Making Everything Easier!™

2nd Edition

Statistics FOR DUMMIES®

Learn to:

- Grasp statistical ideas, techniques, formulas, and calculations
- Interpret and critique graphs and charts, determine probability, and work with confidence intervals
- Critique and analyze data from polls and experiments

Deborah J. Rumsey, PhD

Professor of Statistics, The Ohio State University

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a2"><div class="calibre1">
<img alt="cover.eps" src="images/00001.jpg" class="calibre2"/>
</div>  <div class="mbppagebreak" id="a1"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a2" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a1" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a5">
<h1 id="a3" class="calibre3">Statistics For Dummies,<sup
class="calibre4">®</sup> 2nd Edition</h1>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold">Visit <a
href="http://www.dummies.com/cheatsheet/statistics">www.dummies
.com/cheatsheet/statistics</a> to view this book's cheat sheet.
</span></span></p></blockquote>
<div class="calibre1"><p class="calibre8">
<span class="bold">Table of Contents</span></p>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a8">Introduction</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a9">About This Book</a>
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a10">Conventions Used in This Book</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a11">What You’re Not to Read</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a12">Foolish Assumptions</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a13">How This Book Is Organized</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a14">Part I: Vital
Statistics about Statistics</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a15">Part II: Number-Crunching Basics</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a16">Part III: Distributions and the Central Limit
Theorem</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a17">Part IV: Guesstimating and Hypothesizing with
Confidence</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a18">Part V: Statistical Studies and the Hunt for a
Meaningful Relationship</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
```

anchor.a#a19">Part VI: The Part of Tens</div>
</blockquote><div class="calibre1">Icons Used in This Book</div>
<div class="calibre1">Where to Go from Here</div>
</blockquote><div class="calibre1">Part I: Vital Statistics about Statistics
</div><blockquote class="calibre9">
<div class="calibre1">Chapter 1: Statistics in a Nutshell</div>
<blockquote class="calibre9"><div class="calibre1">Thriving in a Statistical World</div>
<div class="calibre1">Designing Appropriate Studies</div>
<blockquote class="calibre9"><div class="calibre1">Surveys</div>
<div class="calibre1">Experiments</div>
</blockquote><div class="calibre1">Collecting Quality Data</div>
<blockquote class="calibre9"><div class="calibre1">Selecting a good sample</div>
<div class="calibre1">Avoiding bias in your data</div>
</blockquote><div class="calibre1">Creating Effective Summaries</div>
<blockquote class="calibre9"><div class="calibre1">Descriptive statistics</div>
<div class="calibre1">Charts and graphs</div>
</blockquote><div class="calibre1">Determining Distributions</div>
<div class="calibre1">Performing Proper Analyses</div>
<blockquote class="calibre9"><div class="calibre1">Margin of error and confidence intervals</div>
<div class="calibre1">Hypothesis tests</div>
<div class="calibre1">Correlation, regression, and two-way tables
</div>
</blockquote><div class="calibre1">Drawing Credible Conclusions</div>
<blockquote class="calibre9"><div class="calibre1"><a

<div class="calibre1">Statistics in the office</div>
</blockquote></blockquote><div class="calibre1">Chapter 3: Taking Control: So Many Numbers, So Little Time</div>
<blockquote class="calibre9"><div class="calibre1">Detecting Errors, Exaggerations, and Just Plain Lies</div>
<blockquote class="calibre9"><div class="calibre1">Checking the math
</div>
<div class="calibre1">Uncovering misleading statistics</div>
<div class="calibre1">Looking for lies in all the right places
</div>
</blockquote><div class="calibre1">Feeling the Impact of Misleading Statistics</div>
</blockquote><div class="calibre1">Chapter 4: Tools of the Trade</div>
<blockquote class="calibre9"><div class="calibre1">Statistics: More than Just Numbers</div>
<div class="calibre1">Grabbing Some Basic Statistical Jargon</div>
<blockquote class="calibre9"><div class="calibre1">Data</div>
<div class="calibre1">Data set</div>
<div class="calibre1">Variable</div>
<div class="calibre1">Population</div>
<div class="calibre1">Sample, random, or otherwise</div>
<div class="calibre1">Statistic</div>
<div class="calibre1">Parameter</div>
<div class="calibre1">Bias</div>
<div class="calibre1">Mean (Average)</div>
<div class="calibre1">Median</div>
<div class="calibre1">Standard deviation</div>

<div class="calibre1">Percentile</div>
<div class="calibre1">Standard score</div>
<div class="calibre1">Distribution and normal distribution</div>
<div class="calibre1">Central Limit Theorem</div>
<div class="calibre1">z-values</div>
<div class="calibre1">Experiments</div>
<div class="calibre1">Surveys (Polls)</div>
<div class="calibre1">Margin of error</div>
<div class="calibre1">Confidence interval</div>
<div class="calibre1">Hypothesis testing</div>
<div class="calibre1">p-values</div>
<div class="calibre1">Statistical significance</div>
<div class="calibre1">Correlation versus causation</div>
</blockquote></blockquote></blockquote><div class="calibre1">Part II: Number-Crunching Basics</div><blockquote class="calibre9">
<div class="calibre1">Chapter 5: Means, Medians, and More</div>
<blockquote class="calibre9"><div class="calibre1">Summing Up Data with Descriptive Statistics</div>
<div class="calibre1">Crunching Categorical Data: Tables and Percents</div>
<div class="calibre1">Measuring the Center with Mean and Median</div>
<blockquote class="calibre9"><div class="calibre1">Averaging out to the mean</div>
<div class="calibre1">Splitting your data down the median</div>
<div class="calibre1">Comparing means and medians: Histograms</div>

</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a118">Accounting for Variation</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a119">Reporting the standard
deviation</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a120">Being out of range</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a121">Examining the Empirical Rule (68-95-99.7)
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a122">Measuring Relative Standing with Percentiles
</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a123">Calculating
percentiles</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a124">Interpreting percentiles</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a125">Gathering a five-number summary</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a126">Exploring interquartile range</div>
</blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a129">Chapter 6: Getting the
Picture: Graphing Categorical Data</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a130">Take Another Little
Piece of My Pie Chart</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a131">Tallying personal
expenses</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a132">Bringing in a lotto revenue</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a133">Ordering takeout</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a134">Projecting age trends</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a135">Raising the Bar on Bar Graphs</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a136">Tracking
transportation expenses</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a137">Making a lotto profit</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a138">Tipping the scales on a bar graph</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a139">Pondering pet peeves</div>

```
</blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a142">Chapter 7: Going by
the Numbers: Graphing Numerical Data</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a143">Handling
Histograms</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a144">Making a histogram</a>
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a145">Interpreting a histogram</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a146">Putting numbers with pictures</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a147">Detecting misleading histograms</a></div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a148">Examining Boxplots</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a149">Making a boxplot</a>
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a150">Interpreting a boxplot</a></div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a151">Tackling Time Charts</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a152">Interpreting time
charts</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a153">Understanding variability: Time charts versus
histograms</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a154">Spotting misleading time charts</a></div>
</blockquote></blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a157">Part III:
Distributions and the Central Limit Theorem</a></div>
<blockquote class="calibre9">
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a160">Chapter 8: Random Variables and the Binomial
Distribution</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a161">Defining a Random
Variable</a></div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a162">Discrete versus
continuous</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a163">Probability distributions</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
```

anchor.a#a164">The mean and variance of a discrete random variable</div>
</blockquote><div class="calibre1">Identifying a Binomial</div>
<blockquote class="calibre9"><div class="calibre1">Checking binomial conditions step by step</div>
<div class="calibre1">No fixed number of trials</div>
<div class="calibre1">More than success or failure</div>
<div class="calibre1">Trials are not independent</div>
<div class="calibre1">Probability of success (p) changes</div>
</blockquote><div class="calibre1">Finding Binomial Probabilities Using a Formula</div>
<div class="calibre1">Finding Probabilities Using the Binomial Table</div>
<blockquote class="calibre9"><div class="calibre1">Finding probabilities for specific values of X</div>
<div class="calibre1">Finding probabilities for X greater-than, less-than, or between two values</div>
</blockquote><div class="calibre1">Checking Out the Mean and Standard Deviation of the Binomial</div>
</blockquote><div class="calibre1">Chapter 9: The Normal Distribution</div>
<blockquote class="calibre9"><div class="calibre1">Exploring the Basics of theNormal Distribution</div>
<div class="calibre1">Meeting the Standard Normal (Z-) Distribution</div>
<blockquote class="calibre9"><div class="calibre1">Checking out Z</div>
<div class="calibre1">Standardizing from X to Z</div>
<div class="calibre1">Finding probabilities for Z with the Z-table</div>
</blockquote><div class="calibre1">Finding Probabilities for a Normal

Distribution</div>
<div class="calibre1">Finding X When You Know the Percent</div>
<blockquote class="calibre9"><div class="calibre1">Figuring out a percentile for a normal distribution</div>
<div class="calibre1">Translating tricky wording in percentile problems</div>
</blockquote><div class="calibre1">Normal Approximation to the Binomial
</div>
</blockquote><div class="calibre1">Chapter 10: The t-Distribution</div>
<blockquote class="calibre9"><div class="calibre1">Basics of the t-Distribution</div>
<blockquote class="calibre9"><div class="calibre1">Comparing the t- and Z-distributions</div>
<div class="calibre1">Discovering the effect of variability on t-distributions</div>
</blockquote><div class="calibre1">Using the t-Table</div>
<blockquote class="calibre9"><div class="calibre1">Finding probabilities with the t-table</div>
<div class="calibre1">Figuring percentiles for the t-distribution
</div>
<div class="calibre1">Picking out t*-values for confidence intervals</div>
</blockquote><div class="calibre1">Studying Behavior Using the t-Table
</div>
</blockquote><div class="calibre1">Chapter 11: Sampling Distributions and the Central Limit Theorem</div>
<blockquote class="calibre9"><div class="calibre1">Defining a Sampling Distribution</div>
<div class="calibre1">The Mean of a Sampling Distribution</div>
<div class="calibre1">Measuring Standard Error</div>
<blockquote class="calibre9"><div class="calibre1"><a

href="https://calibre-pdf-anchor.a#a206">Sample size and standard error</div>
<div class="calibre1">Population standard deviation and standard error</div>
</blockquote><div class="calibre1">Looking at the Shape of a Sampling Distribution</div>
<blockquote class="calibre9"><div class="calibre1">Case 1: The distribution of X is normal</div>
<div class="calibre1">Case 2: The distribution of X is not normal — enter the Central Limit Theorem</div>
</blockquote><div class="calibre1">Finding Probabilities for the Sample Mean</div>
<div class="calibre1">The Sampling Distribution of the Sample Proportion</div>
<div class="calibre1">Finding Probabilities for the Sample Proportion</div>
</blockquote></blockquote><div class="calibre1">Part IV: Guesstimating and Hypothesizing with Confidence</div><blockquote class="calibre9">
<div class="calibre1">Chapter 12: Leaving Room for a Margin of Error</div>
<blockquote class="calibre9"><div class="calibre1">Seeing the Importance of That Plus or Minus</div>
<div class="calibre1">Finding the Margin of Error: A General Formula</div>
<blockquote class="calibre9"><div class="calibre1">Measuring sample variability</div>
<div class="calibre1">Calculating margin of error for a sample proportion</div>
<div class="calibre1">Reporting results</div>
<div class="calibre1">Calculating margin of error for a sample mean</div>
<div class="calibre1"><a href="https://calibre-pdf-

anchor.a#a226">Being confident you’re right</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a227">Determining the Impact of Sample Size
</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a228">Sample size and margin
of error</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a229">Bigger isn’t always (that much) better!
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a230">Keeping margin of error in perspective</div>
</blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a233">Chapter 13: Confidence
Intervals: Making Your Best Guesstimate</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a234">Not All Estimates Are
Created Equal</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a235">Linking a Statistic to a Parameter</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a236">Getting with the Jargon</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a237">Interpreting Results with Confidence</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a238">Zooming In on Width</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a239">Choosing a Confidence Level</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a240">Factoring In the Sample Size</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a241">Counting On Population Variability</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a242">Calculating a Confidence Interval for a
Population Mean</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a243">Case 1: Population
standard deviation is known</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a244">Case 2: Population standard deviation is unknown
and/or n is small</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a245">Figuring Out What Sample Size You Need
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a246">Determining the Confidence Interval for One
Population Proportion</div>
<div class="calibre1"><a href="https://calibre-pdf-

anchor.a#a247">Creating a Confidence Interval for the
Difference of Two Means</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a248">Case 1: Population
standard deviations are known</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a249">Case 2: Population standard deviations are
unknown and/or sample sizes are small</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a250">Estimating the Difference of Two
Proportions</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a251">Spotting Misleading Confidence Intervals
</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a254">Chapter 14: Claims, Tests, and
Conclusions</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a255">Setting Up the
Hypotheses</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a256">Defining the null
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a257">What’s the alternative?</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a258">Gathering Good Evidence (Data)</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a259">Compiling the Evidence: The Test Statistic
</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a260">Gathering sample
statistics</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a261">Measuring variability using standard errors
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a262">Understanding standard scores</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a263">Calculating and interpreting the test
statistic</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a264">Weighing the Evidence and Making Decisions:
p-Values</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a265">Connecting test
statistics and p-values</div>
<div class="calibre1"><a href="https://calibre-pdf-

anchor.a#a266">Defining a p-value</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a267">Calculating a p-value</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a268">Making Conclusions</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a269">Setting boundaries for
rejecting Ho</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a270">Testing varicose veins</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a271">Assessing the Chance of a Wrong Decision
</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a272">Making a false alarm:
Type-1 errors</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a273">Missing out on a detection: Type-2 errors
</div>
</blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a276">Chapter 15: Commonly
Used Hypothesis Tests: Formulas and Examples</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a277">Testing One Population
Mean</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a278">Handling Small Samples and Unknown Standard
Deviations: The t-Test</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a279">Putting the t-test to
work</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a280">Relating t to Z</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a281">Handling negative t-values</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a282">Examining the not-equal-to alternative</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a283">Testing One Population Proportion</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a284">Comparing Two (Independent) Population
Averages</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a285">Testing for an Average Difference (The Paired t-
Test)</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a286">Comparing Two Population Proportions</div>
</blockquote></blockquote><div class="calibre1"><a

href="https://calibre-pdf-anchor.a#a289">Part V: Statistical Studies and the Hunt for a Meaningful Relationship</div>
<blockquote class="calibre9">
<div class="calibre1">Chapter 16: Polls, Polls, and More Polls
</div>
<blockquote class="calibre9"><div class="calibre1">Recognizing the Impact of Polls</div>
<blockquote class="calibre9"><div class="calibre1">Getting to the source</div>
<div class="calibre1">Surveying what’s hot</div>
<div class="calibre1">Impacting lives</div>
</blockquote><div class="calibre1">Behind the Scenes: The Ins and Outs of Surveys</div>
<blockquote class="calibre9"><div class="calibre1">Planning and designing a survey</div>
<div class="calibre1">Selecting the sample</div>
<div class="calibre1">Carrying out a survey</div>
<div class="calibre1">Interpreting results and finding problems
</div>
</blockquote></blockquote><div class="calibre1">Chapter 17: Experiments: Medical Breakthroughs or Misleading Results?
</div>
<blockquote class="calibre9"><div class="calibre1">Boiling Down the Basics of Studies</div>
<blockquote class="calibre9"><div class="calibre1">Looking at the lingo of studies</div>
<div class="calibre1">Observing observational studies</div>
<div class="calibre1">Examining experiments</div>
</blockquote><div class="calibre1">Designing a Good Experiment</div>
<blockquote class="calibre9"><div class="calibre1">Designing the experiment to make comparisons</div>

<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a311">Selecting the sample size</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a312">Choosing the subjects</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a313">Making random assignments</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a314">Controlling for confounding variables</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a315">Respecting ethical issues</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a316">Collecting good data</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a317">Analyzing the data properly</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a318">Making appropriate conclusions</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a319">Making Informed Decisions</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a322">Chapter 18: Looking for Links: Correlation
and Regression</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a323">Picturing a
Relationship with a Scatterplot</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a324">Making a
scatterplot</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a325">Interpreting a scatterplot</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a326">Quantifying Linear Relationships Using the
Correlation</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a327">Calculating the
correlation</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a328">Interpreting the correlation</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a329">Examining properties of the correlation
</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a330">Working with Linear Regression</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a331">Figuring out which
variable is X and which is Y</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a332">Checking the conditions</div>
<div class="calibre1"><a href="https://calibre-pdf-

anchor.a#a333">Calculating the regression line</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a334">Interpreting the regression line</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a335">Putting it all together with an example: The
regression line for the crickets</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a336">Making Proper Predictions</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a337">Explaining the Relationship: Correlation versus
Cause and Effect</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a340">Chapter 19: Two-Way Tables and
Independence</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a341">Organizing a Two-Way
Table</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a342">Setting up the
cells</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a343">Figuring the totals</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a344">Interpreting Two-Way Tables</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a345">Singling out variables
with marginal distributions</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a346">Examining all groups — a joint distribution
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a347">Comparing groups with conditional
distributions</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a348">Checking Independence and Describing
Dependence</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a349">Checking for
independence</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a350">Describing a dependent relationship</div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a351">Cautiously Interpreting Results</div>
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a352">Checking for
legitimate cause and effect</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a353">Projecting from sample to population</div>

<div class="calibre1">Making prudent predictions</div>
<div class="calibre1">Resisting the urge to jump to conclusions
</div>
</blockquote></blockquote></blockquote><div class="calibre1">Part VI: The Part of Tens</div><blockquote class="calibre9">
<div class="calibre1">Chapter 20: Ten Tips for the Statistically Savvy Sleuth</div>
<blockquote class="calibre9"><div class="calibre1">Pinpoint Misleading Graphs</div>
<blockquote class="calibre9"><div class="calibre1">Pie charts</div>
<div class="calibre1">Bar graphs</div>
<div class="calibre1">Time charts</div>
<div class="calibre1">Histograms</div>
</blockquote><div class="calibre1">Uncover Biased Data</div>
<div class="calibre1">Search for a Margin of Error</div>
<div class="calibre1">Identify Non-Random Samples</div>
<div class="calibre1">Sniff Out Missing Sample Sizes</div>
<div class="calibre1">Detect Misinterpreted Correlations</div>
<div class="calibre1">Reveal Confounding Variables</div>
<div class="calibre1">Inspect the Numbers</div>
<div class="calibre1">Report Selective Reporting</div>
<div class="calibre1">Expose the Anecdote</div>
</blockquote><div class="calibre1">Chapter 21: Ten Surefire Exam Score Boosters</div>
<blockquote class="calibre9"><div class="calibre1">Know What You Don’t Know, and then Do Something about It</div>
<div class="calibre1">Avoid “Yeah-Yeah” Traps</div>

```
<blockquote class="calibre9"><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a381">Yeah-yeah trap #1</a>
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a382">Yeah-yeah trap #2</a></div>
</blockquote><div class="calibre1"><a href="https://calibre-
pdf-anchor.a#a383">Make Friends with Formulas</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a384">Make an “If-Then-How” Chart</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a385">Figure Out What the Question Is Asking</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a386">Label What You’re Given</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a387">Draw a Picture</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a388">Make the Connection and Solve the Problem</a>
</div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a389">Do the Math — Twice</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a390">Analyze Your Answers</a></div>
</blockquote></blockquote><div class="calibre1"><a
href="https://calibre-pdf-anchor.a#a393">Appendix: Tables for
Reference</a></div>
<div class="calibre1"><a href="https://calibre-pdf-
anchor.a#a396">Cheat Sheet</a></div>
</div>  <div class="mbppagebreak" id="a4"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a3" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a4" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a5" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a7">
<div class="calibre1">
<p class="calibre10"><span class="calibre11"><span
class="bold"><span class="italic">Statistics For Dummies,<sup
class="calibre12">®</sup> 2nd Edition</span></span></span></p>
<blockquote class="calibre5"><p class="calibre10"><span
class="calibre11"><span class="bold">by Deborah J. Rumsey,
PhD</span></span></p></blockquote>
<div class="calibre13">
<img alt="WileyTitlePageLogo.eps" src="images/00002.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span class="bold">Statistics For Dummies,
<sup class="calibre4">®</sup> 2nd Edition</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Published by<br class="calibre1"/><span><span
class="calibre16"><span class="bold">Wiley Publishing, Inc.
</span></span></span><br class="calibre1"/>111 River St.<br
class="calibre1"/>Hoboken, NJ 07030-5774<br class="calibre1"/>
<a href="http://www.wiley.com">www.wiley.com</a></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Copyright © 2011 by Wiley Publishing, Inc.,
Indianapolis, Indiana</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Published simultaneously in Canada</span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">No part of this publication may be
reproduced, stored in a retrieval system, or transmitted in any
form or by any means, electronic, mechanical, photocopying,
recording, scanning, or otherwise, except as permitted under
Sections 107 or 108 of the 1976 United States Copyright Act,
without either the prior written permission of the Publisher,
or authorization through payment of the appropriate per-copy
fee to the Copyright Clearance Center, 222 Rosewood Drive,
```

<blockquote class="calibre9"><p class="calibre14">For technical support, please visit <a
href="http://www.wiley.com/techsupport">www.wiley.com/techsuppo
rt.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Wiley also publishes its books in a variety
of electronic formats. Some content that appears in print may
not be available in electronic books.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Library of Congress Control Number:
2011921775</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">ISBN: 978-0-470-91108-2</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Manufactured in the United States of
America</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">10 9 8 7 6 5 4 3 2 1</p></blockquote>
<div class="calibre13">
<blockquote class="calibre9"><p class="calibre14">
<img alt="WileycopyrightLogo.eps" src="images/00003.jpg"
class="calibre2"/></p></blockquote>
</div>
<p class="calibre6">About the Author</p>
<blockquote class="calibre9"><p class="calibre14">Deborah J. Rumsey, PhD, is a
Statistics Education Specialist and Auxiliary Professor in the
Department of Statistics at The Ohio State University.
Dr. Rumsey is a Fellow of the American Statistical Association.
She has won the Presidential Teaching Award from Kansas State
University and has been inducted into the Wall of Inspiration
at her high school alma mater, Burlington High School, in
Burlington, Wisconsin. She is also the author of Statistics II For
Dummies, Statistics Workbook For Dummies, Probability For
Dummies, and Statistics Essentials
For Dummies. She has published numerous
papers and given many professional presentations and workshops
on the subject of statistics education. She is the original
conference designer of the biennial United States Conference on
Teaching Statistics (USCOTS). Her passions include being with
her family, camping and bird watching, getting seat time on her
Kubota tractor, and cheering the Ohio State Buckeyes on to

their next national championship.</p></blockquote>
<p class="calibre6">Dedication</p>
<blockquote class="calibre9"><p class="calibre14">To my husband Eric: My sun rises and sets
with you. To my son Clint: I love you up to the moon and back.
</p></blockquote>
<p class="calibre6">Author’s Acknowledgments</p>
<blockquote class="calibre9"><p class="calibre14">My heartfelt thanks to Lindsay Lefevere and
Kathy Cox for the opportunity to write For Dummies
 books for Wiley; to my project editors Georgette
Beatty, Corbin Collins, and Tere Drenth for their unwavering
support and vision; to Marjorie Bond, Monmouth College, for
agreeing to be my technical editor (again!); to Paul
Stephenson, who also provided technical editing; and to Caitie
Copple and Janet Dunn for great copy editing.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Special thanks to Elizabeth Stasny, Joan
Garfield, Kythrie Silva, Kit Kilen, Peg Steigerwald, Mike
O’Leary, Tony Barkauskas, Ken Berk, and Jim Higgins for
inspiration and support along the way; and to my entire family
for their steadfast love and encouragement.</p>
</blockquote>
<p class="calibre6">Publisher’s Acknowledgments</p>
<blockquote class="calibre9"><p class="calibre14">We’re proud of this book; please send us your
comments through our online registration form located at <a
href="http://dummies.custhelp.com">http://dummies.custhelp.com<
/a>. For other comments, please contact our Customer Care
Department within the U.S. at 877-762-2974, outside the U.S. at
317-572-3993, or fax 317-572-4002.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Some of the people who helped bring this book
to market include the following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Acquisitions, Editorial, and
Media Development</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Project Editor: Corbin
Collins</p></blockquote>

```html
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="italic">(Previous Edition: Tere Drenth)</span></span>
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Senior Project Editor:</span></span></span>
Georgette Beatty</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Executive Editor:</span></span></span> Lindsay
Sandman</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Copy Editor:</span></span></span> Caitlin
Copple</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="italic">(Previous Edition: Janet S. Dunn, PhD)</span>
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Assistant Editor:</span></span></span> David
Lutton</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Technical Editors:</span></span></span> Marjorie
E. Bond, Paul L. Stephenson III</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Editorial Manager:</span></span></span> Michelle
Hacker</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Editorial Supervisor and Reprint Editor:</span>
</span></span> Carmen Krikorian</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Editorial Assistant:</span></span></span> Jennette
ElNaggar</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Cover Photos:</span></span></span> ©
iStockphoto.com/Norebbo</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">Cartoons:</span></span></span> Rich Tennant (<a
href="http://www.the5thwave.com">www.the5thwave.com</a>)</span>
```

</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Composition Services
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Project Coordinator: Sheree
Montgomery</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Layout and Graphics: Carrie
A. Cesavice, Corrie Socolovitch</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Proofreaders: Dwight Ramsey,
Shannon Ramsey</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Indexer: Christine
Karpeles</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Publishing and Editorial
for Consumer Dummies</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Diane Graves Steele, Vice
President and Publisher, Consumer Dummies</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Kristin Ferguson-Wagstaffe,
Product Development Director, Consumer Dummies</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Ensley Eikenburg, Associate
Publisher, Travel</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Kelly Regan, Editorial
Director, Travel</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Publishing for Technology
Dummies</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="bold">Andy Cummings, Vice President and Publisher, Dummies Technology/General User</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Composition Services</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Debbie Stailey, Director of Composition Services</p></blockquote>
</div> <div class="mbppagebreak" id="a6"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

    </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a23">
<div class="calibre1">
<div class="calibre1">
<p id="a8" class="calibre10"><span class="calibre11"><span
class="bold"><span>Introduction</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Y</span><span>ou get hit with an
incredible amount of statistical information on a daily basis.
You know what I’m talking about: charts, graphs, tables, and
headlines that talk about the results of the latest poll,
survey, experiment, or other scientific study. The purpose of
this book is to develop and sharpen your skills in sorting
through, analyzing, and evaluating all that info, and to do so
in a clear, fun, and pain-free way. You also gain the ability
to decipher and make important decisions about statistical
results (for example, the results of the latest medical
studies), while being ever aware of the ways that people can
mislead you with statistics. And you see how to do it right
when it’s your turn to design the study, collect the data,
crunch the numbers, and/or draw the conclusions.</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>This book is also designed to help
those of you out there who are taking an introductory
statistics class and can use some back-up. You’ll gain a
working knowledge of the big ideas of statistics and gather a
boatload of tools and tricks of the trade that’ll help you get
ahead of the curve when you take your exams.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>This book is chock-full of real
examples from real sources that are relevant to your everyday
life — from the latest medical breakthroughs, crime studies,
and population trends to the latest U.S. government reports. I
even address a survey on the worst cars of the millennium! By
reading this book, you’ll understand how to collect, display,
and analyze data correctly and effectively, and you’ll be ready
to critically examine and make informed decisions about the
```

latest polls, surveys, experiments, and reports that bombard
you every day. You even find out how to use crickets to gauge
temperature!</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You also get to enjoy poking a little
fun at statisticians (who take themselves too seriously at
times). After all, with the right skills and knowledge, you
don’t have to be a statistician to understand introductory
statistics.</p></blockquote>
<p id="a9" class="calibre6">About This Book</p>
<blockquote class="calibre9"><p class="calibre14">This book departs from traditional
statistics texts, references, supplemental books, and study
guides in the following ways:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It includes practical and intuitive explanations of
statistical concepts, ideas, techniques, formulas, and
calculations found in an introductory statistics course.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It shows you clear and concise step-by-step procedures
that explain how you can intuitively work through statistics
problems.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It includes interesting real-world examples relating to
your everyday life and workplace.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It gives you upfront and honest answers to your questions
like, “What does this really mean?” and “When and how will I
ever use this?”</p></blockquote><div
class="calibre19"> </div>
<p id="a10" class="calibre6">Conventions Used in This Book
</p>
<blockquote class="calibre9"><p class="calibre14">You should be aware of three
conventions as you make your way through this book:

</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Definition of
sample size (n):
 When I refer to the size of a sample, I mean the final
number of individuals who participated in and provided
information for the study. In other words, n
 stands for the size of the final data set.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Dual-use of
the word
statistics:

In some situations, I refer to
statistics as a subject of study or as a field of research, so
the word is a singular noun. For example, “Statistics is really
quite an interesting subject.” In other situations, I refer to
statistics as the plural of statistic,
 in a numerical sense. For example, “The most
common statistics are the mean and the standard
deviation.”</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Use of the
word data:
 You’re probably unaware of the debate raging
amongst statisticians about whether the word data
 should be singular (“data is . . .”) or plural
(“data are . . .”). It got so bad that recently one group of
statisticians had to develop two different versions of a
statistics T-shirt: “Messy Data Happens” and “Messy Data
Happen.” At the risk of offending some of my colleagues, I go
with the plural version of the word <span><span

```
class="calibre16"><span class="italic">data</span></span>
</span><span> in this book.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Use of the
term </span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">standard deviation:</span>
</span></span></span><span> When I use the term </span><span>
<span class="calibre16"><span class="italic">standard
deviation,</span></span></span><span> I mean </span><span><span
class="calibre16"><span class="italic">s,</span></span></span>
<span> the sample standard deviation. (When I refer to the
population standard deviation, I let you know.)</span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Here are a few other basic conventions
to help you navigate this book:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>I use </span><span><span class="calibre16"><span
class="italic">italics</span></span></span><span> to let you
know a new statistical term is appearing on the scene.</span>
</span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If you see a </span><span><span class="calibre16"><span
class="bold">boldfaced</span></span></span><span> term or
phrase in a bulleted list, it’s been designated as a keyword or
key phrase.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Addresses for Web sites appear in </span><code
class="calibre20">monofont</code><span>.</span></span></p>
</blockquote><div class="calibre19"> </div>
<p id="a11" class="calibre6"><span class="calibre17"><span
class="bold"><span>What You’re Not to Read</span></span></span>
</p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I like to think that you won’t skip
anything in this book, but I also know you’re a busy person. So
to save time, feel free to skip anything marked with the
Technical Stuff icon as well as text in sidebars (the shaded
```

gray boxes that appear throughout the book). These items
feature information that’s interesting but not crucial to your
basic knowledge of statistics.</p></blockquote>
<p id="a12" class="calibre6">Foolish Assumptions</p>
<blockquote class="calibre9"><p class="calibre14">I don’t assume that you’ve had any
previous experience with statistics, other than the fact that
you’re a member of the general public who gets bombarded every
day with statistics in the form of numbers, percents, charts,
graphs, “statistically significant” results, “scientific”
studies, polls, surveys, experiments, and so on.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What I do assume is that you can do
some of the basic mathematical operations and understand some
of the basic notation used in algebra, such as the variables
x and
y,
 summation signs, taking the square root, squaring
a number, and so on. If you need to brush up on your algebra
skills, check out Algebra I For Dummies,
2nd Edition, by Mary Jane Sterling (Wiley).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">I don’t want to mislead you: You do
encounter formulas in this book, because statistics does
involve a bit of number crunching. But don’t let that worry
you. I take you slowly and carefully through each step of any
calculations you need to do. I also provide examples for you to
work along with this book, so that you can become familiar and
comfortable with the calculations and make them your own.
</p></blockquote>
<p id="a13" class="calibre6">How This Book Is Organized
</p>
<blockquote class="calibre9"><p class="calibre14">This book is organized into five parts
that explore the major areas of introductory statistics, along
with a final part that offers some quick top-ten nuggets for
your information and enjoyment. Each part contains chapters
that break down each major area of statistics into
understandable pieces.</p></blockquote>
<blockquote class="calibre5"><p id="a14" class="calibre6">Part I: Vital
Statistics about Statistics</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre14">This part helps you become aware of the quantity and quality of statistics you encounter in your workplace and your everyday life. You find out that a great deal of that statistical information is incorrect, either by accident or by design. You take a first step toward becoming statistically savvy by recognizing some of the tools of the trade, developing an overview of statistics as a process for getting and interpreting information, and getting up to speed on some statistical jargon.</p></blockquote>
<blockquote class="calibre5"><p id="a15" class="calibre6">Part II: Number-Crunching Basics</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This part helps you become more familiar and comfortable with making, interpreting, and evaluating data displays (otherwise known as charts, graphs, and so on) for different types of data. You also find out how to summarize and explore data by calculating and combining some commonly used statistics as well as some statistics you may not know about yet.</p></blockquote>
<blockquote class="calibre5"><p id="a16" class="calibre6">Part III: Distributions and the Central Limit Theorem
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this part, you get into all the details of the three most common statistical distributions: the binomial distribution, the normal (and standard normal, also known as Z-distribution), and the t-distribution. You discover the characteristics of each distribution and how to find and interpret probabilities, percentiles, means, and standard deviations. You also find measures of relative standing (like percentiles).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Finally, you discover how statisticians measure variability from sample to sample and why a measure of precision in your sample results is so important. And you get the lowdown on what some statisticians describe as the “Crowning Jewel of all Statistics”: the Central Limit Theorem (CLT). I don’t use quite this level of flourishing language to describe the CLT; I just tell my students it’s an MDR (“Mighty Deep Result”; coined by my PhD adviser). As for how my students describe their feelings about the CLT, I’ll leave that to your

imagination.</p></blockquote>
<blockquote class="calibre5"><p id="a17" class="calibre6">Part IV:
Guesstimating and Hypothesizing with Confidence
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This part focuses on the two methods
for taking the results from a sample and generalizing them to
make conclusions about an entire population. (Statisticians
call this process statistical inference.
)
These two methods are confidence
intervals and hypothesis tests.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this part, you use confidence
intervals to come up with good estimates for one or two
population means or proportions, or for the difference between
them (for example, the average number of hours teenagers spend
watching TV per week or the percentage of men versus women in
the United States who take arthritis medicine every day). You
get the nitty-gritty on how confidence intervals are formed,
interpreted, and evaluated for correctness and credibility. You
explore the factors that influence the width of a confidence
interval (such as sample size) and work through formulas, step-
by-step calculations, and examples for the most commonly used
confidence intervals.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The hypothesis tests in this part show
you how to use your data to test someone’s claim about one or
two population means or proportions, or the difference between
them. (For example, a company claims their packages are
delivered in two days on average — is this true?) You discover
how researchers (should) go about forming and testing
hypotheses and how you can evaluate their results for accuracy
and credibility. You also get detailed step-by-step directions
and examples for carrying out and interpreting the results of
the most commonly used hypothesis tests.</p>
</blockquote>
<blockquote class="calibre5"><p id="a18" class="calibre6">Part V: Statistical
Studies and the Hunt for a Meaningful Relationship
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This part gives an overview of surveys,
experiments, and observational studies. You find out what these
studies do, how they are conducted, what their limitations are,

and how to evaluate them to determine whether you should
believe the results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You also get all the details on how to
examine pairs of numerical variables and categorical variables
to look for relationships; this is the object of a great number
of studies. For pairs of categorical variables, you create two-
way tables and find joint, conditional, and marginal
probabilities and distributions. You check for independence,
and if a dependent relationship is found, you describe the
nature of the relationship using probabilities. For numerical
variables you create scatterplots, find and interpret
correlation, perform regression analyses, study the fit of the
regression line and the impact of outliers, describe the
relationship using the slope, and use the line to make
predictions. All in a day’s work!</p>
</blockquote>
<blockquote class="calibre5"><p id="a19" class="calibre6">Part VI: The Part of
Tens</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This quick and easy part shares ten
ways to be a statistically savvy sleuth and root out suspicious
studies and results, as well as ten surefire ways to boost your
statistics exam score.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Some statistical calculations involve
the use of statistical tables, and I provide quick and easy
access to all the tables you need for this book in the
appendix. These tables are the Z
-table (for the standard normal, also called the
Z
-distribution), the t
-table (for the
t-
distribution), and the binomial table (for — you guessed it —
the binomial distribution). Instructions and examples for using
these three tables are provided in their corresponding sections
of this book.</p></blockquote>
<p id="a20" class="calibre6">Icons Used in This Book
</p>
<blockquote class="calibre9"><p class="calibre14">Icons are used in this book to draw
your attention to certain features that occur on a regular
basis. Here’s what they mean:</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> This icon refers to helpful hints,
ideas, or shortcuts that you can use to save time. It also
highlights alternative ways to think about a particular
concept.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> This icon is
reserved for particular ideas that I hope you’ll remember long
after you read this book.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> This icon
refers to specific ways that researchers or the media can
mislead you with statistics and tells you what you can do about
it. It also points out potential problems and cautions to keep
an eye out for on exams.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> This icon is a
sure bet if you have a special interest in understanding the
more technical aspects of statistical issues. You can skip this
icon if you don’t want to get into the gory details.
</p></blockquote>
<p id="a21" class="calibre6">Where to Go from Here
</p>
<blockquote class="calibre9"><p class="calibre14">This book is written in such a way that
you can start anywhere and still be able to understand what’s
going on. So you can take a peek at the table of contents or
the index, look up the information that interests you, and flip
to the page listed. However if you have a specific topic in
mind and are eager to dive into it, here are some directions:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
To work on finding and interpreting graphs, charts, means
or medians, and the like, head to Part II.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
To find info on the normal, Z
-, <span><span class="calibre16"><span

```
class="italic">t</span></span></span><span>-, or binomial
distributions or the Central Limit Theorem, see Part III.
</span></span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>To focus on confidence intervals and hypothesis tests of
all shapes and sizes, flip to Part IV.</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>To delve into surveys, experiments, regression, and two-
way tables, see Part V.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Or if you aren’t sure where you want to
start, you may just go with Chapter 1 for the big picture and
then plow your way through the rest of the book. Happy reading!
</span></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a22"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a20" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a18" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a19" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a23"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a17" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a14" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a15" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a12" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a13" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a10" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a11" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a16"
```

```
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a8" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a9" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a22" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a21" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#5KzRDSqba4c6hbtT4ckd63"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a>
</div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a26">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part I</span></span>
</span></span></p>
<p id="a24" class="calibre10"><span class="calibre11"><span
class="bold"><span>Vital Statistics about Statistics</span>
</span></span></p>
<div class="calibre13">
<img alt="9780470911082 pp0101.eps" src="images/00009.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">When you turn on the TV or open a newspaper,
you’re bombarded with numbers, charts, graphs, and statistical
results. From today’s poll to the latest major medical
breakthroughs, the numbers just keep coming. Yet much of the
statistical information you’re asked to consume is actually
wrong — by accident or even by design. How is a person to know
what to believe? By doing a lot of good detective work.</span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">This part helps awaken the statistical sleuth
that lies within you by exploring how statistics affect your
everyday life and your job, how bad much of the information out
there really is, and what you can do about it. This part also
helps you get up to speed with some useful statistical jargon.
</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a25"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">
```


      </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a48">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 1</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a27" class="calibre10"><span class="calibre11"><span
class="bold"><span>Statistics in a Nutshell</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Finding out what the process of statistics is all
about</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Gaining success with statistics in your everyday life,
your career, and in the classroom</span></span></p>
</blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>he world today is
overflowing with data to the point where anyone (even me!) can
be overwhelmed. I wouldn’t blame you if you were cynical right
now about statistics you read about in the media — I am too at
times. The good news is that while a great deal of misleading
and incorrect information is lying out there waiting for you, a
lot of great stuff is also being produced; for example, many
studies and techniques involving data are helping improve the
quality of our lives. Your job is to be able to sort out the
good from the bad and be confident in your ability to do that.
Through a strong understanding of statistics and statistical
procedures, you gain power and confidence with numbers in your
everyday life, in your job, and in the classroom. That’s what
```

this book is all about.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, I give you an overview
of the role statistics plays in today’s data-packed society and
what you can do to not only survive but thrive. You get a much
broader view of statistics as a partner in the scientific
method — designing effective studies, collecting good data,
organizing and analyzing the information, interpreting the
results, and making appropriate conclusions. (And you thought
statistics was just number-crunching!)</p>
</blockquote>
<p id="a28" class="calibre6">Thriving in a Statistical World
</p>
<blockquote class="calibre9"><p class="calibre14">It’s hard to get a handle on the flood
of statistics that affect your daily life in large and small
ways. It begins the moment you wake up in the morning and check
the news and listen to the meteorologist give you her
predictions for the weather based on her statistical analyses
of past data and present weather conditions. You pore over
nutritional information on the side of your cereal box while
you eat breakfast. At work you pull numbers from charts and
tables, enter data into spreadsheets, run diagnostics, take
measurements, perform calculations, estimate expenses, make
decisions using statistical baselines, and order inventory
based on past sales data.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">At lunch you go to the No. 1 restaurant
based on a survey of 500 people. You eat food that was priced
based on marketing data. You go to your doctor’s appointment
where they take your blood pressure, temperature, weight, and
do a blood test; after all the information is collected, you
get a report showing your numbers and how you compare to the
statistical norms.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You head home in your car that’s been
serviced by a computer running statistical diagnostics. When
you get home, you turn on the news and hear the latest crime
statistics, see how the stock market performed, and discover
how many people visited the zoo last week.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">At night, you brush your teeth with
toothpaste that’s been statistically proven to fight cavities,
read a few pages of your
New York Times
Best-Seller (based on statistical sales estimates), and go to

sleep — only to start it all over again the next morning. But how can you be sure that all those statistics you encounter and depend on each day are correct? In Chapter 2, I discuss in more depth a few examples of how statistics is involved in our lives and workplaces, what its impact is, and how you can raise your awareness of it.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Some statistics are vague, inappropriate, or just plain wrong. You need to become more aware of the statistics you encounter each day and train your mind to stop and say “wait a minute!”, sift through the information, ask questions, and raise red flags when something’s not quite right. In Chapter 3, you see ways in which you can be misled by bad statistics and develop skills to think critically and identify problems before automatically believing results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Like any other field, statistics has its own set of jargon, and I outline and explain some of the most commonly used statistical terms in Chapter 4. Knowing the language increases your ability to understand and communicate statistics at a higher level without being intimidated. It raises your credibility when you use precise terms to describe what’s wrong with a statistical result (and why). And your presentations involving statistical tables, graphs, charts, and analyses will be informational and effective. (Heck, if nothing else, you need the jargon because I use it throughout this book; don’t worry though, I always review it.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the next sections, you see how statistics is involved in each phase of the scientific method.
</p></blockquote>
<p id="a29" class="calibre6">Designing Appropriate Studies
</p>
<blockquote class="calibre9"><p class="calibre14">Everyone’s asking questions, from drug companies to biologists; from marketing analysts to the U.S. government. And ultimately, everyone will use statistics to help them answer their questions. In particular, many medical and psychological studies are done because someone wants to know the answer to a question. For example,</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18">

Will this vaccine be effective in preventing the flu?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
What do Americans think about the state of the economy?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Does an increase in the use of social networking Web
sites cause depression in teenagers?</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">The first step after a research
question has been formed is to design an effective study to
collect data that will help answer that question. This step
amounts to figuring out what process you’ll use to get the data
you need. In this section, I give an overview of the two major
types of studies — surveys and experiments — and explore why
it’s so important to evaluate how a study was designed before
you believe the results.</p></blockquote>
<blockquote class="calibre5"><p id="a30" class="calibre6">Surveys
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">An observational
study is one in which data is
collected on individuals in a way that doesn’t affect them. The
most common observational study is the survey.
Surveys
 are questionnaires that are presented to
individuals who have been selected from a population of
interest. Surveys take many different forms: paper surveys sent
through the mail, questionnaires on Web sites, call-in polls
conducted by TV networks, phone surveys, and so on.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If conducted
properly, surveys can be very useful tools for getting
information. However, if not conducted properly, surveys can
result in bogus information. Some problems include improper
wording of questions, which can be misleading, lack of response
by people who were selected to participate, or failure to
include an entire group of the population. These potential
problems mean a survey has to be well thought out before it’s

given.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Many
researchers spend a great deal of time and money to do good
surveys, and you’ll know (by the criteria I discuss in Chapter
16) that you can trust them. However, as you are besieged with
so many different types of surveys found in the media, in the
workplace, and in many of your classes, you need to be able to
quickly examine and critique how a survey was designed and
conducted and be able to point out specific problems in a well-
informed way. The tools you need for sorting through surveys
are found in Chapter 16.</p></blockquote>
<blockquote class="calibre5"><p id="a31" class="calibre6">Experiments
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">An experiment
 imposes one or more treatments on the
participants in such a way that clear comparisons can be made.
After the treatments are applied, the responses are recorded.
For example, to study the effect of drug dosage on blood
pressure, one group may take 10 mg of the drug, and another
group may take 20 mg. Typically, a control group is also
involved, in which subjects each receive a fake treatment (a
sugar pill, for example), or a standard, nonexperimental
treatment (like the existing drugs given to AIDS patients.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Good and
credible experiments are designed to minimize bias, collect
lots of good data, and make appropriate comparisons (treatment
group versus control group). Some potential problems that occur
with experiments include researchers and/or subjects who know
which treatment they got, factors not controlled for in the
study that affect the outcome (such as weight of the subject
when studying drug dosage), or lack of a control group (leaving
no baseline to compare the results with).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">But when designed correctly, an
experiment can help a researcher establish a cause-and-effect
relationship if the difference in responses between the
treatment group and the control group is statistically
significant (unlikely to have occurred just by chance).
</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Experiments are
credited with helping to create and test drugs, determining
best practices for making and preparing foods, and evaluating
whether a new treatment can cure a disease, or at least reduce
its impact. Our quality of life has certainly been improved
through the use of well-designed experiments. However, not all
experiments are well-designed, and your ability to determine
which results are credible and which results are incredible
(pun intended) is critical, especially when the findings are
very important to you. All the info you need to know about
experiments and how to evaluate them is found in Chapter 17.
</p></blockquote>
<p id="a32" class="calibre6">Collecting Quality Data
</p>
<blockquote class="calibre9"><p class="calibre14">After a study has been designed, be it
a survey or an experiment, the individuals who will participate
have to be selected, and a process must be in place to collect
the data. This phase of the process is critical to producing
credible data in the end, and this section hits the highlights.
</p></blockquote>
<blockquote class="calibre5"><p id="a33" class="calibre6">Selecting a good
sample</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Statisticians
have a saying, “Garbage in equals garbage out.” If you select
your subjects (the
individuals who will participate in your study) in a way that
is biased — that is,
favoring certain individuals or groups of individuals — then
your results will also be biased. It’s that simple.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose Bob wants to know the opinions
of people in your city regarding a proposed casino. Bob goes to
the mall with his clipboard and asks people who walk by to give
their opinions. What’s wrong with that? Well, Bob is only going
to get the opinions of a) people who shop at that mall; b) on
that particular day; c) at that particular time; d) and who
take the time to respond.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Those circumstances are too restrictive — those folks don’t represent a cross section of the city. Similarly, Bob could put up a Web site survey and ask people to use it to vote. However, only people who know about the site, have Internet access, and want to respond will give him data, and typically only those with strong opinions will go to such trouble. In the end, all Bob has is a bunch of biased data on individuals that don’t represent the city at all.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To minimize bias in a survey, the key word is random.
 You need to select your sample of individuals randomly — that is, with some type of “draw names out of a hat” process. Scientists use a variety of methods to select individuals at random, and you see how they do it in Chapter 16.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Note that in designing an experiment, collecting a random sample of people and asking them to participate often isn’t ethical because experiments impose a treatment on the subjects. What you do is send out requests for volunteers to come to you. Then you make sure the volunteers you select from the group represent the population of interest and that the data is well collected on those individuals so the results can be projected to a larger group. You see how that’s done in Chapter 17.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After going through Chapters 16 and 17, you’ll know how to dig down and analyze others’ methods for selecting samples and even be able to design a plan you can use to select a sample. In the end, you’ll know when to say “Garbage in equals garbage out.”</p></blockquote>
<blockquote class="calibre5"><p id="a34" class="calibre6">Avoiding bias in your data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Bias is the systematic favoritism of certain individuals or certain responses. Bias is the nemesis of statisticians, and they do everything they can to minimize it. Want an example of bias? Say you’re conducting a phone survey on job satisfaction of Americans; if you call people at home during the day between 9

a.m. and 5 p.m., you miss out on everyone who works during the day. Maybe day workers are more satisfied than night workers.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You have to watch for bias when collecting survey data. For instance: Some surveys are too long — what if someone stops answering questions halfway through? Or what if they give you misinformation and tell you they make $100,000 a year instead of $45,000? What if they give you answers that aren’t on your list of possible answers? A host of problems can occur when collecting survey data, and you need to be able to pinpoint those problems.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Experiments are sometimes even more challenging when it comes to bias and collecting data. Suppose you want to test blood pressure; what if the instrument you’re using breaks during the experiment? What if someone quits the experiment halfway through? What if something happens during the experiment to distract the subjects or the researchers? Or they can’t find a vein when they have to do a blood test exactly one hour after a dose of a drug is given? These problems are just some examples of what can go wrong in data collection for experiments, and you have to be ready to look for and find these problems.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After you go through Chapter 16 (on samples and surveys) and Chapter 17 (on experiments), you’ll be able to select samples and collect data in an unbiased way, being sensitive to little things that can really influence the results. And you’ll have the ability to evaluate the credibility of statistical results and to be heard, because you’ll know what you’re talking about.</p>
</blockquote>
<p id="a35" class="calibre6">Creating Effective Summaries
</p>
<blockquote class="calibre9"><p class="calibre14">After good data have been collected, the next step is to summarize them to get a handle on the big picture. Statisticians describe data in two major ways: with numbers (called descriptive statistics
) and with pictures (that is, charts and graphs).
</p></blockquote>
<blockquote class="calibre5"><p id="a36" class="calibre6"><span

