible results, 35**
- visualizing statistical power, 260-261, 333-335**
- VLOOKUP() function, 368-370**

---

## W

---

**when to avoid t-tests, 261-262**

**worksheet formulas**

- mode, calculating, 47-48
- recalculating, 53-54

**worksheet functions**

- defining, 31-32
- LINEST(), 397-428

---

## X-Y-Z

---

**XY charts, correlation analysis, 84**

**Yule Simpson effect, 137-139**

**zero constant, forcing (LINEST() function), 421-422**

**z-scores, 60, 99, 200**

- charting for statistical tests, 209-210
- comparing means between two groups, 201-204
- predicting, 92-93

**z-tests, finding critical value, 220**