```
class="calibre21"><span class="bold"><span>Descriptive
statistics</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span>
</span><span><span class="calibre16"><span
class="italic">Descriptive statistics</span></span></span>
<span> are numbers that describe a data set in terms of its
important features:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If the data are </span><span><span class="calibre16">
<span class="italic">categorical</span></span></span><span>
(where individuals are placed into groups, such as gender or
political affiliation), they are typically summarized using the
number of individuals in each group (called the </span><span>
<span class="calibre16"><span class="italic">frequency</span>
</span></span><span>) or the percentage of individuals in each
group (called the </span><span><span class="calibre16"><span
class="italic">relative frequency</span></span></span><span>).
</span></span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="italic">Numerical
data</span></span></span><span> represent measurements or
counts, where the actual numbers have meaning (such as height
and weight). With numerical data, more features can be
summarized besides the number or percentage in each group. Some
of these features include</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"><span
class="calibre15"><span> • Measures of center (in other words,
where is the “middle” of the data?)</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"><span
class="calibre15"><span> • Measures of spread (how diverse or
how concentrated are the data around the center?)</span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"><span
class="calibre15"><span> • If appropriate, numbers that measure
the relationship between two variables (such as height and
weight)</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Some
```

descriptive statistics are more appropriate than others in
certain situations; for example, the average isn’t always the
best measure of the center of a data set; the median is often a
better choice. And the standard deviation isn’t the only
measure of variability on the block; the interquartile range
has excellent qualities too. You need to be able to discern,
interpret, and evaluate the types of descriptive statistics
being presented to you on a daily basis and to know when a more
appropriate statistic is in order.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The descriptive statistics you see most
often are calculated, interpreted, compared, and evaluated in
Chapter 5. These commonly used descriptive statistics include
frequencies and relative frequencies (counts and percents) for
categorical data; and the mean, median, standard deviation,
percentiles, and their combinations for numerical data.
</p></blockquote>
<blockquote class="calibre5"><p id="a37" class="calibre6">Charts and
graphs</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Data is summarized in a visual way
using charts and/or graphs. These are displays that are
organized to give you a big picture of the data in a flash
and/or to zoom in on a particular result that was found. In
this world of quick information and mini-sound bites, graphs
and charts are commonplace. Most graphs and charts make their
points clearly, effectively, and fairly; however, they can
leave room for too much poetic license, and as a result, can
expose you to a high number of misleading and incorrect graphs
and charts.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> In Chapters 6
and 7, I cover the major types of graphs and charts used to
summarize both categorical and numerical data (see the
preceding section for more about these types of data). You see
how to make them, what their purposes are, and how to interpret
the results. I also show you lots of ways that graphs and
charts can be made to be misleading and how you can quickly
spot the problems. It’s a matter of being able to say “Wait a
minute here! That’s not right!” and knowing why not. Here are
some highlights:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Some of the basic graphs used for categorical data

include pie charts and bar graphs, which break down variables, such as gender or which applications are used on teens’ cellphones. A bar graph, for example, may display opinions on an issue using five bars labeled in order from “Strongly Disagree” up through “Strongly Agree.” Chapter 6 gives you all the important info on making, interpreting, and, most importantly, evaluating these charts and graphs for fairness. You may be surprised to see how much can go wrong with a simple bar chart.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
For numerical data such as height, weight, time, or amount, a different type of graph is needed. Graphs called histograms and boxplots are used to summarize numerical data, and they can be very informative, providing excellent on-the-spot information about a data set. But of course they also can be misleading, either by chance or even by design. (See Chapter 7 for the scoop.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> You’re going to run across charts and graphs every day — you can open a newspaper and probably find several graphs without even looking hard. Having a statistician’s magnifying glass to help you interpret the information is critical so that you can spot misleading graphs before you draw the wrong conclusions and possibly act on them. All the tools you need are ready for you in Chapter 6 (for categorical data) and Chapter 7 (for numerical data).</p></blockquote>
<p id="a38" class="calibre6">Determining Distributions</p>
<blockquote class="calibre9"><p class="calibre14">A variable is a characteristic that’s being counted, measured, or categorized. Examples include gender, age, height, weight, or number of pets you own. A distribution is a listing of the possible values of a variable (or intervals of values), and how often (or at what density) they occur. For example, the distribution of gender at birth in the United States has been estimated at 52.4% male and 47.6% female.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Different types
of distributions exist for different variables. The following
three distributions are the most commonly occurring
distributions in an introductory statistics course, and they
have many applications in the real world:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If a variable is counting the number of successes in a
certain number of trials (such as the number of people who got
well by taking a certain drug), it has a binomial
 distribution.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the variable takes on values that occur according to a
“bell-shaped curve,” such as national achievement test scores,
then that variable has a
normal
distribution.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the variable is based on sample averages and you have
limited data, such as in a test of only ten subjects to see if
a weight-loss program works, the t
-distribution may be in order.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">When it comes to distributions, you
need to know how to decide which distribution a particular
variable has, how to find probabilities for it, and how to
figure out what the long-term average and standard deviation of
the outcomes would be. To get you squared away on these issues,
I’ve got three chapters for you, one dedicated to each
distribution: Chapter 8 is all about the binomial, Chapter 9
handles the normal, and Chapter 10 focuses on the
t
-distribution.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> For those of you taking an

introductory statistics course (or any statistics course, for that matter), you know that one of the most difficult topics to understand is sampling distributions and the Central Limit Theorem (these two things go hand in hand). Chapter 11 walks you through these topics step by step so you understand what a sampling distribution is, what it’s used for, and how it provides the foundation for data analyses like hypothesis tests and confidence intervals (see the next section for more about analyzing data). When you understand the Central Limit Theorem, it actually helps you solve difficult problems more easily, and all the keys to this information are there for you in Chapter 11.</p></blockquote>
<p id="a39" class="calibre6">Performing Proper Analyses</p>
<blockquote class="calibre9"><p class="calibre14">After the data have been collected and described using numbers and pictures, then comes the fun part: navigating through that black box called the statistical analysis. If the study has been designed properly, the original questions can be answered using the appropriate analysis — the operative word here being appropriate.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Many types of analyses exist, and choosing the right analysis for the right situation is critical, as is interpreting results properly, being knowledgeable of the limitations, and being able to evaluate others’ choice of analyses and the conclusions they make with them.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this book, you get all the information and tools you need to analyze data using the most common methods in introductory statistics: confidence intervals, hypothesis tests, correlation and regression, and the analysis of two-way tables. This section gives you a basic overview of those methods.</p></blockquote>
<blockquote class="calibre5"><p id="a40" class="calibre6">Margin of error and confidence intervals</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You often see statistics that try to estimate numbers pertaining to an entire population; in fact, you see them almost every day in the form of survey results.

The media tells you what the average gas price is in the U.S., how Americans feel about the job the president is doing, or how many hours people spend on the Internet each week.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But no one can give you a single-number result and claim it’s an accurate estimate of the entire population unless he collected data on every single member of the population. For example, you may hear that 60 percent of the American people support the president’s approach to healthcare, but you know they didn’t ask you, so how could they have asked everybody? And since they didn’t ask everybody, you know that a one-number answer isn’t going to cut it.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What’s really happening is that data is collected on a sample from the population (for example, the Gallup Organization calls 2,500 people at random), the results from that sample are analyzed, and conclusions are made regarding the entire population (for example, all Americans) based on those sample results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The bottom line is, sample results vary from sample to sample, and this amount of variability needs to be reported (but it often isn’t). The statistic used to measure and report the level of precision in someone’s sample results is called the margin of error.
 In this context, the word
error
 doesn’t mean a mistake was made; it just means that because you didn’t sample the entire population, a gap will exist between your results and the actual value you are trying to estimate for the population.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, someone finds that 60% of the 1,200 people surveyed support the president’s approach to healthcare and reports the results with a margin of error of plus or minus 2%. This final result, in which you present your findings as a range of likely values between 58% and 62%, is called a confidence interval.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Everyone is

exposed to results including a margin of error and confidence intervals, and with today’s data explosion, many people are also using them in the workplace. Be sure you know what factors affect margin of error (like sample size) and what the makings of a good confidence interval are and how to spot them. You should also be able to find your own confidence intervals when you need to.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In Chapter 12, you find out everything you need to know about the margin of error: All the components of it, what it does and doesn’t measure, and how to calculate it for a number of situations. Chapter 13 takes you step by step through the formulas, calculations, and interpretations of confidence intervals for a population mean, population proportion, and the difference between two means and proportions.</p></blockquote>
<blockquote class="calibre5"><p id="a41" class="calibre6">Hypothesis tests</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One main staple of research studies is called hypothesis testing. A hypothesis test
 is a technique for using data to validate or invalidate a claim about a population. For example, a politician may claim that 80% of the people in her state agree with her — is that really true? Or, a company may claim that they deliver pizzas in 30 minutes or less; is that really true? Medical researchers use hypothesis tests all the time to test whether or not a certain drug is effective, to compare a new drug to an existing drug in terms of its side effects, or to see which weight-loss program is most effective with a certain group of people.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The elements about a population that are most often tested are
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
The population mean (Is the average delivery time of 30 minutes really true?)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The population proportion (Is it true that 80% of the

voters support this candidate, or is it less than that?)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The difference in two population means or proportions (Is
it true that the average weight loss on this new program is 10
pounds more than the most popular program? Or, is it true that
this drug decreases blood pressure by 10% more than the current
drug?)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Hypothesis
tests are used in a host of areas that affect your everyday
life, such as medical studies, advertisements, polling data,
and virtually anywhere that comparisons are made based on
averages or proportions. And in the workplace, hypothesis tests
are used heavily in areas like marketing, where you want to
determine whether a certain type of ad is effective or whether
a certain group of individuals buys more or less of your
product now compared to last year.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Often you only hear the conclusions of
hypothesis tests (for example, this drug is significantly more
effective and has fewer side effects than the drug you are
using now); but you don’t see the methods used to come to these
conclusions. Chapter 14 goes through all the details and
underpinnings of hypothesis tests so you can conduct and
critique them with confidence. Chapter 15 cuts right to the
chase of providing step-by-step instructions for setting up and
carrying out hypothesis tests for a host of specific situations
(one population mean, one population proportion, the difference
of two population means, and so on).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">After reading Chapters 14 and 15,
you’ll be much more empowered when you need to know things like
which group you should be marketing a product to; which brand
of tires will last the longest; whether a certain weight-loss
program is effective; and bigger questions like which surgical
procedure you should opt for.</p></blockquote>
<blockquote class="calibre5"><p id="a42" class="calibre6">Correlation,
regression, and two-way tables</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">One of the most common goals of
research is to find links between variables. For example,
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Which lifestyle behaviors increase or decrease the risk
of cancer?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
What side effects are associated with this new drug?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Can I lower my cholesterol by taking this new herbal
supplement?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Does spending a large amount of time on the Internet
cause a person to gain weight?</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Finding links between variables is what
helps the medical world design better drugs and treatments,
provides marketers with info on who is more likely to buy their
products, and gives politicians information on which to build
arguments for and against certain policies.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> In the mega-
business of looking for relationships between variables, you
find an incredible number of statistical results — but can you
tell what’s correct and what’s not? Many important decisions
are made based on these studies, and it’s important to know
what standards need to be met in order to deem the results
credible, especially when a cause-and-effect relationship is
being reported.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Chapter 18 breaks down all the details
and nuances of plotting data from two numerical variables (such
as dosage level and blood pressure), finding and interpreting
<span><span class="calibre16"><span

```
class="italic">correlation</span></span></span><span> (the
strength and direction of the linear relationship between
</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span> and </span><span>
<span class="calibre16"><span class="italic">y</span></span>
</span><span>), finding the equation of a line that best fits
the data (and when doing so is appropriate), and how to use
these results to make predictions for one variable based on
another (called </span><span><span class="calibre16"><span
class="italic">regression</span></span></span><span>). You also
gain tools for investigating when a line fits the data well and
when it doesn’t, and what conclusions you can make (and
shouldn’t make) in the situations where a line does fit.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I cover methods used to look for and
describe links between two categorical variables (such as the
number of doses taken per day and the presence or absence of
nausea) in detail in Chapter 19. I also provide info on
collecting and organizing data into </span><span><span
class="calibre16"><span class="italic">two-way tables</span>
</span></span><span> (where the possible values of one variable
make up the rows and the possible values for the other variable
make up the columns), interpreting the results, analyzing the
data from two-way tables to look for relationships, and
checking for independence. And, as I do throughout this book, I
give you strategies for critically examining results of these
kinds of analyses for credibility.</span></span></p>
</blockquote>
<p id="a43" class="calibre6"><span class="calibre17"><span
class="bold"><span>Drawing Credible Conclusions</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> To perform
statistical analyses, researchers use statistical software that
depends on formulas. But formulas don’t know whether they are
being used properly, and they don’t warn you when your results
are incorrect. At the end of the day, computers can’t tell you
what the results mean; you have to figure it out. Throughout
this book you see what kinds of conclusions you can and can’t
make after the analysis has been done. The following sections
provide an introduction to drawing appropriate conclusions.
</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a44" class="calibre6"><span
class="calibre21"><span class="bold"><span>Reeling in
overstated results</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
```

class="calibre15">Some of the most common mistakes made in conclusions are overstating the results or generalizing the results to a larger group than was actually represented by the study. For example, a professor wants to know which Super Bowl commercials viewers liked best. She gathers 100 students from her class on Super Bowl Sunday and asks them to rate each commercial as it is shown. A top-five list is formed, and she concludes that all Super Bowl viewers liked those five commercials the best. But she really only knows which ones her students liked best — she didn’t study any other groups, so she can’t draw conclusions about all viewers.</p></blockquote>
<blockquote class="calibre5"><p id="a45" class="calibre6">Questioning claims of cause and effect</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One situation in which conclusions cross the line is when researchers find that two variables are related (through an analysis such as regression; see the earlier section “Correlation, regression, and two-way tables” for more info) and then automatically leap to the conclusion that those two variables have a cause-and-effect relationship.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose a researcher conducted a health survey and found that people who took vitamin C every day reported having fewer colds than people who didn’t take vitamin C every day. Upon finding these results, she wrote a paper and gave a press release saying vitamin C prevents colds, using this data as evidence.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now, while it may be true that vitamin C does prevent colds, this researcher’s study can’t claim that. Her study was observational, which means she didn’t control for any other factors that could be related to both vitamin C and colds. For example, people who take vitamin C every day may be more health conscious overall, washing their hands more often, exercising more, and eating better foods; all these behaviors may be helpful in reducing colds.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Until you do a controlled experiment, you can’t make a cause-and-effect conclusion based on relationships you find. (I discuss experiments in more detail earlier in this chapter.)

</p></blockquote>
<p id="a46" class="calibre6">Becoming a Sleuth, Not a Skeptic
</p>
<blockquote class="calibre9"><p class="calibre14">Statistics is about much more than
numbers. To really “get” statistics, you need to understand how
to make appropriate conclusions from studying data and be savvy
enough to not believe everything you hear or read until you
find out how the information came about, what was done with it,
and how the conclusions were drawn. That’s something I discuss
throughout the book, but I really zoom in on it in Chapter 20,
which gives you ten ways to be a statistically savvy sleuth by
recognizing common mistakes made by researchers and the media.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> For you students out there, Chapter 21
brings good statistical practice into the exam setting and
gives you tips on increasing your scores. Much of my advice is
based on understanding the big picture as well as the details
of tackling statistical problems and coming out a winner on the
other side.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Becoming
skeptical or cynical about statistics is very easy, especially
after finding out what’s going on behind the scenes; don’t let
that happen to you. You can find lots of good information out
there that can affect your life in a positive way. Find a good
channel for your skepticism by setting two personal goals:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> To become a well-informed consumer of
the statistical information you see every day</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> To establish job security by being the
statistics “go-to” person who knows when and how to help others
and when to find a statistician</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Through reading and using the
information in this book, you’ll be confident in knowing you
can make good decisions about statistical results. You’ll
conduct your own statistical studies in a credible way. And
you’ll be ready to tackle your next office project, critically
evaluate that annoying political ad, or ace your next exam!

```
</span></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a47"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a38" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a39" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a30" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a31"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a32" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a33" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a34" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a35" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a36" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a37" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a29" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a28"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a27" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a48" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a45" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a44" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a47" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a46" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a41" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a40"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a43" style="min-
width: 10px !important; min-height: 10px !important; border:
```

```
solid 1px !important;"> </a> <a href="#a42" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#5KzRDSqba4c6hbtT4ckd63"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a70">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 2</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a49" class="calibre10"><span class="calibre11"><span
class="bold"><span>The Statistics of Everyday Life</span>
</span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Raising questions about statistics you see in everyday
life</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Encountering statistics in the workplace</span></span>
</p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>oday’s society is
completely taken over by numbers. Numbers are everywhere you
look, from billboards showing the on-time statistics for a
particular airline, to sports shows discussing the Las Vegas
odds for upcoming football games. The evening news is filled
with stories focusing on crime rates, the expected life span of
junk-food junkies, and the president’s approval rating. On a
normal day, you can run into 5, 10, or even 20 different
statistics (with many more on election night). Just by reading
a Sunday newspaper all the way through, you come across
literally hundreds of statistics in reports, advertisements,
and articles covering everything from soup (how much does an
average person consume per year?) to nuts (almonds are known to
have positive health effects — what about other types of
```

nuts?).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter I discuss the
statistics that often appear in your life and work and talk
about how statistics are presented to the general public. After
reading this chapter, you’ll realize just how often the media
hits you with numbers and how important it is to be able to
unravel the meaning of those numbers. Like it or not,
statistics are a big part of your life. So, if you can’t beat
’em, join ’em. And if you don’t want to join ’em, at least try
to understand ’em.</p></blockquote>
<p id="a50" class="calibre6">Statistics and the Media: More Questions
than Answers?</p>
<blockquote class="calibre9"><p class="calibre14">Open a newspaper and start looking for
examples of articles and stories involving numbers. It doesn’t
take long before numbers begin to pile up. Readers are
inundated with results of studies, announcements of
breakthroughs, statistical reports, forecasts, projections,
charts, graphs, and summaries. The extent to which statistics
occur in the media is mind-boggling. You may not even be aware
of how many times you’re hit with numbers nowadays.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This section looks at just a few
examples from one Sunday paper’s worth of news that I read the
other day. When you see how frequently statistics are reported
in the news without providing all the information you need, you
may find yourself getting nervous, wondering what you can and
can’t believe anymore. Relax! That’s what this book is for — to
help you sort out the good information from the bad (the
chapters in Part II give you a great start on that).
</p></blockquote>
<blockquote class="calibre5"><p id="a51" class="calibre6">Probing popcorn
problems</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The first article I came across that
dealt with numbers was “Popcorn plant faces health probe,” with
the subheading: “Sick workers say flavoring chemicals caused
lung problems.” The article describes how the Centers for
Disease Control (CDC) expressed concern about a possible link
between exposure to chemicals in microwave popcorn flavorings
and some cases of fixed obstructive lung disease. Eight people
from one popcorn factory alone contracted this lung disease,
and four of them were awaiting lung transplants.
</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">According to the article, similar cases
were reported at other popcorn factories. Now, you may be
wondering, what about the folks who eat microwave popcorn?
According to the article, the CDC finds “no reason to believe
that people who eat microwave popcorn have anything to fear.”
(Stay tuned.) The next step is to evaluate employees more in-
depth, including conducting surveys to determine health and
possible exposures to the said chemicals, checks of lung
capacity, and detailed air samples. The question here is: How
many cases of this lung disease constitute a real pattern,
compared to mere chance or a statistical anomaly? (You find out
more about this in Chapter 14.)</p></blockquote>
<blockquote class="calibre5"><p id="a52" class="calibre6">Venturing into
viruses</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The second article discussed a recent
cyber attack: A wormlike virus made its way through the
Internet, slowing down Web browsing and e-mail delivery across
the world. How many computers were affected? The experts quoted
in the article said that 39,000 computers were infected, and
they in turn affected hundreds of thousands of other systems.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Questions: How did the experts get that
number? Did they check each computer out there to see whether
it was affected? The fact that the article was written less
than 24 hours after the attack suggests the number is a guess.
Then why say 39,000 and not 40,000 – to make it seem less like
a guess? To find out more on how to guesstimate with confidence
(and how to evaluate someone else’s numbers), see Chapter 13.
</p></blockquote>
<blockquote class="calibre5"><p id="a53" class="calibre6">Comprehending
crashes</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Next in the paper was an alert about
the soaring number of motorcycle fatalities. Experts said that
the fatality rate – the
number of fatalities per 100,000 registered vehicles – for
motorcyclists has been steadily increasing, as reported by the
National Highway Traffic Safety Administration (NHTSA). In the
article, many possible causes for the increased motorcycle
death rate are discussed, including age, gender, size of
engine, whether the driver had a license, alcohol use, and
state helmet laws (or lack thereof). The report is very

comprehensive, showing various tables and graphs with the following titles:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Motorcyclists killed and injured, and fatality and injury rates by year, per number of registered vehicles, and per millions of vehicle miles traveled</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Motorcycle rider fatalities by state, helmet use, and blood alcohol content</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Occupant fatality rates by vehicle type (motorcycles, passenger cars, light trucks), per 10,000 registered vehicles and per 100 million vehicle miles traveled</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Motorcyclist fatalities by age group</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Motorcyclist fatalities by engine size (displacement)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Previous driving records of drivers involved in fatal traffic crashes by type of vehicle (including previous crashes, DUI convictions, speeding convictions, and license suspensions and revocations)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Percentage of alcohol-impaired motorcycle riders killed in traffic crashes by time of day, for single-vehicle, multiple-vehicle, and total crashes</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">This article is very informative and provides a wealth of detailed information regarding motorcycle fatalities and injuries in the U.S. However, the onslaught of so many tables, graphs, rates, numbers, and conclusions can be overwhelming and confusing and allow you to miss the big picture. With a little practice, and help from Part II, you'll be better able to sort out graphs, tables, and charts and all the statistics that go along with them. For example, some important statistical issues come up when you see rates versus counts (such as death rates versus number of deaths). As I address in Chapter 3, counts can give you misleading information if they're used when rates would be more appropriate.</p></blockquote>
<blockquote class="calibre5"><p id="a54" class="calibre6">Mulling malpractice</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Further along in the newspaper was a report about a recent medical malpractice insurance study: Malpractice cases affect people in terms of the fees doctors charge and the ability to get the healthcare they need. The article indicates that 1 in 5 Georgia doctors have stopped doing risky procedures (such as delivering babies) because of the ever-increasing malpractice insurance rates in the state. This is described as a “national epidemic” and a “health crisis” around the country. Some brief details of the study are included, and the article states that of the 2,200 Georgia doctors surveyed, 2,800 of them — which they say represents about 18% of those sampled — were expected to stop providing high-risk procedures.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Wait a minute! That can't be right. Out of 2,200 doctors, 2,800 don't perform the procedures, and that is supposed to represent 18%? That's impossible! You can't have a bigger number on the top of a fraction, and still have the fraction be under 100%, right? This is one of many examples of errors in media reporting of statistics. So what's the real percentage? There's no way to tell from the article. Chapter 5 nails down the particulars of calculating statistics so that you can know what to look for and immediately tell when something's not right.</p></blockquote>
<blockquote class="calibre5"><p id="a55" class="calibre6">Belaboring the loss of land</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the same Sunday paper was an article about the extent of land development and speculation across the United States. Knowing how many homes are likely to be built in

your neck of the woods is an important issue to get a handle on. Statistics are given regarding the number of acres of farmland being lost to development each year. To further illustrate how much land is being lost, the area is also listed in terms of football fields. In this particular example, experts said that the mid-Ohio area is losing 150,000 acres per year, which is 234 square miles, or 115,385 football fields (including end zones). How do people come up with these numbers, and how accurate are they? And does it help to visualize land loss in terms of the corresponding number of football fields? I discuss the accuracy of data collected in more detail in Chapter 16.</p></blockquote>
<blockquote class="calibre5"><p id="a56" class="calibre6">Scrutinizing schools</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The next topic in the paper was school proficiency — specifically, whether extra school sessions help students perform better. The article states that 81.3% of students in this particular district who attended extra sessions passed the writing proficiency test, whereas only 71.7% of those who didn’t participate in the extra school sessions passed it. But is this enough of a difference to account for the $386,000 price tag per year? And what’s happening in these sessions to cause an improvement? Are students in these sessions spending more time just preparing for those exams rather than learning more about writing in general? And here’s the big question: Were the participants in the extra sessions student volunteers who may be more motivated than the average student to try to improve their test scores? The article doesn’t say.</p></blockquote>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">Studying surveys of all shapes and sizes</p></blockquote><div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">Surveys and polls are among the most visible mechanisms used by today’s media to grab your attention. It seems that everyone wants to do a survey, including market managers, insurance companies, TV stations, community groups, and even students in high school classes. Here are just a few examples of survey results that are part of today’s news:</p></blockquote>
<blockquote class="calibre5"><p class="calibre34">With the aging of the American workforce, companies are planning for their future leadership.

(How do they know that the American workforce is aging, and if it is, by how much is it aging?) A recent survey shows that nearly 67% of human-resources managers polled said that planning for succession had become more important in the past five years than it had been in the past. The survey also says that 88% of the 210 respondents said they usually or often fill senior positions with internal candidates. But how many managers did not respond, and is 210 respondents really enough people to warrant a story on the front page of the business section? Believe it or not, when you start looking for them, you’ll find numerous examples in the news of surveys based on far fewer participants than 210. (To be fair, however, 210 can actually be a good number of subjects in some situations. The issues of what sample size is large enough and what percentage of respondents is big enough are addressed in full detail in Chapter 16.)</p></blockquote>
<blockquote class="calibre5"><p class="calibre34">Some surveys are based on current interests and trends. For example, a recent Harris-Interactive survey found that nearly half (47%) of U.S. teens say their social lives would end or be worsened without their cellphones, and 57% go as far as to say that their cellphones are the key to their social life. The study also found that 42% of teens say that they can text while blindfolded (how do you really test this?). Keep in perspective, though, that the study did not tell you what percentage of teens actually have cellphones or what demographic characteristics those teens have compared to teens who do not have cellphones. And remember that data collected on topics like this aren’t always accurate, because the individuals who are surveyed may tend to give biased answers (who wouldn’t want to say they can text blindfolded?). For more information on how to interpret and evaluate the results of surveys, see Chapter 16.
</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<blockquote class="calibre9"><p class="calibre14">Studies like this appear all the time, and the only way to know what to believe is to understand what questions to ask and to be able to critique the quality of the study. That’s all part of statistics! The good news is, with a few clarifying questions, you can quickly critique statistical studies and their results. Chapter 17 helps you do just that.
</p></blockquote>
<blockquote class="calibre5"><p id="a57" class="calibre6">Studying sports</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">The sports section is probably the most
numerically jampacked section of the newspaper. Beginning with
game scores, the win/loss percentages for each team, and the
relative standing for each team, the specialized statistics
reported in the sports world are so deep they require wading
boots to get through. For example, basketball statistics are
broken down by team, by quarter, and by player. For each
player, you get minutes played, field goals, free throws,
rebounds, assists, personal fouls, turnovers, blocks, steals,
and total points.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Who needs to know this stuff, besides
the players’ mothers? Apparently many fans do. Statistics are
something that sports fans can never get enough of and players
often can’t stand to hear about. Stats are the substance of
water-cooler debates and the fuel for armchair quarterbacks
around the world.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Fantasy sports have also made a huge
impact on the sports money-making machine. Fantasy sports are
games where participants act as owners to build their own teams
from existing players in a professional league. The fantasy
team owners then compete against each other. What is the
competition based on? Statistical performance of the players
and teams involved, as measured by rules set up by a “league
commissioner” and an established point system. According to the
Fantasy Sports Trade Association, the number of people age 12
and up who are involved in fantasy sports is more than 30
million, and the amount of money spent is $3–4 billion per
year. (And even here you can ask how the numbers were
calculated — the questions never end, do they?)
</p></blockquote>
<blockquote class="calibre5"><p id="a58" class="calibre6">Banking on business
news</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The business section of the newspaper
provides statistics about the stock market. In one week the
market went down 455 points; is that decrease a lot or a
little? You need to calculate a percentage to really get a
handle on that.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The business section of my paper
contained reports on the highest yields nationwide on every
kind of certificate of deposit (CD) imaginable. (By the way,
how do they know those yields are the highest?) I also found
reports about rates on 30-year fixed loans, 15-year fixed

loans, 1-year adjustable rate loans, new car loans, used car
loans, home equity loans, and loans from your grandmother (well
actually no, but if grandma read these statistics, she might
increase her cushy rates).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Finally, I saw numerous ads for those
beloved credit cards — ads listing the interest rates, the
annual fees, and the number of days in the billing cycle. How
do you compare all the information about investments, loans,
and credit cards in order to make a good decision? What
statistics are most important? The real question is: Are the
numbers reported in the paper giving the whole story, or do you
need to do more detective work to get at the truth? Chapters 16
and 17 help you start tearing apart these numbers and making
decisions about them.</p></blockquote>
<blockquote class="calibre5"><p id="a59" class="calibre6">Touring the travel
news</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You can’t even escape the barrage of
numbers by heading to the travel section. For example, there I
found that the most frequently asked question coming in to the
Transportation Security Administration’s response center (which
receives about 2,000 telephone calls, 2,500 e-mail messages,
and 200 letters per week on average — would you want to be the
one counting all of those?) is, “Can I carry this on a plane?”
This can refer to
anything from an animal to a wedding dress to a giant tin of
popcorn. (I wouldn’t recommend the tin of popcorn. You have to
put it in the overhead compartment horizontally, and because
things shift during flight, the cover will likely open; and
when you go to claim your tin at the end of the flight, you and
your seatmates will be showered. Yes, I saw it happen once.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The number of reported responses in
this case leads to an interesting statistical question: How
many operators are needed at various times of the day to field
those calls, e-mails, and letters coming in? Estimating the
number of anticipated calls is your first step, and being wrong
can cost you money (if you overestimate it) or a lot of bad PR
(if you underestimate it). These kinds of statistical
challenges are tackled in Chapter 13.</p>
</blockquote>
<blockquote class="calibre5"><p id="a60" class="calibre6">Surveying sexual
stats</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">In today’s age of info-overkill, it’s very easy to find out what the latest buzz is, including the latest research on people’s sex lives. An article in my paper reported that married people have 6.9 more sexual encounters per year than people who have never been married. That’s nice to know, I guess, but how did someone come up with this number? The article I’m looking at doesn’t say (maybe some statistics are better left unsaid?).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If someone conducted a survey by calling people on the phone asking for a few minutes of their time to discuss their sex lives, who will be the most likely to want to talk about it? And what are they going to say in response to the question, “How many times a week do you have sex?” Are they going to report the honest truth, tell you to mind your own business, or exaggerate a little? Self-reported surveys can be a real source of bias and can lead to misleading statistics. But how would you recommend people go about finding out more about this very personal subject? Sometimes, research is more difficult than it seems. (Chapter 16 discusses biases that come up when collecting certain types of survey data.) </p></blockquote>
<blockquote class="calibre5"><p id="a61" class="calibre6">Breaking down weather reports</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Weather reports provide another mass of statistics, with forecasts of the next day’s high and low temperatures (how do they decide it’ll be 16 degrees and not 15 degrees?) along with reports of the day’s UV factor, pollen count, pollution standard index, and water quality and quantity. (How do they get these numbers — by taking samples? How many samples do they take, and where do they take them?) You can find out what the weather is right now anywhere in the world. You can get a forecast looking ahead three days, a week, a month, or even a year! Meteorologists collect and record tons and tons of data on the weather each day. Not only do these numbers help you decide whether to take your umbrella to work, but they also help weather researchers to better predict longer term forecasts and even global climate changes over time. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Even with all the information and technologies available to weather researchers, how accurate are weather reports these days? Given the number of times you get rained on when you were told it was going to be sunny, it seems they still have work to do on those forecasts. What the

abundance of data really shows though, is that the number of variables affecting weather is almost overwhelming, not just to you, but for meteorologists, too.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Statistical computer models play an important role in making predictions about major weather-related events, such as hurricanes, earthquakes, and volcano eruptions. Scientists still have some work to do before they can predict tornados before they begin to form or tell you exactly where and when a hurricane is going to hit land, but that's certainly their goal, and they continue to get better at it. For more on modeling and statistics, see Chapter 18.</p></blockquote>
<blockquote class="calibre5"><p id="a62" class="calibre6">Musing about movies</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Moving on to the arts section, I saw several ads for current movies. Each movie ad contains quotes from certain movie critics: "Two thumbs up!" "The supreme adventure of our time," "Absolutely hilarious," or "One of the top ten films of the year!" Do you pay attention to the critics? How do you determine which movies to go to? Experts say that although the popularity of a movie may be affected by the critics' comments (good or bad) in the beginning of a film's run, word of mouth is the most important determinant of how well a film does in the long run.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Studies also show that the more dramatic a movie is, the more popcorn is sold. Yes, the entertainment business even keeps tabs on how much crunching you do at the movies. How do they collect all this information, and how does it impact the types of movies that are made? This, too, is part of statistics: designing and carrying out studies to help pinpoint an audience and find out what they like, and then using the information to help guide the making of the product. So the next time someone with a clipboard asks if you have a minute, you may want to stand up and be counted.
</p></blockquote>
<blockquote class="calibre5"><p id="a63" class="calibre6">Highlighting horoscopes</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Those horoscopes: You read them, but do you believe them? Should you? Can people predict what will

happen more often than just by chance? Statisticians have a way of finding out, by using something they call a hypothesis test (see Chapter 14). So far they haven’t found anyone who can read minds, but people still keep trying!</p></blockquote>
<p id="a64" class="calibre6">Using Statistics at Work </p>
<blockquote class="calibre9"><p class="calibre14">Now put down the Sunday newspaper and move on to the daily grind of the workplace. If you’re working for an accounting firm, of course numbers are part of your daily life. But what about people like nurses, portrait studio photographers, store managers, newspaper reporters, office staff, or construction workers? Do numbers play a role in those jobs? You bet. This section gives you a few examples of how statistics creep into every workplace.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You don’t have to go far to see how statistics weaves its way in and out of your life and work. The secret is being able to determine what it all means and what you can believe, and to be able to make sound decisions based on the real story behind numbers so you can handle and become used to the statistics of everyday life.</p> </blockquote>
<blockquote class="calibre5"><p id="a65" class="calibre6">Delivering babies — and information</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Sue works as a nurse during the night shift in the labor and delivery unit at a university hospital. She takes care of several patients in a given evening, and she does her best to accommodate everyone. Her nursing manager has told her that each time she comes on shift she should identify herself to the patient, write her name on the whiteboard in the patient’s room, and ask whether the patient has any questions. Why? Because a few days after each mother leaves with her baby, the hospital gives her a phone call asking about the quality of care, what was missed, what it could do to improve its service and quality of care, and what the staff could do to ensure that the hospital is chosen over other hospitals in town. For example, surveys show that patients who know the names of their nurses feel more comfortable, ask more questions, and have a more positive experience in the hospital than those who don’t

know the names of their nurses. Sue’s salary raises depend on her ability to follow through with the needs of new mothers. No doubt the hospital has also done a lot of research to determine the factors involved in quality of patient care well beyond nurse-patient interactions. (See Chapter 17 for in-depth info concerning medical studies.)</p></blockquote>
<blockquote class="calibre5"><p id="a66" class="calibre6">Posing for pictures</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Carol recently started working as a photographer for a department store portrait studio; one of her strengths is working with babies. Based on the number of photos purchased by customers over the years, this store has found that people buy more posed pictures than natural-looking ones. As a result, store managers encourage their photographers to take posed shots.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A mother comes in with her baby and has a special request: “Could you please not pose my baby too deliberately? I just like his pictures to look natural.” If Carol says, “Can’t do that, sorry. My raises are based on my ability to pose a child well,” you can bet that the mother is going to fill out that survey on quality service after this session — and not just to get $2.00 off her next sitting (if she ever comes back). Instead, Carol should show her boss the information in Chapter 16 about collecting data on customer satisfaction.</p></blockquote>
<blockquote class="calibre5"><p id="a67" class="calibre6">Poking through pizza data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Terry is a store manager at a local pizzeria that sells pizza by the slice. He is in charge of determining how many workers to have on staff at a given time, how many pizzas to make ahead of time to accommodate the demand, and how much cheese to order and grate, all with minimal waste of wages and ingredients. Friday night at midnight, the place is dead. Terry has five workers left and has five large pans of pizza he could throw in the oven, making about 40 slices of pizza each. Should he send two of his workers home? Should he put more pizza in the oven or hold off?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The store owner has been tracking the demand for weeks now, so Terry knows that every Friday night things slow down between 10 and 12 p.m., but then the bar crowd starts pouring in around midnight and doesn’t let up until the

doors close at 2:30 a.m. So Terry keeps the workers on, puts in the pizzas in 30-minute intervals from midnight on, and is rewarded with a profitable night, with satisfied customers and with a happy boss. For more information on how to make good estimates using statistics, see Chapter 13.</p>
</blockquote>
<blockquote class="calibre5"><p id="a68" class="calibre6">Statistics in the office</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">D.J. is an administrative assistant for a computer company. How can statistics creep into her office workplace? Easy. Every office is filled with people who want to know answers to questions, and they want someone to “Crunch the numbers,” to “Tell me what this means,” to “Find out if anyone has any hard data on this,” or to simply say, “Does this number make any sense?” They need to know everything from customer satisfaction figures to changes in inventory during the year; from the percentage of time employees spend on e-mail to the cost of supplies for the last three years. Every workplace is filled with statistics, and D.J.’s marketability and value as an employee could go up if she’s the one the head honchos turn to for help. Every office needs a resident statistician — why not let it be you?</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a69"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

            </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a78">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 3</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a71" class="calibre10"><span class="calibre11"><span
class="bold"><span>Taking Control: So Many Numbers, So Little
Time</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Examining the extent of statistics abuse</span></span>
</p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Feeling the impact of statistics gone wrong</span></span>
</p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>he sheer amount of
statistics in daily life can leave you feeling overwhelmed and
confused. This chapter gives you a tool to help you deal with
statistics: skepticism! Not radical skepticism like “I can’t
believe anything anymore,” but healthy skepticism like “Hmm, I
wonder where that number came from?” and “I need to find out
more information before I believe these results.” To develop
healthy skepticism, you need to understand how the chain of
statistical information works.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Statistics end up on your TV and in
your newspaper as a result of a process. First, the researchers
who study an issue generate results; this group is composed of
pollsters, doctors, marketing researchers, government
```

researchers, and other scientists. They are considered the original sources of the statistical information.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After they get their results, these researchers naturally want to tell people about it, so they typically either put out a press release or publish a journal article. Enter the journalists or reporters, who are considered the media sources of the information. Journalists hunt for interesting press releases and sort through journals, basically searching for the next headline. When reporters complete their stories, statistics are immediately sent out to the public through all forms of media. Now the information is ready to be taken in by the third group — the consumers of the information (you). You and other consumers of information are faced with the task of listening to and reading the information, sorting through it, and making decisions about it. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">At any stage in the process of doing research, communicating results, or consuming information, errors can take place, either unintentionally or by design. The tools and strategies you find in this chapter give you the skills to be a good detective.</p></blockquote>
<p id="a72" class="calibre6">Detecting Errors, Exaggerations, and Just Plain Lies</p>
<blockquote class="calibre9"><p class="calibre14">Statistics can go wrong for many different reasons. First, a simple, honest error can occur. This can happen to anyone, right? Other times, the error is something other than a simple, honest mistake. In the heat of the moment, because someone feels strongly about a cause and because the numbers don’t quite bear out the point that the researcher wants to make, statistics get tweaked, or, more commonly, exaggerated, either in their values or how they’re represented and discussed.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Another type of error is an error of omission<span> — information that is missing that would have made a big difference in terms of getting a handle on the real story behind the numbers. That omission makes the issue of correctness difficult to address,

```
because you’re lacking information to go on.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You may even encounter situations in
which the numbers have been completely fabricated and can’t be
repeated by anyone because they never happened. This section
gives you tips to help you spot errors, exaggerations, and
lies, along with some examples of each type of error that you,
as an information consumer, may encounter.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a73" class="calibre6"><span
class="calibre21"><span class="bold"><span>Checking the
math</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The first thing you want to do when you
come upon a statistic or the result of a statistical study is
to ask, “Is this number correct?” Don’t assume it is! You’d
probably be surprised at the number of simple arithmetic errors
that occur when statistics are collected, summarized, reported,
or interpreted.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> To spot arithmetic errors or omissions
in statistics:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Check to be
sure everything adds up.</span></span></span><span> In other
words, do the percents in the pie chart add up to 100 (or close
enough due to rounding)? Do the number of people in each
category add up to the total number surveyed?</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Double-check
even the most basic calculations.</span></span></span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Always look
for a total so you can put the results into proper perspective.
</span></span></span><span> Ignore results based on tiny sample
sizes.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
```

class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Examine
whether the projections are reasonable.
 For example, if three deaths due to a certain condition
are said to happen per minute, that adds up to over 1.5 million
such deaths in a year. Depending on what condition is being
reported, this number may be unreasonable.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p id="a74" class="calibre6">Uncovering
misleading statistics</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">By far, the most common abuses of
statistics are subtle, yet effective, exaggerations of the
truth. Even when the math checks out, the underlying statistics
themselves can be misleading if they exaggerate the facts.
Misleading statistics are harder to pinpoint than simple math
errors, but they can have a huge impact on society, and,
unfortunately, they occur all the time.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Breaking down
statistical debates</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Crime statistics are a great example of
how statistics are used to show two sides of a story, only one
of which is really correct. Crime is often discussed in
political debates, with one candidate (usually the incumbent)
arguing that crime has gone down during her tenure, and the
challenger often arguing that crime has gone up (giving the
challenger something to criticize the incumbent for). How can
two candidates make such different conclusions based on the
same data set? Turns out, depending on the way you measure
crime, getting either result can be possible.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Table 3-1 shows the population of the
United States for 1998 to 2008, along with the number of
reported crimes and the crime rates
 (crimes per 100,000 people), calculated by taking
the number of crimes divided by the population size and
multiplying by 100,000.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 3-1" src="images/00013.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Now compare the number of crimes and the crime rates for 2001 and 2002 in Table 3-1. In column 2, you see that the number of crimes increased by 2,285 from 2001 to 2002 (11,878,954 – 11,876,669). This represents an increase of 0.019% (dividing the difference, 2,285, by the number of crimes in 2001, 11,876,669). Note the population size (column 3) also increased from 2001 to 2002, by 2,656,365 people (287,973,924 – 285,317,559), or 0.931% (dividing this difference by the population size in 2001). However, in column 4, you see the crime rate decreased from 2001 to 2002 from 4,162.6 (per 100,000 people) in 2001 to 4,125.0 (per 100,000) in 2002. How did the crime rate decrease? Although the number of crimes and the number of people both went up, the number of crimes increased at a slower rate than the increase in population size did (0.019% compare to 0.931%).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So how should the crime trend be reported? Did crime actually go up or down from 2001 to 2002? Based on the crime rate — which is a more accurate gauge — you can conclude that crime decreased during that year. But be watchful of the politician who wants to show that the incumbent didn’t do his job; he will be tempted to look at the number of crimes and claim that crime went up, creating an artificial controversy and resulting in confusion (not to mention skepticism) on behalf of the voters. (Aren’t election years fun?)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To create an even playing field when measuring how often an event occurs, you convert each number to a percent by dividing by the total to get what statisticians call a rate . Rates are usually better than count data because rates allow you to make fair comparisons when the totals are different.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Untwisting tornado statistics</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Which state has the most tornados? It depends on how you look at it. If you just count the number of tornados in a given year (which is how I’ve seen the media report it most often), the top state is Texas. But think about it. Texas is the second biggest state (after Alaska). Yes,

Texas is in that part of the U.S. called “Tornado Alley,” and yes, it gets a lot of tornados, but it also has a huge surface area for those tornados to land and run.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A more fair comparison, and how meteorologists look at it, is to look at the number of tornados per 10,000 square miles. Using this statistic (depending on your source), Florida comes out on top, followed by Oklahoma, Indiana, Iowa, Kansas, Delaware, Louisiana, Mississippi, and Nebraska, and finally Texas weighs in at number 10. (Although I’m sure this is one statistic they are happy to rank low on; as opposed to their AP rankings in NCAA football.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Other tornado statistics measured and reported include the state with the highest percentage of killer tornadoes as a percentage of all tornados (Tennessee); and the total length of tornado paths per 10,000 square miles (Mississippi). Note each of these statistics is reported appropriately as a rate (amount per unit).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Before believing statistics indicating “the highest XXX” or “the lowest XXX,” take a look at how the variable is measured to see whether it’s fair and whether there are other statistics that should be examined too to get the whole picture. Also make sure the units are appropriate for making fair comparisons.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Zeroing in on what the scale tells you</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Charts and graphs are useful for making a quick and clear point about your data. Unfortunately, many times the charts and graphs accompanying everyday statistics aren’t done correctly and/or fairly. One of the most important elements to watch for is the way that the chart or graph is scaled. The scale of a graph is the quantity used to represent each tick mark on the axis of the graph. Do the tick marks increase by 1s, 10s, 20s, 100s, 1,000s, or what? The scale can make a big difference in terms of the way the graph or chart looks.</p>
</blockquote>

<blockquote class="calibre9"><p class="calibre14">For example, the Kansas Lottery routinely shows its recent results from the Pick 3 Lottery. One of the statistics reported is the number of times each number (0 through 9) is drawn among the three winning numbers. Table 3-2 shows a chart of the number of times each number was drawn during 1,613 total Pick 3 games (4,839 single numbers drawn). It also reports the percentage of times that each number was drawn. Depending on how you choose to look at these results, you can again make the statistics appear to tell very different stories.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The way lotteries typically display results like those in Table 3-2 is shown in Figure 3-1a. Notice that in this chart, it seems that the number 1 doesn’t get drawn nearly as often (only 468 times) as number 2 does (513 times). The difference in the height of these two bars appears to be very large, exaggerating the difference in the number of times these two numbers were drawn. However, to put this in perspective, the actual difference here is 513 – 468 = 45 out of a total of 4,839 numbers drawn. In terms of percentages, the difference between the number of times the number 1 and the number 2 are drawn is 45 ÷ 4,839 = 0.009, or only nine-tenths of one percent.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 3-1: Bar charts showing a) number of times each number was drawn; and b) percentage of times each number was drawn.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">What makes this chart exaggerate the differences? Two issues come to mind. First, notice that the vertical axis, which shows the number of times (or frequency) that each number is drawn, goes up by 5s. So a difference of 5 out of a total of 4,839 numbers drawn appears significant. Stretching the scale so that differences appear larger than they really are is a common trick used to exaggerate results. Second, the chart starts counting at 465, not at 0. Only the top part of each bar is shown, which also exaggerates the results. In comparison, Figure 3-1b graphs the

percentage
 of times each number was drawn. Normally
the shape of a graph wouldn’t change when going from counts to
percentages; however, this chart uses a more realistic scale
than the one in Figure 3-1a (going by 2% increments) and starts
at 0, both of which make the differences appear as they really
are — not much different at all. Boring, huh?</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Maybe the lottery folks thought so too.
In fact, maybe they use Figure 3-1a rather than Figure 3-1b
because they want you to think that some “magic” is involved in
the numbers, and you can’t blame them; that’s their business.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Looking at the
scale of a graph or chart can really help you keep the reported
results in proper perspective. Stretching the scale out or
starting the y-axis at the highest
possible number makes differences appear larger; squeezing down
the scale or starting the
y-axis at a
much lower value than needed makes differences appear smaller
than they really are.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Checking your
sources</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When examining the results of any
study, check the source of the information. The best results
are often published in reputable journals that are well known
by the experts in the field. For example, in the world of
medical science, the Journal of the American Medical
Association (JAMA), the
New England
Journal of Medicine, The Lancet, and
the British Medical Journal
 are all reputable journals doctors use to publish
results and read about new findings.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Consider the source and who
financially supported the research. Many companies finance

research and use it for advertising their products. Although
that in itself isn’t necessarily a bad thing, in some cases a
conflict of interest on the part of researchers can lead to
biased results. And if the results are very important to you,
ask whether more than one study was conducted, and if so, ask
to examine all the studies that were conducted, not just those
whose results were published in journals or appeared in
advertisements.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Counting on sample
size</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Sample size isn’t everything, but it
does count for a great deal in surveys and studies. If the
study is designed and conducted correctly, and if the
participants are selected randomly (that is, with no bias; see
Chapter 16 for more on random samples), sample size is an
important factor in determining the accuracy and repeatability
of the results. (See Chapters 16 and 17 for more information on
designing and carrying out studies.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Many surveys are based on large numbers
of participants, but that isn’t always true for other types of
research, such as carefully controlled experiments. Because of
the high cost of some types of research in terms of time and
money, some studies are based on a small number of participants
or products. Researchers have to find the appropriate balance
when determining sample size.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The most
unreliable results are those based on anecdotes,
 stories that talk about a single incident in an
attempt to sway opinion. Have you ever told someone not to buy
a product because you had a bad experience with it? Remember
that an anecdote (or story) is really a nonrandom sample whose
size is only one.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Considering cause and
effect</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Headlines often simplify or skew the
“real” information, especially when the stories involve
statistics and the studies that generated the statistics.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">A study conducted a few years back evaluated videotaped sessions of 1,265 patient appointments with 59 primary-care physicians and 6 surgeons in Colorado and Oregon. This study found that physicians who had not been sued for malpractice spent an average of 18 minutes with each patient, compared to 16 minutes for physicians who
had
 been sued for malpractice. The study was reported by the media with the headline, “Bedside manner fends off malpractice suits.” However, this study seemed to say that if you are a doctor who gets sued, all you have to do is spend more time with your patients, and you’re off the hook. (Now when did bedside manner get characterized as time spent?)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Beyond that, are we supposed to believe that a doctor who has been sued needs only add a couple more minutes of time with each patient to avoid being sued in the future? Maybe what the doctor does during that time counts much more than how much time the doctor actually spends with each patient. You tackle the issues of cause-and-effect relationships between variables in Chapter 18.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Finding what you wanted to find</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You may wonder how two political candidates can discuss the same topic and get two opposing conclusions, both based on “scientific surveys.” Even small differences in a survey can create big differences in results. (See Chapter 16 for the full scoop on surveys.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One common source of skewed survey results comes from question wording. Here are three different questions that are trying to get at the same issue — public opinion regarding the line-item veto option available to the president:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Should the line-item veto be available to the president to eliminate waste (yes/no/no opinon)?</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">

Does the line-item veto give the president too much individual power (yes/no/no opinion)?</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
 What is your opinion on the presidential line-item veto? Choose 1–5, with 1 = strongly opposed and 5 = strongly support. </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">The first two questions are misleading and will lead to biased results in opposite directions. The third version will draw results that are more accurate in terms of what people really think. However, not all surveys are written with the purpose of finding the truth; many are written to support a certain viewpoint.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Research shows that even small changes in wording affect survey outcomes, leading to results that conflict when different surveys are compared. If you can tell from the wording of the question how they want you to respond to it, you know you’re looking at a leading question; and leading questions lead to biased results. (See Chapter 16 for more on spotting problems with surveys.) </p></blockquote>
<blockquote class="calibre5"><p id="a75" class="calibre6">Looking for lies in all the right places</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Every once in a while, you hear about someone who faked his data, or “fudged the numbers.” Probably the most commonly committed lie involving statistics and data is when people throw out data that don’t fit their hypothesis, don’t fit the pattern, or appear to be outliers. In cases when someone has clearly made an error (for example, someone’s age is recorded as 200), removing that erroneous data point or trying to correct the error makes sense. Eliminating data for any other reason is ethically wrong; yet it happens. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Regarding missing data from experiments, a commonly used phrase is “Among those who completed the study. . . .” What about those who didn’t complete the study, especially a medical one? Did they get tired of the side effects of the experimental drug and quit? If so, the loss of this person will create results that are biased toward positive outcomes.</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Before
believing the results of a study, check out how many people
were chosen to participate, how many finished the study, and
what happened to all the participants, not just the ones who
experienced a positive result.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Surveys are not immune to problems from
missing data, either. For example, it’s known by statisticians
that the opinions of people who respond to a survey can be very
different from the opinions of those who don’t. In general, the
lower the percentage of people who respond to a survey (the
response rate), the less credible the results will be. For more
about surveys and missing data, see Chapter 16.
</p></blockquote>
<p id="a76" class="calibre6">Feeling the Impact of Misleading
Statistics</p>
<blockquote class="calibre9"><p class="calibre14">You make decisions every day based on
statistics and statistical studies that you’ve heard about or
seen, many times without even realizing it. Misleading
statistics affect your life in small or large ways, depending
on the type of statistics that cross your path and what you
choose to do with the information you’re given. Here are some
little everyday scenarios where statistics slip in:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
“Gee, I hope Rex doesn’t chew up my rugs again while I’m
at work. I heard somewhere that dogs on Prozac deal better with
separation anxiety. How did they figure that out? And what
would I tell my friends?”</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
“I thought everyone was supposed to drink eight glasses
of water a day, but now I hear that too much water could be bad
for me; what should I believe?”</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
“A study says people spend two hours a day at work
checking and sending personal e-mails. How is that possible? No

wonder my boss is paranoid.”</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">You may run into other situations
involving statistics that can have a larger impact on your
life, and you need to be able to sort it all out. Here are some
examples:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
 A group lobbying for a new skateboard park tells you 80%
of the people surveyed agree that taxes should be raised to pay
for it, so you should too. Will you feel the pressure to say
yes?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The radio news at the top of the hour says cellphones
cause brain tumors. Your spouse uses his cellphone all the
time. Should you panic and throw away all cellphones in your
house?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
You see an advertisement that tells you a certain drug
will cure your particular ill. Do you run to your doctor and
demand a prescription?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Although not
all statistics are misleading and not everyone is out to get
you, you do need to be vigilant. By sorting out the good
information from the suspicious and bad information, you can
steer clear of statistics that go wrong. The tools and
strategies in this chapter are designed to help you to stop and
say, “Wait a minute!” so you can analyze and critically think
about the issues and make good decisions.</p>
</blockquote>
</div>
</div> <div class="mbppagebreak" id="a77"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a74" style="min-width: 10px !important; min-height:

```
10px !important; border: solid 1px !important;"> </a>
<a href="#a75" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a76" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a77"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a71" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a72" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a73" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a78" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#5KzRDSqba4c6hbtT4ckd63"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a>
</div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a107">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 4</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a79" class="calibre10"><span class="calibre11"><span
class="bold"><span>Tools of the Trade</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Seeing statistics as a process, not just as
numbers</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Getting familiar with some basic statistical
jargon</span></span></p></blockquote><div
class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I</span><span>n today’s world, the
buzzword is </span><span><span class="calibre16"><span
class="italic">data,</span></span></span><span> as in, “Do you
have any data to support your claim?” “What data do you have on
this?” “The data supported the original hypothesis that . . .
,” “Statistical data show that . . . ,” and “The data bear this
out . . . .” But the field of statistics is not just about
data.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Statistics is
the entire process involved in gathering evidence to answer
questions about the world, in cases where that evidence happens
to be data.</span></span></p></blockquote>
```

<blockquote class="calibre9"><p class="calibre14">In this chapter, you see firsthand how
statistics works as a process and where the numbers play their
part. You’re also introduced to the most commonly used forms of
statistical jargon, and you find out how these definitions and
concepts all fit together as part of that process. So the next
time you hear someone say, “This survey had a margin of error
of plus or minus 3 percentage points,” you’ll have a basic idea
of what that means.</p></blockquote>
<p id="a80" class="calibre6">Statistics: More than Just Numbers
</p>
<blockquote class="calibre9"><p class="calibre14">Statisticians don’t just “do
statistics.” Although the rest of the world views them as
number crunchers, they think of themselves as the keepers of
the scientific method. Of course, statisticians work with
experts in other fields to satisfy their need for data, because
man cannot live by statistics alone, but crunching someone’s
data is only a small part of a statistician’s job. (In fact, if
that’s all we did all day, we’d quit our day jobs and moonlight
as casino consultants.) In reality, statistics is involved in
every aspect of the scientific method —
formulating good questions, setting up studies, collecting good
data, analyzing the data properly, and making appropriate
conclusions. But aside from analyzing the data properly, what
do any of these aspects have to do with statistics? In this
chapter you find out.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">All research starts with a question,
such as:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Is it possible to drink too much water?</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
What’s the cost of living in San Francisco?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Who will win the next presidential election?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span

class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Do herbs really help maintain good health?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Will my favorite TV show get renewed for next year?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">None of these questions asks anything
directly about numbers. Yet each question requires the use of
data and statistical processes to come up with the answer.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose a researcher wants to determine
who will win the next U.S. presidential election. To answer
with confidence, the researcher has to follow several steps:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Determine the population to be studied.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> In this case, the researcher intends
to study registered voters who plan to vote in the next
election.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Collect the data.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> This step is a challenge, because you
can’t go out and ask every person in the United States whether
they plan to vote, and if so, for whom they plan to vote.
Beyond that, suppose someone says, “Yes, I plan to vote.” Will
that person really vote come
Election Day? And will that same person tell you whom he
actually plans to vote for? And what if that person changes his
mind later on and votes for a different candidate?
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Organize, summarize, and analyze the data.
</p></blockquote><div

class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> After the researcher has gone out and
collected the data she needs, getting it organized, summarized,
and analyzed helps the researcher answer her question. This
step is what most people recognize as the business of
statistics.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Take all the data summaries, charts, graphs,
and analyses and draw conclusions from them to try to answer
the researcher’s original question.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Of course, the researcher will not be
able to have 100% confidence that her answer is correct,
because not every person in the United States was asked. But
she can get an answer that she is nearly
 100% sure is correct. In fact, with a sample of
about 2,500 people who are selected in a fair and
 unbiased
way (that is, every possible sample of size
2,500 had an equal chance of being selected), the researcher
can get accurate results within plus or minus 2.5% (if all the
steps in the research process are done correctly).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> In making
conclusions, the researcher has to be aware that every study
has limits and that — because the chance for error always
exists — the results could be wrong. A numerical value can be
reported that tells others how confident the researcher is
about the results and how accurate these results are expected
to be. (See Chapter 12 for more information on margin of
error.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> After the
research is done and the question has been answered, the
results typically lead to even more questions and even more
research. For example, if men appear to favor one candidate but
women favor the opponent, the next questions may be: “Who goes
to the polls more often on Election Day — men or women — and
what factors determine whether they will vote?”
</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">The field of statistics is really the business of using the scientific method to answer research questions about the world. Statistical methods are involved in every step of a good study, from designing the research to collecting the data, organizing and summarizing the information, doing an analysis, drawing conclusions, discussing limitations, and, finally, designing the next study in order to answer new questions that arise. Statistics is more than just numbers — it’s a process.</p></blockquote>
<p id="a81" class="calibre6">Grabbing Some Basic Statistical Jargon</p>
<blockquote class="calibre9"><p class="calibre14">Every trade has a basic set of tools, and statistics is no different. If you think about the statistical process as a series of stages that you go through to get from question to answer, you may guess that at each stage you’ll find a group of tools and a set of terms (or jargon) to go along with it. Now if the hair is beginning to stand up on the back of your neck, don’t worry. No one is asking you to become a statistics expert and plunge into the heavy-duty stuff, or to turn into a statistics nerd who uses this jargon all the time. Hey, you don’t even have to carry a calculator and pocket protector in your shirt pocket (because statisticians really don’t do that; it’s just an urban myth). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But as the world becomes more numbers-conscious, statistical terms are thrown around more in the media and in the workplace, so knowing what the language really means can give you a leg up. Also, if you’re reading this book because you want to find out more about how to calculate some statistics, understanding basic jargon is your first step. So, in this section, you get a taste of statistical jargon; I send you to the appropriate chapters elsewhere in the book to get details.</p></blockquote>
<blockquote class="calibre5"><p id="a82" class="calibre6">Data </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Data are the actual pieces of information that you collect through your study. For example, I asked five of my friends how many pets they own, and the data they gave me are the following: 0, 2, 1, 4, 18. (The fifth friend counted each of her aquarium fish as a separate pet.) Not all data are numbers; I also recorded the gender of

each of my friends, giving me the following data: male, male, female, male, female.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Most data fall into one of two groups: numerical or categorical. (I present the main ideas about these variables here; see Chapter 5 for more details.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Numerical data: These data have meaning as a measurement, such as a person’s height, weight, IQ, or blood pressure; or they’re a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. (Statisticians also call numerical data quantitative data.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Numerical data can be further broken into two types: discrete and continuous.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite). For example, the number of heads in 100 coin flips takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the fastest scenario) on up to infinity. Its possible values are listed as 100, 101, 102, 103, . . . (representing the countably infinite case).</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks represents nearly-continuous data from 0.00 gallons to 20.00 gallons, represented

by the interval [0, 20], inclusive. (Okay, you
can
 count all these values, but why would you want
to? In cases like these, statisticians bend the definition of
continuous a wee bit.) The lifetime of a C battery can be
anywhere from 0 to infinity, technically, with all possible
values in between. Granted, you don’t expect a battery to last
more than a few hundred hours, but no one can put a cap on how
long it can go (remember the Energizer Bunny?).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Categorical
data: Categorical data represent
characteristics such as a person’s gender, marital status,
hometown, or the types of movies they like. Categorical data
can take on numerical values (such as “1” indicating male and
“2” indicating female), but those numbers don’t have meaning.
You couldn’t add them together, for example. (Other names for
categorical data are qualitative data, or
Yes/No data.)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/>
Ordinal data mixes
numerical and categorical data. The data fall into categories,
but the numbers placed on the categories have meaning. For
example, rating a restaurant on a scale from 0 to 4 stars gives
ordinal data. Ordinal data are often treated as categorical,
where the groups are ordered when graphs and charts are made. I
don’t address them separately in this book.</p>
</blockquote>
<blockquote class="calibre5"><p id="a83" class="calibre6">Data set
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
data
set
<span>is the collection of all the data
taken from your sample. For example, if you measured the
weights of five packages, and those weights were 12, 15, 22,
68, and 3 pounds, those five numbers (12, 15, 22, 68, 3)

```
constitute your data set. If you only record the general size
of the package (for example, small, medium, or large), your
data set may look like this: medium, medium, medium, large,
small.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a84" class="calibre6"><span
class="calibre21"><span class="bold"><span>Variable</span>
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>A </span><span><span class="calibre16">
<span class="italic">variable</span></span></span><span> is any
characteristic or numerical value that varies from individual
to individual. A variable can represent a count (for example,
the number of pets you own); or a measurement (the time it
takes you to wake up in the morning). Or the variable can be
categorical, where each individual is placed into a group (or
category) based on certain criteria (for example, political
affiliation, race, or marital status). Actual pieces of
information recorded on individuals regarding a variable are
the data.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a85" class="calibre6"><span
class="calibre21"><span class="bold"><span>Population</span>
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For virtually any question you may want
to investigate about the world, you have to center your
attention on a particular group of individuals (for example, a
group of people, cities, animals, rock specimens, exam scores,
and so on). For example:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>What do Americans think about the president’s foreign
policy?</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>What percentage of planted crops in Wisconsin did deer
destroy last year?</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>What’s the prognosis for breast cancer patients taking a
new experimental drug?</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
```

class="calibre2"/>
What percentage of all cereal boxes get filled according to specification?</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In each of these examples, a question is posed. And in each case, you can identify a specific group of individuals being studied: the American people, all planted crops in Wisconsin, all breast cancer patients, and all cereal boxes that are being filled, respectively. The group of individuals you want to study in order to answer your research question is called a population. Populations, however, can be hard to define. In a good study, researchers define the population very clearly, whereas in a bad study, the population is poorly defined.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The question of whether babies sleep better with music is a good example of how difficult defining the population can be. Exactly how would you define a baby? Under three months old? Under a year? And do you want to study babies only in the United States, or all babies worldwide? The results may be different for older and younger babies, for American versus European versus African babies, and so on.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Many times researchers want to study and make conclusions about a broad population, but in the end — to save time, money, or just because they don’t know any better — they study only a narrowly defined population. That shortcut can lead to big trouble when conclusions are drawn. For example, suppose a college professor wants to study how TV ads persuade consumers to buy products. Her study is based on a group of her own students who participated to get five points extra credit. This test group may be convenient, but her results can’t be generalized to any population beyond her own students, because no other population was represented in her study.</p></blockquote>
<blockquote class="calibre5"><p id="a86" class="calibre6">Sample, random, or otherwise</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When you sample some soup, what do you do? You stir the pot, reach in with a spoon, take out a little bit of the soup, and taste it. Then you draw a conclusion about the whole pot of soup, without actually having tasted all of

it. If your sample is taken in a fair way (for example, you didn't just grab all the good stuff) you will get a good idea how the soup tastes without having to eat it all. Taking a sample works the same way in statistics. Researchers want to find out something about a population, but they don't have time or money to study every single individual in the population. So they select a subset of individuals from the population, study those individuals, and use that information to draw conclusions about the whole population. This subset of the population is called a sample.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Although the idea of a selecting a sample seems straightforward, it's anything but. The way a sample is selected from the population can mean the difference between results that are correct and fair and results that are garbage. Example: Suppose you want a sample of teenagers' opinions on whether they're spending too much time on the Internet. If you send out a survey using text messaging, your results won't represent the opinions of all teenagers,
 which is your intended population. They will represent only those teenagers who have access to text messages. Does this sort of statistical mismatch happen often? You bet.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Some of the biggest culprits of statistical misrepresentation caused by bad sampling are surveys done on the Internet. You can find thousands of surveys on the Internet that are done by having people log on to a particular Web site and give their opinions. But even if 50,000 people in the U.S. complete a survey on the Internet, it doesn't represent the population of all Americans. It represents only those folks who have Internet access, who logged on to that particular Web site, and who were interested enough to participate in the survey (which typically means that they have strong opinions about the topic in question). The result of all these problems is bias
 — systematic favoritism of certain individuals or certain outcomes of the study.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> How do you select a sample in a way that avoids bias? The key word is <span

class="italic">random. A
random
sample is a sample selected by equal
opportunity; that is, every possible sample the same size as
yours had an equal chance to be selected from the population.
What random really means
is that no group in the population is favored in or excluded
from the selection process.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Non-random (in other
words bad)
samples
 are samples that were selected in such a
way that some type of favoritism and/or automatic exclusion of
a part of the population was involved. A classic example of a
non-random sample comes from polls for which the media asks you
to phone in your opinion on a certain issue (“call-in” polls).
People who choose to participate in call-in polls do not
represent the population at large because they had to be
watching that program, and they had to feel strongly enough to
call in. They technically don’t represent a sample at all, in
the statistical sense of the word, because no one selected them
beforehand — they selected themselves to participate, creating
a volunteer or
self-
selected sample. The results will be
skewed toward people with strong opinions.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">To take an authentic random sample, you
need a randomizing mechanism to select the individuals. For
example, the Gallup Organization starts with a computerized
list of all telephone exchanges in America, along with
estimates of the number of residential households that have
those exchanges. The computer uses a procedure called
random digit
dialing (RDD) to randomly create
phone numbers from those exchanges, and then selects samples of
telephone numbers from those. So what really happens is that
the computer creates a list of all possible
 household phone numbers in America and
then selects a subset of numbers from that list for Gallup to
call.</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">Another example of random sampling involves the use of random number generators. In this process, the items in the sample are chosen using a computer-generated list of random numbers, where each sample of items has the same chance of being selected. Researchers may use this type of randomization to assign patients to a treatment group versus a control group in an experiment. This process is equivalent to drawing names out of a hat or drawing numbers in a lottery. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> No matter how large a sample is, if it’s based on non-random methods, the results will not represent the population that the researcher wants to draw conclusions about. Don’t be taken in by large samples — first check to see how they were selected. Look for the term random sample. If you see that term, dig further into the fine print to see how the sample was actually selected and use the preceding definition to verify that the sample was, in fact, selected randomly. A small random sample is better than a large non-random one. </p></blockquote>
<blockquote class="calibre5"><p id="a87" class="calibre6">Statistic </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A statistic is a number that summarizes the data collected from a sample. People use many different statistics to summarize data. For example, data can be summarized as a percentage (60% of U.S. households sampled own more than two cars), an average (the average price of a home in this sample is . . .), a median (the median salary for the 1,000 computer scientists in this sample was . . .), or a percentile (your baby’s weight is at the 90th percentile this month, based on data collected from over 10,000 babies). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span class="calibre15"><span>The type of statistic calculated depends on the type of data. For example, percentages are used to summarize categorical data, and means are used to summarize numerical data. The price of a home is a numerical variable, so you can calculate its mean or standard deviation. However, the color of a home is a categorical variable; finding the standard deviation or median of color makes no sense. In this case, the important statistics are the percentages of homes of each

```
color.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Not all
statistics are correct or fair, of course. Just because someone
gives you a statistic, nothing guarantees that the statistic is
scientific or legitimate. You may have heard the saying,
“Figures don’t lie, but liars figure.”</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a88" class="calibre6"><span
class="calibre21"><span class="bold"><span>Parameter</span>
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Statistics are based on sample data,
not on population data. If you collect data from the entire
population, that process is called a </span><span><span
class="calibre16"><span class="italic">census.</span></span>
</span><span> If you then summarize the entire census
information from one variable into a single number, that number
is a </span><span><span class="calibre16"><span
class="italic">parameter,</span></span></span><span> not a
statistic. Most of the time, researchers are trying to estimate
the parameters using statistics. The U.S. Census Bureau wants
to report the total number of people in the U.S., so it
conducts a census. However, due to logistical problems in doing
such an arduous task (such as being able to contact homeless
folks), the census numbers can only be called </span><span>
<span class="calibre16"><span class="italic">estimates</span>
</span></span><span> in the end, and they’re adjusted upward to
account for people the census missed.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a89" class="calibre6"><span
class="calibre21"><span class="bold"><span>Bias</span></span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="italic">Bias</span></span></span><span> is a word you
hear all the time, and you probably know that it means
something bad. But what really constitutes bias? </span><span>
<span class="calibre16"><span class="italic">Bias</span></span>
</span><span> is systematic favoritism that is present in the
data collection process, resulting in lopsided, misleading
results. Bias can occur in any of a number of ways:</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">In the way the
```

sample is selected: For example, if you want to estimate how much holiday shopping people in the United States plan to do this year, and you take your clipboard and head out to a shopping mall on the day after Thanksgiving to ask customers about their shopping plans, you have bias in your sampling process. Your sample tends to favor those die-hard shoppers at that particular mall who were braving the massive crowds on that day known to retailers and shoppers as “Black Friday.”</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
In the way data are collected: Poll questions are a major source of bias. Because researchers are often looking for a particular result, the questions they ask can often reflect and lead to that expected result. For example, the issue of a tax levy to help support local schools is something every voter faces at one time or another. A poll question asking, “Don’t you think it would be a great investment in our future to support the local schools?” has a bit of bias. On the other hand, so does “Aren’t you tired of paying money out of your pocket to educate other people’s children?” Question wording can have a huge impact on results.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Other issues that result in bias with polls are timing, length, level of question difficulty, and the manner in which the individuals in the sample were contacted (phone, mail, house-to-house, and so on). See Chapter 16 for more information on designing and evaluating polls and surveys.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When examining polling results that are important to you or that you’re particularly interested in, find out what questions were asked and exactly how the questions were worded before drawing your conclusions about the results.</p></blockquote>
<blockquote class="calibre5"><p id="a90" class="calibre6">Mean (Average)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The mean, also referred to by statisticians as the average, is the most common statistic used to measure the center, or middle, of a

numerical data set. The
mean is the sum
of all the numbers divided by the total number of numbers. The
mean of the entire population is called the population mean,
 and the mean of a sample is called the
sample mean. (See
Chapter 5 for more on the mean.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The mean may
not be a fair representation of the data, because the average
is easily influenced by
outliers (very
small or large values in the data set that are not typical).
</p></blockquote>
<blockquote class="calibre5"><p id="a91" class="calibre6">Median
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The median is another way to measure
the center of a numerical data set. A statistical median is
much like the median of an interstate highway. On many
highways, the median is the middle, and an equal number of
lanes lay on either side of it. In a numerical data set, the
median is the point
at which there are an equal number of data points whose values
lie above and below the median value. Thus, the median is truly
the middle of the data set. See Chapter 5 for more on the
median.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The next time
you hear an average reported, look to see whether the median is
also reported. If not, ask for it! The average and the median
are two different representations of the middle of a data set
and can often give two very different stories about the data,
especially when the data set contains outliers (very large or
small numbers that are not typical).</p>
</blockquote>
<blockquote class="calibre5"><p id="a92" class="calibre6">Standard
deviation</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Have you heard anyone report that a
certain result was found to be “two standard deviations above

the mean”? More and more, people want to report how significant their results are, and the number of standard deviations above or below average is one way to do it. But exactly what is a standard deviation?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard deviation is a measurement statisticians use for the amount of variability (or spread) among the numbers in a data set. As the term implies, a standard deviation is a standard (or typical) amount of deviation (or distance) from the average (or mean, as statisticians like to call it). So the standard deviation, in very rough terms, is the average distance from the mean.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for standard deviation (denoted by s) is as follows, where n equals the number of values in the data set, each x represents a number in the data set, and is the average of all the data:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">For detailed instructions on calculating the standard deviation, see Chapter 5.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The standard deviation is also used to describe where most of the data should fall, in a relative sense, compared to the average. For example, if your data have the form of a bell-shaped curve (also known as a normal distribution), about 95% of the data lie within two standard deviations of the mean. (This result is called the empirical rule, or the 68–95–99.7% rule.

 See Chapter 5 for more on this.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The standard
deviation is an important statistic, but it is often absent
when statistical results are reported. Without it, you’re
getting only part of the story about the data. Statisticians
like to tell the story about the man who had one foot in a
bucket of ice water and the other foot in a bucket of boiling
water. He said on average he felt just great! But think about
the variability in the two temperatures for each of his feet.
Closer to home, the average house price, for example, tells you
nothing about the range of house prices you may encounter when
house-hunting. The average salary may not fully represent
what’s really going on in your company, if the salaries are
extremely spread out.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Don’t be
satisfied with finding out only the average — be sure to ask
for the standard deviation as well. Without a standard
deviation, you have no way of knowing how spread out the values
may be. (If you’re talking starting salaries, for example, this
could be very important!)</p></blockquote>
<blockquote class="calibre5"><p id="a93" class="calibre6">Percentile
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You’ve probably heard references to
percentiles before. If you’ve taken any kind of standardized
test, you know that when your score was reported, it was
presented to you with a measure of where you stood compared to
the other people who took the test. This comparison measure was
most likely reported to you in terms of a percentile. The
percentile reported
for a given score is the percentage of values in the data set
that fall below that certain score. For example, if your score
was reported to be at the 90th percentile, that means that 90%
of the other people who took the test with you scored lower
than you did (and 10% scored higher than you did). The median
is right in the middle of a data set, so it represents the 50th
percentile. For more specifics on percentiles, see Chapter 5.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Percentiles are

used in a variety of ways for comparison purposes and to determine relative standing (that is, how an individual data value compares to the rest of the group). Babies' weights are often reported in terms of percentiles, for example. Percentiles are also used by companies to see where they stand compared to other companies in terms of sales, profits, customer satisfaction, and so on. </p></blockquote>
<blockquote class="calibre5"><p id="a94" class="calibre6">Standard score</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard score is a slick way to put results in perspective without having to provide a lot of details — something that the media loves. The standard score represents the number of standard deviations above or below the mean (without caring what that standard deviation or mean actually are).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose Bob took his statewide 10th-grade test recently and scored 400. What does that mean? Not much, because you can't put 400 into perspective. But knowing that Bob's standard score on the test is +2 tells you everything. It tells you that Bob's score is two standard deviations above the mean. (Bravo, Bob!) Now suppose Emily's standard score is –2. In this case, this is not good (for Emily), because it means her score is two standard deviations below the mean. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The process of taking a number and converting it to a standard score is called standardizing. For the details on calculating and interpreting standard scores when you have a normal (bell-shaped) distribution, see Chapter 9.</p></blockquote>
<blockquote class="calibre5"><p id="a95" class="calibre6">Distribution and normal distribution</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The distribution of a data set (or a population) is a

listing or function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individuals in each group. When a distribution of numerical data is organized, they’re often ordered from smallest to largest, broken into reasonably sized groups (if appropriate), and then put into graphs and charts to examine the shape, center, and amount of variability in the data.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The world of statistics includes dozens of different distributions for categorical and numerical data; the most common ones have their own names. One of the most well-known distributions is called the normal distribution, also known as the bell-shaped curve. The normal distribution is based on numerical data that is continuous; its possible values lie on the entire real number line. Its overall shape, when the data are organized in graph form, is a symmetric bell-shape. In other words, most (around 68%) of the data are centered around the mean (giving you the middle part of the bell), and as you move farther out on either side of the mean, you find fewer and fewer values (representing the downward sloping sides on either side of the bell).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The mean (and hence the median) is directly in the center of the normal distribution due to symmetry, and the standard deviation is measured by the distance from the mean to the inflection point (where the curvature of the bell changes from concave up to concave down). Figure 4-1 shows a graph of a normal distribution with mean 0 and standard deviation 1 (this distribution has a special name, the standard normal distribution or Z-distribution).The shape of the curve resembles the outline of a bell.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 4-1: A standard normal (Z-) distribution has a bell-shaped curve with mean 0 and standard deviation 1.</p>
</div>

<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Because every distinct population of data has a different mean and standard deviation, an infinite number of different normal distributions exist, each with its own mean and its own standard deviation to characterize it. See Chapter 9 for plenty more on the normal and standard normal distributions.</p></blockquote>
<blockquote class="calibre5"><p id="a96" class="calibre6">Central Limit Theorem</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The normal distribution is also used to help measure the accuracy of many statistics, including the mean, using an important result in statistics called the
Central Limit Theorem.
 This theorem gives you the ability to measure how much your sample mean will vary, without having to take any other sample means to compare it with (thankfully!). By taking this variability into account, you can now use your data to answer questions about the population, such as “What’s the mean household income for the whole U.S.?”; or “This report said 75% of all gift cards go unused; is that really true?” (These two particular analyses made possible by the Central Limit Theorem are called confidence intervals
and hypothesis tests,
respectively, and are described in Chapters 13 and 14, respectively.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The Central Limit Theorem (
CLT
 for short) basically says that for non-normal data, your sample mean has an approximate normal distribution, no matter what the distribution of the original data looks like (as long as your sample size was large enough). And it doesn’t just apply to the sample mean; the CLT is also true for other sample statistics, such as the sample proportion (see Chapters 13 and 14). Because statisticians know so much about the normal distribution (see the preceding section), these analyses are much easier. See Chapter 11 for more on the Central Limit Theorem, known by statisticians as the “Crown

jewel in the field of all statistics.” (Should you even bother
to tell them to get a life?)</p></blockquote>
<blockquote class="calibre5"><p id="a97" class="calibre6">z-values
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If a data set
has a normal distribution, and you standardize all the data to
obtain standard scores, those standard scores are called
z-values. All
z
-values have what is known as a standard
normal distribution (or
Z-
distribution). The standard normal distribution
 is a special normal distribution with a mean
equal to 0 and a standard deviation equal to 1.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard normal distribution is
useful for examining the data and determining statistics like
percentiles, or the percentage of the data falling between two
values. So if researchers determine that the data have a normal
distribution, they usually first standardize the data (by
converting each data point into a z
-value) and then use the standard normal distribution to
explore and discuss the data in more detail. See Chapter 9 for
more details on z-values.
</p></blockquote>
<blockquote class="calibre5"><p id="a98" class="calibre6">Experiments
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">An experiment
 is a study that imposes a treatment (or control)
to the subjects (participants), controls their environment (for
example, restricting their diets, giving them certain dosage
levels of a drug or placebo, or asking them to stay awake for a
prescribed period of time), and records the responses. The
purpose of most experiments is to pinpoint a cause-and-effect
relationship between two factors (such as alcohol consumption
and impaired vision; or dosage level of a drug and intensity of

side effects). Here are some typical questions that experiments try to answer:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Does taking zinc help reduce the duration of a cold? Some studies show that it does.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Does the shape and position of your pillow affect how well you sleep at night? The Emory Spine Center in Atlanta says yes.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Does shoe heel height affect foot comfort? A study done at UCLA says up to one-inch heels are better than flat soles. </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In this section, I discuss some additional definitions of words that you may hear when someone is talking about experiments. Chapter 17 is entirely dedicated to the subject. For now, just concentrate on basic experiment lingo.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Treatment group versus control group</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Most experiments try to determine whether some type of experimental treatment (or important factor) has a significant effect on an outcome. For example, does zinc help to reduce the length of a cold? Subjects who are chosen to participate in the experiment are typically divided into two groups: a treatment group and a control group. (More than one treatment group is possible.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18">
The treatment group consists of participants who receive the experimental treatment whose effect is being studied (in this case, zinc tablets). </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span

class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The control group
consists of participants who do not receive the experimental
treatment being studied. Instead, they get a placebo (a fake
treatment; for example, a sugar pill); a standard,
nonexperimental treatment (such as vitamin C, in the zinc
study); or no treatment at all, depending on the situation.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In the end, the responses of those in
the treatment group are compared with the responses from the
control group to look for differences that are statistically
significant (unlikely to have occurred just by chance).
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Placebo
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
placebo is a
fake treatment, such as a sugar pill. Placebos are given to the
control group to account for a psychological phenomenon called
the placebo effect, in
which patients receiving a fake treatment still report having a
response, as if it were the real treatment. For example, after
taking a sugar pill a patient experiencing the placebo effect
might say, “Yes, I feel better already,” or “Wow, I
am
 starting to feel a bit dizzy.” By
measuring the placebo effect in the control group, you can
tease out what portion of the reports from the treatment group
were real and what portion were likely due to the placebo
effect. (Experimenters assume that the placebo effect affects
both the treatment and control groups.)</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Blind and double-
blind</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
blind experiment
 is one in which the subjects who are participating in
the study are not aware of whether they’re in the treatment
group or the control group. In the zinc example, the vitamin C
tablets and the zinc tablets would be made to look exactly

alike and patients would not be told which type of pill they were taking. A blind experiment attempts to control for bias on the part of the participants.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A double-blind experiment controls for potential bias on the part of both the patients and the researchers. Neither the patients nor the researchers collecting the data know which subjects received the treatment and which didn’t. So who does know what’s going on as far as who gets what treatment? Typically a third party (someone not otherwise involved in the experiment) puts together the pieces independently. A double-blind study is best, because even though researchers may claim to be unbiased, they often have a special interest in the results — otherwise they wouldn’t be doing the study!</p></blockquote>
<blockquote class="calibre5"><p id="a99" class="calibre6">Surveys (Polls)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A survey (more commonly known as a poll) is a questionnaire; it’s most often used to gather people’s opinions along with some relevant demographic information. Because so many policymakers, marketers, and others want to “get at the pulse of the American public” and find out what the average American is thinking and feeling, many people now feel that they cannot escape the barrage of requests to take part in surveys and polls. In fact, you’ve probably received many requests to participate in surveys, and you may even have become numb to them, simply throwing away surveys received in the mail or saying “no” when asked to participate in a telephone survey.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If done properly, a good survey can really be informative. People use surveys to find out what TV programs Americans (and others) like, how consumers feel about Internet shopping, and whether the United States should allow someone under 35 to become president. Surveys are used by companies to assess the level of satisfaction their customers feel, to find out what products their customers want, and to determine who is buying their products. TV stations use surveys to get instant reactions to news stories and events, and movie producers use them to determine how to end their movies.

</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, if I had to choose one word to
assess the general state of surveys in the media today, I’d say
it’s quantity rather than
quality. In other
words, you’ll find no shortage of bad surveys. But in this book
you find no shortage of good tips and information for
analyzing, critiquing, and understanding survey results, and
for designing your own surveys to do the job right. (To take
off with surveys, head to Chapter 16.)</p>
</blockquote>
<blockquote class="calibre5"><p id="a100" class="calibre6">
Margin of
error</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You’ve probably heard or seen results
like this: “This survey had a margin of error of plus or minus
3 percentage points.” What does this mean? Most surveys (except
a census) are based on information collected from a sample of
individuals, not the entire population. A certain amount of
error is bound to occur — not in the sense of calculation error
(although there may be some of that, too) but in the sense of
sampling error, which
is the error that occurs simply because the researchers aren’t
asking everyone. The margin of error is
supposed to measure the maximum amount by which the sample
results are expected to differ from those of the actual
population. Because the results of most survey questions can be
reported in terms of percentages, the margin of error most
often appears as a percentage, as well.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">How do you interpret a margin of error?
Suppose you know that 51% of people sampled say that they plan
to vote for Ms. Calculation in the upcoming election. Now,
projecting these results to the whole voting population, you
would have to add and subtract the margin of error and give a
range of possible results in order to have sufficient
confidence that you’re bridging the gap between your sample and
the population. Supposing a margin of error of plus or minus 3
percentage points, you would be pretty confident that between
48% (51% – 3%) and 54% (51% + 3%) of the population will vote
for Ms. Calculation in the election, based on the sample

results. In this case, Ms. Calculation may get slightly more or slightly less than the majority of votes and could either win or lose the election. This has become a familiar situation in recent years when the media want to report results on Election Night, but based on early exit polling results, the election is “too close to call.” For more on the margin of error, see Chapter 12.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The margin of error measures accuracy; it does not measure the amount of bias that may be present (find a discussion of bias earlier in this chapter). Results that look numerically scientific and precise don’t mean anything if they were collected in a biased way. </p></blockquote>
<blockquote class="calibre5"><p id="a101" class="calibre6"> Confidence interval</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One of the biggest uses of statistics is to estimate a population parameter using a sample statistic. In other words, use a number that summarizes a sample to help you guesstimate the corresponding number that summarizes the whole population (the definitions of parameter and statistic appear earlier in this chapter). You’re looking for a population parameter in each of the following questions: </p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
What’s the average household income in America? (Population = all households in America; parameter = average household income.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
What percentage of all Americans watched the Academy Awards this year? (Population = all Americans; parameter = percentage who watched the Academy Awards this year.) </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
What’s the average life expectancy of a baby born today? (Population = all babies born today; parameter = average life expectancy.)</p></blockquote><div class="calibre19"> </div>

<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
How effective is this new drug on adults with
Alzheimer’s? (Population = all people who have Alzheimer’s;
parameter = percentage of these people who see improvement when
taking this drug.)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">It’s not possible to find these
parameters exactly; they each require an estimate based on a
sample. You start by taking a random sample from a population
(say a sample of 1,000 households in America) and then finding
the corresponding statistic from that sample (the sample’s mean
household income). Because you know that sample results vary
from sample to sample, you need to add a “plus or minus
something” to your sample results if you want to draw
conclusions about the whole population (all households in
America). This “plus or minus” that you add to your sample
statistic in order to estimate a parameter is the margin of
error.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When you take a sample statistic (such
as the sample mean or sample percentage) and add/subtract a
margin of error, you come up with what statisticians call a
confidence interval.
A confidence interval represents a range of likely values for
the population parameter, based on your sample statistic. For
example, suppose the average time it takes you to drive to work
each day is 35 minutes, with a margin of error of plus or minus
5 minutes. You estimate that the average time to work would be
anywhere from 30 to 40 minutes. This estimate is a confidence
interval.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Some confidence
intervals are wider than others (and wide isn’t good, because
it equals less accuracy). Several factors influence the width
of a confidence interval, such as sample size, the amount of
variability in the population being studied, and how confident
you want to be in your results. (Most researchers are happy
with a 95% level of confidence in their results.) For more on
factors that influence confidence intervals, as well as
instructions for calculating and interpreting confidence
intervals, see Chapter 13.</p></blockquote>
<blockquote class="calibre5"><p id="a102" class="calibre6">
Hypothesis

testing</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Hypothesis test is a term you probably haven’t run across in your everyday dealings with numbers and statistics. But I guarantee that hypothesis tests have been a big part of your life and your workplace, simply because of the major role they play in industry, medicine, agriculture, government, and a host of other areas. Any time you hear someone talking about their study showing a “statistically significant result,” you’re encountering a hypothesis test. (A statistically significant result is one that is unlikely to have occurred by chance. See Chapter 14 for the full scoop.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Basically, a hypothesis test is a statistical procedure in which data are collected from a sample and measured against a claim about a population parameter. For example, if a pizza delivery chain claims to deliver all pizzas within 30 minutes of placing the order, on average, you could test whether this claim is true by collecting a random sample of delivery times over a certain period and looking at the average delivery time for that sample. To make your decision, you must also take into account the amount by which your sample results can change from sample to sample (which is related to the margin of error).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Because your decision is based on a sample and not the entire population, a hypothesis test can sometimes lead you to the wrong conclusion. However, statistics are all you have, and if done properly, they can give you a good chance of being correct. For more on the basics of hypothesis testing, see Chapter 14.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A variety of hypothesis tests are done in scientific research, including t-tests (comparing two population means), paired t-tests (looking at before/after data), and tests of claims made about proportions or means for one or more populations. For specifics on these hypothesis tests, see Chapter 15.</p></blockquote>
<blockquote class="calibre5"><p id="a103" class="calibre6">

```
<span class="calibre21"><span class="bold"><span>p-
values</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Hypothesis tests are used to test the
validity of a claim that is made about a population. This claim
that’s on trial, in essence, is called the </span><span><span
class="calibre16"><span class="italic">null hypothesis. </span>
</span></span><span>The </span><span><span class="calibre16">
<span class="italic">alternative hypothesis </span></span>
</span><span>is the one you would believe if the null
hypothesis is concluded to be untrue. The evidence in the trial
is your data and the statistics that go along with it. All
hypothesis tests ultimately use a </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>-value to weigh the strength of the evidence (what the
data are telling you about the population). The </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>-value is a number between 0 and 1 and interpreted
in the following way:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>A small </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>-value (typically
</span><span><span class="calibre40">≤</span></span><span>
0.05) indicates strong evidence against the null hypothesis, so
you reject it.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>A large </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>-value (&gt; 0.05)
indicates weak evidence against the null hypothesis, so you
fail to reject it.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="italic">p</span>
</span></span><span>-values very close to the cutoff (0.05) are
considered to be marginal (could go either way). Always report
the </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>-value so your
readers can draw their own conclusions.</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For example, suppose a pizza place
```

claims their delivery times are 30 minutes or less on average
but you think it’s more than that. You conduct a hypothesis
test because you believe the null hypothesis, H
_o
, that the mean delivery time is 30 minutes max,
is incorrect. Your alternative hypothesis (H_a
) is that the mean time is greater than 30 minutes. You
randomly sample some delivery times and run the data through
the hypothesis test, and your p
-value turns out to be 0.001, which is much less than
0.05. You conclude that the pizza place is wrong; their
delivery times are in fact more than 30 minutes on average, and
you want to know what they’re gonna do about it! (Of course you
could be wrong by having sampled an unusually high number of
late pizzas just by chance; but whose side am I on?) For more
on p-values, head to
Chapter 14.</p></blockquote>
<blockquote class="calibre5"><p id="a104" class="calibre6">
Statistical
significance</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Whenever data are collected to perform
a hypothesis test, the researcher is typically looking for
something out of the ordinary. (Unfortunately, research that
simply confirms something that was already well known doesn’t
make headlines.) Statisticians measure the amount by which a
result is out of the ordinary using hypothesis tests (see
Chapter 14). They define a statistically
significant result as a result with
a very small probability of happening just by chance, and
provide a number called a
p-value to
reflect that probability (see the previous section on
p
-values).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, if a drug is found to be
more effective at treating breast cancer than the current
treatment is, researchers say that the new drug shows a
statistically significant improvement in the survival rate of
patients with breast cancer. That means that based on their
data, the difference in the overall results from patients on
the new drug compared to those using the old treatment is so
big that it would be hard to say it was just a coincidence.

However, proceed with caution: You can’t say that these results necessarily apply to each individual or to each individual in the same way. For full details on statistical significance, see Chapter 14.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When you hear that a study’s results are statistically significant, don’t automatically assume that the study’s results are important. Statistically significant means the results were unusual, but unusual doesn’t always mean important. For example, would you be excited to learn that cats move their tails more often when lying in the sun than when lying in the shade, and that those results are statistically significant? This result may not even be important to the cat, much less anyone else!</p> </blockquote>
<blockquote class="calibre9"><p class="calibre14">Sometimes statisticians make the wrong conclusion about the null hypothesis because a sample doesn’t represent the population (just by chance). For example, a positive effect that’s experienced by a sample of people who took the new treatment may have just been a fluke; or in the example in the preceding section, the pizza company really was delivering those pizzas on time and you just got an unlucky sample of slow ones. However, the beauty of research is that as soon as someone gives a press release saying that she found something significant, the rush is on to try to replicate the results, and if the results can’t be replicated, this probably means that the original results were wrong for some reason (including being wrong just by chance). Unfortunately, a press release announcing a “major breakthrough” tends to get a lot of play in the media, but follow-up studies refuting those results often don’t show up on the front page.</p> </blockquote>
<blockquote class="calibre9"><p class="calibre22"> One statistically significant result shouldn’t lead to quick decisions on anyone’s part. In science, what most often counts is not a single remarkable study, but a body of evidence that is built up over time, along with a variety of well-designed follow-up studies. Take any major breakthroughs you hear about with a grain of salt and wait until the follow-up work has been done before using the information from a single study to make important decisions in your life. The results may not be replicable, and even if they are, you can’t know if they

necessarily apply to each individual.</p>
</blockquote>
<blockquote class="calibre5"><p id="a105" class="calibre6">
Correlation
versus causation</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Of all of the
misunderstood statistical issues, the one that's perhaps the
most problematic is the misuse of the concepts of correlation
and causation.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Correlation, as a
statistical term, is the extent to which two numerical
variables have a linear relationship (that is, a relationship
that increases or decreases at a constant rate). Following are
three examples of correlated variables:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The number of times a cricket chirps per second is
strongly related to temperature; when it's cold outside, they
chirp less frequently, and as the temperature warms up, they
chirp at a steadily increasing rate. In statistical terms, you
say number of cricket chirps and temperature have a strong
positive correlation.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The number of crimes (per capita) has often been found to
be related to the number of police officers in a given area.
When more police officers patrol the area, crime tends to be
lower, and when fewer police officers are present in the same
area, crime tends to be higher. In statistical terms we say the
number of police officers and the number of crimes have a
strong negative correlation.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The consumption of ice cream (pints per person) and the
number of murders in New York are positively correlated. That
is, as the amount of ice cream sold per person increases, the
number of murders increases. Strange but true!
</p></blockquote><div class="calibre19"> </div>

```
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>But correlation as a statistic isn’t
able to explain </span><span><span class="calibre16"><span
class="italic">why</span></span></span><span> or </span><span>
<span class="calibre16"><span class="italic">how </span></span>
</span><span>the relationship between two variables, </span>
<span><span class="calibre16"><span class="italic">x</span>
</span></span><span> and </span><span><span class="calibre16">
<span class="italic">y, </span></span></span><span>exists; only
that it does exist.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span><span class="calibre16"><span
class="italic">Causation </span></span></span><span>goes a step
further than correlation, stating that a change in the value of
the </span><span><span class="calibre16"><span
class="italic">x</span></span></span><span> variable </span>
<span><span class="calibre16"><span class="italic">will
cause</span></span></span><span> a change in the value of the
</span><span><span class="calibre16"><span class="italic">y
</span></span></span><span>variable. Too many times in
research, in the media, or in the public consumption of
statistical results, that leap is made when it shouldn’t be.
For instance, you can’t claim that consumption of ice cream
</span><span><span class="calibre16"><span
class="italic">causes </span></span></span><span>an increase in
murder rates just because they are correlated. In fact, the
study showed that temperature was positively correlated with
both ice cream sales and murders. (For more on correlation and
causation, see Chapter 18.) When can you make the causation
leap? The most compelling case is when a well-designed
experiment is conducted that rules out other factors that could
be related to the outcomes (see Chapter 17 for information on
experiments showing cause-and-effect).</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> You may find
yourself wanting to jump to a cause-and-effect relationship
when a correlation is found; researchers, the media, and the
general public do it all the time. However, before making any
conclusions, look at how the data were collected and/or wait to
see if other researchers are able to replicate the results (the
first thing they try to do after someone else’s “groundbreaking
result” hits the airwaves).</span></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a106"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
```

```
always !important; white-space: pre-wrap !important">

<a href="#a89" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a88" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a81" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a80"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a83" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a82" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a85" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a84" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a87" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a86" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a79" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a96"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a97" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a94" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a95" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a92" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a93" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a90" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a91" style="min-width: 10px !important; min-height: 10px
!important; border: solid 1px !important;"> </a> <a href="#a98"
style="min-width: 10px !important; min-height: 10px !important;
border: solid 1px !important;"> </a> <a href="#a99" style="min-
width: 10px !important; min-height: 10px !important; border:
solid 1px !important;"> </a> <a href="#a105" style="min-width:
10px !important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a104" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a107" style="min-width: 10px
```

```
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> <a href="#a106" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a>
<a href="#a101" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a100" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a103" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a102" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a110">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part II</span>
</span></span></span></p>
<p id="a108" class="calibre10"><span class="calibre11"><span
class="bold"><span>Number-Crunching Basics</span></span></span>
</p>
<div class="calibre13">
<img alt="9780470911082 pp0201.eps" src="images/00019.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Number crunching: It’s a dirty job, but
somebody has to do it. Why not let it be you? Even if you
aren’t a numbers person and calculations aren’t your thing, the
step-by-step approach in this part may be just what you need to
boost your confidence in doing and really understanding
statistics.</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">In this part, you get down to the basics of
number crunching, from making and interpreting charts and
graphs to cranking out and understanding means, medians,
standard deviations, and more. You also develop important
skills for critiquing someone else’s statistical information
and getting at the real truth behind the data.</span></p>
</blockquote>
</div>
</div>  <div class="mbppagebreak" id="a109"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a110" style="min-width: 10px !important; min-height:
```

10px !important; border: solid 1px !important;">
      </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a128">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 5</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a111" class="calibre10"><span class="calibre11"><span
class="bold"><span>Means, Medians, and More</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Summarizing data effectively</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Interpreting commonly used statistics</span></span></p>
</blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Realizing what statistics do and don’t say</span></span>
</p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>E</span><span>very data set has a
story, and if statistics are used properly, they do a good job
of uncovering and reporting that story. Statistics that are
improperly used can tell a different story, or only part of it,
so knowing how to make good decisions about the information
you’re given is very important.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>A </span><span><span class="calibre16">
<span class="italic">descriptive statistic </span></span>
```

(or statistic for short) is a number that summarizes or describes some characteristic about a set of data. In this chapter, you see some of the most common descriptive statistics and how they are used, and you find out how to calculate them, interpret them, and put them together to get a good picture of a data set. You also find out what these statistics say and what they don't say about the data.</p></blockquote>
<p id="a112" class="calibre6">Summing Up Data with Descriptive Statistics</p>
<blockquote class="calibre9"><p class="calibre14">Descriptive statistics take a data set and boil it down to a set of basic information. Summarized data are often used to provide people with information that is easy to understand and that helps answer their questions. Picture your boss coming to you and asking, "What's our client base like these days, and who's buying our products?" How would you like to answer that question — with a long, detailed, and complicated stream of numbers that are sure to glaze her eyes over? Probably not. You want clean, clear, and concise statistics that sum up the client base for her, so that she can see how brilliant you are and then send you off to collect even more data to see how she can include more people in the client base. (That's what you get for being efficient.) </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Summarizing data has other purposes, as well. After all the data have been collected from a survey or some other kind of study, the next step is for the researcher to try to make sense out of the data. Typically, the first step researchers take is to run some basic statistics on the data to get a rough idea about what's happening in it. Later in the process, researchers can do more analyses to formulate or test claims made about the population the data came from, estimate certain characteristics about the population (like the mean), look for links between variables they measured, and so on. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Another big part of research is reporting the results, not only to your peers, but also to the media and the general public. Although a researcher's peers may be anxiously waiting to hear about all the complex analyses that were done on a data set, the general public is neither ready for nor interested in that. What does the public want? Basic information. Statistics that make a point clearly and concisely are usually used to relay information to the media

and to the public.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If you really
need to learn more from data, a quick statistical overview
isn’t enough. In the statistical world, less is not more, and
sometimes the real story behind the data can get lost in the
shuffle. To be an informed consumer of statistics, you need to
think about which statistics are being reported, what these
statistics really mean, and what information is missing. This
chapter focuses on these issues.</p></blockquote>
<p id="a113" class="calibre6">Crunching Categorical Data: Tables and
Percents</p>
<blockquote class="calibre9"><p class="calibre14">Categorical data
(also known as qualitative data)
capture qualities or characteristics about the individual, such
as a person’s eye color, gender, political party, or opinion on
some issue (using categories such as Agree, Disagree, or No
opinion). Categorical data tend to fall into groups or
categories pretty naturally. “Political party,” for example,
typically has four groups in the United States: Democrat,
Republican, Independent, and Other. Categorical data often come
from survey data, but they can also be collected in
experiments. For example, in an experimental test of a new
medical treatment, researchers may use three categories to
assess the outcome of the experiment: Did the patient get
better, worse, or stay the same while undergoing the treatment?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Categorical data are often summarized
by reporting the percentage of individuals falling into each
category. For example, pollsters may report political
affiliation statistics by giving the percentage of Republicans,
Democrats, Independents, and Others. To calculate the
percentage of individuals in a certain category, find the
number of individuals in that category, divide by the total
number of people in the study, and then multiply by 100%. For
example, if a survey of 2,000 teenagers included 1,200 females
and 800 males, the resulting percentages would be (1,200 ÷
2,000) ∗ 100% = 60% female and (800
÷ 2,000) ∗ 100% = 40% male.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">You can break down categorical data
further by creating something called two-way tables.
Two-way
tables (also called
crosstabs
) are tables with rows and columns. They
summarize the information from two categorical variables at
once, such as gender and political party, so you can see (or
easily calculate) the percentage of individuals in each
combination of categories and use them to make comparisons
between groups.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, if you had data about the
gender and political party of your respondents, you would be
able to look at the percentage of Republican females,
Republican males, Democratic females, Democratic males, and so
on. In this example, the total number of possible combinations
in your table would be 2 ∗ 4 = 8, or
the total number of gender categories times the total number of
party affiliation categories. (See Chapter 19 for the full
scoop, and then some, on two-way tables.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The U.S. government calculates and
summarizes loads of categorical data using crosstabs. Typical
age and gender data, reported by the U.S. Census Bureau for a
survey conducted in 2009, are shown in Table 5-1. (Normally age
would be considered a numerical variable, but the way the U.S.
government reports it, age is broken down into categories,
making it a categorical variable.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 5-1a" src="images/00020.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 5-1b" src="images/00021.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You can examine many different facets
of the U.S. population by looking at and working with different
numbers from Table 5-1. For example, looking at gender, you
notice that women slightly outnumber men — the population in
2009 was 50.67% female (divide total number of females by total
population size and multiply by 100%) and 49.33% male (divide
total number of males by total population size and multiply by
100%). You can also look at age: The percentage of the entire
population that is under 5 years old was 6.94% (divide the

total number under age 5 by the total population size and multiply by 100%). The largest group belongs to the 45–49 year olds, who made up 7.44% of the population.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Next, you can explore a possible relationship between gender and age by comparing various parts of the table. You can compare, for example, the percentage of females to males in the 80-and-over age group. Because these data are reported in 5-year increments, you have to do a little math in order to get your answer, though. The percentage of the population that’s female and aged 80 and above (looking at column 7 of Table 5-1) is 2.27% + 1.54% + 0.69% + 0.21% + 0.04% = 4.75%. The percentage of males aged 80 and over (looking at column 5 of Table 5-1) is 1.52% + 0.84% + 0.28% + 0.05% + 0.01% = 2.70%. This shows that the 80-and-over age group for the females is about 76% larger than the males (because [4.75 – 2.70] ÷ 2.70 = 0.76).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">These data confirm the widely accepted notion that women tend to live longer than men. However, the gap between men and women is narrowing over time. According to the U.S. Census Bureau, back in 2001 the percentage of women who were 80 years old and over was 4.36, compared to 2.31 for the men. The females in this age group outnumbered the males by a whopping 89% back in 2001 (note that [4.36 – 2.31] ÷ 2.31 = 0.89).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> After you have the crosstabs that show the breakdown of two categorical variables, you can conduct hypothesis tests to determine whether a significant relationship or link between the two variables exists, taking into account the fact that data vary from sample to sample. Chapter 14 gives you all the details on hypothesis tests.</p></blockquote>
<p id="a114" class="calibre6">Measuring the Center with Mean and Median</p>
<blockquote class="calibre9"><p class="calibre14">With numerical data,
 measurable characteristics such as height, weight, IQ, age, or income are represented by numbers that make sense within the context of the problem (for example in units of feet, dollars, or people). Because the data have numerical meaning, you can summarize them in more ways than is possible with categorical data. The most common way to summarize a

numerical data set is to describe where the center is. One way of thinking about what the center of a data set means is to ask, “What’s a typical value?” Or, “Where is the middle of the data?” The center of a data set can actually be measured in different ways, and the method chosen can greatly influence the conclusions people make about the data. This section hits on measures of center.</p></blockquote>
<blockquote class="calibre5"><p id="a115" class="calibre6">
Averaging out to the mean</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">NBA players make a lot of money, right? You often hear about players like Kobe Bryant or LeBron James who make tens of millions of dollars a year. But is that what the typical NBA player makes? Not really (although I don’t exactly feel sorry for the others, given that they still make more money than most of us will ever make). Tens of millions of dollars is the kind of money you can command when you are a superstar among superstars, which is what these elite players are.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So how much money does the typical NBA player make? One way to answer this is to look at the average (the most commonly used statistic of all time).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The average
, also called the mean
 of a data set, is denoted . The formula for finding the mean is:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">where each value in the data set is denoted by an x with a subscript
i that goes from 1 (the first number) to
n (the last number).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

```
class="calibre15"><span>Here’s how you calculate the mean of a
data set:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">1. Add up all the numbers in the data set.</span>
</span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Divide by the number of numbers in the data
set, </span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">n</span></span></span></span>
<span><span class="calibre16"><span class="bold">.</span>
</span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> The mean I
discuss here applies to a sample of data and is technically
called the </span><span><span class="calibre16"><span
class="italic">sample mean.</span></span></span><span> The mean
of an</span><span><span class="calibre16"><span class="italic">
</span></span></span><span>entire population</span><span><span
class="calibre16"><span class="italic">
</span></span></span><span>of data is denoted with the Greek
letter </span><span>μ</span><span> and is called the </span>
<span><span class="calibre16"><span class="italic">population
mean.</span></span></span><span> It’s found by summing up all
the values in the population and dividing by the population
size, denoted </span><span><span class="calibre16"><span
class="italic">N</span></span></span><span> (to distinguish it
from a sample size, </span><span><span class="calibre16"><span
class="italic">n</span></span></span><span>). Typically the
population mean is unknown, and you use a sample mean to
estimate it (plus or minus a margin of error; see all the
details in Chapter 13).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For example, player salary data for the
13 players on the 2010 NBA Champion Los Angeles Lakers is shown
in Table 5-2.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table 5-2a" src="images/00024.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table 5-2b" src="images/00025.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The mean of all the salaries on this
```

team is $91,378,064 ÷ 13 = $7,029,082. That’s a pretty nice average salary, isn’t it? But notice that Kobe Bryant really stands out at the top of this list, and he should — his salary was the second highest in the entire league that season (just behind Tracy McGrady). If you remove Kobe from the equation (literally), the average salary of all the Lakers players besides Kobe becomes $68,343,689 ÷ 12 = $5,695,307 — a difference of around 1.3 million.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">This new mean is still a hefty amount, but it’s significantly lower than the mean salary of all the players including Kobe. (Fans would tell you that this reflects his importance to the team, and others would say no one is worth that much money; this issue is but the tip of the iceberg of the never-ending debates that sports fans — me included — love to have about statistics.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Bottom line: The mean doesn’t always tell the whole story. In some cases it may be a bit misleading, and this is one of those cases. That’s because every year a few top-notch players (like Kobe) make much more money than anybody else, and their salaries pull up the overall average salary.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Numbers in a data set that are extremely high or extremely low compared to the rest of the data are called outliers
. Because of the way the average is calculated, high outliers tend to drive the average upward (as Kobe’s salary did in the preceding example). Low outliers tend to drive the average downward.</p></blockquote>
<blockquote class="calibre5"><p id="a116" class="calibre6">
Splitting your data down the median</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Remember in school when you took an exam, and you and most of the rest of the class did badly, but a couple of nerds got 100? Remember how the teacher didn’t curve the scores to reflect the poor performance of most of the class? Your teacher was probably using the average, and the average in that case didn’t really represent what statisticians might consider the best measure of center for the students’ scores.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What can you report, other than the

average, to show what the salary of a “typical” NBA player would be or what the test score of a “typical” student in your class was? Another statistic used to measure the center of a data set is called the median. The median is still an unsung hero of statistics in the sense that it isn’t used nearly as often as it should be, although people are beginning to report it more nowadays.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The median of a data set is the value that lies exactly in the middle when the data have been ordered. It’s denoted in different ways; some people use M and some use . Here are the steps for finding the median of a data set:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Order the numbers from smallest to largest.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. If the data set contains an odd number of numbers, choose the one that is exactly in the middle. You’ve found the median.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. If the data set contains an even number of numbers, take the two numbers that appear in the middle and average them to find the median.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">The salaries for the Los Angeles Lakers during the 2009–2010 season (refer to Table 5-2) are ordered from smallest (at the bottom) to largest (at the top). Because the list contains the names and salaries of 13 players, the middle salary is the seventh one from the bottom: Derek Fisher, who earned $5.048 million that season from the Lakers. Derek is at the median.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> This median salary ($5.048 million) is well below the average of $7.029 million for the 2009–2010 Lakers team. Notice that only 4

players of the 13 earned more than the average Lakers salary of $7.029 million. Because the average includes outliers (like the salary of Kobe Bryant), the median salary is more representative of center for the team salaries. The median isn’t affected by the salaries of those players who are way out there on the high end the way the average is.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Note:
 By the way, the lowest Lakers’ salary for the 2009–2010 season was $959,111 — a lot of money by most people’s standards, but peanuts compared to what you imagine when you think of an NBA player’s salary!</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The U.S. government most often uses the median to represent the center with respect to income data again because the median is not affected by outliers. For example, the U.S. Census Bureau reported that in 2008, the median household income was $50,233 while the mean was found to be $68,424. That’s quite a difference!</p>
</blockquote>
<blockquote class="calibre5"><p id="a117" class="calibre6">
Comparing means and medians: Histograms</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Sometimes the mean versus median debate can get quite interesting. Suppose you’re part of an NBA team trying to negotiate salaries. If you represent the owners, you want to show how much everyone is making and how much money you’re spending, so you want to take into account those superstar players and report the average. But if you’re on the side of the players, you would want to report the median, because that’s more representative of what the players in the middle are making. Fifty percent of the players make a salary above the median, and 50 percent make a salary below the median. To sort it all out, it’s best to find and compare both the mean and the median. A graph showing the shape of the data is a great place to start.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> One of the graphs you can make to illustrate the shape of numerical data (how many values are close to/far from the mean, where the center is, how many outliers there might be) is a histogram. A <span

class="italic">histogram is a graph
that organizes and displays numerical data in picture form,
showing groups of data and the number or percentage of the data
that fall into each group. It gives you a nice snapshot of the
data set. (See Chapter 7 for more information on histograms and
other types of data displays.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Data sets can have many different
possible shapes; here is a sampling of three shapes that are
commonly discussed in introductory statistics courses:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If most of the data are on the left side of the histogram
but a few larger values are on the right, the data are said to
be skewed to the right.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Histogram A in Figure 5-1 shows an
example of data that are skewed to the right. The few larger
values bring the mean upwards but don’t really affect the
median. So when data are skewed right, the mean is larger than
the median. An example of such data
is NBA salaries.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If most of the data are on the right, with a few smaller
values showing up on the left side of the histogram, the data
are skewed to the left.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Histogram B in Figure 5-1 shows an
example of data that are skewed to the left. The few smaller
values bring the mean down, and again the median is minimally
affected (if at all). An example of skewed-left data is the
amount of time students use to take an exam; some students
leave early, more of them stay later, and many stay until the
bitter end (some would stay forever if they could!). When data
are skewed left, the mean is smaller than the median
.</p></blockquote><div
class="calibre19"> </div>

<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the data are
symmetric, they
have about the same shape on either side of the middle. In
other words, if you fold the histogram in half, it looks about
the same on both sides.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Histogram C in Figure 5-1 shows an
example of symmetric data in a histogram. With symmetric data,
the mean and median are close together.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> By looking at Histogram A in Figure 5-
1 (whose shape is skewed right), you can see that the “tail” of
the graph (where the bars are getting shorter) is to the right,
while the “tail” is to the left in Histogram B (whose shape is
skewed left). By looking at the direction of the tail of a
skewed distribution, you determine the direction of the
skewness. Always add the direction when describing a skewed
distribution.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 5-1: A) Data skewed right; B)
data skewed left; and C) symmetric data.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0501.eps"
src="images/00027.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Histogram C is symmetric (it has about
the same shape on each side). However, not all symmetric data
has a bell shape like Histogram C does. As long as the shape is
approximately the same on both sides, then you say that the
shape is symmetric.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The average (or
mean) of a data set is affected by outliers, but the median is
not. In statistical lingo, if a statistic is not affected by a
certain characteristic of the data (such as outliers, or
skewness), then you say that statistic is resistant
<span> to that characteristic. In this case the median

```
is resistant to outliers; the mean is not. If someone reports
the average value, also ask for the median so that you can
compare the two statistics and get a better feel for what’s
actually going on in the data and what’s truly typical.</span>
</span></p></blockquote>
<p id="a118" class="calibre6"><span class="calibre17"><span
class="bold"><span>Accounting for Variation</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Variation always exists in a data set,
regardless of which characteristics you’re measuring, because
not every individual is going to have the same exact value for
every variable. Variation is what makes the field of statistics
what it is. For example, the price of homes varies from house
to house, from year to year, and from state to state. The
amount of time it takes you to get to work varies from day to
day. The trick to dealing with variation is to be able to
measure that variation in a way that best captures it.</span>
</span></p></blockquote>
<blockquote class="calibre5"><p id="a119" class="calibre6">
<span class="calibre21"><span class="bold"><span>Reporting the
standard deviation</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>By far the most common measure of
variation for numerical data is the standard deviation. The
</span><span><span class="calibre16"><span
class="italic">standard deviation</span></span></span><span>
measures how concentrated the data are around the mean; the
more concentrated, the smaller the standard deviation. It’s not
reported nearly as often as it should be, but when it is, you
often see it in parentheses: (</span><span><span
class="calibre16"><span class="italic">s</span></span></span>
<span> = 2.68).</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Calculating standard
deviation</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The formula for the sample standard
deviation of a data set (</span><span><span class="calibre16">
<span class="italic">s</span></span></span><span>) is</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq05004.eps"
src="images/00028.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To calculate </span><span><span
class="calibre16"><span class="italic">s</span></span></span>
```

, do the following steps:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Find the average of the data set,
<img alt="9780470911082-eq05005.eps"
src="images/00029.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Take each number in the data set (

x) and subtract the mean
from it to get <img alt="9780470911082-
eq05006.eps" src="images/00030.jpg" class="calibre2"/>
.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Square each of the differences,
<img alt="9780470911082-eq05007.eps"
src="images/00031.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Add up all of the results from Step 3 to get
the sum of squares: <img
alt="9780470911082-eq05008.eps" src="images/00032.jpg"
class="calibre2"/>.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Divide the sum of squares (found in Step 4) by
the number of numbers in the data set minus one; that is,
(n
 – 1). Now you
have:</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq05009.eps"
src="images/00033.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 6. Take the square root to get
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq05010.eps"
src="images/00034.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"> which is the sample standard deviation,

s. Whew!
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> At the end of
Step 5 you have found a statistic called the sample variance,
 denoted by s
²
. The variance is another way to measure
variation in a data set; its downside is that it’s in square
units. If your data are in dollars, for example, the variance
would be in square dollars — which makes no sense. That’s why
we proceed to Step 6. Standard deviation has the same units as
the original data.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Look at the following small example:
Suppose you have four quiz scores: 1, 3, 5, and 7. The mean is
16 ÷ 4 = 4 points. Subtracting the mean from each number, you
get (1 – 4) = –3, (3 – 4) = –1, (5 – 4) = +1, and (7 – 4) = +3.
Squaring each of these results, you get 9, 1, 1, and 9. Adding
these up, the sum is 20. In this example, n
= 4, and therefore
n – 1 = 3, so
you divide 20 by 3 to get 6.67. The units here are “points
squared,” which obviously makes no sense. Finally, you take the
square root of 6.67, to get 2.58. The standard deviation for
these four quiz scores is 2.58 points.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Because calculating the standard
deviation involves many steps, in most cases you have a
computer calculate it for you. However, knowing how to
calculate the standard deviation helps you better interpret
this statistic and can help you figure out when the statistic

may be wrong.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Statisticians
divide by n – 1 instead of
by n
in the formula for s
so the results have nicer properties that operate on a
theoretical plane that’s beyond the scope of this book (not the
Twilight Zone but
close; trust me, that’s more than you want to know about
that!
).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The standard
deviation of an entire population of data
 is denoted with the Greek letter
σ. When I use the term standard deviation,
 I mean s,
 the sample standard deviation. (When I refer to the
population standard deviation, I let you know.)
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Interpreting standard
deviation</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Standard deviation can be difficult to
interpret as a single number on its own. Basically, a small
standard deviation means that the values in the data set are
close to the mean of the data set, on average, and a large
standard deviation means that the values in the data set are
farther away from the mean, on average.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A small standard deviation can be a
goal in certain situations where the results are restricted,
for example, in product manufacturing and quality control. A
particular type of car part that has to be 2 centimeters in
diameter to fit properly had better not have a very big
standard deviation during the manufacturing process. A big
standard deviation in this case would mean that lots of parts

end up in the trash because they don’t fit right; either that or the cars will have problems down the road.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">But in situations where you just observe and record data, a large standard deviation isn’t necessarily a bad thing; it just reflects a large amount of variation in the group that is being studied. For example, if you look at salaries for everyone in a certain company, including everyone from the student intern to the CEO, the standard deviation may be very large. On the other hand, if you narrow the group down by looking only at the student interns, the standard deviation is smaller, because the individuals within this group have salaries that are less variable. The second data set isn’t better, it’s just less variable.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Similar to the mean, outliers affect the standard deviation (after all, the formula for standard deviation includes the mean). In the NBA salaries example, the salaries of the L.A. Lakers in the 2009–2010 season (shown in Table 5-2) range from the highest, $23,034,375 (Kobe Bryant) down to $959,111 (Didier Ilunga-Mbenga and Josh Powell). Lots of variation, to be sure! The standard deviation of the salaries for this team turns out to be $6,567,405; it’s almost as large as the average. However, as you may guess, if you remove Kobe Bryant’s salary from the data set, the standard deviation decreases because the remaining salaries are more concentrated around the mean. The standard deviation becomes $4,671,508.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Watch for the units when determining whether a standard deviation is large. For example, a standard deviation of 2 in units of years is equivalent to a standard deviation of 24 in units of months. Also look at the value of the mean when putting standard deviation into perspective. If the average number of Internet newsgroups that a user posts to is 5.2 and the standard deviation is 3.4, that’s a lot of variation, relatively speaking. But if you’re talking about the age of the newsgroup users where the mean is 25.6 years, that same standard deviation of 3.4 would be comparatively smaller.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Understanding properties of standard deviation</p>
</blockquote>

<blockquote class="calibre9"><p class="calibre14">Here are some properties that can help
you when interpreting a standard deviation:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The standard deviation can never be a negative number,
due to the way it’s calculated and the fact that it measures a
distance (distances are never negative numbers).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The smallest possible value for the standard deviation is
0, and that happens only in contrived situations where every
single number in the data set is exactly the same (no
deviation).</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The standard deviation is affected by outliers (extremely
low or extremely high numbers in the data set). That’s because
the standard deviation is based on the distance

from the mean.
 And remember, the mean is also affected by
outliers.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The standard deviation has the same units as the original
data.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre6">Lobbying for standard
deviation</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard deviation is a commonly
used statistic, but it doesn’t often get the attention it
deserves. Although the mean and median are out there in common
sight in the everyday media, you rarely see them accompanied by
any measure of how diverse that data set was, and so you are
getting only part of the story. In fact, you could be missing
the most interesting part of the story.</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre14">Without standard deviation, you can’t
get a handle on whether the data are close to the average (as
are the diameters of car parts that come off of a conveyor belt
when everything is operating correctly) or whether the data are
spread out over a wide range (as are house prices and income
levels in the U.S.).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example if someone told you that
the average starting salary for someone working at Company
Statistix is $70,000, you may think, “Wow! That’s great.” But
if the standard deviation for starting salaries at Company
Statistix is $20,000, that’s a lot of variation in terms of how
much money you can make, so the average starting salary of
$70,000 isn’t as informative in the end, is it?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">On the other hand, if the standard
deviation was only $5,000, you would have a much better idea of
what to expect for a starting salary at that company. Which is
more appealing? That’s a decision each person has to make;
however it’ll be a much more informed decision once you realize
standard deviation matters.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Without the standard deviation, you
can’t compare two data sets effectively. Suppose two sets of
data have the same average; does that mean that the data sets
must be exactly the same? Not at all. For example, the data
sets 199, 200, 201; and 0, 200, 400 both have the same average
(200) yet they have very different standard deviations. The
first data set has a very small standard
deviation (s=1) compared to the
second data set (s=200).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">References to the standard deviation
may become more commonplace in the media as more and more
people (like you, for example) discover what the standard
deviation can tell them about a set of results and start asking
for it. In your career, you are likely to see the standard
deviation reported and used as well.</p>
</blockquote>
<blockquote class="calibre5"><p id="a120" class="calibre6">
Being out of

range</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The range is another statistic that
some folks use to measure diversity in a data set. The
range
 is the largest value in the data set minus
the smallest value in the data set. It’s easy to find; all you
do is put the numbers in order (from smallest to largest) and
do a quick subtraction. Maybe that’s why the range is used so
often; it certainly isn’t because of its interpretative value.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The range of a
data set is almost meaningless. It depends on only two numbers
in the data set, both of which may reflect extreme values
(outliers). My advice is to ignore the range and find the
standard deviation, which is a more informative measure of the
variation in the data set because it involves all the values.
Or you can also calculate another statistic called the
interquartile range,
which is similar to the range with an important difference — it
eliminates outlier and skewness issues by only looking at the
middle 50% of the data and finding the range for those values.
The section “Exploring interquartile range” at the end of this
chapter gives you more details.</p></blockquote>
<p id="a121" class="calibre6">Examining the Empirical Rule (68-95-99.7)
</p>
<blockquote class="calibre9"><p class="calibre14">Putting a measure of center (such as
the mean or median) together with a measure of variation (such
as standard deviation or interquartile range) is a good way to
describe the values in a population. In the case where the data
are in the shape of a bell curve (that is, they have a normal
distribution; see Chapter 9), the population mean and standard
deviation are the combination of choice, and a special rule
links them together to get some pretty detailed information
about the population as a whole.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The Empirical Rule
 says that if a population has a normal
distribution with population mean μ
and standard deviation σ, then:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span

```
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>About 68% of the values lie within 1 standard deviation
of the mean (or between the mean minus 1 times the standard
deviation, and the mean plus 1 times the standard deviation).
In statistical notation, this is represented as </span>
<span>μ</span><span>
</span><span><span class="calibre40">±</span></span><span>
1</span><span>σ</span><span>.</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>About 95% of the values lie within 2 standard deviations
of the mean (or between the mean minus 2 times the standard
deviation, and the mean plus 2 times the standard deviation).
The statistical notation for this is </span><span>μ</span>
<span>
</span><span><span class="calibre40">±</span></span><span>
2</span><span>σ</span><span>.</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>About 99.7% of the values lie within 3 standard
deviations of the mean (or between the mean minus 3 times the
standard deviation and the mean plus 3 times the standard
deviation). Statisticians use the following notation to
represent this: </span><span>μ</span><span>
</span><span><span class="calibre40">±</span></span><span>
3</span><span>σ</span><span>.</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> The Empirical Rule is also known as
the </span><span><span class="calibre16"><span
class="italic">68-95-99.7 Rule,</span></span></span><span> in
correspondence with those three properties. It’s used to
describe a population rather than a sample, but you can also
use it to help you decide whether a sample of data came from a
normal distribution. If a sample is large enough and you can
see that its histogram looks close to a bell-shape, you can
check to see whether the data follow the 68-95-99.7 percent
specifications. If yes, it’s reasonable to conclude the data
came from a normal distribution. This is huge because the
normal distribution has lots of perks, as you can see in
Chapter 9.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
```

class="calibre15">Figure 5-2 illustrates all three
components of the Empirical Rule.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The reason that so many (about 68%) of
the values lie within 1 standard deviation of the mean in the
Empirical Rule is because when the data are bell-shaped, the
majority of the values are mounded up in the middle, close to
the mean (as Figure 5-2 shows).</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 5-2: The Empirical Rule (68%,
95%, and 99.7%).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9784070911082-fg0502.eps"
src="images/00035.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Adding another standard deviation on
either side of the mean increases the percentage from 68 to 95,
which is a big jump and gives a good idea of where “most” of
the data are located. Most researchers stay with the 95% range
(rather than 99.7%) for reporting their results, because
increasing the range to 3 standard deviations on either side of
the mean (rather than just 2) doesn’t seem worthwhile, just to
pick up that last 4.7% of the values.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The Empirical
Rule tells you about what percentage of values are within a
certain range of the mean, and I need to stress the word

about. These results are
approximations only, and they only apply if the data follow a
normal distribution. However, the Empirical Rule is an
important result in statistics because the concept of “going
out about two standard deviations to get about 95% of the
values” is one that you see mentioned often with confidence
intervals and hypothesis tests (see Chapters 13 and 14).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s an example of using the
Empirical Rule to better describe a population whose values
have a normal distribution: In a study of how people make
friends in cyberspace using newsgroups, the age of the users of
an Internet newsgroup was reported to have a mean of 31.65

years, with a standard deviation of 8.61 years. Suppose the data were graphed using a histogram and were found to have a bell-shaped curve similar to what’s shown in Figure 5-2.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">According to the Empirical Rule, about 68% of the newsgroup users had ages within 1 standard deviation (8.61 years) of the mean (31.65 years). So about 68% of the users were between ages 31.65 – 8.61 years and 31.65 + 8.61 years, or between 23.04 and 40.26 years. About 95% of the newsgroup users were between the ages of 31.65 – 2(8.61), and 31.65 + 2(8.61), or between 14.43 and 48.87 years. Finally, about 99.7% of the newsgroup users’ ages were between 31.65 – 3(8.61) and 31.65 + 3(8.61), or between 5.82 and 57.48 years.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This application of the rule gives you a much better idea about what’s happening in this data set than just looking at the mean, doesn’t it? As you can see, the mean and standard deviation used together add value to your results; plugging these values into the Empirical Rule allows you to report ranges for “most” of the data yourself.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Remember, the condition for being able to use the Empirical Rule is that the data have a normal distribution. If that’s not the case (or if you don’t know what the shape actually is), you can’t use it. To describe your data in these cases, you can use percentiles, which represent certain cutoff points in the data (see the later section “Gathering a five-number summary”).</p></blockquote>
<p id="a122" class="calibre6">Measuring Relative Standing with Percentiles</p>
<blockquote class="calibre9"><p class="calibre14">Sometimes the precise values of the mean, median, and standard deviation just don’t matter, and all you are interested in is where you stand compared to the rest of the herd. In this situation, you need a statistic that reports relative standing, and that statistic is called a percentile. The kth<span><span class="calibre16"><span

```
class="italic">percentile </span></span></span><span>is a
number in the data set that splits the data into two pieces:
The lower piece contains </span><span><span class="calibre16">
<span class="italic">k</span></span></span><span> percent of
the data, and the upper piece contains the rest of the data
(which amounts to [100 – </span><span><span class="calibre16">
<span class="italic">k</span></span></span><span>] percent,
because the total amount of data is 100%). </span><span><span
class="calibre16"><span class="bold"><span class="italic">Note:
</span></span></span></span><span>
</span><span><span class="calibre16"><span
class="italic">k</span></span></span><span> is any number
between 1 and 100.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> The median is the 50th percentile: The
point in the data where 50% of the data fall below that point,
and 50% fall above it.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In this section, you find out how to
calculate, interpret, and put together percentiles to help you
uncover the story behind a data set.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a123" class="calibre6">
<span class="calibre21"><span class="bold"><span>Calculating
percentiles</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To calculate the </span><span><span
class="calibre16"><span class="italic">k</span></span></span>
<span><span class="calibre41"><sup class="calibre4">th</sup>
</span></span><span> percentile (where </span><span><span
class="calibre16"><span class="italic">k</span></span></span>
<span> is any number between one and one hundred), do the
following steps:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 1. Order all the numbers in the data set from
smallest to largest.</span></span></span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Multiply </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">k</span></span></span></span><span><span
class="calibre16"><span class="bold"> percent times the total
number of numbers, </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">n</span></span></span></span><span><span
```

class="calibre16">.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3a. If your result from Step 2 is a whole number,
go to Step 4. If the result from Step 2 is not a whole number,
round it up to the nearest whole number and go to Step 3b.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3b. Count the numbers in your data set from left
to right (from the smallest to the largest number) until you
reach the value indicated by Step 3a. The corresponding value
in your data set is the kth
 percentile.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Count the numbers in your data set from left
to right until you reach the one indicated by Step 2. The
k
th

percentile is the average of that corresponding value in your
data set and the value that directly follows it.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you have 25 test
scores, and in order from lowest to highest they look like
this: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78,
79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. To find the 90th
percentile for these (ordered) scores, start by multiplying 90%
times the total number of scores, which gives 90%
∗ 25 = 0.90 ∗ 25
= 22.5. Rounding up to the nearest whole number, you get 23.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Counting from left to right (from the
smallest to the largest number in the data set), you go until
you find the 23rd number in the data set. That number is 98,
and it’s the 90th percentile for this data set.

</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now say you want to find the 20th percentile. Start by taking 0.20 ∗ 25 = 5; this is a whole number, so proceed from Step 3a to Step 4, which tells us the 20th percentile is the average of the 5th and 6th numbers in the ordered data set (62 and 66). The 20th percentile then comes to (62 + 66) ÷ 2 = 64. The median (the 50th percentile) for the test scores is the 13th score: 77. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> There is no single definitive formula for calculating percentiles. The formula here is designed to make finding the percentile easier and more intuitive, especially if you're doing the work by hand; however, other formulas are used when you're working with technology. The results you get using various methods may differ but not by much.</p></blockquote>
<blockquote class="calibre5"><p id="a124" class="calibre6">Interpreting percentiles</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Percentiles report the relative standing of a particular value within a data set. If that's what you're most interested in, the actual mean and standard deviation of the data set are not important, and neither is the actual data value. What's important is where you stand — not in relation to the mean, but in relation to everyone else: That's what a percentile gives you.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, in the case of exam scores, who cares what the mean is, as long as you scored better than most of the class? Who knows, it may have been an impossible exam and 40 points out of 100 was a great score (that happened to me in an advanced math class once; heaven forbid this should ever happen to you!). In this case, your score itself is meaningless, but your percentile tells you everything.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose your exam score is better than 90% of the rest of the class. That means your exam score is at the 90th percentile (so k = 90), which hopefully gets you an A. Conversely, if your score is at the 10th percentile (which would never happen to you, because you're such an excellent student), then <span

class="calibre16">k
= 10; that means only 10% of the other scores are below
yours, and 90% of them are above yours; in this case an A is
not in your future.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A nice property of percentiles is they
have a universal interpretation: Being at the 95th percentile
means the same thing no matter if you are looking at exam
scores or weights of packages sent through the postal service;
the 95th percentile always means 95% of the other values lie
below yours, and 5% lie above it. This also allows you to
fairly compare two data sets that have different means and
standard deviations (like ACT scores in reading versus math).
It evens the playing field and gives you a way to compare
apples to oranges, so to speak.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> A percentile is
not
a percent; a percentile is a number
(or the average of two numbers) in the data set that marks a
certain percentage of the way through the data. Suppose your
score on the GRE was reported to be the 80th percentile. This
doesn’t mean you scored 80% of the questions correctly. It
means that 80% of the students’ scores were lower than yours
and 20% of the students’ scores were higher than yours.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> A high
percentile doesn’t always constitute a good thing. For example,
if your city is at the 90th percentile in terms of crime rate
compared to cities of the same size, that means 90% of cities
similar to yours have a crime rate that is lower than yours,
which is not good for you. Another example is golf scores; a
low score in golf is a good thing, so being at the 80th
percentile with your score wouldn’t qualify you for the PGA
tour, let’s just say that.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Comparing household
incomes</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The U.S. government often reports
percentiles among its data summaries. For example, the U.S.
Census Bureau reported the median (the 50th percentile)
household income for 2001 to be $42,228, and in 2007 it was
reported to be $50,233. The Bureau also reports various
percentiles for household income for each year, including the

10th, 20th, 50th, 80th, 90th, and 95th. Table 5-3 shows the values of each of these percentiles for both 2001 and 2007. </p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Looking at the percentiles for 2001 in Table 5-3, you can see that the bottom half of the incomes are closer together than the top half of the incomes are. The difference between the 20th percentile and the 50th percentile is about $24,000, whereas the spread between the 50th percentile and the 80th percentile is more like $41,000. The difference between the 10th and 50th percentiles is only about $31,000, whereas the difference between the 50th and the 90th percentiles is a whopping $74,000.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The percentiles for 2007 are all higher than the percentiles for 2001 (which is a good thing!). They are also more spread out. For 2007, the difference between the 20th and 50th percentiles is around $30,000, and from the 50th to the 80th it’s approximately $50,000; both of these differences are larger than for 2001. Similarly, the 10th percentile is farther from the 50th (about $38,000 difference) in 2007 compared to 2001, and the 50th is farther from the 90th (by about $86,000) in 2007, compared to 2001.These results tell us that incomes are increasing in general at all levels between 2001 and 2007, but the gap is widening between those levels. For example, the 10th percentile for income in 2001 was $10,913 (as seen in Table 5-3), compared to $12,162 in 2007; this represents about an 11 percent increase (subtract the two and divide by 10,913). Now compare the 95th percentiles for 2007 versus 2001; the increase is almost 18%. Now, technically, you may want to adjust the 2001 values for inflation, but you get the basic idea. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Percentage changes affect the variability in a data set. For example, when salary raises are given on a percentage basis, the diversity in the salaries also increases; it’s the “rich get richer” idea. The guy making $30,000 gets a 10 percent raise and his salary goes up to $33,000 (an increase of $3,000); but the guy making $300,000 gets a 10 percent raise and now makes $330,000 (a difference of $30,000). So when you first get hired for a new job, negotiate the highest possible salary you can because your raises that follow will also net a higher amount.

</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Examining ACT
Scores</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Each year millions of U.S. high school
students take a nationally administered ACT exam as part of the
process of applying for colleges. The test is designed to
assess college readiness in the areas of English, Math,
Reading, and Science. Each test has a possible score of 36
points.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">ACT does not release the average or
standard deviation of the test scores for a given exam. (That
would be a real hassle if they did, because these statistics
can change from exam to exam, and people would complain that
this exam was harder than that exam when the actual scores are
not relevant.) To avoid these issues, and for other reasons,
ACT reports test results using percentiles.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Percentiles are usually reported in the
form of a predetermined list. For example, the U.S. Census
Bureau reports the 10th, 20th, 50th, 80th, 90th, and 95th
percentiles for household income (as shown in Table 5-3).
However, ACT uses percentiles in a different way. Rather than
reporting the exam scores corresponding to a premade list of
percentiles, they list each possible exam score and report its
corresponding percentile, whatever that turns out to be. That
way, to find out where you stand, you just look up your score
and you’ll find out your percentile.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Table 5-4 shows the 2009 percentiles
for the scores on the Mathematics and Reading ACT exams. To
interpret an exam score, find the row corresponding to the
score and the column for the exam area (for example, Reading).
Intersect row and column and you find out which percentile your
score represents; in other words, you see what percentage of
your fellow exam-taking comrades scored lower than you.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 5-4" src="images/00037.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you scored 30 on
the Math exam; in Table 5-4 you look at the row for 30 in the
column for Math; you see your score is at the 95th percentile.

In other words 95% of the students scored lower than you, and only 5% scored higher than you.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now suppose you also scored a 30 on the Reading exam. Just because a score of 30 represents the 95th percentile for Math doesn’t necessarily mean a score of 30 is at the 95th percentile for Reading as well. (It’s probably reasonable to expect that fewer people score 30 or higher on the Math exam than on the Reading exam.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">To test my theory, look at column 3 of Table 5-4 in the row for a score of 30. You see that a score of 30 on the Reading exam puts you at the 91st percentile — not quite as great as your position on the Math exam, but certainly not a bad score.</p></blockquote>
<blockquote class="calibre5"><p id="a125" class="calibre6">
Gathering a five-number summary</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Beyond reporting a single measure of center and/or a single measure of spread, you can create a group of statistics and put them together to get a more detailed description of a data set. The Empirical Rule (as seen in “Examining the Empirical Rule (68-95-99.7)” earlier in this chapter) uses the mean and standard deviation in tandem to describe a bell-shaped data set. In the case where your data are not bell-shaped, you use a different set of statistics (based on percentiles) to describe the big picture of your data. This method involves cutting the data into four pieces (with an equal amount of data in each piece) and reporting the resulting five cutoff points that separate these pieces. These cutoff points are represented by a set of five statistics that describe how the data are laid out.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The five-number summary is a set of five descriptive statistics that divide the data set into four equal sections. The five numbers in a five-number summary are:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. The minimum
 (smallest) number in the data set
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="calibre15"> 2. The 25th percentile
 (also known as the first quartile,
 or Q
<sub
class="calibre42">1</sub>₎
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The median
 (50th percentile)</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. The 75th percentile
 (also known as the third quartile,
 or Q
<sub
class="calibre42">3</sub>₎
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. The maximum
 (largest) number in the data set
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you want to find
the five-number summary of the following 25 (ordered) exam
scores: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78,
79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. The minimum is 43,
the maximum is 99, and the median is the number directly in the
middle, 77.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To find Q
₁
 and
Q₃

you use the steps shown in the
section “Calculating percentiles,” with <span

class="calibre16">n
 = 25. Step 1 is done because the data are ordered. For
Step 2, since Q₁
 is the 25th percentile, multiply 0.25
∗ 25 = 6.25. This is not a whole number, so
Step 3a says to round it up to 7 and proceed to Step 3b.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Following Step 3b, you count from left
to right in the data set until you reach the 7th number, 68;
this is Q₁
. For Q₃
 (the 75th percentile) you multiply 0.75
∗ 25 = 18.75, which you round up to 19. The
19th number on the list is 89, so that’s Q
₃
. Putting it all together, the five-number
summary for these 25 test scores is 43, 68, 77, 89, and 99. To
best interpret a five-number summary, you can use a boxplot;
see Chapter 7 for details.</p></blockquote>
<blockquote class="calibre5"><p id="a126" class="calibre6">
Exploring
interquartile range</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The purpose of the five-number summary
is to give descriptive statistics for center, variation, and
relative standing all in one shot. The measure of center in the
five-number summary is the median, and the first quartile,
median, and third quartiles are measures of relative standing.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To obtain a measure of variation based
on the five-number summary, you can find what’s called the
interquartile range
(or
IQR). The IQR
 equals Q<span><span

```
class="calibre41"><sub class="calibre42">3</sub></span></span>
<span> – </span><span><span class="calibre16"><span
class="italic">Q</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
<span> (that is, the 75th percentile minus the 25th percentile)
and reflects the distance taken up by the innermost 50% of the
data. If the </span><span><span class="calibre16"><span
class="italic">IQR</span></span></span><span> is small, you
know a lot of data are close to the median. If the </span>
<span><span class="calibre16"><span class="italic">IQR</span>
</span></span><span> is large, you know the data are more
spread out from the median. The </span><span><span
class="calibre16"><span class="italic">IQR</span></span></span>
<span> for the test scores data set is 89 – 68 = 21, which is
fairly large, seeing as how test scores only go from 0 to 100.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> The
interquartile range is a much better measure of variation than
the regular range (maximum value minus minimum value; see the
section “Being out of range” earlier in this chapter). That’s
because the interquartile range doesn’t take outliers into
account; it cuts them out of the data set by only focusing on
the distance within the middle 50 percent of the data (that is,
between the 25th and 75th percentiles).</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Descriptive
statistics that are well chosen and used correctly can tell you
a great deal about a data set, such as where the center is
located, how diverse the data are, and where a good portion of
the data lies. However, descriptive statistics can’t tell you
everything about the data, and in some cases they can be
misleading. Be on the lookout for situations where a different
statistic would be more appropriate (for example, the median
describes center more fairly than the mean when the data is
skewed), and keep your eyes peeled for situations where
critical statistics are missing (for example, when a mean is
reported without a corresponding standard deviation).</span>
</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a127"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">
```

```
<a href="#a118" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a119" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a112" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a113" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a111" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a116" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a117" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a114" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a115" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a123" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a122" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a121" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a120" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a127" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a126" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a125" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a124" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a128" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a141">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 6</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a129" class="calibre10"><span class="calibre11"><span
class="bold"><span>Getting the Picture: Graphing Categorical
Data</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Making data displays for categorical data</span></span>
</p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Interpreting and critiquing charts and graphs</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>D</span><span>ata displays, especially
charts and graphs, seem to be everywhere, showing everything
from election results, broken down by every conceivable
characteristic, to how the stock market has fared over the past
few years (months, weeks, days, minutes). We’re living in an
instant gratification, fast-information society; everyone wants
to know the bottom line and be spared the details.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The abundance of graphs and charts is
not necessarily a bad thing, but you have to be careful; some
of them are incorrect or even misleading (sometimes
intentionally and sometimes by accident), and you have to know
what to look for.</span></span></p></blockquote>
```

<blockquote class="calibre9"><p class="calibre14">This chapter is about graphs involving
categorical data
(data that places individuals into groups or categories, such
as gender, opinion, or whether a patient takes medication every
day. Here you find out how to read and make sense of these data
displays and get some tips for evaluating them and spotting
problems. (Note:
 Data displays for numerical data,
 such as weight, exam score, or the
number of pills taken
by a patient each day, come in Chapter 7.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The most common types of data displays
for categorical data are pie charts and bar graphs. In this
chapter, I present examples of each type of data display and
share some thoughts on interpretation and tips for critically
evaluating each type.</p></blockquote>
<p id="a130" class="calibre6">Take Another Little Piece of My Pie
Chart</p>
<blockquote class="calibre9"><p class="calibre14">A pie chart takes categorical data and
breaks them down by group, showing the percentage of
individuals that fall into each group. Because a pie chart
takes on the shape of a circle, the “slices” that represent
each group can easily be compared and contrasted.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Because each individual in the study
falls into one and only one category, the sum of all the slices
of the pie should be 100% or close to it (subject to a bit of
rounding off). However, just in case, keep your eyes open for
pie charts whose percentages just don’t add up.
</p></blockquote>
<blockquote class="calibre5"><p id="a131" class="calibre6">
Tallying
personal expenses</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When you spend your money, what do you
spend it on? What are your top three expenses? According to the
U.S. Bureau of Labor Statistics 2008 Consumer Expenditure

Survey, the top six sources of consumer expenditures in the U.S. were housing (33.9%), transportation (17.0%), food (12.8%), personal insurance and pensions (11.1%), healthcare (5.9%), and entertainment (5.6%). These six categories make up over 85% of average consumer expenses. (Although the exact percentages change from year to year, the list of the top six items remains the same.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 6-1 summarizes the 2008 U.S. expenditures in a pie chart. Notice that the “Other” category is a bit large in this chart (13.7%). However, with so many other possible expenditures out there (including this book), each one would only get a tiny slice of the pie for itself, and the resulting pie chart would be a mess. In this case, it is too difficult to break “Other” down further. (But in many other cases you can.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Ideally, a pie chart shouldn’t have too many slices because a large number of slices distracts the reader from the main point(s) the pie chart is trying to relay. However, lumping the remaining categories into one slice that’s one of the largest in the whole pie chart leaves readers wondering what’s included in that particular slice. With charts and graphs, doing it right is a delicate balance.</p></blockquote>
<blockquote class="calibre5"><p id="a132" class="calibre6">Bringing in a lotto revenue</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">State lotteries bring in a great deal of revenue, and they also return a large portion of the money received, with some of the revenues going to prizes and some being allocated to state programs such as education. Where does lottery revenue come from? Figure 6-2 is a pie chart showing the types of games and their percentage of revenue as recently reported by Ohio’s state lottery. (Note the slices don’t sum to 100% exactly due to slight rounding error.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You can see by the pie chart in Figure 6-2 that 49.3% of the lottery sales revenue comes from the instant (scratch-off) games. The rest come from various lottery-type games in which players choose a set of numbers and win if a certain number of their numbers match those chosen by the lottery.</p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span

```html
class="bold">Figure 6-1:</span><span> Pie chart showing how
people in the U.S. spend their money.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0601.eps"
src="images/00038.jpg" class="calibre2"/></span></p>
</blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 6-2:</span><span> Pie chart breaking down a
state’s lottery revenue.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0602.eps"
src="images/00039.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Notice that this pie chart doesn’t tell
you </span><span><span class="calibre16"><span
class="italic">how much</span></span></span><span> money came
in, only</span><span><span class="calibre16"><span
class="italic"> what percentage</span></span></span><span> of
the money came from each type of game. About half the money
(49.3%) came from instant scratch-off games; does this revenue
represent a million dollars, two million dollars, ten million
dollars, or more? You can’t answer these questions without
knowing the total amount of revenue dollars.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I was, however, able to find this
information on another chart provided by the lottery Web site:
The total revenue (over a 10-year period) was reported as
“1,983.1 million dollars” — which you also know as 1.9831
billion dollars. Because 49.3% of sales came from instant
games, they therefore represent sales revenue of $977,668,300
over a 10-year period. That’s a lot of (or dare I say a
“lotto”) scratching.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a133" class="calibre6">
<span class="calibre21"><span class="bold"><span>Ordering
takeout</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>It’s also important to watch for totals
when examining a pie chart from a survey. A newspaper I read
reported the latest results of a “people poll.” They asked,
“What is your favorite night to order takeout for dinner?” The
results are shown in a pie chart (see Figure 6-3).</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
```

class="calibre15">You can clearly see that Friday night
is the most popular night for ordering takeout (and that result
makes sense) with decreasing demand moving from Saturday
through Monday. The actual percentages shown in Figure 6-3
really only apply to the people who were surveyed; how close
these results mimic the population depends on many factors, one
of which is sample size. But unfortunately, sample size is not
included as part of this graph. (For example, it would be nice
to see “n = XXX” below the
title; where n represents sample
size.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Without knowing the sample size, you
can’t tell how accurate the information is. Which results would
you find to be more accurate — those based on 25 people, 250
people, or 2,500 people? When you see the number 10%, you don’t
know if it’s 10 out of 100, 100 out of 1,000, or even 1 out of
10. To statisticians, 1 ÷ 10 is not the same as 100 ÷ 1,000,
even though they both represent 10%. (Don’t tell that to
mathematicians — they’ll think you’re nuts!)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Pie charts
often don’t include mention of the total sample size. Always
check for the sample size, especially if the results are very
important to you; don’t assume it’s large! If you don’t see the
sample size, go to the source of the data and ask for it.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 6-3: Pie chart for takeout
food survey results.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0603.eps"
src="images/00040.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre5"><p id="a134" class="calibre6">
Projecting age
trends</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The U.S. Census Bureau provides an
almost unlimited amount of data, statistics, and graphics about
the U.S. population, including the past, present, and
projections for the future. It often makes comparisons between

years in order to look for changes and trends.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One recent Census Bureau population
report looked at what it calls the “older U.S. population” (by
the government’s definition, this means people 65 years old or
over). Age was broken into the following groups: 65–69 years,
70–74 years, 75–79 years, 80–84 years, and 85 and over. The
Bureau calculated and reported the percentage in each age group
for the year 2010 and made projections for the percentage in
each age group for the year 2050.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">I made side-by-side pie charts for the
years 2010 versus 2050 (projections) to make comparisons; you
can see the results in Figure 6-4. The percentage of the older
population in each age group for 2010 is shown in one pie
chart, and alongside it is a pie chart of the projected
percentage for each age group for 2050 (based on the current
age of the entire U.S. population, birth and death rates, and
other variables).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you compare the sizes of the slices
from one graph to the other in Figure 6-4, you see that the
slices for corresponding age groups are larger for the 2050
projections (compared to 2010) as the age groups get older, and
the slices are smaller for the 2050 projections (compared to
2010) as the age groups get younger. For example the 65–69 age
group decreases from 30% in 2010 to a projected 25% in 2050;
while the 85-and-over age group increases from 14% in 2010 to
19% projected for 2050.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 6-4: Side-by-side pie charts
on the aging population, 2010 versus 2050 projections.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0604.eps"
src="images/00041.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The results from Figure 6-4 indicate a
shift in the ages of the population toward the older
categories. From there, the medical and social research
communities can examine the ramifications of this trend in
terms of healthcare, assisted living, social security, and so
on.</p></blockquote>

```
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> The operative
words here are </span><span><span class="calibre16"><span
class="italic">if the trend continues.</span></span></span>
<span> As you know, many variables affect population size, and
you need to take those into account when interpreting these
projections into the future. The U.S. government always points
out caveats like this in their reports; it is very diligent
about that.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> The pie charts in Figure 6-4 work well
for comparing groups because they are side-by-side on the same
graph, using the same coding for the age groups in each, and
their slices are in the same order for both as you move
clockwise around the graphs. They aren’t all scrambled up on
each graph so you have to hunt for a certain age group on each
graph separately.</span></span></p></blockquote>
<img alt="SB-Begin" src="images/00011.jpg" class="calibre2"/>
<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10"><span
class="calibre23"><span class="bold"><span>Evaluating a pie
chart</span></span></span></p></blockquote><div
class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34"><span
class="calibre35"><span>The following tips help you taste test
a pie chart for statistical correctness:</span></span></p>
</blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Check to be sure the percentages add up to 100% or very
close to it (any round-off error should be very small).</span>
</span></p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Beware of slices of the pie called “Other” that are
larger than many of the other slices.</span></span></p>
</blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Look for a reported total number of units (people, dollar
amounts, and so on) so that you can determine (in essence) how
“big” the pie was before being divided up into the slices that
you’re looking at.</span></span></p></blockquote>
```

<blockquote class="calibre44"><p class="calibre45"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Avoid three-dimensional pie charts; they don’t show the
slices in their proper proportions. The slices in front look
larger than they should.</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<p id="a135" class="calibre6">Raising the Bar on Bar Graphs
</p>
<blockquote class="calibre9"><p class="calibre14">A
bar graph (or
bar
chart) is perhaps the most common
data display used by the media. Like a pie chart, a bar graph
breaks categorical data down by group. Unlike a pie chart, it
represents these amounts by using bars of different lengths;
whereas a pie chart most often reports the amount in each group
as percentages, a bar graph uses either the number of
individuals in each group (also called the frequency
) or the percentage in each group (called the
relative frequency).
</p></blockquote>
<blockquote class="calibre5"><p id="a136" class="calibre6">
Tracking
transportation expenses</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How much of their income do people in
the United States spend on transportation to get back and forth
to work? It depends on how much money they make. The Bureau of
Transportation Statistics (did you know such a department
existed?) conducted a study on transportation in the U.S.
recently, and many of its findings are presented as bar graphs
like the one shown in Figure 6-5.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">This particular bar graph shows how
much money is spent on transportation for people in different
household-income groups. It appears that as household income
increases, the total expenditures on transportation also
increase. This makes sense, because the more money people have,
the more they have available to spend.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">But would the bar graph change if you looked at transportation expenditures not in terms of total dollar amounts, but as the percentage of household income? The households in the first group make less than $5,000 a year and have to spend $2,500 of it on transportation. (
Note: The label reads “2.5,” but because the units are in thousands of dollars, the 2.5 translates into $2,500.)</p>
</blockquote>
<div class="calibre1">
<p class="calibre38">Figure 6-5: Bar graph showing transportation expenses by household income group.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0605.eps"
src="images/00042.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">This $2,500 represents 50% of the annual income of those who make $5,000 per year; the percentage of the total income is even higher for those who make less than $5,000 per year. The households earning $30,000–$40,000 per year pay $6,000 per year on transportation, which is between 15% and 20% of their household income. So, although the people making more money spend more dollars on transportation, they don’t spend more as a percentage of their total income. Depending on how you look at expenditures, the bar graph can tell two somewhat different stories.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Another point to check out is the groupings on the graph. The categories for household income as shown aren’t equivalent. For example, each of the first four bars represents household incomes in intervals of $5,000, but the next three groups increase by $10,000 each, and the last group contains every household making more than $50,000 per year. Bar graphs using different-sized intervals to represent numerical values (such as Figure 6-5) make true comparisons between groups more difficult. (However, I’m sure the government has its reasons for reporting the numbers this way; for example, this may be the way income is broken down for tax-related purposes.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One last thing: Notice that the numerical groupings in Figure 6-5 overlap on the boundaries.

For example, $30,000 appears in both the 5th and 6th bars of the graph. So, if you have a household income of $30,000, which bar do you fall into? (You can’t tell from Figure 6-5, but I’m sure the instructions are buried in a huge report in the basement of some building in Washington, D.C.) This kind of overlap appears quite frequently in graphs, but you need to know how the borderline values are being treated. For example, the rule may be “Any data lying exactly on a boundary value automatically goes into the bar to its immediate right.” (Looking at Figure 6-5, that puts a household with a $30,000 income into the 6th bar rather than the 5th.) As long as they are being consistent for each boundary, that’s okay. The alternative, describing the income boundaries for the 5th bar as “20,000 to $29,999.99,” is not an improvement. Along those lines, income data can also be presented using a histogram (see Chapter 7), which has a slightly different look to it.
</p></blockquote>
<blockquote class="calibre5"><p id="a137" class="calibre6">
Making a lotto profit</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">That lotteries rake in the bucks is a well-known fact; but they also shell it out. How does it all shake out in terms of profits? Figure 6-6 shows the recent sales and expenditures of a certain state lottery.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 6-6: Bar graph of lottery sales and expenditures for a certain state.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In my opinion, this bar graph needs some additional info from behind the scenes to make it more understandable. The bars in Figure 6-6 don’t represent similar types of entities. The first bar represents sales (a form of revenue), and the other bars represent expenditures. The graph would be much clearer if the first bar weren’t included; for example, the total sales could be listed as a footnote.
</p></blockquote>
<blockquote class="calibre5"><p id="a138" class="calibre6">
Tipping the scales on a bar graph</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"> Another way a graph can be misleading is through its choice of scale on the frequency/relative frequency axis (that is, the axis where the amounts in each group are reported), and/or its starting value. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">By using a “stretched out” scale (for example, having each half inch of a bar represent 10 units versus 50 units), you can stretch the truth, make differences look more dramatic, or exaggerate values. Truth-stretching can also occur if the frequency axis starts out at a number that’s very close to where the differences in the heights of the bars start; you are in essence chopping off the bottom of the bars (the less exciting part) and just showing their tops; emphasizing (in a misleading way) where the action is. Not every frequency axis has to start at zero, but watch for situations that elevate the differences.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A good example of a graph with a stretched out scale is seen in Chapter 3, regarding the results of numbers drawn in the “Pick 3” lottery. (You choose three one-digit numbers and if they all match what’s drawn, you win.) In Chapter 3, the percentage of times each number (from 0–9) was drawn is shown in Table 3-2, and the results are displayed in a bar graph in Figure 3-1a. The scale on the graph is stretched and starts at 465, making the differences in the results look larger than they really are; for example, it looks like the number 1 was drawn much less often, whereas the number 2 was drawn much more often, when in reality there is no statistical difference between the percentage of times each number was drawn. (I checked.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Why was the graph in Figure 3-1a made this way? It might lead people to think they’ve got an inside edge if they choose the number 2 because it’s “on a hot streak”; or they might be led to choose the number 1 because it’s “due to come up.” Both of these theories are wrong, by the way; because the numbers are chosen at random, what happened in the past doesn’t matter. In Figure 3-1b you see a graph that’s been made correctly. (For more examples of where our intuition can go wrong with probability and what the scoop really is, see another of my books, Probability For Dummies
, also published by Wiley.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

```
class="calibre15"><span>Alternatively, by using a “squeezed
down” scale (for example, having each half inch of a bar
represent 50 units versus 10 units), you can downplay
differences, making results look less dramatic than they
actually are. For example, maybe a politician doesn’t want to
draw attention to a big increase in crime from the beginning to
the end of her term, so she may have the number of crimes of
each type shown where each half inch of a bar represents 500
crimes, versus 100 crimes. This squeezes the numbers together
and makes differences less noticeable. Her opponent in the next
election would go the other way and use a stretched-out scale
to emphasize a crime increase in dramatic fashion, and voilà!
(Now you know the answer to the question “How can two people
talk about the same data and get two different conclusions?”
Welcome to the world of politics.)</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> With a pie chart, however, the scale
can’t be changed to over-emphasize (or downplay) the results.
No matter how you slice up a pie chart, you’re always slicing
up a circle, and the proportion of the total pie belonging to
any given slice won’t change, even if you make the pie bigger
or smaller.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a139" class="calibre6">
<span class="calibre21"><span class="bold"><span>Pondering pet
peeves</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>A recent survey of 100 people with
office jobs asked them to report their biggest pet peeves in
the workplace. (Before going on, you may want to jot down a
couple of yours, just for fun.) A bar graph of the results of
the survey is shown in Figure 6-7. Poor time management looks
to be the number-one issue for these workers (I hope they
didn’t do this survey on company time).</span></span></p>
</blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 6-7:</span><span> Bar graph for survey data
with multiple responses.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0607.eps"
src="images/00044.jpg" class="calibre2"/></span></p>
</blockquote>
<img alt="SB-Begin" src="images/00011.jpg" class="calibre2"/>
<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10"><span
```

class="calibre23">Evaluating a bar graph</p></blockquote><div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">To raise the statistical bar on bar graphs, check out these tips:</p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
Bars that divide up values of a numerical variable (such as income) should be equal in width (if possible) for fair comparison.</p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
Be aware of the scale of the bar graph and determine whether it’s an appropriate representation of the information. </p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
Some bar graphs don’t sum to one because they are showing the results of more than one variable; make sure it’s clear what’s being summarized.</p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
Check whether the results are shown as the percentage within each group (relative frequencies) or the number in each group (frequencies).</p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
 If you see relative frequencies, check for the total sample size — it matters. If you see frequencies, divide each one by the total sample size to get percentages, which are easier to compare.</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<blockquote class="calibre9"><p class="calibre22"><span class="calibre15"><span> If you take a look at the percentages shown for each pet peeve listed, you see they don’t sum to one. That tells you that each person surveyed was allowed to choose more than one pet peeve (like that would be hard to do); perhaps they were asked to name their top three pet peeves, for example. For this data set and others like it that allow for multiple responses, a pie chart wouldn’t be possible (unless

```
you made one for every single pet peeve on the list).</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Note that Figure 6-7 is a </span><span>
<span class="calibre16"><span class="italic">horizontal bar
graph</span></span></span><span> (its bars go side to side) as
opposed to a </span><span><span class="calibre16"><span
class="italic">vertical bar graph</span></span></span><span>
(in which bars go up and down, as in Figure 6-6). Either
orientation is fine; use whichever one you prefer when you make
a bar graph. Do, however, make sure that you label the axes
appropriately and include proper units (such as gender,
opinion, or day of the week) where appropriate.</span></span>
</p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a140"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a140" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a138" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a130" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a131" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a132" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a133" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a134" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a135" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a136" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a137" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a141" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a139" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a129" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
```

!important; min-height: 10px !important; border: solid 1px !important;"> </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a156">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 7</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a142" class="calibre10"><span class="calibre11"><span
class="bold"><span>Going by the Numbers: Graphing Numerical
Data</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Making and interpreting histograms and boxplots for
numerical data</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Examining time charts for numerical data collected over
time</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Strategies for spotting misleading and incorrect
graphs</span></span></p></blockquote><div
class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>he main purpose of charts
and graphs is to summarize data and display the results to make
your point clearly, effectively, and correctly. In this
chapter, I present data displays used to summarize </span>
<span><span class="calibre16"><span
class="italic">numerical</span></span></span><span> data — data
that represent </span><span><span class="calibre16"><span
class="italic">counts</span></span></span><span> (such as the
```

number of pills a patient with diabetes takes per day, or the number of accidents at an intersection per year) or
measurements (the time it takes you to get to work/school each day, or your blood pressure).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You see examples of how to make, interpret, and evaluate the most common data displays for numerical data: time charts, histograms, and boxplots. I also point out many potential problems that can occur in these graphs, including how people often misread what’s there. This information will help you develop important detective skills for quickly spotting misleading graphs.</p>
</blockquote>
<p id="a143" class="calibre6">Handling Histograms</p>
<blockquote class="calibre9"><p class="calibre14">A histogram provides a snapshot of all the data broken down into numerically ordered groups, making it a quick way to get the big picture of the data, in particular, its general shape. In this section you find out how to make and interpret histograms, and how to critique them for correctness and fairness.</p></blockquote>
<blockquote class="calibre5"><p id="a144" class="calibre6">
Making a histogram</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
histogram is a special graph applied to data broken down into numerically ordered groups; for example, age groups such as 10–20, 21–30, 31–40, and so on. The bars connect to each other in a histogram — as opposed to a bar graph (Chapter 6) for categorical data, where the bars represent categories that don’t have a particular order, and are separated. The height of each bar of a histogram represents either the number of individuals (called the
frequency) in each group or the percentage of individuals (the relative frequency) in each group. Each individual in the data set falls into exactly one bar.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You can make a histogram from any numerical data set; however, you can’t

determine the actual values of the data set from a histogram
because all you know is which group each data value falls into.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">An award winning
example</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s an example of how to create a
histogram for all you movie lovers out there (especially those
who love old movies). The Academy Awards started in 1928, and
one of the most popular categories for this award is Best
Actress in a Motion Picture. Table 7-1 shows the winners of the
first eight Best Actress Oscars, the years they won (1928–
1935), their ages at the time of winning their awards, and the
movies they were in. From the table you see the ages range from
22 to 62 — much wider than you may have thought it would be.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 7-1" src="images/00045.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To find out more about the ages of Best
Actresses, I expanded my data set to the period 1928–2009. The
age variable for this data set is numerical, so you can graph
it using a histogram. From there you can answer questions like:
What do the ages of these actresses look like? Are they mostly
young, old, in between? Are their ages all spread out, or are
they similar? Are most of them in a certain age range, with a
few outliers (either very young or very old actresses, compared
to the others)? To investigate these questions, a histogram of
ages of the Best Award actresses is shown in Figure 7-1.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-1: Histogram of Best Actress
Academy Award winners’ ages, 1928–2009.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0701.eps"
src="images/00046.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Notice that the age groups are shown on
the horizontal (x) axis. They go by
groups of 5 years each: 20–25, 25–30, 30–35, . . . 80–85. The
percentage (relative frequency) of actresses in each age group
appears on the vertical (

y) axis. For example, about 27 percent of the actresses were between 30 and 35 years of age when they won their Oscars.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Creating appropriate groups</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> For Figure 7-1, I used groups of 5 years each in the above example because increments of 5 create natural breaks for years and because it provides enough bars to look for general patterns. You don’t have to use this particular grouping, however; you have a bit of poetic license when making a histogram. (However, this freedom allows others to deceive you as you see in the later section “Detecting misleading histograms.”) Here are some tips for setting up your histogram:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Each data set requires different ranges for its groupings, but you want to avoid ranges that are too wide or too narrow.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • If a histogram has really wide ranges for its groups, it places all the data into a very small number of bars that make meaningful comparisons impossible.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • If the histogram has very narrow ranges for its groups, it looks like a big series of tiny bars that cloud the big picture. This can make the data look very choppy with no real pattern.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Make sure your groups have equal widths. If one bar is wider than the others, it may contain more data than it should.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">One idea that may be appropriate for your histogram is to take the range of the data (largest minus smallest) and divide by 10 to get 10 groupings.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span

class="calibre7">Handling borderline values</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the Academy Award example, what happens if an actress’s age lies right on a borderline? For example, in Table 7-1 Norma Shearer was 30 years old in 1930 when she won the Oscar for The Divorcee. Does she belong in the 25–30 age group (the lower bar) or the 30–35 age group (the upper bar)? </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> As long as you are consistent with all the data points, you can either put all the borderline points into their respective lower bars or put all of them into their respective upper bars. The important thing is to pick a direction and be consistent. In Figure 7-1, I went with the convention of putting all borderline values into their respective upper bars — which puts Norma Shearer’s age in the 3rd bar, the 30–35 age group of Figure 7-1. </p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Clarifying the axes</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The most complex part of interpreting a histogram for the reader is to get a handle on what’s being shown on the x and y axes. Having good descriptive labels on the axes will help. Most statistical software packages label the x -axis using the variable name you provided when you entered your data (for example “age” or “weight”). However, the label for the y -axis isn’t as clear. Statistical software packages often label the y-axis of a histogram by writing “frequency” or “percent” by default. These terms can be confusing: frequency or percentage of what? </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Clarify the y

-axis label on your histogram by changing “frequency” to “number of” and adding the variable name. To modify a label that simply reads “percent,” clarify by writing “percentage of” and the variable. For example, in the histogram of ages of the Best Actress winners shown in Figure 7-1, I labeled the
y
-axis “Percentage of actresses in each age group.” In the next section you see how to interpret the results from a histogram. How old are those actresses anyway?
</p></blockquote>
<blockquote class="calibre5"><p id="a145" class="calibre6">
Interpreting a histogram</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A histogram tells you three main features of numerical data:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
How the data are distributed among the groups (statisticians call this the shape
 of the data)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The amount of variability in the data (statisticians call this the amount of spread in the data)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Where the center of the data is (statisticians use different measures)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre6">Checking out the shape of the data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One of the features that a histogram can show you is the shape of the data — in other words, the manner in which the data fall into the groups. For example, all the data may be exactly the same, in

which case the histogram is just one tall bar; or the data might have an equal number in each group; in which case the shape is flat.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Some data sets have a distinct shape. Here are three shapes that stand out:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18">
Symmetric:

A histogram is symmetric if you cut it down the middle and the left-hand and right-hand sides resemble mirror images of each other.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Figure 7-2a shows a symmetric data set; it represents the amount of time each of 50 survey participants took to fill out a certain survey. You see that the histogram is close to symmetric.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Skewed right:
A skewed right histogram looks like a lopsided mound, with a tail going off to the right.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Figure 7-1, showing the ages of the Best Actress Award winners, is skewed right. You see on the right side there are a few actresses whose ages are older than the rest.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Skewed left:

If a histogram is skewed left, it looks like a lopsided mound with a tail going off to the left.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span class="calibre15"><span> Figure 7-2b shows a histogram of 17 exam scores. The shape is skewed left; you see a few students

```html
who scored lower than everyone else.</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Following are
some particulars about classifying the shape of a data set:
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Don’t expect
symmetric data to have an exact and perfect shape.</span>
</span></span><span> Data hardly ever fall into perfect
patterns, so you have to decide whether the data shape is close
enough to be called symmetric.</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><span> If the shape is close enough to
symmetric that another person would notice it, and the
differences aren’t enough to write home about, I’d classify it
as symmetric or roughly symmetric. Otherwise, you classify the
data as non-symmetric. (More sophisticated statistical
procedures exist that actually test data for symmetry, but
they’re beyond the scope of this book.)</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Don’t assume
that data are skewed if the shape is non-symmetric.</span>
</span></span><span> Data sets come in all shapes and sizes,
and many of them don’t have a distinct shape at all. I include
skewness on the list here because it’s one of the more common
non-symmetric shapes, and it’s one of the shapes included in a
standard introductory statistics course.</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><span> If a data set does turn out to be
skewed (or close to it), make sure to denote the direction of
the skewness (left or right).</span></span></p></blockquote>
<div class="calibre19"> </div>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-2:</span><span> Comparing the shape of a)
a symmetric histogram and b) a skewed left histogram.</span>
</span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
```

class="calibre15"><img alt="9780470911082-fg0702.eps"
src="images/00047.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">As you know from Figure 7-1, the
actresses' ages in Figure 7-1 are skewed right. Most of the
actresses were between 20 and 50 years of age when they won,
with about 27% of them between the ages of 30–35. A few
actresses were older when they won their Oscars; about 6
percent were between 60–65 years of age, and less than 4%
(total) were 70 years old or over (if you add the percentages
from the last two bars in the histogram). The last three bars
are what make the data have a shape that is skewed right.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Measuring center:
Mean versus median</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A histogram gives you a rough idea of
where the "center" of the data lies. The word
center
 is in quotes because many different
statistics are used to designate center. The two most common
measures of center are the average (the mean) and the median.
(For details on measures of center, see Chapter 5.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> To visualize the average age (the
mean), picture the data as people sitting on a teeter-totter.
Your objective is to balance it. Because data don't move
around, assume the people stay where they are and you move the
pivot point (which you can also think of as the hinge or
fulcrum) anywhere you want. The mean is the place the pivot
point has to be in order to balance the weight on each side of
the teeter-totter.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The balancing point of the teeter-
totter is affected by the weights of the people on each side,
not by the number of people on each side. So the mean is
affected by the actual values of the data, rather than the
amount of data.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The median is the place where you put
the pivot point so you have an equal number of people on each
side of the teeter-totter, regardless of their weights. With
the same number of people on each side, the teeter-totter
wouldn't balance in terms of weight unless the teeter-totter

had people with the same total weight on each side. So the
median isn’t affected by the values of the data, just their
location within the data set.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/>The mean is
affected by outliers, values in
the data set that are away from the rest of the data, on the
high end and/or the low end. The median, being the middle
number, is not affected by outliers.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Viewing variability:
Amount of spread around the mean</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You also get a sense of variability in
the data by looking at a histogram. For example, if the data
are all the same, they are all placed into a single bar, and
there is no variability. If an equal amount of data is in each
group, the histogram looks flat with the bars close to the same
height; this means a fair amount of variability.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The idea of a
flat histogram indicating some variability may go against your
intuition, and if it does you’re not alone. If you’re thinking
a flat histogram means no variability, you’re probably thinking
about a time chart, where single numbers are plotted over time
(see the section “Tackling Time Charts” later in this chapter).
Remember, though, that a histogram doesn’t show data over time
— it shows all the data at one point in time.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Equally confusing is the idea that a
histogram with a big lump in the middle and tails sloping
sharply down on each side actually has less variability than a
histogram that’s straight across. The curves looking like hills
in a histogram represent clumps of data that are close
together; a flat histogram shows data equally dispersed, with
more variability.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Variability in
a histogram is higher when the taller bars are more spread out
around the mean and lower when the taller bars are close to the

mean.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For the Best Actress Award winners' ages shown in Figure 7-1, you see many actresses are in the age range from 30–35, and most of the ages are between 20–50 years in age, which is quite diverse; then you have those outliers, those few older actresses (I count 7 of them) that spread the data out farther, increasing its overall variability.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The most common statistic used to measure variability in a data set is the standard deviation,
 which in a rough sense measures the average distance that the data lie from the mean. The standard deviation for the Best Actress age data is 11.35 years. (See Chapter 5 for all the details on standard deviation.) A standard deviation of 11.35 years is fairly large in the context of this problem, but the standard deviation is based on average distance from the mean, and the mean is influenced by outliers, so the standard deviation will be as well (see Chapter 5 for more information).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the later section "Interpreting a boxplot," I discuss another measure of variability, called the
interquartile range
(IQR), which is a more appropriate measure of variability when you have skewed data.
</p></blockquote>
<blockquote class="calibre5"><p id="a146" class="calibre6">
Putting numbers with pictures</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You can't actually calculate measures of center and variability from the histogram itself because you don't know the exact data values. To add detail to your findings, you should always calculate the basic statistics of center and variation along with your histogram. (All the descriptive statistics you need, and then some, appear in Chapter 5.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 7-1 is a histogram for the Best Actress ages; you can see it is skewed right. Then for Figure 7-3, I calculated some basic (that is, descriptive) statistics from the data set. Examining these numbers, you find the median

age is 33.00 years and the mean age is 35.69 years.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The mean age is higher than the median
age because of a few actresses that were quite a bit older than
the rest when they won their awards. For example, Jessica Tandy
won for her role in Driving Miss Daisy
when she was 81, and Katharine Hepburn won the Oscar for
On
Golden Pond when she was 74. The
relationship between the median and mean confirms the skewness
(to the right) found in Figure 7-1.</p>
</blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-3: Descriptive statistics
for Best Actress ages (1928–2009).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0703.eps"
src="images/00048.jpg" class="calibre2"/></p>
</blockquote>
<br class="calibre1"/>
<br class="calibre1"/>
<br class="calibre1"/>
<blockquote class="calibre9"><p class="calibre14">Here are some tips for connecting the
shape of the histogram (discussed in the previous section) with
the mean and median:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the
histogram is skewed right, the mean is greater than the median.
</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> This is the case because skewed-right
data have a few large values that drive the mean upward but do
not affect where the exact middle of the data is (that is, the
median). Looking at the histogram of ages of the Best Actress
Award winners in Figure 7-1, you see they’re skewed right.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the

histogram is close to symmetric, then the mean and median are close to each other.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39">
Close to symmetric means it’s almost the same on either side; it doesn’t need to be exact.
Close
 is defined in the context of the data; for example, the numbers 50 and 55 are said to be close if all the values lie between 0 and 1,000, but they are considered to be farther apart if all the values lie between 49 and 56.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> The histogram shown in Figure 7-2a is close to symmetric. Its mean and median are both equal to 3.5.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
If the histogram is skewed left, the mean is less than the median.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> This is the case because skewed-left data have a few small values that drive the mean downward but do not affect where the exact middle of the data is (that is, the median).</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Figure 7-2b represents the exam scores of 17 students, and the data are skewed left. I calculated the mean and median of the original data set to be 70.41 and 74.00, respectively. The mean is lower than the median due to a few students who scored quite a bit lower than the others. These findings match the general shape of the histogram shown in Figure 7-2b.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> The tips for interpreting histograms found in the previous section can also be used the other way around. If for some reason you don’t have a histogram of the data, and you only have the mean and median to go by, you compare them to each other to get a rough idea as

to the shape of the data set.</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the mean is much larger than the median, the data are
generally skewed right; a few values are larger than the rest.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the mean is much smaller than the median, the data are
generally skewed left; a few smaller values bring the mean
down.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the mean and median are close, you know the data is
fairly balanced, or symmetric, on each side.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Under certain
conditions, you can put together the mean and standard
deviation to describe a data set in quite a bit of detail. If
the data have a normal distribution (a bell-shaped hill in the
middle, sloping down at the same rate on each side; see Chapter
5), the Empirical Rule can be applied.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The Empirical Rule (also in Chapter 5)
says that if the data have a normal distribution, about 68% of
the data lie within 1 standard deviation of the mean, about 95%
of the data lie within 2 standard deviations from the mean, and
99.7% of the data lie within 3 standard deviations of the mean.
These percentages are custom-made for the normal distribution
(bell-shaped data) only and can’t be used for data sets of
other shapes.</p></blockquote>
<blockquote class="calibre5"><p id="a147" class="calibre6">
Detecting
misleading histograms</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">There are no hard and fast rules for
how to create a histogram; the person making the graph gets to
choose the groupings on the x
-axis as well as the scale and starting and ending points
on the <span><span class="calibre16"><span

```
class="italic">y</span></span></span><span>-axis. Just because
there is an element of choice, however, doesn’t mean every
choice is appropriate; in fact, a histogram can be made to be
misleading in many ways. In the following sections, you see
examples of misleading histograms and how to spot them.</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Missing the mark with
too few groups</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Although the number of groups you use
for a histogram is up to the discretion of the person making
the graph, there is such a thing as going overboard, either by
having way too few bars, with everything lumped together, or by
having way too many bars, where every little difference is
magnified.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> To decide how many bars a histogram
should have, I take a good look at the groupings used to form
the bars on the </span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>-axis and see if
they make sense. For example, it doesn’t make sense to talk
about exam scores in groups of 2 points; that’s too much detail
— too many bars. On the other hand, it doesn’t make sense to
group actresses’ ages by intervals of 20 years; that’s not
descriptive enough.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figures 7-4 and 7-5 illustrate this
point. Each histogram summarizes </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span> = 222 observations of the amount of time between
eruptions of the Old Faithful geyser in Yellowstone Park.
Figure 7-4 uses six bars that group the data by 10-minute
intervals. This histogram shows a general skewed left pattern,
but with 222 observations you are cramming an awful lot of data
into only six groups; for example, the bar for 75–85 minutes
has more than 90 pieces of data in it. You can break it down
further than that.</span></span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-4:</span><span> Histogram #1 showing time
between eruptions for Old Faithful geyser (</span><span><span
class="calibre36"><span class="italic">n</span></span></span>
<span> = 222).</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0704.eps"
```

```html
src="images/00049.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figure 7-5 is a histogram of the same
data set, where the time between eruptions is broken into
groups of 3 minutes each, resulting in 19 bars. Notice the
distinct pattern in the data that shows up with this histogram
which wasn’t uncovered in Figure 7-4. You see two distinct
peaks in the data; one peak around the 50-minute mark, and one
around the 75-minute mark. A data set with two peaks is called
</span><span><span class="calibre16"><span
class="italic">bimodal; </span></span></span><span>Figure 7-5
shows a clear example.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Looking at Figure 7-5, you can conclude
that the geyser has two categories of eruptions; one group that
has a shorter waiting time, and another group that has a longer
waiting time. Within each group you see the data are fairly
close to where the peak is located. Looking at Figure 7-4, you
couldn’t say that.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> If the interval
for the groupings of the numerical variable is really small,
you see too many bars in the histogram; the data may be hard to
interpret because the heights of the bars look more variable
than they should be. On the other hand, if the ranges are
really large, you see too few bars, and you may miss something
interesting in the data. </span></span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-5:</span><span> Histogram #2 showing time
between eruptions for Old Faithful geyser (</span><span><span
class="calibre36"><span class="italic">n</span></span></span>
<span> = 222).</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0705.eps"
src="images/00050.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Watching the scale
and start/finish lines</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The </span><span><span
class="calibre16"><span class="italic">y</span></span></span>
<span>-axis of a histogram shows how many individuals are in
each group, using counts or percents. A histogram can be
```

misleading if it has a deceptive scale and/or inappropriate starting and ending points on the y-axis.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Watch the scale on the y-axis of a histogram. If it goes by large increments and has an ending point that’s much higher than needed, you see a great deal of white space above the histogram. The heights of the bars are squeezed down, making their differences look more uniform than they should. If the scale goes by small increments and ends at the smallest value possible, the bars become stretched vertically, exaggerating the differences in their heights and suggesting a bigger difference than really exists.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">An example comparing scales on the vertical (y) axes is shown in Figures 7-5 and 7-6. I took the Old Faithful data (time between eruptions) and made a histogram with vertical increments of 20 minutes, from 0 to 100; see Figure 7-6. Compare this to Figure 7-5, with vertical increments of 5 minutes, from 0 to 35. Figure 7-6 has a lot of white space and gives the appearance that the times are more evenly distributed among the groups than they really are. It also makes the data set look smaller, if you don’t pay attention to what’s on the y-axis. Of the two graphs, Figure 7-5 is more appropriate.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-6: Histogram #3 of Old Faithful geyser eruption times.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<p id="a148" class="calibre6">Examining Boxplots</p>
<blockquote class="calibre9"><p class="calibre14">A boxplot is a one-dimensional graph of numerical data based on the five-

number summary, which includes the minimum value, the 25th percentile (known as Q₁
), the median, the 75th percentile (Q
₃
), and the maximum value. In essence, these five descriptive statistics divide the data set into four parts; each part contains 25% of the data. (See Chapter 5 for a full discussion of the five-number summary.)</p>
</blockquote>
<blockquote class="calibre5"><p id="a149" class="calibre6">
Making a boxplot</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To make a boxplot, follow these steps:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Find the five-number summary of your data set. (Use the steps outlined in Chapter 5.)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Create a vertical (or horizontal) number line whose scale includes the numbers in the five-number summary and uses appropriate units of equal distance from each other.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Mark the location of each number in the five-number summary just above the number line (for a horizontal boxplot) or just to the right of the number line (for a vertical boxplot).</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Draw a box around the marks for the 25th percentile and the 75th percentile.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Draw a line in the box where the median is located.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="calibre15"> 6. Determine whether or not outliers are present.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> To make this determination, calculate
the IQR (by
subtracting Q₃
 – Q
<sub
class="calibre42">1</sub>); then multiply
by 1.5. Add this amount to the value of Q
₃

and subtract this amount from
Q₁
. This gives you a wider boundary around the median than
the box does. Any data points that fall outside this boundary
are determined to be outliers.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 7. If there are no outliers (according to your
results of Step 6), draw lines from the upper and lower edges
of the box out to the minimum and maximum values in the data
set.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 8. If there are outliers (according to your
results of Step 6), indicate their location on the boxplot with
* signs. Instead of drawing a line from the edge of the box all
the way to the most extreme outlier, stop the line at the last
data value that isn’t an outlier.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Many if not most software packages
indicate outliers in a data set by using an asterisk (*) or
star symbol and use the procedure outlined in Step 6 to
identify outliers. However, not all packages use these symbols

and procedures; check to see what your package does before
analyzing your data with a boxplot. </p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A horizontal boxplot for ages of the
Best Actress Oscar award winners from 1928–2009 is shown in
Figure 7-7. You can see the numbers separating sections of the
boxplot match the five-number summary statistics shown in
Figure 7-3.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Boxplots can be
vertical (straight up and down) with the values on the axis
going from bottom (lowest) to top (highest); or they can be
horizontal, with the values on the axis going from left
(lowest) to right (highest). The next section shows you how to
interpret a boxplot.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-7: Boxplot of Best Actress
ages (1928–2009; n = 83 actresses).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0707.eps"
src="images/00052.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre5"><p id="a150" class="calibre6">
Interpreting a
boxplot</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Similar to a histogram (see the section
“Interpreting a histogram”), a boxplot can give you information
regarding the shape, center, and variability of a data set.
Boxplots differ from histograms in terms of their strengths and
weaknesses, as you see in the upcoming sections, but one of
their biggest strengths is how they handle skewed data.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Checking the shape
with caution!</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A boxplot can show whether a data set
is symmetric (roughly the same on each side when cut down the
middle) or skewed (lopsided). A symmetric data set shows the
median roughly in the middle of the box. Skewed data show a
lopsided boxplot, where the median cuts the box into two
unequal pieces. If the longer part of the box is to the right
(or above) the median, the data is said to be

skewed
right. If the longer part is to the
left (or below) the median, the data is skewed left
.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">As shown in the boxplot of the data in
Figure 7-7, the ages are skewed right. The part of the box to
the left of the median (representing the younger actresses) is
shorter than the part of the box to the right of the median
(representing the older actresses). That means the ages of the
younger actresses are closer together than the ages of the
older actresses. Figure 7-3 shows the descriptive statistics of
the data and confirms the right skewness: the median age (33
years) is lower than the mean age (35.69 years).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> If one side of
the box is longer than the other, it does not mean that side
contains more data. In fact, you can’t tell the sample size by
looking at a boxplot; it’s based on percentages, not counts.
Each section of the boxplot (the minimum to Q
₁
,
Q₁
 to the median, the median to Q
₃
, and
Q₃
 to the maximum) contains 25% of the data no matter what.
If one of the sections is longer than another, it indicates a
wider range in the values of data in that section (meaning the
data are more spread out). A smaller section of the boxplot
indicates the data are more condensed (closer together).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Although a
boxplot can tell you whether a data set is symmetric (when the
median is in the center of the box), it can’t tell you the
shape of the symmetry the way a histogram can. For example,
Figure 7-8 shows histograms from two different data sets, each
one containing 18 values that vary from 1 to 6. The histogram

```
on the left has an equal number of values in each group, and
the one on the right has two peaks at 2 and 5. Both histograms
show the data are symmetric, but their shapes are clearly
different.</span></span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-8:</span><span> Histograms of two
symmetric data sets.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0708.eps"
src="images/00053.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figure 7-9 shows the corresponding
boxplots for these same two data sets; notice they are exactly
the same. This is because the data sets both have the same
five-number summaries — they’re both symmetric with the same
amount of distance between </span><span><span
class="calibre16"><span class="italic">Q</span></span></span>
<span><span class="calibre41"><sub class="calibre42">1</sub>
</span></span><span>, the median, and </span><span><span
class="calibre16"><span class="italic">Q</span></span></span>
<span><span class="calibre41"><sub class="calibre42">3</sub>
</span></span><span>. However, if you just saw the boxplots and
not the histograms, you might think the shapes of the two data
sets are the same, when indeed they are not.</span></span></p>
</blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-9:</span><span> Boxplots of the two
symmetric data sets from Figure 7-8.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0709.eps"
src="images/00054.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Despite its weakness in detecting the
type of symmetry (you can add in a histogram to your analyses
to help fill in that gap), a boxplot has a great upside in that
you can identify actual measures of spread and center directly
from the boxplot, where on a histogram you can’t. A boxplot is
also good for comparing data sets by showing them on the same
graph, side by side.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> All graphs have
```

strengths and weaknesses; it’s always a good idea to show more than one graph of your data for that reason.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Measuring variability with IQR</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Variability in a data set that is described by the five-number summary is measured by the interquartile range (IQR). The
IQR
 is equal to Q
₃
 –
Q₁
, the difference between the 75th percentile and the 25th percentile (the distance covering the middle 50% of the data). The larger the IQR, the more variable the data set is.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">From Figure 7-3, the variability in age of the Best Actress winners as measured by the
IQR
 is Q₃
 – Q₁
 = 39 – 28 = 11 years. Of the group of actresses whose ages were closest to the median, half of them were within 11 years of each other when they won their awards.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Notice that the IQR ignores data below the 25th percentile or above the 75th, which may contain outliers that could inflate the measure of variability of the entire data set. So if data is skewed, the IQR
 is a more appropriate measure of variability than the standard deviation.</p></blockquote>

```
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Picking out the
center using the median</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The median, part of the five-number
summary, is shown by the line that cuts through the box in the
boxplot. This makes it very easy to identify. The mean,
however, is not part of the boxplot and can’t be determined
accurately by just looking at the boxplot.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You don’t see the mean on a boxplot
because boxplots are based completely on percentiles. If data
are skewed, the median is the most appropriate measure of
center. Of course you can calculate the mean separately and add
it to your results; it’s never a bad idea to show both.</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Investigating Old
Faithful’s boxplot</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The relevant descriptive statistics for
the Old Faithful geyser data are found in Figure 7-10.</span>
</span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 7-10:</span><span> Descriptive statistics
for Old Faithful data.</span></span></p>
</div>
<br class="calibre1"/>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0710.eps"
src="images/00055.jpg" class="calibre2"/></span></p>
</blockquote>
<br class="calibre1"/>
<br class="calibre1"/>
<br class="calibre1"/>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You can predict from the data set that
the shape will be skewed left a bit because the mean is lower
than the median by about 4 minutes. The </span><span><span
class="calibre16"><span class="italic">IQR</span></span></span>
<span> is </span><span><span class="calibre16"><span
class="italic">Q</span></span></span><span><span
class="calibre41"><sub class="calibre42">3</sub></span></span>
<span> – </span><span><span class="calibre16"><span
class="italic">Q</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
```

 = 81 – 60 = 21 minutes, which shows the amount of overall variability in the time between eruptions; 50% of the eruptions are within 21 minutes of each other.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A vertical boxplot for length of time between eruptions of the Old Faithful geyser is shown in Figure 7-11. You confirm that the data are skewed left because the lower part of the box (where the small values are) is longer than the upper part of the box.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You see the values of the boxplot in Figure 7-11 that mark the five-number summary and the information shown in Figure 7-10, including the
IQR
 of 21 minutes to measure variability. The center as marked by the median is 75 minutes; this is a better measure of center than the mean (71 minutes), which is driven down a bit by the left skewed values (the few that are shorter times than the rest of the data).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Looking at the boxplot (Figure 7-11), you see there are no outliers denoted by stars. However, note that the boxplot doesn’t pick up on the bimodal shape of the data that you see in Figure 7-5. You need a good histogram for that.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-11: Boxplot of eruption times for Old Faithful geyser (n
 = 222).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Denoting outliers</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Looking at the boxplot in Figure 7-7 for the Best Actress ages data, you see a set of outliers (seven in all) on the right side of the data set, marked by a group of stars (as described in Step 8 in the earlier section “Making a boxplot”). Three of the stars lie on top of one another because three actresses were the same age, 61, when they won their Oscars.</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">You verify these outliers by applying
the rule described in Step 6 of the section “Making a Boxplot.”
The IQR is 11 (from
Figure 7-3), so you take 11 ∗ 1.5 =
16.5 years. Add this amount to Q
₃
 and you get 39 + 16.5 = 55.5 years;
subtracting this amount from Q
₁
 you get 28 – 16.5 = 11.5 years. So an
actress whose age was below 11.5 years (that is, 11 years old
and under) or above 55.5 years (that is, 56 years old or over)
is considered to be an outlier.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Of course, the lower end of this
boundary (11.5 years) isn’t relevant because the youngest
actress was 21 (Figure 7-3 shows the minimum is 21). So you
know there aren’t any outliers on the low end of this data set.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, seven outliers are on the high
end of the data set, where the 56-and-over actresses’ ages are.
Table 7-2 shows the information on all seven outliers in the
Best Actress ages data set.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 7-2" src="images/00057.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The youngest of the outliers is 60
years old (Katharine Hepburn, 1967). Just to compare, the next
youngest age in the data set is 49 (Susan Sarandon, 1995). This
indicates a clear break in this data set.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Making mistakes when
interpreting a boxplot</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">It’s a common mistake to associate the
size of the box in a boxplot with the amount of data in the
data set. Remember that each of the four sections shown in the
boxplot contains an equal percentage (25%) of the data; the
boxplot just marks off the places in the data set that separate
those sections.</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> In particular,
if the median splits the box into two unequal parts, the larger
part contains data that’s more variable than the other part, in
terms of its range of values. However, there is still the same
amount of data (25%) in the larger part of the box as there is
in the smaller part.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Another common error involves sample
size. A boxplot is a one-dimensional graph with only one axis
representing the variable being measured. There is no second
axis that tells you how many data points are in each group. So
if you see two boxplots side-by-side and one of them has a very
long box and the other has a very short one, don’t conclude
that the longer one has more data in it. The length of the box
represents the variability in the data, not the number of data
values.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> When viewing or
making a boxplot, always make sure the sample size (
n
) is included as part of the title. You
can’t figure out the sample size otherwise.</p>
</blockquote>
<p id="a151" class="calibre6">Tackling Time Charts
</p>
<blockquote class="calibre9"><p class="calibre14">A
time chart
(also called a line graph) is a data
display used to examine trends in data over time (also known as
time series data). Time charts show time on the
x
-axis (for example, by month, year, or day) and
the values of the variable being measured on the
y
-axis (like birth rates, total sales, or
population size). Each point on the time chart summarizes all
the data collected at that particular time; for example, the
average of all pepper prices for January or the total revenue
for 2010.</p></blockquote>
<blockquote class="calibre5"><p id="a152" class="calibre6">
Interpreting
time charts</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> To interpret a
time chart, look for patterns and trends as you move across the
chart from left to right.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The time chart in Figure 7-12 shows the
ages of the Best Actress winners, in order of year won, from
1928–2009. Each dot indicates the age of a single actress, the
one that won the Oscar that year. You see a bit of a cyclical
pattern across time; that is, the ages go up, down, up, down,
up, down with at least some regularity. It’s hard to say what
may be going on here; many variables go into determining an
Oscar winner, including the type of movie, type of female role,
mood of the voters, and so forth, and some of these variables
may have a cyclical pattern to them.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 7-12 also shows a very faint
trend in age that is tending uphill; indicating that the Best
Actress Award winners may be winning their awards increasingly
later in life. Again, I wouldn’t make too many assumptions from
this result because the data has a great deal of variability.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">As far as variability goes, you see
that the ages represented by the dots do fluctuate quite a bit
on the y-axis (representing
age); all the dots basically fall between 20 and 80 years, with
most of them between 25 and 45 years, I’d say. This goes along
with the descriptive statistics found in Figure 7-3.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-12: Time Chart #1 for ages
of Best Actress Academy Award winners, 1928–2009.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0712.eps"
src="images/00058.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre5"><p id="a153" class="calibre6">
Understanding
variability: Time charts versus histograms
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Variability in
a histogram should not be confused with variability in a time
chart. If values change over time, they’re shown on a time
chart as highs and lows, and many changes from high to low
(over time) indicate lots of variability. So a flat line on a
time chart indicates no change and no variability in the values
across time. For example, if the price of a product stays the
same for 12 months in a row, the time chart for price would be
flat.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But when the heights of a histogram’s
bars appear flat, the data is spread out uniformly across all
the groups, indicating a great deal of variability in the data.
(For an example, refer to Figure 7-2a.)</p>
</blockquote>
<blockquote class="calibre5"><p id="a154" class="calibre6">
Spotting
misleading time charts</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">As with any graph, you have to evaluate
the units of the numbers being plotted. For example, it’s
misleading to chart the
 number of
crimes over time, rather than the crime rate
 (crimes per capita) — because the population size
of a city changes over time, crime rate is the appropriate
measure. Make sure you understand what numbers are being
graphed and examine them for fairness and appropriateness.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Watching the scale
and start/end points</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The scale on the vertical axis can make
a big difference in the way the time chart looks. Refer to
Figure 7-12 to see my original time chart of the ages for the
Best Actress Academy Award winners from 1928–2009 in increments
of 10 years. You see a fair amount of variability, as discussed
previously.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In Figure 7-12, the starting and ending
points on the vertical axis are 0 to 100, which creates a
little bit of extra white space on the top and bottom of the
picture. I could have used 10 and 90 as my start/end points,
but this graph looks reasonable.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Now what happens if I change the vertical axis? Figure 7-13 shows the same data, with start/end points of 20 and 80. The increments of 10 years appear longer than the increments of 10 years shown in Figure 7-12. Both of these changes in the graph exaggerate the differences in ages even more.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-13: Time Chart #2 for ages of Best Actress Oscar Award winners, 1928–2009.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> How do you decide which graph is the best one for your data? There is no perfect graph; there is no right or wrong answer; but there are limits. You can quickly spot problems just by zooming in on the scale and start/end points.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Simplifying excess data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A time chart of the time between eruptions for the Old Faithful data is shown in Figure 7-14. You see 222 dots on this graph; each one represents the time between one eruption and the next, for every eruption during a 16-day period.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This figure looks very complex; data are everywhere, there are too many points to really see anything, and you can’t find the forest for the trees. There is such a thing as having too much data, especially nowadays when you can measure data continuously and meticulously using all kinds of advanced technology. I’m betting they didn’t have a student standing by the geyser recording eruption times on a clipboard, for example!</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To get a clearer picture of the Old Faithful data, I combined all the observations from a single day and found its mean; I did this for all 16 days, and then I plotted all the means on a time chart in order. This reduced the data from 222 points to 16 points. The time chart is shown in Figure 7-15.</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">From this time chart I see a little bit
of a cyclical pattern to the data; every day or two it appears
to shift from short times between eruptions to longer times
between eruptions. While these changes are not definitive, it
does provide important information for scientists to follow up
on when studying the behavior of geysers like Old Faithful.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-14: Time chart showing time
between eruptions for Old Faithful Geyser (n
 = 222 consecutive observations).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0714.eps"
src="images/00060.jpg" class="calibre2"/></p>
</blockquote>
<div class="calibre1">
<p class="calibre38">Figure 7-15: Time chart showing daily
average time between eruptions for Old Faithful geyser (
n
 = 16 consecutive days).</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0715.eps"
src="images/00061.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> A time chart
condenses all the data for one unit of time into a single
point. By contrast, a histogram displays the entire sample of
data that was collected at that one unit of time. For example,
Figure 7-15 shows the daily average time between eruptions for
16 days. For any given day, you can make a histogram of all the
eruptions observed on that particular day. Displaying a time
chart of average times over 16 days accompanied by a histogram
summarizing all the eruptions for a particular day would be a
great one-two punch.</p></blockquote>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">Evaluating time
charts</p></blockquote><div
class="calibre33"> </div>

```html
<blockquote class="calibre5"><p class="calibre34"><span
class="calibre35"><span>Here is a checklist for evaluating time
charts, with a couple more thoughts added in:</span></span></p>
</blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Examine the scale and start/end points on the vertical
axis (the one showing the values of the data). Large increments
and/or lots of white space make differences look less dramatic;
small increments and/or a plot that totally fills the page
exaggerate differences.</span></span></p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If the amount of data you have is overwhelming, consider
boiling it down by finding means/medians for blocks of time and
plotting those instead.</span></span></p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Watch for gaps in the timeline on a time chart. For
example, it’s misleading to show equally spaced points on the
horizontal (time) axis for 1990, 2000, 2005, and 2010. This
happens when years are just treated like labels, rather than
real numbers.</span></span></p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><span
class="calibre35"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>As with any graph, take the units into account; be sure
they’re appropriate for comparison over time. For example, are
dollar amounts adjusted for inflation? Are you looking at
number of crimes, or the crime rate?</span></span></p>
</blockquote>
</blockquote></div><div class="calibre37"> </div>
<img alt="SB-End" src="images/00012.jpg" class="calibre2"/>
</div>
</div>  <div class="mbppagebreak" id="a155"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a149" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a156" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a148" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
```

```
href="#a154" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a155" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a152" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a153" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a150" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a151" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a143" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a142" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a145" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a144" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a147" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a146" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a159">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part III</span>
</span></span></span></p>
<p id="a157" class="calibre10"><span class="calibre11"><span
class="bold"><span>Distributions and the Central Limit
Theorem</span></span></span></p>
<div class="calibre13">
<img alt="9780470911082 pp0301.eps" src="images/00062.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Statisticians study populations; that’s their
bread and butter. They measure, count, or classify
characteristics of a population (using random variables); find
probabilities and proportions; and create (or estimate)
numerical summaries for the population (that is, param-eters
for the population). Sometimes you know a great deal about a
population from the start; sometimes it’s hazier. This part
studies populations under both scenarios.</span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">If a population fits a specific distribution,
tools are available for studying it. In Chapters 8 through 10,
you see three commonly used distributions: the binomial
distribution (for categorical data) and the normal and t-
distributions (for numerical data).</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">If the specifics about a population are
unknown (as happens most of the time), you take a sample and
generalize its results to the population. However, sample
results vary, and you need to take that into account. In
Chapter 11 you investigate sample variability, measure the
```

precision of your sample results, and find probabilities for their likelihood. From there you’ll be able to properly estimate parameters and test claims made about them, but that’s another Part — IV, to be exact.</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a158"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

      </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a177">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 8</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a160" class="calibre10"><span class="calibre11"><span
class="bold"><span>Random Variables and the Binomial
Distribution</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Identifying a binomial random variable</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Finding probabilities using a formula or table</span>
</span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Calculating the mean and variance</span></span></p>
</blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>S</span><span>cientists and engineers
often build models for the phenomena they are studying to make
predictions and decisions. For example, where and when is this
hurricane going to hit when it makes landfall? How many
accidents will occur at this intersection this year if it’s not
redone? Or, what will the deer population be like in a certain
region five years from now?</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To answer these questions, scientists
```

(usually working with statisticians) define a characteristic they are measuring or counting (such as number of intersections, location and time when a hurricane hits, population size, and so on) and treat it as a variable that changes in some random way, according to a certain pattern. They cleverly call them — you guessed it — random variables. In this chapter, you find out more about random variables, their types and characteristics, and why they are important. And you look at the details of one of the most common random variables: the binomial.</p></blockquote>
<p id="a161" class="calibre6">Defining a Random Variable
</p>
<blockquote class="calibre9"><p class="calibre14">A
random variable
is a characteristic, measurement, or count that changes randomly according to a certain set or pattern. Its notation is
X, Y
, Z, and so on. In this section, you see how different random variables are characterized and how they behave in the long term in terms of their means and standard deviations.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> In math you have variables like X and
Y
 that take on certain values depending on the problem (for example, the width of a rectangle), but in statistics the variables change in a random way. By
random,
 statisticians mean that you don’t know exactly what the next outcome will be but you do know that certain outcomes happen more frequently than others; everything’s not 50-50. (Like when I try to shoot baskets; it’s definitely not a 50% chance I’ll make one and 50% chance I’ll miss. It’s more like 5% chance of making it and a 95% chance of missing it.) You can use that information to better study data and populations and make good decisions. (For example, don’t put me in your basketball game to shoot free throws.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Data have different types: categorical
and numerical (see Chapter 4). While both types of data are
associated with random variables, I discuss only numerical
random variables here (this falls in line with most intro stat
courses as well). For information on analyzing categorical
variables, see Chapters 6 and 19.</p>
</blockquote>
<blockquote class="calibre5"><p id="a162" class="calibre6">
Discrete
versus continuous</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Numerical random variables represent
counts and measurements. They come in two different flavors:
discrete and continuous, depending on the type of outcomes that
are possible.</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Discrete
random variables: If the possible
outcomes of a random variable can be listed out using whole
numbers (for example, 0, 1, 2 . . . . , 10; or 0, 1, 2, 3), the
random variable is discrete.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Continuous
random variables: If the possible
outcomes of a random variable can only be described using an
interval of real numbers (for example, all real numbers from
zero to infinity), the random variable is continuous
.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Discrete random variables typically
represent counts — for example, the number of people who voted
yes for a smoking ban out of a random sample of 100 people
(possible values are 0, 1, 2, . . . , 100); or the number of
accidents at a certain intersection over one year’s time
(possible values are 0, 1, 2, . . .).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Discrete random
variables have two classes: finite and countably infinite. A

discrete random variable is finite
 if its list of possible values has a fixed
(finite) number of elements in it (for example, the number of
smoking ban supporters in a random sample of 100 voters has to
be between 0 and 100). One very common finite random variable
is the binomial, which is discussed in this chapter in detail.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A discrete random variable is
countably
infinite if its possible values can
be specifically listed out but they have no specific end. For
example, the number of accidents occurring at a certain
intersection over a 10-year period can take on possible values:
0, 1, 2, . . . (you know they end somewhere but you can’t say
where, so you list them all).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Continuous random variables typically
represent measurements, such as time to complete a task (for
example 1 minute 10 seconds, 1 minute 20 seconds, and so on) or
the weight of a newborn. What separates continuous random
variables from discrete ones is that they are
uncountably
infinite; they have too many
possible values to list out or to count and/or they can be
measured to a high level of precision (such as the level of
smog in the air in Los Angeles on a given day, measured in
parts per million).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Examples of commonly used continuous
random variables can be found in Chapter 9 (the normal
distribution) and Chapter 10 (the t
-distribution).</p></blockquote>
<blockquote class="calibre5"><p id="a163" class="calibre6">
Probability
distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A discrete random variable
X
 can take on a certain set of possible
outcomes, and each of those outcomes has a certain probability
of occurring. The notation used for any specific outcome is a
lowercase x. For example, say
you roll a die and look at the outcome. The random variable
<span

class="italic">X is the outcome of
the die (which takes on possible values of 1, 2, . . . , 6).
Now if you roll the die and get a 1, that’s a specific outcome,
so you write “x = 1.”
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The probability of any specific outcome
occurring is denoted p(x
), which you pronounce “p
of x.” It signifies the
probability that the random variable X
 takes on a specific value, which you call “little
x.
” For example, to denote the
probability of getting a 1 on a die, you write
p
(1).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Statisticians
use an uppercase X when they talk
about random variables in their general form; for example, “Let
X be the outcome of
the roll of a single die.” They use lowercase
x
 when they talk about specific outcomes of the
random variable, like
x = 1 or
x = 2.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A list or function showing all possible
values of a discrete random variable, along with their
probabilities, is called a probability
distribution, p
(x). For example, when

you roll a single die, the possible outcomes are 1, 2, 3, 4, 5, and 6, and each has a probability of 1/6 (if the die is fair). As another example, suppose 40% of renters living in an apartment complex own one dog, 7% own two dogs, 3% own three dogs, and 50% own zero dogs. For X = the number of dogs owned, the probability distribution for X is shown in Table 8-1.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre5"><p id="a164" class="calibre6">The mean and variance of a discrete random variable</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The mean of a random variable is the average of all the outcomes you would expect in the long term (over all possible samples). For example, if you roll a die a billion times and record the outcomes, the average of those outcomes is 3.5. (Each outcome happens with equal chance, so you average the numbers 1 through 6 to get 3.5.) However, if the die is loaded and you roll a 1 more often than anything else, the average outcome from a billion rolls is closer to 1 than to 3.5.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The notation for the mean of a random variable X is (pronounced “mu sub x”; or just “mu x”). Because you are looking at all the outcomes in the long term, it’s the same as looking at the mean of an entire population of values, which is why you denote it and not . (The latter represents the mean of a sample of values [see Chapter 5].) You put the

X in the
subscript to remind you that the variable this mean belongs to
is the X variable (as
opposed to a Y variable or some
other letter).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The variance
 of a random variable is roughly interpreted as
the average squared distance from the mean for all the outcomes
you would get in the long term, over all possible samples. This
is the same as the variance of the population of all possible
values. The notation for variance of a random variable
X
 is <img alt="9780470911082-
eq08004.eps" src="images/00067.jpg" class="calibre2"/>.
You say “sigma sub x, squared” or just
“sigma squared.”</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard deviation of a random
variable X is the square root
of the variance, denoted by <img alt="9780470911082-
eq08005.eps" src="images/00068.jpg" class="calibre2"/>
(say “sigma x”
or just “sigma”). It roughly
represents the average distance from the mean.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Just like for the mean, you use the
Greek notation to denote the variance and standard deviation of
a random variable. The English notation s
<sup
class="calibre4">2</sup>
and s
 represent the variance and standard deviation of a
sample of
individuals, not the entire population (see Chapter 5).
</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The variance is
in square units, so it can’t be easily interpreted. You use
standard deviation for interpretation because it is in the
original units of X. The standard
deviation can be roughly interpreted as the average distance
away from the mean.</p></blockquote>
<p id="a165" class="calibre6">Identifying a Binomial
</p>
<blockquote class="calibre9"><p class="calibre14">The most well-known and loved discrete
random variable is the binomial. Binomial
 means two names and is
associated with situations involving two outcomes; for example
yes/no, or success/failure (hitting a red light or not,
developing a side effect or not). This section focuses on the
binomial random variable — when you can use it, finding
probabilities for it, and finding its mean and variance.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A random variable is binomial (that is,
it has a binomial distribution) if the following four
conditions are met:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. There are a fixed number of trials
(n).</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Each trial has two possible
outcomes: success or failure.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The probability of success (call it
p) is the same for
each trial.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. The trials are independent, meaning
the outcome of one trial doesn’t influence that of any other.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Let X
equal the total number of successes in n
trials; if all four conditions are met,
X
 has a binomial distribution with probability of
success (on each trial) equal to p
.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The lowercase p
here stands for the probability of getting a success on
one single (individual) trial. It’s not the same as
p
(x),
which means the probability of
getting x successes
in n
trials.</p>
</blockquote>
<blockquote class="calibre5"><p id="a166" class="calibre6">
Checking
binomial conditions step by step</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You flip a fair coin 10 times and count
the number of heads (X). Does
X
 have a binomial distribution? You can
check by reviewing your responses to the questions and
statements in the list that follows:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Are there a fixed number of trials?
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> You’re flipping the coin 10 times,
which is a fixed number. Condition 1 is met, and
n
 = 10.</p></blockquote><div

class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Does each trial have only two possible
outcomes — success or failure?</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The outcome of each flip is either
heads or tails, and you’re interested in counting the number of
heads. That means success = heads, and failure = tails.
Condition 2 is met.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Is the probability of success the same for
each trial?</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Because the coin is fair, the
probability of success (getting a head) is p
 = 1/2 for each trial. You also know that 1 – 1/2 = 1/2
is the probability of failure (getting a tail) on each trial.
Condition 3 is met.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Are the trials independent?
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> You assume the coin is being flipped
the same way each time, which means the outcome of one flip
doesn’t affect the outcome of subsequent flips. Condition 4 is
met.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Because the random variable
X
 (the number of successes [heads] that
occur in 10 trials [flips]) meets all four conditions, you
conclude it has a binomial distribution with n
= 10 and p = 1/2.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But not every situation that appears
binomial actually is. Read on to see some examples of what I

mean.</p></blockquote>
<blockquote class="calibre5"><p id="a167" class="calibre6">
No fixed
number of trials</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose that you’re going to flip a
fair coin until you get four heads and you’ll count how many
flips it takes to get there; in this case X
 = number of flips. This certainly sounds like a binomial
situation: Condition 2 is met because you have success (heads)
and failure (tails) on each flip; condition 3 is met with the
probability of success (heads) being the same (0.5) on each
flip; and the flips are independent, so condition 4 is met.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, notice that X
isn’t counting the number of heads, it counts the number
of trials needed to get 4 heads. The number of successes
(X) is fixed rather
than the number of trials (n
). Condition 1 is not met, so X
 does not have a binomial distribution in this case.
</p></blockquote>
<blockquote class="calibre5"><p id="a168" class="calibre6">
More than
success or failure</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Some situations involve more than two
possible outcomes, yet they can appear to be binomial. For
example, suppose you roll a fair die 10 times and let
X
 be the outcome of each roll (1, 2, 3, . .
. , 6). You have a series of n
= 10 trials, they are independent, and the probability of
each outcome is the same for each roll. However, on each roll
you’re recording the outcome on a six-sided die, a number from
1 to 6. This is not a success/failure situation, so condition 2
is not met.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>However, depending on what you’re
recording, situations originally having more than two outcomes
can fall under the binomial category. For example, if you roll

```
a fair die 10 times and each time you record whether or not you
get a 1, then condition 2 is met because your two outcomes of
interest are getting a 1 (“success”) and not getting a 1
(“failure”). In this case,</span><span><span class="calibre16">
<span class="italic"> p</span></span></span><span> (the
probability of success) = 1/6, and 5/6 is the probability of
failure. So if </span><span><span class="calibre16"><span
class="italic">X</span></span></span><span> is counting the
number of 1s you get in 10 rolls, </span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> is a binomial random variable.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a169" class="calibre6">
<span class="calibre21"><span class="bold"><span>Trials are not
independent</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The independence condition is violated
when the outcome of one trial affects another trial. Suppose
you want to know opinions of adults in your city regarding a
proposed casino. Instead of taking a random sample of, say, 100
people, to save time you select 50 married couples and ask each
of them what their opinion is. In this case it’s reasonable to
say couples have a higher chance of agreeing on their opinions
than individuals selected at random, so the independence
condition 4 is not met.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a170" class="calibre6">
<span class="calibre21"><span class="bold"><span>Probability of
success (p) changes</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You have 10 people — 6 women and 4 men
— and you want to form a committee of 2 people at random. Let
</span><span><span class="calibre16"><span
class="italic">X</span></span></span><span> be the number of
women on the committee of 2. The chance of selecting a woman at
random on the first try is 6/10. Because you can’t select this
same woman again, the chance of selecting another woman is now
5/9. The value of</span><span><span class="calibre16"><span
class="italic"> p </span></span></span><span>has changed, and
condition 3 is not met.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> If the
population is very large (for example all U.S. adults), </span>
<span><span class="calibre16"><span class="italic">p </span>
</span></span><span>still changes every time you choose
someone, but the change is negligible, so you don’t worry about
it. You still say the trials are independent with the same
probability of success, </span><span><span class="calibre16">
```

```
<span class="italic">p.</span></span></span><span> (Life is so
much easier that way!)</span></span></p></blockquote>
<p id="a171" class="calibre6"><span class="calibre17"><span
class="bold"><span>Finding Binomial Probabilities Using a
Formula</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>After you identify that</span><span>
<span class="calibre16"><span class="italic"> X </span></span>
</span><span>has a binomial distribution (the four conditions
from the section “Checking binomial conditions step by step”
are met), you’ll likely want to find probabilities for </span>
<span><span class="calibre16"><span class="italic">X.</span>
</span></span><span> The good news is that you don’t have to
find them from scratch; you get to use established formulas for
finding binomial probabilities, using the values of </span>
<span><span class="calibre16"><span class="italic">n </span>
</span></span><span>and </span><span><span class="calibre16">
<span class="italic">p</span></span></span><span> unique to
each problem. Probabilities for a binomial random
variable</span><span><span class="calibre16"><span
class="italic"> X </span></span></span><span>can be found using
the following formula for </span><span><span class="calibre16">
<span class="italic">p</span></span></span><span>(</span><span>
<span class="calibre16"><span class="italic">x</span></span>
</span><span>):</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq08006.eps"
src="images/00069.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>where</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="italic">n</span>
</span></span><span> is the fixed number of trials.</span>
</span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="italic">x</span>
</span></span><span> is the specified number of successes.
</span></span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="italic">n</span>
</span></span><span> – </span><span><span class="calibre16">
```

x is the number of failures.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
p is the probability of success on any given trial.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
1 – p is the probability of failure on any given trial. (Note: Some textbooks use the letter q to denote the probability of failure rather than 1 – p.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">These probabilities hold for any value of X between 0 (lowest number of possible successes in n trials) and n (highest number of possible successes).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The number of ways to rearrange x successes among n trials is called “n choose x, ” and the notation is . It’s important to note that this math expression is not a fraction; it’s math shorthand to represent the number of ways to do these types of

rearrangements.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In general, to calculate “
n
 choose x,” you use the
following formula:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08008.eps"
src="images/00071.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">The notation n
! stands for n-factorial, the
number of ways to rearrange n
items. To calculate
n!, you
multiply n(n
 – 1)(n – 2) . . . (2)(1).
For example 5! is 5(4)(3)(2)(1) = 120; 2! is 2(1) = 2; and 1!
is 1. By convention, 0! equals 1.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you have to cross three traffic
lights on your way to work. Let X
 be the number of red lights you hit out of the three.
How many ways can you hit two red lights on your way to work?
Well, you could hit a green one first, then the other two red;
or you could hit the green one in the middle and have red ones
for the first and third lights, or you could hit red first,
then another red, then green. Letting G = green and R=red, you
can write these three possibilities as: GRR, RGR, RRG. So you
can hit two red lights on your way to work in three ways,
right?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Check the math. In this example, a
“trial” is a traffic light; and a “success” is a red light. (I
know, that seems weird, but a success is whatever you are
interested in counting, good or bad.) So you have
n
 = 3 total traffic lights, and you’re interested

in the situation where you get x
 = 2 red ones. Using the fancy notation,
<img alt="9780470911082-eq08009.eps"
src="images/00072.jpg" class="calibre2"/> means “3 choose
2” and stands for the number of ways to rearrange
2 successes in 3 trials.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To calculate “3 choose 2,” you do the
following:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08010.eps"
src="images/00073.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">This confirms the three possibilities
listed for getting two red lights.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now suppose the lights operate
independently of each other and each one has a 30% chance of
being red. Suppose you want to find the probability
distribution for X. (That is, a list
of all possible values of
X — 0,1,2,3 —
and their probabilities.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Before you dive into the calculations,
you first check the four conditions (from the section “Checking
binomial conditions step by step”) to see if you have a
binomial situation here. You have n
 = 3 trials (traffic lights) — check. Each trial is
success (red light) or failure (yellow or green light; in other
words, “non-red” light) — check. The lights operate
independently, so you have the independent trials taken care
of, and because each light is red 30% of the time, you know
p = 0.30 for each
light. So X = number of red
traffic lights has a binomial distribution. To fill in the
nitty gritties for the formulas, 1 – p
= probability of a non-red light = 1 – 0.30 = 0.70; and
the number of non-red lights is 3 – X

.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Using the formula for
p
(x), you obtain the
probabilities for x = 0, 1, 2, and 3
red lights:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08011.eps"
src="images/00074.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08012.eps"
src="images/00075.jpg" class="calibre2"/>;
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08013.eps"
src="images/00076.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08014.eps"
src="images/00077.jpg" class="calibre2"/>;
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08015.eps"
src="images/00078.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08016.eps"
src="images/00079.jpg" class="calibre2"/>; and
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08017.eps"
src="images/00080.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq08018.eps"
src="images/00081.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">The final probability distribution
for X
is shown in Table 8-2. Notice these
probabilities all sum to 1 because every possible value
of X
is listed and accounted for.

```html
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table 8-2" src="images/00082.jpg"
class="calibre2"/></span></p></blockquote>
<p id="a172" class="calibre6"><span class="calibre17"><span
class="bold"><span>Finding Probabilities Using the Binomial
Table</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The previous section deals with values
of </span><span><span class="calibre16"><span class="italic">n
</span></span></span><span>that are pretty small, but you may
wonder how you are going to handle the formula for calculating
binomial probabilities when </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span> gets large. No worries! A large range of binomial
probabilities are provided in the binomial table in the
appendix. Here’s how to use it:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Within the binomial table you see
several mini-tables; each one corresponds with a
different</span><span><span class="calibre16"><span
class="italic"> n </span></span></span><span>for a binomial
(</span><span><span class="calibre16"><span
class="italic">n</span></span></span><span> = 1, 2, 3, ..., 15,
and 20 are available). Each mini-table has rows and columns.
Running down the side of any mini-table, you see all the
possible values of</span><span><span class="calibre16"><span
class="italic"> X </span></span></span><span>from 0 through
</span><span><span class="calibre16"><span class="italic">n,
</span></span></span><span> each with its own row. The columns
of the binomial table represent various values of </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span> from 0.10 through 0.90.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p id="a173" class="calibre6">
<span class="calibre21"><span class="bold"><span>Finding
probabilities for specific values of X</span></span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To use the binomial table in the
appendix to find probabilities for </span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> = total number of successes in </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span> trials where </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span> is the probability
of success on any individual trial, follow these steps:</span>
</span></p></blockquote>
```

<blockquote class="calibre9"><p class="calibre30">1. Find the mini-table associated with your particular value of n (the number of trials). </p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the column that represents your particular value of p (or the one closest to it, if appropriate).</p> </blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Find the row that represents the number of successes (x) you are interested in.</p></blockquote> <div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Intersect the row and column from Steps 2 and 3. This gives you the probability for x successes, written as p(x).</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For the traffic light example from “Finding Binomial Probabilities Using a Formula,” you can use the binomial table (Table A-3 in the appendix) to verify the results found by the binomial formula shown back in Table 8-2. Go to the mini-table where n =3 and look in the column where p = 0.30. You see four probabilities listed for this mini-table: 0.343, 0.441, 0.189, and 0.027; these are the probabilities for <span>

```
<span class="calibre16"><span class="italic">X</span></span>
</span><span> = 0, 1, 2, and 3 red lights, respectively,
matching those from Table 8-2.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a174" class="calibre6">
<span class="calibre21"><span class="bold"><span>Finding
probabilities for X greater-than, less-than, or between two
values</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The binomial table (Table A-3 in the
appendix) shows probabilities for</span><span><span
class="calibre16"><span class="italic"> X </span></span></span>
<span>being equal to any value from 0 to </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span>, for a variety of </span><span><span class="calibre16">
<span class="italic">p</span></span></span><span>s. To find
probabilities for</span><span><span class="calibre16"><span
class="italic"> X </span></span></span><span>being less-than,
greater-than, or between two values, just find the
corresponding values in the table and add their probabilities.
For the traffic light example, you count the number of times
(</span><span><span class="calibre16"><span
class="italic">X</span></span></span><span>) that you hit a red
light (out of 3 possible lights). Each light has a 0.30 chance
of being red, so you have a binomial distribution with </span>
<span><span class="calibre16"><span class="italic">n </span>
</span></span><span>= 3 and </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span> = 0.30. If you want the probability that you hit more
than one red light, you find </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span> &gt; 1) by adding
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(2) + </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>(3) from Table A-3 to get 0.189 + 0.027 = 0.216.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The probability that you hit between 1
and 3 (inclusive) red lights is </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(1 </span><span><span class="calibre40">≤</span></span>
<span>
</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>
</span><span><span class="calibre40">≤</span></span><span> 3) =
0.441 + 0.189 + 0.027 = 0.657.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
```

```
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> You have to
distinguish between a </span><span><span class="calibre16">
<span class="italic">greater-than</span></span></span><span>
(&gt;) and a </span><span><span class="calibre16"><span
class="italic">greater-than-or-equal-to</span></span></span>
<span> (</span><span><span class="calibre40">≥</span></span>
<span>) probability when working with discrete random
variables. Repackaging the previous two examples, you see
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">x</span></span></span>
<span> &gt; 1) = 0.216 but </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>
</span><span><span class="calibre40">≥</span></span><span> 1) =
0.657. This is a non-issue for continuous random variables (see
Chapter 9).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> Other phrases to remember: </span>
<span><span class="calibre16"><span class="italic">at
least</span></span></span><span> means that number or higher,
and </span><span><span class="calibre16"><span
class="italic">at most</span></span></span><span> means that
number or lower. For example, the probability that</span><span>
<span class="calibre16"><span class="italic"> X </span></span>
</span><span>is at least 2 is </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">x </span></span></span><span><span
class="calibre40">≥</span></span><span> 2); the probability
that</span><span><span class="calibre16"><span class="italic">
X </span></span></span><span>is at most 2 is </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>
</span><span><span class="calibre40">≤</span></span><span> 2).
</span></span></p></blockquote>
<p id="a175" class="calibre6"><span class="calibre17"><span
class="bold"><span>Checking Out the Mean and Standard Deviation
of the Binomial</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Because the binomial distribution is so
commonly used, statisticians went ahead and did all the grunt
work to figure out nice, easy formulas for finding its mean,
variance, and standard deviation. (That is, they’ve already
```

applied the methods from the section “Defining a Random Variable” to the binomial distribution formulas, crunched everything out, and presented the results to us on a silver platter — don’t you love it when that happens?) The following results are what came out of it.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If X
 has a binomial distribution with n
 trials and probability of success p
 on each trial, then:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. The mean of X
 is .
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. The variance of X
 is .
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The standard deviation of
X
 is .
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you flip a fair coin 100 times and let
X be the number of heads; then X has a binomial distribution with n = 100 and
 p
= 0.50. Its mean is heads (which makes sense, because heads and tails are 50-50). The variance of X
is , which is in square units (so you can’t interpret it); and the standard

deviation is the square root of the variance, which is 5. That
means when you flip a coin 100 times, and do that over and
over, the average number of heads you’ll get is 50, and you can
expect that to vary by about 5 heads on average.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> The formula for the mean of a binomial
distribution has intuitive meaning. The p
 in the formula represents the probability of a success,
yes, but it also represents the proportion
 of successes you can expect in n
 trials. Therefore, the total number
 of successes you can expect — that is, the mean
of X — is <img
alt="9780470911082-eq08027.eps" src="images/00088.jpg"
class="calibre2"/>.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for variance has intuitive
meaning as well. The only variability in the outcomes of each
trial is between success (with probability p
) and failure (with probability 1 – p
). Over n trials, the
variance of the number of successes/failures is measured

by
<img alt="9780470911082-eq08028.eps"
src="images/00089.jpg" class="calibre2"/>. The standard
deviation is just the square root.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> If the value of
n is too large to use the binomial formula or the binomial
table to calculate probabilities (see the earlier sections in
this chapter), there’s an alternative. It turns out that if n
is large enough, you can use the normal distribution to get an
approximate answer for a binomial probability. The mean and
standard deviation of the binomial are involved in this

```
process. All the details are in Chapter 9.</span></span></p>
</blockquote>
</div>
</div>  <div class="mbppagebreak" id="a176"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a177" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a175" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a170" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a173" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a171" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a174" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a176" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a172" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a169" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a168" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a167" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a166" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a165" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a164" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a163" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a162" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a161" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a160" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a190">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 9</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a178" class="calibre10"><span class="calibre11"><span
class="bold"><span>The Normal Distribution</span></span></span>
</p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Understanding the normal and standard normal
distributions</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Going from start to finish when finding normal
probabilities</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Working backward to find percentiles</span></span></p>
</blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I</span><span>n your statistical
travels you’ll come across two major types of random variables:
discrete and continuous. </span><span><span class="calibre16">
<span class="italic">Discrete random variables</span></span>
</span><span> basically count things (number of heads on 10
coin flips, number of female Democrats in a sample, and so on).
The most well-known discrete random variable is the binomial.
(See Chapter 8 for more on discrete random variables and
binomials). A </span><span><span class="calibre16"><span
```

class="italic">continuous random variable
 is typically based on measurements; it either takes on
an uncountably infinite number of values (values within an
interval on the real line), or it has so many possible values
that it may as well be deemed continuous (for example, time to
complete a task, exam scores, and so on).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, you understand and
calculate probabilities for the most famous continuous random
variable of all time — the normal distribution. You also find
percentiles for the normal distribution, where you are given a
probability as a percent and you have to find the value of
X that’s associated
with it. And you can think how funny it would be to see a
statistician wearing a T-shirt that said “I’d rather be
normal.”</p></blockquote>
<p id="a179" class="calibre6">Exploring the Basics of theNormal
Distribution</p>
<blockquote class="calibre9"><p class="calibre14">A continuous random variable
X
 has a normal distribution if its values
fall into a smooth (continuous) curve with a bell-shaped
pattern. Each normal distribution has its own mean, denoted by
the Greek letter μ (say “mu”); and
its own standard deviation, denoted by the Greek letter
σ (say “sigma”). But no matter what their
means and standard deviations are, all normal distributions
have the same basic bell shape. Figure 9-1 shows some examples
of normal distributions.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 9-1: Three normal
distributions, with means and standard deviations of a) 90 and
30; b) 120 and 30; and c) 90 and 10, respectively.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0901.eps"
src="images/00090.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Every normal distribution has certain
properties. You use these properties to determine the relative
standing of any particular result on the distribution, and to

find probabilities. The properties of any normal distribution are as follows:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Its shape is symmetric (that is, when you cut it in half the two pieces are mirror images of each other).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Its distribution has a bump in the middle, with tails going down and out to the left and right.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The mean and the median are the same and lie directly in the middle of the distribution (due to symmetry).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Its standard deviation measures the distance on the distribution from the mean to the inflection point
 (the place where the curve changes from an “upside-down-bowl” shape to a “right-side-up-bowl” shape).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Because of its unique bell shape, probabilities for the normal distribution follow the Empirical Rule (full details in Chapter 5), which says the following:</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • About 68 percent of its values lie within one standard deviation of the mean. To find this range, take the value of the standard deviation, then find the mean plus this amount, and the mean minus this amount.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre29"> • About 95 percent of its values lie within two standard deviations of the mean. (Here you take 2 times the standard deviation, then add it to and subtract it from the mean.)</p></blockquote><div class="calibre19"> </div>

<blockquote class="calibre9"><p class="calibre29"> • Almost all of its values (about 99.7 percent of them) lie within three standard deviations of the mean. (Take 3 times the standard deviation and add it to and subtract it from the mean.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Precise probabilities for all possible intervals of values on the normal distribution (not just for those within 1, 2, or 3 standard deviations from the mean) are found using a table with minimal (if any) calculations. (The next section gives you all the info on this table.)</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Take a look again at Figure 9-1. To compare and contrast the distributions shown in Figure 9-1a, b, and c, you first see they are all symmetric with the signature bell shape. The examples in Figure 9-1a and Figure 9-1b have the same standard deviation, but their means are different; Figure 9-1b is located 30 units to the right of Figure 9-1a because its mean is 120 compared to 90. Figures 9-1a and c have the same mean (90), but Figure 9-1a has more variability than Figure 9-1c due to its higher standard deviation (30 compared to 10). Because of the increased variability, the values in Figure 9-1a stretch from 0 to 180 (approximately), while the values in Figure 9-1c only go from 60 to 120.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Finally, Figures 9-1b and c have different means and different standard deviations entirely; Figure 9-1b has a higher mean which shifts it to the right, and Figure 9-1c has a smaller standard deviation; its values are the most concentrated around the mean.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Noting the mean and standard deviation is important so you can properly interpret numbers located on a particular normal distribution. For example, you can compare where the number 120 falls on each of the normal distributions in Figure 9-1. In Figure 9-1a, the number 120 is one standard deviation above the mean (because the standard deviation is 30, you get 90 + 1
∗ 30 = 120). So on this first distribution, the number 120 is the upper value for the range where about 68%

of the data are located, according to the Empirical Rule (see
Chapter 5).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In Figure 9-1b, the number 120 lies
directly on the mean, where the values are most concentrated.
In Figure 9-1c, the number 120 is way out on the rightmost
fringe, 3 standard deviations above the mean (because the
standard deviation this time is 10, you get 90 + 3[10]=120). In
Figure 9-1c, values beyond 120 are very unlikely to occur
because they are beyond the range where about 99.7% of the
values should be, according to the Empirical Rule.
</p></blockquote>
<p id="a180" class="calibre6">Meeting the Standard Normal (Z-)
Distribution</p>
<blockquote class="calibre9"><p class="calibre14">One very special member of the normal
distribution family is called the standard normal distribution,
or Z-distribution. The
Z-distribution is
used to help find probabilities and percentiles for regular
normal distributions (
X). It serves
as the standard by which all other normal distributions are
measured.</p></blockquote>
<blockquote class="calibre5"><p id="a181" class="calibre6">
Checking out
Z</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The Z-
distribution is a normal distribution with mean zero and
standard deviation 1; its graph is shown in Figure 9-2. Almost
all (about 99.7%) of its values lie between –3 and +3 according
to the Empirical Rule. Values on the Z-
distribution are called z
-values, z-scores, or standard
scores. A z-value represents
the number of standard deviations that a particular value lies
above or below the mean. For example, z
 = 1 on the <span

class="italic">Z-distribution
represents a value that is 1 standard deviation above the mean.
Similarly, z = –1 represents a
value that is one standard deviation below the mean (indicated
by the minus sign on the
z-value). And a
z-
value of 0 is — you guessed it —
right on the mean. All
z-values are
universally understood.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you refer back to Figure 9-1 and the
discussion regarding where the number 120 lies on each normal
distribution in “Exploring the Basics of the Normal
Distribution,” you can now calculate z-
values to get a much clearer picture. In Figure 9-1a, the
number 120 is located one standard deviation above the mean, so
its z-value is 1. In
Figure 9-1b, 120 is equal to the mean, so its
z-
value is 0. Figure 9-1c shows that 120 is 3
standard deviations above the mean, so its z-
value is 3.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 9-2: The Z-
distribution has a mean of 0 and standard deviation of 1.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0902.eps"
src="images/00091.jpg" class="calibre2"/></p>
</blockquote>
<br class="calibre1"/>
<br class="calibre1"/>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> High standard
scores (z-values) aren’t always the best. For example, if
you’re measuring the amount of time needed to run around the
block, a standard score of +2 is a bad thing because your time
was two standard deviations above (more than) the overall

average time. In this case, a standard score of –2 would be much better, indicating your time was two standard deviations below (less than) the overall average time.</p>
</blockquote>
<blockquote class="calibre5"><p id="a182" class="calibre6">
Standardizing from X to Z</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Probabilities for any continuous distribution are found by finding the area under a curve (if you’re into calculus, you know that means integration; if you’re not into calculus, don’t worry about it). Although the bell-shaped curve for the normal distribution looks easy to work with, calculating areas under its curve turns out to be a nightmare requiring high-level math procedures (believe me, I won’t be going there in this book!). Plus, every normal distribution is different, causing you to repeat this process over and over each time you have to find a new probability.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To help get over this obstacle, statisticians worked out all the math gymnastics for one particular normal distribution, made a table of its probabilities, and told the rest of us to knock ourselves out. Can you guess which normal distribution they chose to crank out the table for?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Yes, all the basic results you need to find probabilities for any normal distribution (
X
) can be boiled down into one table based on the standard normal (Z-) distribution. This table is called the
Z-table and is found in the appendix. Now all you need is one formula that transforms values from your normal distribution (X) to the
Z-
distribution; from there you can use the Z-table to find any probability you need.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Changing an x-
value to a z-value is called
<span

class="italic">standardizing. The
so-called “z-formula” for standardizing an x-
value to a z-value is:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq09001.eps"
src="images/00092.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">You take your x-
value, subtract the mean of X,
 and divide by the standard deviation of
X.
 This gives you the corresponding standard score
(z-
value or z-
score).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Standardizing
is just like changing units (for example, from Fahrenheit to
Celsius). It doesn’t affect probabilities for X; that’s why you
can use the Z-table to find them!</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> You can standardize an
x-
value from any distribution (not just the normal)
using the z-formula. Similarly,
not all standard scores come from a normal distribution.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Because you
subtract the mean from your x-
values and divide everything by the standard deviation
when you standardize, you are literally taking the mean and
standard deviation of
X out of the
equation. This is what allows you to compare everything on the

scale from –3 to +3 (the
Z-distribution)
where negative values indicate being below the mean, positive
values indicate being above the mean, and a value of 0
indicates you’re right on the mean.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Standardizing also allows you to
compare numbers from different distributions. For example,
suppose Bob scores 80 on both his math exam (which has a mean
of 70 and standard deviation of 10) and his English exam (which
has a mean of 85 and standard deviation of 5). On which exam
did Bob do better, in terms of his relative standing in the
class?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Bob’s math exam score of 80
standardizes to a z-value of
<img alt="9780470911082-eq09018.eps" src="images/00093.jpg"
class="calibre2"/>. That tells us his math score is one
standard deviation above the class average. His English
exam score of 80 standardizes to a z
-value of <img alt="9780470911082-eq09019.eps"
src="images/00094.jpg" class="calibre2"/>, putting him
one standard deviation below the class average. Even though Bob
scored 80 on both exams, he actually did better on the math
exam than the English exam, relatively speaking.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> To interpret a
standard score, you don’t need to know the original score, the
mean, or the standard deviation. The standard score gives you
the relative standing of a value, which in most cases is what
matters most. In fact, on most national achievement tests, they
won’t even tell you what the mean and standard deviation were
when they report your results; they just tell you where you
stand on the distribution by giving you your z
-score.</p></blockquote>
<blockquote class="calibre5"><p id="a183" class="calibre6">
Finding
probabilities for Z with the Z-table</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>A full set of less-than probabilities
for a wide range of z-values is in the Z-table (Table A-1 in

```
the appendix). To use the </span><span><span class="calibre16">
<span class="italic">Z-</span></span></span><span>table to find
probabilities for the standard normal (</span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>) distribution, do the following:</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">1. Go to the row that represents the first digit
of your </span></span></span><span><span class="calibre16">
<span class="bold"><span class="italic">z</span></span></span>
</span><span><span class="calibre16"><span class="bold">-value
and the first digit after the decimal point.</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Go to the column that represents the second
digit after the decimal point of your </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">z</span></span></span></span><span><span
class="calibre16"><span class="bold">-value.</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 3. Intersect the row and column.</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> This result represents </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>(</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> &lt; </span><span>
<span class="calibre16"><span class="italic">z</span></span>
</span><span>), the probability that the random variable
</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> is less than the
number </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span> (also known as the
percentage of </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>-values that are
less than yours).</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For example, suppose you want to find
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span> &lt; 2.13). Using the Z-table, find the row for 2.1 and
the column for 0.03. Intersect that row and column to find the
```

probability: 0.9834. You find that p
(Z < 2.13) =
0.9834.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you want to look for
p
(Z < –2.13). You
find the row for –2.1 and the column for 0.03. Intersect the
row and column and you find 0.0166; that means
p
(Z < –2.13) equals
0.0166. (This happens to be one minus the probability that
Z is less than 2.13
because p(Z
 < +2.13) equals 0.9834. That’s true because the
normal distribution is symmetric; more on that in the following
section.)</p></blockquote>
<p id="a184" class="calibre6">Finding Probabilities for a Normal
Distribution</p>
<blockquote class="calibre9"><p class="calibre14">Here are the steps for finding a
probability when X has any normal
distribution:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Draw a picture of the distribution.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Translate the problem into one of the
following:
p

(X
 <

a<span><span

```
class="calibre16"><span class="bold">), </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold">(</span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">X</span></span></span></span><span><span
class="calibre16"><span class="bold"> &gt; </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">b</span></span></span></span><span><span
class="calibre16"><span class="bold">), or </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold">(</span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">a</span></span></span></span><span><span
class="calibre16"><span class="bold"> &lt; </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">X</span></span></span></span><span><span
class="calibre16"><span class="bold"> &lt; </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">b</span></span></span></span><span><span
class="calibre16"><span class="bold">). Shade in the area on
your picture.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 3. Standardize </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">a</span></span></span></span><span><span
class="calibre16"><span class="bold"> (and/or </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">b</span></span></span></span><span><span
class="calibre16"><span class="bold">) to a </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">z</span></span></span></span><span><span
class="calibre16"><span class="bold">-score using the </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">z</span></span></span></span><span><span
class="calibre16"><span class="bold">-formula:</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq09002.eps"
src="images/00095.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 4. Look up the </span></span></span><span><span
class="calibre16"><span class="bold"><span
```

class="italic">z-score on the

Z-table (Table A-1 in the
appendix) and find its corresponding probability.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> (See the section “Standardizing from
X to
Z
” for more on the Z
-table).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5a. If you need a “less-than” probability — that
is, p
(

X <
a) — you’re done.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5b. If you want a “greater-than” probability —
that is,
p

(X
>
b
) — take one
minus the result from Step 4.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5c. If you need a “between-two-values”
probability — that is, <span><span

```
class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold">(</span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">a </span></span></span></span><span><span
class="calibre16"><span class="bold">&lt; </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">X </span></span></span></span><span><span
class="calibre16"><span class="bold">&lt; </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">b</span></span></span></span><span><span
class="calibre16"><span class="bold">) — do Steps 1–4 for
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">b</span></span></span></span>
<span><span class="calibre16"><span class="bold"> (the larger
of the two values) and again for </span></span></span><span>
<span class="calibre16"><span class="bold"><span
class="italic">a</span></span></span></span><span><span
class="calibre16"><span class="bold"> (the smaller of the two
values), and subtract the results.</span></span></span></span>
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> The probability
that </span><span><span class="calibre16"><span
class="italic">X</span></span></span><span> is equal to any
single value is 0 for any continuous random variable (like the
normal). That’s because continuous random variables consider
probability as being area under the curve, and there’s no area
under a curve at one single point. This isn’t true of discrete
random variables.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Suppose, for example, that you enter a
fishing contest. The contest takes place in a pond where the
fish lengths have a normal distribution with mean </span>
<span>μ</span><span> = 16 inches and standard deviation </span>
<span>σ</span><span> = 4 inches.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Problem 1: What’s the chance of catching a small fish —
say, less than 8 inches?</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Problem 2: Suppose a prize is offered for any fish over
24 inches. What’s the chance of winning a prize?</span></span>
```

```
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Problem 3: What’s the chance of catching a fish between
16 and 24 inches?</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To solve these problems using the steps
that I just listed, first draw a picture of the normal
distribution at hand. Figure 9-3 shows a picture of </span>
<span><span class="calibre16"><span class="italic">X</span>
</span></span><span>’s distribution for fish lengths. You can
see where the numbers of interest (8, 16, and 24) fall.</span>
</span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 9-3:</span><span> The distribution of fish
lengths in a pond.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0903.eps"
src="images/00096.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Next, translate each problem into
probability notation. Problem 1 is really asking you to find
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> &lt; 8). For Problem 2, you want </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">X</span></span></span><span>
</span><span><span class="calibre16"><span class="bold">&gt;
</span></span></span><span> 24). And Problem 3 is looking for
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(16 &lt; </span>
<span><span class="calibre16"><span class="italic">X</span>
</span></span><span> &lt; 24).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Step 3 says change the </span><span>
<span class="calibre16"><span class="italic">x-</span></span>
</span><span>values to </span><span><span class="calibre16">
<span class="italic">z-</span></span></span><span>values using
the </span><span><span class="calibre16"><span
class="italic">z-</span></span></span><span>formula:</span>
</span></p></blockquote>
```

```
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq09003.eps"
src="images/00097.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For Problem 1 of the fish example, you
have the following:</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq09004.eps"
src="images/00098.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Similarly for Problem 2, </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>(</span><span><span class="calibre16"><span
class="italic">X</span></span></span><span> &gt; 24)
becomes</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq09005.eps"
src="images/00099.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>And Problem 3 translates from </span>
<span><span class="calibre16"><span class="italic">p</span>
</span></span><span>(16 &lt; </span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> &lt; 24) to</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq09006.eps"
src="images/00100.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figure 9-4 shows a comparison of the
</span><span><span class="calibre16"><span class="italic">X-
</span></span></span><span>distribution and </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>distribution for the values </span><span><span
class="calibre16"><span class="italic">x</span></span></span>
<span> = 8, 16, and 24, which standardize to </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span> = –2, 0, and +2, respectively.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Now that you have changed </span><span>
<span class="calibre16"><span class="italic">x-</span></span>
</span><span>values to </span><span><span class="calibre16">
<span class="italic">z-</span></span></span><span>values, you
move to Step 4 and find (or calculate) probabilities for those
```

```html
</span><span><span class="calibre16"><span class="italic">z-
</span></span></span><span>values using the </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>table (in the appendix). In Problem 1 of the fish
example, you want </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span> &lt; –2); go to the </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>table and look at the row for –2.0 and the column for
0.00, intersect them, and you find 0.0228 — according to Step
5a, you’re done. The chance of a fish being less than 8 inches
is equal to 0.0228.</span></span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 9-4:</span><span> Standardizing numbers
from a normal distribution (</span><span><span
class="calibre36"><span class="italic">X</span></span></span>
<span>) to numbers on the </span><span><span class="calibre36">
<span class="italic">Z-</span></span></span><span>distribution.
</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg0904.eps"
src="images/00101.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For Problem 2, find </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span>
</span><span><span class="calibre16"><span class="bold">&gt;
</span></span></span><span> 2.00). Because it’s a “greater-
than” problem, this calls for Step 5b. To be able to use the
</span><span><span class="calibre16"><span class="italic">Z-
</span></span></span><span>table, you need to rewrite this in
terms of a “less-than” statement. Because the entire
probability for the </span><span><span class="calibre16"><span
class="italic">Z-</span></span></span><span>distribution equals
1, we know </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>
</span><span><span class="calibre16"><span class="bold">&gt;
</span></span></span><span> 2.00) = 1 – </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> &lt; 2.00) = 1 –
```

```
0.9772 = 0.0228 (using the </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>table). So, the chance that a fish is greater than 24
inches is also 0.0228. (Note: The answers to Problems 1 and 2
are the same because the </span><span><span class="calibre16">
<span class="italic">Z-</span></span></span><span>distribution
is symmetric; refer to Figure 9-3.)</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In Problem 3, you find </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>(0 &lt; </span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> &lt; 2.00); this
requires Step 5c. First find </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>(</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> &lt; 2.00), which
is 0.9772 from the </span><span><span class="calibre16"><span
class="italic">Z-</span></span></span><span>table. Then find
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z </span></span></span>
<span>&lt; 0), which is 0.5000 from the </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>table. Subtract them to get 0.9772 – 0.5000 = 0.4772. The
chance of a fish being between 16 and 24 inches is 0.4772.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> The </span>
<span><span class="calibre16"><span class="italic">Z</span>
</span></span><span>-table does not list every possible value
of </span><span><span class="calibre16"><span class="italic">Z;
</span></span></span><span> it just carries them out to two
digits after the decimal point. Use the one closest to the one
you need. And just like in an airplane where the closest exit
may be behind you, the closest </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>-value may be the one that is lower than the one you
need.</span></span></p></blockquote>
<p id="a185" class="calibre6"><span class="calibre17"><span
class="bold"><span>Finding X When You Know the Percent</span>
</span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Another popular normal distribution
problem involves finding percentiles for </span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> (see Chapter 5 for a detailed rundown on percentiles).
```

That is, you are given the percentage or probability of being at or below a certain
x-value, and you have to find the x-value that corresponds to it. For example, if you know that the people whose golf scores were in the lowest 10% got to go to the tournament, you may wonder what the cutoff score was; that score would represent the 10th percentile.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> A percentile isn’t a percent. A percent is a number between 0 and 100; a percentile is a value of
X (a height, an IQ, a test score, and so on).</p></blockquote>
<blockquote class="calibre5"><p id="a186" class="calibre6">
Figuring out a percentile for a normal distribution</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Certain percentiles are so popular that they have their own names and their own notation. The three “named” percentiles are
Q₁
 — the first quartile, or the 25th percentile;
Q
₂ — the 2nd quartile (also known as the median
 or the 50th percentile); and Q
₃
 — the 3rd quartile or the 75th percentile. (See Chapter 5 for more information on quartiles.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are the steps for finding any percentile for a normal distribution X:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1a. If you’re given the probability (percent) less than <span><span class="calibre16"><span

```
class="bold"><span class="italic">x</span></span></span></span>
<span><span class="calibre16"><span class="bold"> and you need
to find </span></span></span><span><span class="calibre16">
<span class="bold"><span class="italic">x</span></span></span>
</span><span><span class="calibre16"><span class="bold">, you
translate this as: Find </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">a</span></span></span></span><span><span
class="calibre16"><span class="bold"> where </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold">(</span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">X</span></span></span></span><span><span
class="calibre16"><span class="bold"> &lt; </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">a</span></span></span></span><span><span
class="calibre16"><span class="bold">) = </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold"> (and </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre16"><span class="bold"> is the given
probability). That is, find the </span></span></span><span>
<span class="calibre16"><span class="bold"><span
class="italic">p</span></span></span></span><span><span
class="calibre41"><sup class="calibre4">th</sup></span></span>
<span><span class="calibre16"><span class="bold">percentile for
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">X</span></span></span></span>
<span><span class="calibre16"><span class="bold">. Go to Step
2.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 1b. If you’re given the probability (percent)
greater than </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">x</span></span></span></span><span><span
class="calibre16"><span class="bold"> and you need to find
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">x,</span></span></span>
</span><span><span class="calibre16"><span class="bold"> you
translate this as: Find </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">b</span></span></span></span><span><span
class="calibre16"><span class="bold"> where </span></span>
```

p(
X >
b) =
p (and
p is given). Rewrite this
as a percentile (less-than) problem: Find
b where
p(
X <
b) = 1 –
p. This means find the (1 –
p
)th percentile
for X.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the corresponding percentile for

Z by looking in the body of
the Z
-table (in the
appendix) and finding the probability that is closest to
<span

class="bold">p
 (from Step
1a) or 1 –
p
 (from
Step 1b)
.
 Find
the row and column this probability is in (using the table
backwards). This is the desired
z-value.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Change the z-value back into an
x
-value
(original units) by using <img
alt="9780470911082-eq09007.eps" src="images/00102.jpg"
class="calibre2"/>.
You’ve
(finally!) found the desired percentile for
X.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The formula in this step is just a
rewriting of the z-formula,
<img alt="9780470911082-eq09008.eps" src="images/00103.jpg"
class="calibre2"/>, so it’s solved for x
.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre6">Doing a low
percentile problem</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Look at the fish example used
previously in “Finding Probabilities for a Normal
Distribution,” where the lengths (<span

class="calibre16">X
) of fish in a pond have a normal distribution with mean
16 inches and standard deviation 4 inches. Suppose you want to
know what length marks the bottom 10 percent of all the fish
lengths in the pond. What percentile are you looking for?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Being at the bottom 10 percent means
you have a “less-than” probability that’s equal to 10 percent,
and you are at the 10th percentile.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now go to Step 1a in the preceding
section and translate the problem. In this case, because you’re
dealing with a “less-than” situation, you want to find
x
 such that p
(X <
x
) = 0.10. This represents the 10th percentile for
X. Figure 9-5 shows a
picture of this situation.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 9-5: Bottom 10 percent of fish
in the pond, according to length.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg0905.eps"
src="images/00104.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now go to Step 2, which says to find
the 10th percentile for
Z. Looking in
the body of the Z-table (in the
appendix), the probability closest to 0.10 is 0.1003, which
falls in the row for z = –1.2 and the
column for 0.08. That means the 10th percentile for
Z
 is –1.28; so a fish whose length is 1.28
standard deviations below the mean marks the bottom 10 percent

of all fish lengths in the pond.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But exactly how long is that fish, in
inches? In Step 3, you change the z-
value back to an
x-value (fish
length in inches) using the z
-formula solved for
x; you get
x = 16 + –1.28
∗ 4 = 10.88 inches. So 10.88 inches
marks the lowest 10 percent of fish lengths. Ten percent of the
fish are shorter than that.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Working with a higher
percentile</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now suppose you want to find the length
that marks the top 25 percent of all
the fish in the pond. This problem calls for Step 1b (in
“Finding a percentile for a normal distribution”) because being
in the top part of the distribution means you’re dealing with a
greater-than probability. The number you are looking for is
somewhere in the right tail (upper area) of the
X-
distribution, with p
 = 25 percent of the probability to its right and 1 –
p = 75 percent to its
left. Thinking in terms of the Z-
table and how it only uses less-than probabilities, you
need to find the 75th percentile for Z,
then change it to an
 x-value
.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Step 2: The 75th percentile of
Z
 is the <span

class="calibre16">z-
value where p(Z
 < z) = 0.75. Using the
Z-table (in the
appendix), you find the probability closest to 0.7500 is
0.7486, and its corresponding z-
value is in the row for 0.6 and column for 0.07. Put
these together and you get a z-
value of 0.67. This is the 75th percentile for
Z.
 In Step 3, change the z-
value back to an
x-value (length
in inches) using the z-formula solved for
x to get
x
 = 16 + 0.67 ∗ 4
= 18.68 inches. So, 75% of the fish are shorter than 18.68
inches. And to answer the original question, the top 25% of the
fish in the pond are longer than 18.68 inches.
</p></blockquote>
<blockquote class="calibre5"><p id="a187" class="calibre6">
Translating
tricky wording in percentile problems</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Some percentile
problems are especially challenging to translate. For example,
suppose the amount of time for a racehorse to run around a
track in a qualifying round has a normal distribution with mean
120 seconds and standard deviation 5 seconds. The best 10
percent of the times qualify; the rest don’t. What’s the cutoff
time for qualifying?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because “best times” mean “lowest
times” in this case, the percentage of times that lie
below

```
</span></span><span> the cutoff must be 10, and the percentage
</span><span><span class="calibre16"><span
class="italic">above</span></span></span><span> the cutoff must
be 90. (It’s an easy mistake to think it’s the other way
around.) The percentile of interest is therefore the 10th,
which is down on the left tail of the distribution. You now
work this problem the same way I worked Problem 1 regarding
fish lengths (see the section, “Finding Probabilities for a
Normal Distribution”). The standard score for the 10th
percentile is </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span> = –1.28 looking at
the</span><span><span class="calibre16"><span class="italic">
Z</span></span></span><span>-table (in the appendix).
Converting back to original units, you get </span><img
alt="9780470911082-eq09009.eps" src="images/00105.jpg"
class="calibre2"/><span> seconds. So the cutoff time needed for
a racehorse to qualify (that is, to be among the fastest 10%)
is 113.6 seconds. (Notice this number is less than the average
time of 120 seconds, which makes sense; a negative </span>
<span><span class="calibre16"><span class="italic">z</span>
</span></span><span>-value is what makes this happen.)</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> The 50</span><span><span
class="calibre41"><sup class="calibre4">th</sup></span></span>
<span> percentile for the normal distribution is the mean
(because of symmetry) and its </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>-score is zero. Smaller percentiles, like the 10th, lie
below the mean and have negative </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>-scores. Larger percentiles, like the 75th, lie above the
mean and have positive </span><span><span class="calibre16">
<span class="italic">z</span></span></span><span>-scores.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Here’s another style of wording that
has a bit of a twist: Suppose times to complete a statistics
exam have a normal distribution with a mean of 40 minutes and
standard deviation of 6 minutes. Deshawn’s time comes in at the
90th percentile. What percentage of the students are still
working on their exams when Deshawn leaves? Because Deshawn is
at the 90th percentile, 90 percent of the students have exam
times lower than hers. That means 90% of the students left
before Deshawn, so 100 – 90 = 10 percent of the students are
still working when Deshawn leaves.</span></span></p>
</blockquote>
```

<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> To be able to
decipher the language used to imply a percentile problem, look
for clues like the bottom 10% (also
known as the 10th percentile) and the top 10%
 (also known as the 90th percentile). For
the
best 10%, you must determine whether
low or high numbers qualify as “best.”</p>
</blockquote>
<p id="a188" class="calibre6">Normal Approximation to the Binomial
</p>
<blockquote class="calibre9"><p class="calibre14">Suppose you flip a fair coin 100 times
and you let X equal the number of
heads. What’s the probability that X
 is greater than 60? In Chapter 8, you solve problems
like this (involving fewer flips) using the binomial
distribution. For binomial problems where n
 (the number of trials) is small, you can either use the
direct formula (found in Chapter 8), the binomial table (found
in the appendix), or you can use technology if available (such
as a graphing calculator or Microsoft Excel).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">However, if n
 is large the calculations get unwieldy and the binomial
table runs out of numbers. If there’s no technology available
(like when taking an exam), what can you do to find a binomial
probability? Turns out, if n
 is large enough, you can use the normal distribution to
find a very close approximate answer with a lot less work.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But what do I mean by
n
 being “large enough”? To determine whether
n is large enough to

use what statisticians call the normal approximation to
the binomial, both of the following
conditions must hold:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
n
∗
p
≥ 10
(at least 10), where p is the probability
of success</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
n
∗ (1 – p
) ≥
 10 (at least 10), where 1 – p
 is the probability of failure</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">To find the normal approximation to the
binomial distribution when n
 is large, use the following steps:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Verify whether
n is large enough to use
the normal approximation by checking the two appropriate
conditions.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For the coin-flipping question, the
conditions are met because n
∗

p = 100
∗ 0.50 = 50, and n
∗ (1 – p
) = 100 ∗ (1 – 0.50) = 50,
both of which are at least 10. So go ahead with the normal
approximation.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Translate the problem into a probability
statement about X.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For the coin-flipping example, you
need to find p(X
 > 60).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Standardize the
x-value to a
z-value, using the

z-formula:
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-eq09010.eps"
src="images/00106.jpg" class="calibre2"/></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For the mean of the normal
distribution, use <img alt="9780470911082-eq09011.eps"
src="images/00107.jpg" class="calibre2"/> (the mean of
the binomial), and for the standard deviation

<img alt="9780470911082-eq09012.eps"
src="images/00108.jpg" class="calibre2"/>, use
<img alt="9780470911082-eq09013.eps" src="images/00109.jpg"
class="calibre2"/> (the standard deviation of the
binomial; see Chapter 8).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For the coin-flipping example, use
<img alt="9780470911082-eq09014.eps"
src="images/00110.jpg" class="calibre2"/> and <img
alt="9780470911082-eq09015.eps" src="images/00111.jpg"
class="calibre2"/> = <img alt="9780470911082-
eq09016.eps" src="images/00112.jpg" class="calibre2"/>.
Then put these values into the z
-formula to get <img alt="9780470911082-
eq09017.eps" src="images/00113.jpg" class="calibre2"/>.
To solve the problem, you need to find
p
(Z > 2).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> On an exam, you
won’t see µ and σ
 in the problem when you have a binomial distribution.
However, you know the formulas that allow you to calculate both
of them using n and
p
 (both of which will be given in the problem).
Just remember you have to do that extra step to calculate the
µ and σ
needed for the z-formula.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Proceed as you usually would for any normal
distribution. That is, do Steps 4 and 5 described in the
earlier section “Finding Probabilities for a Normal
Distribution
.
”</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

```
class="calibre15"><span> Continuing the example, </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span>(</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> &gt; 2.00) = 1 –
0.9772 = 0.0228 from the </span><span><span class="calibre16">
<span class="italic">Z</span></span></span><span>-table
(appendix). So the chance of getting more than 60 heads in 100
flips of a coin is only about 2.28 percent. (I wouldn’t bet on
it.)</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> When using the
normal approximation to find a binomial probability, your
answer is an </span><span><span class="calibre16"><span
class="italic">approximation</span></span></span><span> (not
exact) — be sure to state that. Also show that you checked both
necessary conditions for using the normal approximation.</span>
</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a189"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a187" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a185" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a184" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a190" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a186" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a181" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a179" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a183" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a182" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a178" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a189" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a188" style="min-width: 10px !important; min-height:
```

10px !important; border: solid 1px !important;">   </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a201">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 10</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a191" class="calibre10"><span class="calibre11"><span
class="bold"><span>The </span><span><span class="calibre36">
<span class="bold"><span class="italic">t</span></span></span>
</span><span>-Distribution</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Characteristics of the </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>distribution</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Relationship between </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span> and </span><span><span class="calibre16"><span
class="italic">t-</span></span></span>
<span>distributions</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Understanding and using the </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>table</span></span></p></blockquote><div
class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>he </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
```

distribution is one of the mainstays of data analysis.
You may have heard of the “t-
test” for example, which is often used to compare two
groups in medical studies and scientific experiments.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This short chapter covers the basic
characteristics and uses of the t-
distribution. You find out how it compares to the normal
distribution (more on that in Chapter 9) and how to use the
t-
table to find probabilities and
percentiles.</p></blockquote>
<p id="a192" class="calibre6">Basics of the t-Distribution
</p>
<blockquote class="calibre9"><p class="calibre14">In this section, you get an overview of
the t-distribution, its
main characteristics, when it’s used, and how it’s related to
the Z-distribution (see
Chapter 9).</p></blockquote>
<blockquote class="calibre5"><p id="a193" class="calibre6">
Comparing the
t- and Z-distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The normal distribution is that well-
known bell-shaped distribution whose mean is
μ and whose standard deviation is
σ (see Chapter 9 for more on the normal
distribution). The most common normal distribution is the
standard normal (also called the Z-
distribution), whose mean is 0 and standard deviation is
1.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The t-
distribution can be thought of as a cousin of the
standard normal distribution — it looks similar in that it’s
centered at zero and has a basic bell-shape, but it’s shorter
and flatter than the Z-<span>distribution. Its
standard deviation is proportionally larger compared to the

```
</span><span><span class="calibre16"><span class="italic">Z,
</span></span></span><span> which is why you see the fatter
tails on each side.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figure 10-1 compares the </span><span>
<span class="calibre16"><span class="italic">t-</span></span>
</span><span> and standard normal (</span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>) distributions in their most general forms.</span>
</span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 10-1:</span><span> Comparing the standard
normal (</span><span><span class="calibre36"><span
class="italic">Z-</span></span></span><span>) distribution to a
generic </span><span><span class="calibre36"><span
class="italic">t-</span></span></span><span>distribution.
</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg1001.eps"
src="images/00114.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>distribution is typically used to study the mean of a
population, rather than to study the individuals within a
population. In particular, it is used in many cases when you
use data to estimate the population mean — for example, to
estimate the average price of all the new homes in California.
Or when you use data to test someone’s claim about the
population mean — for example, is it true that the mean price
of all the new homes in California is $500,000?</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> These procedures are called </span>
<span><span class="calibre16"><span class="italic">confidence
intervals</span></span></span><span> and </span><span><span
class="calibre16"><span class="italic">hypothesis tests</span>
</span></span><span> and are discussed in Chapters 13 and 14,
respectively.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The connection between the normal
distribution and the </span><span><span class="calibre16"><span
class="italic">t-</span></span></span><span>distribution is
that the </span><span><span class="calibre16"><span
```

class="italic">t-distribution is
often used for analyzing the mean of a population if the
population has a normal distribution (or fairly close to it).
Its role is especially important if your data set is small or
if you don’t know the standard deviation of the population
(which is often the case).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When statisticians use the term
t-
distribution, they aren’t talking
about just one individual distribution. There is an entire
family of specific t-distributions,
depending on what sample size is being used to study the
population mean.
Each t-
distribution is distinguished by what statisticians call
its degrees of freedom.
In situations where you have one population and your sample
size is n, the degrees of
freedom for the corresponding t-
distribution is
n – 1. For
example, a sample of size 10 uses a t-
distribution with 10 – 1, or 9, degrees of freedom,
denoted t₉
 (pronounced tee sub-nine).
Situations involving two populations use different degrees of
freedom and are discussed in Chapter 15.</p>
</blockquote>
<blockquote class="calibre5"><p id="a194" class="calibre6">
Discovering
the effect of variability on t-distributions
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">t-distributions based
on smaller sample sizes have larger standard deviations than
those based on larger sample sizes. Their shapes are flatter;

their values are more spread out. That’s because results based on smaller data sets are more variable than results based on large data sets.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The larger the sample size is, the larger the degrees of freedom will be, and the more the t-distributions look like the standard normal distribution (Z-
distribution). A rough cutoff point where the
t
- and
Z-distributions become similar (at least similar enough for jazz or government work) is around n = 30.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 10-2 shows what different
t-
distributions look like for different sample sizes and how they all compare to the standard normal (Z-) distribution.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 10-2:
t-distributions for different sample sizes compared to the Z
-distribution.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<p id="a195" class="calibre6">Using the t-Table</p>
<blockquote class="calibre9"><p class="calibre14">Each normal distribution has its own mean and standard deviation that classify it, so finding probabilities for each normal distribution on its own is not the way to go. Thankfully, you can standardize the values of any normal distribution to become values on a standard normal

(Z-
) distribution (whose mean is 0 and
standard deviation is 1) and use a Z-
table (in the appendix) to find probabilities. (Chapter 9
has info on normal distributions.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In contrast, a t-
distribution is not classified by its mean and standard
deviation, but by the sample size of the data set being used
(n). Unfortunately,
there is no single “standard t-
distribution” that you can use to transform the numbers
and find probabilities on a table. Because it wouldn’t be
humanly possible to create a table of probabilities and
corresponding t-values for every
possible t-distribution,
statisticians created one table showing certain values of
t-distributions for a
selection of degrees of freedom and a selection of
probabilities. This table is called the t-table
 (it appears in the appendix). In this section,
you find out how to find probabilities, percentiles, and
critical values (for confidence intervals) using the
 t-
table.</p></blockquote>
<blockquote class="calibre5"><p id="a196" class="calibre6">
Finding
probabilities with the t-table</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Each row of the t-
table (in the appendix) represents a different
 t-
distribution, classified by its degrees of
freedom (df<span>). The columns
represent various common greater-than probabilities, such as
0.40, 0.25, 0.10, and 0.05. The numbers across a row indicate

```
the values on the</span><span><span class="calibre16"><span
class="italic"> t-</span></span></span><span>distribution
(the</span><span><span class="calibre16"><span class="italic">
t-</span></span></span><span>values) corresponding to the
greater-than probabilities shown at the top of the columns.
Rows are arranged by degrees of freedom.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Another term
for greater-than probability is </span><span><span
class="calibre16"><span class="italic">right-tail probability,
</span></span></span><span> which indicates that such
probabilities represent areas on the right-most end (tail) of
the </span><span><span class="calibre16"><span
class="italic">t</span></span></span><span>-distribution.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">For example, the second row of the <span>
<span class="calibre16"><span class="italic">t-</span></span>
</span>table is for the <span><span class="calibre16"><span
class="italic">t</span></span></span><span><span
class="calibre41"><sub class="calibre42">2 </sub></span>
</span>distribution (2 degrees of freedom, pronounced <span>
<span class="calibre16"><span class="italic">tee sub-two</span>
</span></span>). You see that the second number, 0.816, is the
value on the <span><span class="calibre16"><span
class="italic">t</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
distribution whose area to its right (its right-tail
probability) is 0.25 (see the heading for column 2). In other
words, the probability that <span><span class="calibre16"><span
class="italic">t</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
is greater than 0.816 equals 0.25. In probability notation,
that means <span><span class="calibre16"><span
class="italic">p</span></span></span>(<span><span
class="calibre16"><span class="italic">t</span></span></span>
<span><span class="calibre41"><sub class="calibre42">2 </sub>
</span></span>&gt; 0.816) = 0.25.</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The next number in row two of the
</span><span><span class="calibre16"><span
class="italic">t</span></span></span><span>-table is 1.886,
which lies in the 0.10 column. This means the probablity of
being greater than 1.886 on the </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span><span class="calibre41"><sub class="calibre42">2</sub>
```

 distribution is 0.10. Because 1.886 falls to the
right of 0.816, its right-tail probability is lower.</p>
</blockquote>
<blockquote class="calibre5"><p id="a197" class="calibre6">
Figuring
percentiles for the t-distribution</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You can also use the t-
table (in the appendix) to find percentiles for a
t
-distribution. A percentile
 is a number on a distribution whose less-than
probability is the given percentage; for example, the 95th
percentile of the t-distribution with
n – 1 degrees of
freedom is that value of
t<sub class="calibre42">n
– </sub><sub
class="calibre42">1</sub> whose left-tail
(less-than) probability is 0.95 (and whose right-tail
probability is 0.05). (See Chapter 5 for particulars on
percentiles.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you have a sample of size 10
and you want to find the 95th percentile of its corresponding
t-
distribution. You have
n
 – 1= 9 degrees of freedom, so you look at the row
for df = 9. The 95th
percentile is the number where 95% of the values lie below it
and 5% lie above it, so you want the right-tail area to be
0.05. Move across the row, find the column for 0.05, and you
get t₉
 = 1.833. This is the 95th percentile of the
t-
distribution with 9 degrees of freedom.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Now, if you increase the sample size to
n = 20, the value of
the 95th percentile decreases; look at the row for 20 – 1 = 19
degrees of freedom, and in the column for 0.05 (a right-tail
probability of 0.05) you find t
₁₉
 = 1.729. Notice that the 95th percentile
for the t₁₉
distribution is less than the 95th percentile for
the t₉
distribution (1.833). This is because larger degrees of
freedom indicate a smaller standard deviation and the
t
-values are more concentrated about the
mean, so you reach the 95th percentile with a smaller value of
t. (See the section
“Discovering the effect of variability on t-distributions,”
earlier in this chapter.)</p></blockquote>
<blockquote class="calibre5"><p id="a198" class="calibre6">
Picking out
t*-values for confidence intervals</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Confidence intervals
estimate population parameters, such as the population mean, by
using a statistic (for example, the sample mean) plus or minus
a margin of error. (See Chapter 13 for all the information you
need on confidence intervals and more.) To compute the margin
of error for a confidence interval, you need a
critical
value (the number of standard errors
you add and subtract to get the margin of error you want; see
Chapter 13). When the sample size is large (at least 30), you
use critical values on the Z
-distribution (shown in Chapter 13) to build the margin
of error. When the sample size is small (less than 30) and/or
the population standard deviation is unknown, you use the
t-
distribution to find critical

values.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To help you find critical values for
the t-distribution, you
can use the last row of the t-
table, which lists common confidence levels, such as 80%,
90%, and 95%. To find a critical value, look up your confidence
level in the bottom row of the table; this tells you which
column of the t-table you need.
Intersect this column with the row for your df
 (see Chapter 13 for degrees of freedom formulas). The
number you see is the critical value (or the t
*-value) for your confidence interval. For example, if
you want a t*-value for a 90%
confidence interval when you have 9 degrees of freedom, go to
the bottom of the table, find the column for 90%, and intersect
it with the row for df = 9. This gives
you a t*-
value of 1.833 (rounded).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Across the top
row of the t-table, you see
right-tail probabilities for the t-
distribution. But confidence intervals involve both left-
and right-tail probabilities (because you add and subtract the
margin of error). So half of the probability left from the
confidence interval goes into each tail. You need to take that
into account. For example, a t
*-value for a 90% confidence interval has 5% for its
greater-than probability and 5% for its less-than probability
(taking 100% minus 90% and dividing by 2). Using the top row of
the t-table, you would
have to look for 0.05 (rather than 10%, as you might be
inclined to do.) But using the bottom row of the table, you

just look for 90%. (The result you get using either method ends up being in the same column.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When looking for t*-
values for confidence intervals, use the bottom row of the t-table as your guide, rather than the headings at the top of the table.
</p></blockquote>
<p id="a199" class="calibre6">Studying Behavior Using the t-Table
</p>
<blockquote class="calibre9"><p class="calibre14">You can use computer software to calculate any probabilities, percentiles, or critical values you need for any t-distribution (or any other distribution) if it’s available to you. (On exams it may not be available.) However, one of the nice things about using a table to find probabilities (rather than using computer software) is that the table can tell you information about the behavior of the distribution itself — that is, it can give you the big picture. Here are some nuggets of big-picture information about the
t-distribution you can glean by scanning the t-
table (in the appendix).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In Figure 10-2, as the degrees of freedom increase, the values on each t-
distribution become more concentrated around the mean, eventually resembling the
Z-distribution. The t-table confirms this pattern as well. Because of the way the t-
table is set up, if you choose any column and move down through the numbers in the column, you’re increasing the degrees of freedom (and sample size) and keeping the right-tail probability the same. As you do this, you see the
 t-
values getting smaller and smaller, indicating the

t-values are becoming closer to (hence more concentrated around) the mean.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">I labeled the second-to-last row of the t-table with a z in the df column. This indicates the “limit” of the t-values as the sample size (n) goes to infinity. The t-values in this row are approximately the same as the z-values on the Z-table (in the appendix) that correspond to the same greater-than probabilities. This confirms what you already know: As the sample size increases, the t- and the Z-distributions look more and more alike. For example, the t-value in row 30
of the t-table corresponding to a right-tail probability of 0.05 (column 0.05) is 1.697. This lies close to z = 1.645, the value corresponding to a right-tail area of 0.05 on the Z-distribution. (See row Z of the t-table.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> It doesn’t take a super-large sample size for the values on the t-distribution to get close to the values on a Z-distribution. For example, when <span><span class="calibre16"><span

```
class="italic">n</span></span></span><span> = 31 and </span>
<span><span class="calibre16"><span class="italic">df</span>
</span></span><span> = 30, the values in the </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>table are already quite close to the corresponding values
on the </span><span><span class="calibre16"><span
class="italic">Z-</span></span></span><span>table.</span>
</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a200"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a200" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a201" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a192" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a193" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a191" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a196" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a197" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a194" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a195" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a198" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a199" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a215">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 11</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a202" class="calibre10"><span class="calibre11"><span
class="bold"><span>Sampling Distributions and the Central Limit
Theorem</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Understanding the concept of a sampling
distribution</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Putting the Central Limit Theorem to work</span></span>
</p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Determining the factors that affect precision</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>W</span><span>hen you take a sample of
data, it’s important to realize the results will vary from
sample to sample. Statistical results based on samples should
include a measure of how much those results are expected to
vary. When the media reports statistics like the average price
of a gallon of gas in the U.S. or the percentage of homes on
the market that were sold over the last month, you know they
didn’t sample every possible gas station or every possible home
sold. The question is, how much would their results change if
```

another sample was selected?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This chapter addresses this question by studying the behavior of means for all possible samples, and the behavior of proportions from all possible samples. By studying the behavior of all possible samples, you can gauge where your sample results fall and understand what it means when your sample results fall outside of certain expectations. </p></blockquote>
<p id="a203" class="calibre6">Defining a Sampling Distribution </p>
<blockquote class="calibre9"><p class="calibre14">A random variable is a characteristic of interest that takes on certain values in a random manner. For example, the number of red lights you hit on the way to work or school is a random variable; the number of children a randomly selected family has is a random variable. You use capital letters such as X or Y to denote random variables and you use small case letters x or y to denote actual outcomes of random variables. A distribution is a listing, graph, or function of all possible outcomes of a random variable (such as X) and how often each actual outcome (x), or set of outcomes, occurs. (See Chapter 8 for more details on random variables and distributions.) </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose a million of your closest friends each rolls a single die and records each actual outcome (x). A table or graph of all these possible outcomes (one through six) and how often they occurred represents the distribution of the random variable X. A graph of the distribution of X in this case is

shown in Figure 11-1a. It shows the numbers 1–6 appearing with equal frequency (each one occurring 1/6 of the time), which is what you expect over many rolls if the die is fair.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now suppose each of your friends rolls this single die 50 times (
n = 50) and records the average, . The graph of all their averages of all their samples represents the distribution of the random variable . Because this distribution is based on sample averages rather than individual outcomes, this distribution has a special name. It’s called the
sampling distribution of the sample mean,
. Figure 11-1b shows the sampling distribution of , the average of 50 rolls of a die.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 11-1b (average of 50 rolls) shows the same range (1 through 6) of outcomes as Figure 11-1a (individual rolls), but Figure 11-1b has more possible outcomes. You could get an average of 3.3 or 2.8 or 3.9 for 50 rolls, for example, whereas someone rolling a single die can only get whole numbers from 1 to 6. Also, the shape of the graphs are different; Figure 11-1a shows a flat shape, where each outcome is equally likely, and Figure 11-1b has a mound shape; that is, outcomes near the center (3.5) occur with high frequency and outcomes near the edges (1 and 6) occur with extremely low frequency. A detailed look at the differences and similarities in shape, center, and spread for individuals versus averages, and the reasons behind them, is the topic of the following sections. (See Chapter 8 if you need background info on shape, center, and spread of random variables before diving in.)</p></blockquote>
<p id="a204" class="calibre6">The Mean of a Sampling Distribution
</p>
<blockquote class="calibre9"><p class="calibre14">Using the die-rolling example from the preceding section, X is a random

variable denoting the outcome you can get from a single die
(assuming the die is fair). The mean of X
(over all possible outcomes) is denoted by <img
alt="9780470911082-eq11005.eps" src="images/00120.jpg"
class="calibre2"/> (pronounced mu sub-x
); in this case its value is 3.5 (as shown in
Figure 11-1a). If you roll a die 50 times and take the average,
the random variable <img alt="9780470911082-eq11006.eps"
src="images/00121.jpg" class="calibre2"/> represents any
outcome you could get. The mean of <img
alt="9780470911082-eq11007.eps" src="images/00122.jpg"
class="calibre2"/>, denoted <img
alt="9780470911082-eq11008.eps" src="images/00123.jpg"
class="calibre2"/> (pronounced mu sub-x-bar
) equals 3.5 as well. (You can see this
result in Figure 11-1b.)</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 11-1: Distributions of a)
individual rolls of one die; and b) average of 50 rolls of one
die.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1101.eps"
src="images/00124.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">This result is no coincidence! In
general, the mean of the population of all possible sample
means is the same as the mean of the original population.
(Notationally speaking, you write <img
alt="9780470911082-eq11009.eps" src="images/00125.jpg"
class="calibre2"/>.) It’s a mouthful, but it makes sense
that the average of the averages from all possible samples is
the same as the average of the population that the samples came
from. In the die rolling example, the average of the population
of all 50-roll averages equals the average of the population of
all single rolls (3.5).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Using subscripts on <img
alt="9780470911082-eq11010.eps" src="images/00126.jpg"
class="calibre2"/>, you can distinguish which mean you’re
talking about — the mean of X

(all individuals in a population) or the mean of
<img alt="9780470911082-eq11011.eps" src="images/00127.jpg"
class="calibre2"/> (all sample means from the
population).</p></blockquote>
<p id="a205" class="calibre6">Measuring Standard Error
</p>
<blockquote class="calibre9"><p class="calibre14">The values in any population deviate
from their mean; for instance, people’s heights differ from the
overall average height. Variability in a population of
individuals (X) is measured in
standard deviations
(see Chapter 5 for details on standard deviation). Sample means
vary because you’re not sampling the whole population, only a
subset; and as samples vary, so will their means. Variability
in the sample mean (<img alt="9780470911082-eq11012.eps"
src="images/00128.jpg" class="calibre2"/>) is measured in
terms of standard errors.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/>
Error here doesn’t
mean there’s been a mistake — it means there is a gap between
the population and sample results.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The standard error of the sample mean
is denoted by <img alt="9780470911082-eq11013.eps"
src="images/00129.jpg" class="calibre2"/>

(sigma sub-x-bar). Its
formula is <img alt="9780470911082-eq11014.eps"
src="images/00130.jpg" class="calibre2"/>, where
<img alt="9780470911082-eq11015.eps" src="images/00131.jpg"
class="calibre2"/> is population standard deviation

(sigma sub-x) and
n is size of each
sample. In the next sections you see the effect each of these
two components has on the standard error.</p>

</blockquote>
<blockquote class="calibre5"><p id="a206" class="calibre6">
Sample size
and standard error</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The first component of standard error
is the sample size, n. Because
 n
is in the denominator of the standard error
formula, the standard error decreases as n
increases. It makes sense that having more data gives
less variation (and more precision) in your results.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose X
is the time it takes for a clerical worker to type and
send one letter of recommendation, and say X
has a normal distribution with mean 10.5 minutes and
standard deviation 3 minutes. The bottom curve in Figure 11-2
shows the picture of the distribution of X,
 the individual times for all clerical workers in the
population. According to the Empirical Rule (see Chapter 9),
most of the values are within 3 standard deviations of the mean
(10.5) — between 1.5 and 19.5.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now take a random sample of 10 clerical
workers, measure their times, and find the average, <img
alt="9780470911082-eq11017.eps" src="images/00132.jpg"
class="calibre2"/>, each time. Repeat this process over
and over, and graph all the possible results for all possible
samples. The middle curve in Figure 11-2 shows the picture of
the sampling distribution of <img alt="9780470911082-
eq11018.eps" src="images/00133.jpg" class="calibre2"/>.
Notice that it’s still centered at 10.5 (which you expected)
but its variability is smaller; the standard error
in this case is <img alt="9780470911082-eq11019.eps"
src="images/00134.jpg" class="calibre2"/> minutes (quite
a bit less than 3 minutes, the standard deviation
of the individual times). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Looking at Figure 11-2, the average
times for samples of 10 clerical workers are closer to the mean
(10.5) than the individual times are. That’s because average

times don’t change as much from sample to sample as individual times change from person to person.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now take all possible random samples of 50 clerical workers and find their means; the sampling distribution is shown in the tallest curve in Figure 11-2. The standard error of goes down to minutes. You can see the average times for 50 clerical workers are even closer to 10.5 than the ones for 10 clerical workers. By the Empirical Rule, most of the values fall between 10.5 – 3(.42) = 9.24 and 10.5 + 3(.42) = 11.76. Larger samples give even more precision around the mean because they change even less from sample to sample. </p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 11-2: Distributions of times for 1 worker, 10 workers, and 50 workers.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Why is having more precision around the mean important? Because sometimes you don’t know the mean but want to determine what it is, or at least get as close to it as possible. How can you do that? By taking a large random sample from the population and finding its mean. You know that your sample mean will be close to the actual population mean if your sample is large, as Figure 11-2 shows (assuming your data are collected correctly; see Chapter 16 for details on collecting good data).</p>
</blockquote>
<blockquote class="calibre5"><p id="a207" class="calibre6">
Population standard deviation and standard error</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The second component of standard error involves the amount of diversity in the population (measured by standard deviation). In the standard error formula <img alt="9780470911082-eq11022.eps"

src="images/00138.jpg" class="calibre2"/> you see the
population standard deviation, <img alt="9780470911082-
eq11023.eps" src="images/00139.jpg" class="calibre2"/>,
is in the numerator. That means as the population
standard deviation increases, the standard error of the sample
means also increases. Mathematically this makes sense; how
about statistically?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you have two ponds full of fish
(call them pond #1 and pond #2), and you’re interested in the
length of the fish in each pond. Assume the fish lengths in
each pond have a normal distribution (see Chapter 9). You’ve
been told that the fish lengths in pond #1 have a mean of 20
inches and a standard deviation of 2 inches (see Figure 11-3a).
Suppose the fish in pond #2 also average 20 inches but have a
larger standard deviation of 5 inches (see Figure 11-3b).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Comparing Figures 11-3a and 11-3b, you
see the lengths for the two populations of fish have the same
shape and mean, but the distribution in Figure 11-3b (for pond
#2) has more spread, or variability, than the distribution
shown in Figure 11-3a (for pond #1). This spread confirms that
the fish in pond #2 vary more in length than those in pond #1.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now suppose you take a random sample of
100 fish from pond #1, find the mean length of the fish, and
repeat this process over and over. Then you do the same with
pond #2. Because the lengths of individual fish in pond #2 have
more variability than the lengths of individual fish in pond
#1, you know the average lengths of samples from pond #2 will
have more variability than the average lengths of samples from
pond #1 as well. (In fact, you can calculate their standard
errors using the formula earlier in this section to be 0.20 and
0.50, respectively.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Estimating the
population average is harder when the population varies a lot
to begin with — estimating the population average is much
easier when the population values are more consistent. The
bottom line is the standard error of the sample mean is larger
when the population standard deviation is larger.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 11-3: Distributions of fish

lengths a) in pond #1; b) in pond #2.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1103.eps"
src="images/00140.jpg" class="calibre2"/></p>
</blockquote>
<p id="a208" class="calibre6">Looking at the Shape of a Sampling
Distribution</p>
<blockquote class="calibre9"><p class="calibre14">Now that you know about the mean and
standard error of <img alt="9780470911082-eq11024.eps"
src="images/00141.jpg" class="calibre2"/>, the next step
is to determine the shape of the sampling distribution of
<img alt="9780470911082-eq11025.eps"
src="images/00142.jpg" class="calibre2"/>; that is, the
shape of the distribution of all possible sample means (all
possible values of <img alt="9780470911082-eq11026.eps"
src="images/00143.jpg" class="calibre2"/>) from all
possible samples. You proceed differently for different
conditions, which I divide into two cases: 1) the original
distribution for X (the population)
is normal, or has a normal distribution; and 2) the original
distribution for X (the population)
is
not normal, or is unknown.
</p></blockquote>
<blockquote class="calibre5"><p id="a209" class="calibre6">
Case 1: The
distribution of X is normal</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">If
 X has a normal
distribution, then <img alt="9780470911082-eq11027.eps"
src="images/00144.jpg" class="calibre2"/> does too, no
matter what the sample size n
 is. In the example regarding the amount of time (
X
) for a clerical worker to complete a task
(refer to the section “Sample size and standard error”), you
knew X had a normal
distribution (refer to the lowest curve in Figure 11-2). If you
refer to the other curves in Figure 11-2, you see the average

times for samples of n = 10 and
 n
 = 50 clerical workers, respectively, also
have normal distributions. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> When
X
 has a normal distribution, the sample
means also always have a normal distribution, no matter what
size samples you take, even if you take samples of only 2
clerical workers at a time.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The difference between the curves in
Figure 11-2 is not their means or their shapes, but rather
their amount of variability (how close the values in the
distribution are to the mean). Results based on large samples
vary less and will be more concentrated around the mean than
results from small samples or results from the individuals in
the population.</p></blockquote>
<blockquote class="calibre5"><p id="a210" class="calibre6">
Case 2: The
distribution of X is not normal — enter the Central Limit
Theorem</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If
 X has any
distribution that is not normal, or if its
distribution is unknown, you can’t automatically say the sample
mean (<img alt="9780470911082-eq11028.eps"
src="images/00145.jpg" class="calibre2"/>) has a normal
distribution. But incredibly, you can use a normal distribution
to approximate the
distribution of <img alt="9780470911082-eq11029.eps"
src="images/00146.jpg" class="calibre2"/> — if the sample
size is large enough. This momentous result is due to what
statisticians know and love as the Central Limit Theorem.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The
Central
Limit Theorem (abbreviated
CLT
) says that if<span

class="calibre16"> X
does not have a normal
distribution (or its distribution is unknown and hence can’t be
deemed to be normal), the shape of the sampling distribution of
<img alt="9780470911082-eq11030.eps"
src="images/00147.jpg" class="calibre2"/> is
approximately normal,
as long as the sample size, n,
 is large enough. That is, you get an approximate
 normal distribution for the means of large
samples, even if the distribution of the original values
(X) is
 not
normal.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/>Most statisticians agree that if
n is at least 30,
this approximation will be reasonably close in most cases,
although different distribution shapes for X
 have different values of n
 that are needed. The larger the sample size (
n
), the closer the distribution of the
sample means will be to a normal distribution.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Averaging a fair die
is approximately normal</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Consider the die rolling example from
the earlier section “Defining a Sampling Distribution.” Notice
in Figure 11-1a, the distribution of X
(the population of outcomes based on millions of single
rolls) is flat; the individual outcomes of each roll go from 1
to 6, and each outcome is equally likely.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Things change when you look at

averages. When you roll a die a large number of times (say a sample of 50 times) and look at your outcomes, you’ll probably find about the same number of 6s as 1s (note that 6 and 1 average out to 3.5); 5s as 2s (5 and 2 also average out to 3.5); and 4s as 3s (which also average out to 3.5 — do you see a pattern here?). So if you roll a die 50 times, you have a high probability of getting an overall average that’s close to 3.5. Sometimes just by chance things won’t even out as well, but that won’t happen very often with 50 rolls.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Getting an average at the extremes with 50 rolls is a very rare event. To get an average of 1 on 50 rolls, you need all 50 rolls to be 1. How likely is that? (If it happens to you, buy a lottery ticket right away, it’s the luckiest day of your life!) The same is true for getting an average near 6.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So the chance that your average of 50 rolls is close to the middle (3.5) is highest, and the chance of it being at or close to the extremes (1 or 6) is extremely low. As for averages between 1 and 6, the probabilities get smaller as you move farther from 3.5, and the probabilities get larger as you move closer to 3.5; in particular, statisticians show that the shape of the sampling distribution of sample means in Figure 11-1b is
approximately
normal as long as the sample size is large enough. (See Chapter 9 for particulars on the shape of the normal distribution.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Note that if you roll the die even more times, the chance of the average being close to 3.5 increases, and the sampling distribution of the sample means looks more and more like a normal distribution.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Averaging an unfair die is still approximately normal</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">However, sometimes the values of
 X
don’t occur with equal probability like they do when you roll a fair die. What happens then? For example, say the die isn’t fair, and the average value for many individual rolls turns out to be 2 instead of 3.5. This means the distribution of<span

class="italic"> X is skewed right (more low values like 1, 2, and 3, and fewer high values like 4, 5, and 6). But if the distribution of X (millions of individual rolls of this unfair die) is skewed right, how does the distribution of (average of 50 rolls of this unfair die) end up with an approximate normal distribution? </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Say that one person, Bob, is doing 50 rolls. What will the distribution of Bob’s outcomes look like? Bob is more likely to get low outcomes (like 1 and 2) and less likely to get high outcomes (like 5 and 6) — the distribution of Bob’s outcomes will be skewed right as well. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In fact, because Bob rolled his die a large number of times (50), the distribution of his individual outcomes has a good chance of matching the distribution of X (the outcomes from millions of rolls). However, if Bob had only rolled his die a few times (say, 6 times), he would be unlikely to even get the higher numbers like 5 and 6, and hence his distribution wouldn’t look as much like the distribution of X .</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you run through the results of each of a million people like Bob who rolled this unfair die 50 times, each of their million distributions will look very similar to each other and very similar to the distribution of X. The more rolls they make each time, the closer their distributions get to the distribution of X and to each other. And here is the key: If their distributions of outcomes have a similar shape, no matter what that similar shape is, then their averages will be similar as well. Some people will get higher averages than 2 by chance, and some will get lower averages by chance, but these types of averages get less and less likely the farther you get from 2. This means you’re getting an approximate normal distribution centered at 2.</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The big deal
is, it doesn’t matter if you started out with a skewed
distribution, or some totally wacky distribution for
 X.
Because each of them had a large sample
size (number of rolls), the distributions of each person’s
sample results end up looking similar, so their averages will
be similar, close together, and close to a normal distribution.
In fancy lingo, the distribution of <img
alt="9780470911082-eq11035.eps" src="images/00149.jpg"
class="calibre2"/> is approximately
normal as long as n
 is large enough. This is all due to the Central Limit
Theorem.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> In order for
the CLT to work when X does
 not
have a normal distribution, each person
needs to roll their die enough times (that is,
n
 must be large enough) so they have a good chance
of getting all possible values of X
, especially those outcomes that won’t occur as often. If
n
is too small, some folks will not
get the outcomes that have low probabilities and their means
will differ from the rest by more than they should. As a
result, when you put all the means together, they may not
congregate around a single value. In the end, the approximate
normal distribution may not show up.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Clarifying three
major points about the CLT</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">I want to alert you to a few sources of
confusion about the Central Limit Theorem before they happen to
you:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span

class="calibre15">
The CLT is needed only when the distribution of X is not a normal distribution or is unknown. It is not needed if X started out with a normal distribution. </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The formulas for the mean and standard error of   are not due to the CLT. These are just mathematical results that are always true. To see these formulas, check out the sections “The Mean of a Sampling Distribution” and “Measuring Standard Error,” earlier in this chapter. </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The n stated in the CLT refers to the size of the sample you take each time, not the number of samples you take. Bob rolling a die 50 times is one sample of size 50, so n = 50. If 10 people do it, you have 10 samples, each of size 50, and n is still 50. </p></blockquote><div class="calibre19"> </div>
<p id="a211" class="calibre6">Finding Probabilities for the Sample Mean</p>
<blockquote class="calibre9"><p class="calibre14">After you’ve established through the conditions addressed in case 1 or case 2 (see the previous sections) that has a normal or approximately normal distribution, you’re in luck. The normal distribution is a very friendly distribution that has a table for finding

probabilities and anything else you need. For example, you can find probabilities for by converting the
-value to a z-value and finding probabilities using the Z-table (provided in the appendix). (See Chapter 9 for all the details on the normal and Z-distributions.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The general conversion formula from -values to z-values is:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Substituting the appropriate values of the mean and standard error of , the conversion formula becomes:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> Don’t forget to divide by the square root of n in the denominator of z. Always divide by square root of n when the question refers to the average of the x- values.</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">Revisiting the clerical worker example
from the previous section “Sample size and standard error,”
suppose X is the time it
takes a randomly chosen clerical worker to type and send a
standard letter of recommendation. Suppose X
has a normal distribution, and assume the mean is 10.5
minutes and the standard deviation 3 minutes. You take a random
sample of 50 clerical workers and measure their times. What is
the chance that their average time is less than 9.5 minutes?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This question translates to finding
<img alt="9780470911082-eq11044.eps"
src="images/00158.jpg" class="calibre2"/>. As
 X
has a normal distribution to start with,
you know <img alt="9780470911082-eq11045.eps"
src="images/00159.jpg" class="calibre2"/> also has an
exact (not approximate) normal distribution. Converting to
z,
 you get:</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq11046.eps"
src="images/00160.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">So you want P(Z
 < –2.36), which equals 0.0091 (from the
Z
-table in the appendix). So the chance that a
random sample of 50 clerical workers average less than 9.5
minutes to complete this task is 0.91% (very small).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How do you find probabilities for
<img alt="9780470911082-eq11047.eps"
src="images/00161.jpg" class="calibre2"/> if
 X
is not normal, or
unknown? As a result of the CLT
, the distribution of

 X
can be non-normal or even unknown and as long
as n
is large enough, you can still find
approximate
probabilities for <img alt="9780470911082-
eq11048.eps" src="images/00162.jpg" class="calibre2"/>
using the standard normal (Z
-)distribution and the process described earlier. That
is, convert to a z-value and find
approximate probabilities using the Z
-table (in the appendix).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> When you use
the CLT to find a probability for <img
alt="9780470911082-eq11049.eps" src="images/00163.jpg"
class="calibre2"/> (that is, when the distribution of
X is
 not
normal or is unknown), be sure to say that
your answer is an approximation. You
also want to say the approximate answer should be close because
you’ve got a large enough
 n to use the
CLT. (If n is not large
enough for the CLT, you can use the t
-distribution in many cases — see Chapter 10.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Beyond actual
calculations, probabilities about <img
alt="9780470911082-eq11050.eps" src="images/00164.jpg"
class="calibre2"/> can help you decide whether an
assumption or a claim about a population mean is on target,
based on your data. In the clerical workers example, it was
assumed that the average time for all workers to type up a
recommendation letter was 10.5 minutes. Your sample averaged
9.5 minutes. Because the probability that they would average

less than 9.5 minutes was found to be tiny (0.0091), you either got an unusually high number of fast workers in your sample just by chance, or the assumption that the average time for all workers is 10.5 minutes was simply too high. (I’m betting on the latter.) The process of checking assumptions or challenging claims about a population is called hypothesis testing; details are in Chapter 14.</p></blockquote>
<p id="a212" class="calibre6">The Sampling Distribution of the Sample Proportion</p>
<blockquote class="calibre9"><p class="calibre14">The Central Limit Theorem (CLT) doesn’t apply only to sample means for numerical data. You can also use it with other statistics, including sample proportions for categorical data (see Chapter 6). The population proportion, p, is the proportion of individuals in the population who have a certain characteristic of interest (for example, the proportion of all Americans who are registered voters, or the proportion of all teenagers who own cellphones). The sample proportion, denoted (pronounced p-hat), is the proportion of individuals in the sample who have that particular characteristic; in other words, the number of individuals in the sample who have that characteristic of interest divided by the total sample size (n). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, if you take a sample of 100 teens and find 60 of them own cell-phones, the sample proportion of cellphone-owning teens is . This section examines the sampling distribution of all possible sample proportions, , from samples of size n from a population. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The sampling distribution of has the following properties:

</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Its mean, denoted by <img alt="9780470911082-
eq11054.eps" src="images/00169.jpg" class="calibre2"/>
(pronounced mu sub-p-hat), equals
the population proportion, p.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Its standard error, denoted by <img
alt="9780470911082-eq11055.eps" src="images/00170.jpg"
class="calibre2"/> (say sigma sub-p-hat
), equals:</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq11056.eps"
src="images/00171.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> (Note that because n
is in the denominator, the standard error decreases
as n
increases.)</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Due to the CLT, its shape is approximately
normal, provided that the sample size is
large enough. Therefore you can use the normal distribution to
find approximate probabilities for <img
alt="9780470911082-eq11057.eps" src="images/00172.jpg"
class="calibre2"/>. </p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The larger the sample size (n
), the closer the distribution of the sample proportion

is to a normal distribution.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> If you are interested in the number
(rather than the proportion) of individuals in your sample with
the characteristic of interest, you use the binomial
distribution to find probabilities for your results (see
Chapter 8).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> How large is
large enough for the CLT to work for sample proportions? Most
statisticians agree that both np
 and n(1 –
p
) should be greater than or equal to 10. That is,
the average number of successes (np
) and the average number of failures n
(1 – p)
needs to be at least 10.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To help illustrate the sampling
distribution of the sample proportion, consider a student
survey that accompanies the ACT test each year asking whether
the student would like some help with math skills. Assume
(through past research) that 38% of all the students taking the
ACT respond yes. That means p,
 the population proportion, equals 0.38 in this case. The
distribution of responses (yes, no) for this population are
shown in Figure 11-4 as a bar graph (see Chapter 6 for
information on bar graphs).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because 38% applies to all students
taking the exam, I use
p to denote the
population proportion, rather than <img
alt="9780470911082-eq11058.eps" src="images/00173.jpg"
class="calibre2"/>, which denotes sample proportions.
Typically <span

class="italic">p is unknown, but I’m giving it a value here to point out how the sample proportions from samples taken from the population behave in relation to the population proportion.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 11-4: Population percentages for responses to ACT math-help question.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now take all possible samples of n = 1,000 students from this population and find the proportion in each sample who said they need math help. The distribution of these sample proportions is shown in Figure 11-5. It has an approximate normal distribution with mean p = 0.38 and standard error equal to:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">(or about 1.5%).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> The approximate normal distribution works because the two conditions for the CLT are met: 1) np = 1,000(0.38) = 380 (≥ 10); and 2) n(1 – p) = 1,000(0.62) = 620 (also ≥ 10). And because n is so large

(1,000), the approximation is excellent.</p>
</blockquote>
<div class="calibre1">
<p class="calibre38">Figure 11-5: Sampling distribution of
proportion of students responding yes to ACT math-help question
for samples of size 1,000.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1105.eps"
src="images/00176.jpg" class="calibre2"/></p>
</blockquote>
<p id="a213" class="calibre6">Finding Probabilities for the Sample
Proportion</p>
<blockquote class="calibre9"><p class="calibre14">You can find probabilities for
<img alt="9780470911082-eq11060.eps" src="images/00177.jpg"
class="calibre2"/>, the sample proportion, by using the
normal approximation as long as the conditions are met (see the
previous section for those conditions). For the ACT test
example, you assume that 0.38 or 38% of all the students taking
the ACT test would like math help. Suppose you take a random
sample of 100 students. What is the chance that more than 45 of
them say they need math help? In terms of proportions, this is
equivalent to the chance that more than 45 ÷ 100 = 0.45 of them
say they need help; that is, <img alt="9780470911082-
eq11061.eps" src="images/00178.jpg" class="calibre2"/>.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To answer this question, you first
check the conditions: First, is np
 at least 10? Yes, because 100 ∗
 0.38 = 38. Next, is n
(1 – p) at least 10? Again
yes, because 100 ∗ (1 – 0.38) = 62
checks out. So you can go ahead and use the normal
approximation.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You make the conversion of the
<img alt="9780470911082-eq11062.eps" src="images/00179.jpg"
class="calibre2"/>-value to a z
-value using the following general equation:

```
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq11063.eps"
src="images/00180.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>When you plug in the numbers for this
example, you get:</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq11064.eps"
src="images/00181.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>And then you find P(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span> &gt; 1.44) = 1 – 0.9251 = 0.0749 using Table A-1 in the
appendix. So if it’s true that 0.38 percent of all students
taking the exam want math help, the chance of taking a random
sample of 100 students and finding more than 45 needing math
help is </span><span><span class="calibre16"><span
class="italic">approximately</span></span></span><span> 0.0749
(by the CLT).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/><span> As noted in the
previous section on sample means, you can use sample
proportions to check out a claim about a population proportion.
(This procedure is a hypothesis test for a population
proportion; all the details are found in Chapter 15.) In the
ACT example, the probability that more than 45% of the students
in a sample of 100 need math help (when you assumed 38% of the
population needed math help) was found to be 0.0749. Because
this probability is higher than 0.05 (the typical cutoff for
blowing the whistle on a claim about a population value), you
can’t dispute their claim that the percentage in the population
needing math help is only 38%. Our sample result is just not a
rare enough event. (See Chapter 15 for more on hypothesis
testing for a population proportion.)</span></span></p>
</blockquote>
</div>
</div>  <div class="mbppagebreak" id="a214"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a206" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a207" style="min-width: 10px !important; min-height:
```

```
10px !important; border: solid 1px !important;"> </a> <a
href="#a204" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a205" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a202" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a203" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a211" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a210" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a213" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a212" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a215" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a214" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a208" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a209" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a218">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part IV</span>
</span></span></span></p>
<p id="a216" class="calibre10"><span class="calibre11"><span
class="bold"><span>Guesstimating and Hypothesizing with
Confidence</span></span></span></p>
<div class="calibre13">
<img alt="9780470911082 pp0401.eps" src="images/00182.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Anytime you’re given a statistic by itself,
you haven’t really gotten the full story. The statistic alone
is missing the most important part: by how much that statistic
is expected to vary. All good estimates of population
parameters contain not just a statistic but also a margin of
error. This combination of a statistic plus or minus a margin
of error is called a confidence interval.</span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">Now suppose you’re already given a claim,
assumption, or target value for the population parameter, and
you want to test that claim. You do it with a hypothesis test
based on sample statistics. Because sample statistics will
vary, you need techniques that take this into account.</span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">This part gives you a general, intuitive look
at margin of error, confidence intervals, and hypothesis tests:
their function, formulas, calculations, influential factors,
and interpretation. You also get quick references and examples
for the most commonly used confidence intervals and hypothesis
```

```
tests.</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a217"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a218" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a217" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a216" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a232">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 12</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a219" class="calibre10"><span class="calibre11"><span
class="bold"><span>Leaving Room for a Margin of Error</span>
</span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Understanding and calculating margin of error</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Exploring the effect of sample size</span></span></p>
</blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Finding out what margin of error doesn’t measure</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>G</span><span>ood survey and experiment
researchers always include some measure of how accurate their
results are so that consumers of the information can put the
results into perspective. This measure is called the </span>
<span><span class="calibre16"><span class="italic">margin of
error (MOE)</span></span></span><span> — it’s a measure of how
close the sample statistic (one number that summarizes the
sample) is expected to be to the population parameter being
studied. (A population parameter is one number that summarizes
```

the population. Find out more about statistics and parameters in Chapter 4.)Thankfully, many journalists are also realizing the importance of the MOE in assessing information, so reports that include the margin of error are beginning to appear in the media. But what does the margin of error really mean, and does it tell the whole story?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This chapter looks at the margin of error and what it can and can’t do to help you assess the accuracy of statistical information. It also examines the issue of sample size; you may be surprised at how small a sample can be used to get a good handle on the pulse of America — or the world — if the research is done correctly.</p>
</blockquote>
<p id="a220" class="calibre6">Seeing the Importance of That Plus or Minus</p>
<blockquote class="calibre9"><p class="calibre14">Margin of error is probably not a new term to you. You’ve probably heard of it before, most likely in the context of survey results. For example, you may have heard someone report, “This survey had a margin of error of plus or minus three percentage points.” And you may have wondered what you’re supposed to do with that information and how important it really is. The truth is, the survey results themselves (with no MOE) are only a measure of how the sample
 of selected individuals felt about the issue; they don’t reflect how the entire population
 may have felt, had they all
 been asked. The margin of error helps you estimate how close you are to the truth about the population based on your sample data.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Results based on a sample won’t be exactly the same as what you would’ve found for the entire population, because when you take a sample, you don’t get information from everyone in the population. However, if the study is done right (see Chapters 16 and 17 for more about designing good studies), the results from the sample should be close to and representative of the actual values for the entire population, with a high level of confidence.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"

src="images/00008.jpg" class="calibre2"/> The MOE doesn’t mean someone made a mistake; all it means is that you didn’t get to sample everybody in the population, so you expect your sample results to vary from that population by a certain amount. In other words, you acknowledge that your results will change with subsequent samples and are only accurate to within a certain range — which can be calculated using the margin of error.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Consider one example of the type of survey conducted by some of the leading polling organizations, such as the Gallup Organization. Suppose its latest poll sampled 1,000 people from the United States, and the results show that 520 people (52%) think the president is doing a good job, compared to 48% who don’t think so. Suppose Gallup reports that this survey had a margin of error of plus or minus 3%. Now, you know that the majority (more than 50%) of the people in this sample approve of the president, but can you say that the majority of
all Americans approve of the president? In this case, you can’t. Why not?</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You need to include the margin of error (in this case, 3%) in your results. If 52% of
those sampled approve of the president, you can expect that the percent of the population of all Americans who approve of the president will be 52%, plus or minus 3%. Therefore, between 49% and 55% of all Americans approve of the president. That’s as close as you can get with your sample of 1,000. But notice that 49%, the lower end of this range, represents a minority, because it’s less than 50%. So you really can’t say that a majority of the American people support the president, based on this sample. You can only say you’re confident that between 49% and 55% of all Americans support the president, which may or may not be a majority.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span class="calibre15"><span>Think about the sample size for a moment. Isn’t it interesting that a sample of only 1,000 Americans out of a population of well over 310,000,000 can lead you to be within plus or minus only 3% on your survey results? That’s incredible! That means for large populations you only need to sample a tiny portion of the total to get close to the

```
true value (assuming, as always, that you have good data).
Statistics is indeed a powerful tool for finding out how people
feel about issues, which is probably why so many people conduct
surveys and why you’re so often bothered to respond to them as
well.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> When you are working with categorical
variables (those that record certain characteristics that don’t
involve measurements or counts; see Chapter 6), a quick-and-
dirty way to get a rough idea of the margin of error for
proportions, for any given sample size (</span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span>), is simply to find 1 divided by the square root of
</span><span><span class="calibre16"><span
class="italic">n</span></span></span><span>. For the Gallup
poll example,</span><span><span class="calibre16"><span
class="italic"> n </span></span></span><span>= 1,000, and its
square root is roughly 31.62, so the margin of error is roughly
1 divided by 31.62, or about 0.03, which is equivalent to 3%.
In the remainder of this chapter, you see how to get a more
accurate measure of the margin of error.</span></span></p>
</blockquote>
<p id="a221" class="calibre6"><span class="calibre17"><span
class="bold"><span>Finding the Margin of Error: A General
Formula</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The margin of error is the amount of
“plus or minus” that is attached to your sample result when you
move from discussing the sample itself to discussing the whole
population that it represents. Therefore, you know that the
general formula for the margin of error contains a “</span>
<span><span class="calibre40">±</span></span><span>” in front
of it. So, how do you come up with that plus or minus amount
(other than taking a rough estimate, as shown above)? This
section shows you how.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a222" class="calibre6">
<span class="calibre21"><span class="bold"><span>Measuring
sample variability</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Sample results vary, but by how much?
According to the Central Limit Theorem (see Chapter 11), when
sample sizes are large enough, the so-called sampling
distribution of the sample proportions (or the sample means)
follows a bell-shaped curve (or approximate normal distribution
— see Chapter 9). Some of the sample proportions (or sample
means) overestimate the population value and some underestimate
it, but most are close to the middle.</span></span></p>
```

</blockquote>
<blockquote class="calibre9"><p class="calibre14">And what’s in the middle of this
sampling distribution? If you average out the results from all
the possible samples you could take, the average is the actual
population proportion,
 in the case of categorical data, or the actual
population
average, in the case of numerical
data. Normally, you don’t know all the values of the
population, so you can’t look at all of the possible sample
results and average them out — but knowing something about all
the other sample possibilities does help you to measure the
amount by which you expect your own sample proportion (or
average) to vary. (See Chapter 11 for more on sample means and
proportions.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Standard errors
are the basic building blocks of the margin of error. The
standard error of a
statistic is basically equal to the standard deviation of the
population divided by the square root of n
(the sample size). This reflects the fact that the sample
size greatly affects how much that sample statistic is going to
vary from sample to sample. (See Chapter 11 for more about
standard errors.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The number of
standard errors you have to add or subtract to get the MOE
depends on how confident you want to be in your results (this
is called your confidence level).
Typically, you want to be about 95% confident, so the basic
rule is to add or subtract about 2 standard errors (1.96, to be
exact) to get the MOE (you get this from the Empirical Rule;
see Chapter 9). This allows you to account for about 95% of all
possible results that may have occurred with repeated sampling.
To be 99% confident, you add and subtract 2.58 standard errors.
(This assumes a normal distribution on large n; standard
deviation known. See Chapter 11.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You can be more precise about the

number of standard errors you have to add or subtract in order to calculate the MOE for any confidence level; if the conditions are right, you can use values on the standard normal (Z-) distribution. (See Chapter 13 for details.) For any given confidence level, a corresponding value on the standard normal distribution (called a z*-value) represents the number of standard errors to add and subtract to account for that confidence level. For 95% confidence, a more precise z*-value is 1.96 (which is “about” 2), and for 99% confidence, the exact z*-value is 2.58. Some of the more commonly used confidence levels (also known as percentage confidence), along with their corresponding z*-values, are given in Table 12-1.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To find a z*-value like those in Table 12-1, add to the confidence level to make it a less-than probability and find its corresponding z-value on the Z-table. For example, a 95% confidence level means the “between” probability is 95%, so the “less-than” probability is 95% plus 2.5% (half of what’s left), or 97.5%. Look up 0.975 in the body of the Z-table and find z * = 1.96 for a 95% confidence level.</p></blockquote>
<blockquote class="calibre5"><p id="a223" class="calibre6">Calculating margin of error for a sample proportion</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When a polling question asks people to choose from a range of answers (for example, “Do you approve or disapprove the president’s performance?”), the statistic used to report the results is the proportion of people from the sample who fell into a certain group (for example, the “approve” group). This is known as the sample proportion. You find this number by taking the number of people in the sample that fell into the group of interest, divided by the sample size, <span

class="calibre16">n
.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Along with the sample proportion, you
need to report a margin of error. The general formula for
margin of error for the sample proportion (if certain
conditions are met) is <img alt="9780470911082-
EQ12001.eps" src="images/00184.jpg" class="calibre2"/>,
where <img alt="9780470911082-EQ12002.eps"
src="images/00185.jpg" class="calibre2"/> is the sample
proportion, n is
the sample size, and
 z* is the
appropriate z*-value for your
desired level of confidence (from Table 12-1). Here are the
steps for calculating the margin of error for a sample
proportion:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Find the sample size,
n, and

the sample proportion, <img
alt="9780470911082-EQ12003.eps" src="images/00186.jpg"
class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The sample proportion is the number in
the sample with the characteristic of interest, divided by
n.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Multiply the sample proportion by
<img alt="9780470911082-EQ12004.eps"
src="images/00187.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span

```
class="bold">3. Divide the result by </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">n.</span></span></span></span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">4. Take the square root of the calculated value.
</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> You now have the standard error,
</span><img alt="9780470911082-EQ12005.eps"
src="images/00188.jpg" class="calibre2"/><span>.</span></span>
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">5. Multiply the result by the appropriate</span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic"> z*-</span></span></span></span><span>
<span class="calibre16"><span class="bold">value for the
confidence level desired.</span></span></span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> Refer to Table 12-1 for the
appropriate </span><span><span class="calibre16"><span
class="italic">z*</span></span></span><span>-value. If the
confidence level is 95%, the </span><span><span
class="calibre16"><span class="italic">z*</span></span></span>
<span>-value is 1.96.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Looking at the example involving
whether Americans approve of the president, you can find the
actual margin of error. First, assume you want a 95% level of
confidence, so</span><span><span class="calibre16"><span
class="italic"> z* </span></span></span><span>= 1.96. The
number of Americans in the sample who said they approve of the
president was found to be 520. This means that the sample
proportion, </span><img alt="9780470911082-EQ12006.eps"
src="images/00189.jpg" class="calibre2"/><span>, is 520 ÷ 1,000
= 0.52. (The sample size,</span><span><span class="calibre16">
<span class="italic"> n, </span></span></span><span>was 1,000.)
The margin of error for this polling question is calculated in
the following way:</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-EQ12007.eps"
src="images/00190.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
```

<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-EQ12008.eps"
src="images/00191.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">According to this data, you conclude
with 95% confidence that 52% of all Americans approve of the
president, plus or minus 3.1%.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Two conditions
need to be met in order to use a z*
-value in the formula for margin of error for a sample
proportion:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. You need to be sure that
<img alt="9780470911082-EQ12009.eps" src="images/00192.jpg"
class="calibre2"/> is at least 10.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. You need to make sure that
<img alt="9780470911082-EQ12010.eps" src="images/00193.jpg"
class="calibre2"/> is at least 10.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">In the preceding example of a poll on
the president, n = 1,000,
<img alt="9780470911082-EQ12011.eps" src="images/00194.jpg"
class="calibre2"/> = 0.52, and <img
alt="9780470911082-EQ12012.eps" src="images/00195.jpg"
class="calibre2"/> is 1 – 0.52 = 0.48. Now check the
conditions: <img alt="9780470911082-EQ12013.eps"
src="images/00196.jpg" class="calibre2"/> = 1,000
∗ 0.52 = 520, and
<img alt="9780470911082-EQ12014.eps"
src="images/00197.jpg" class="calibre2"/> = 1,000
∗ 0.48 = 480. Both of these numbers are at
least 10, so everything is okay.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Most surveys you come across are based
on hundreds or even thousands of people, so meeting these two
conditions is usually a piece of cake (unless the sample
proportion is very large or very small, requiring a larger
sample size to make the conditions work).</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> A sample proportion is the decimal
version of the sample percentage. In other words, if you have a
sample percentage of 5%, you must use 0.05 in the formula, not
5. To change a percentage into decimal form, simply divide by
100. After all your calculations are finished, you can change
back to a percentage by multiplying your final answer by 100%.
</p></blockquote>
<blockquote class="calibre5"><p id="a224" class="calibre6">
Reporting
results</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Including the margin of error allows
you to make conclusions beyond your sample to the population.
After you calculate and interpret the margin of error, report
it along with your survey results. To report the results from
the president approval poll in the previous section, you say,
“Based on my sample, 52% of all Americans approve of the
president, plus or minus a margin of error of 3.1%. I am 95%
confident in these results.”</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How does a real-life polling
organization report its results? Here’s an example from Gallup:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39">Based on the total random sample of 1,000 adults
in (this) survey, we are 95% confident that the margin of error
for our sampling procedure and its results is no more than
±
3.1
percentage points.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">It sounds sort of like that long list
of disclaimers that comes at the end of a car-leasing
advertisement. But now you can understand the fine print!
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Never accept
the results of a survey or study without the margin of error
for the study. The MOE is the only way to estimate how close
the sample statistics are to the actual population parameters
you’re interested in. Sample results vary, and if a different
sample had been chosen, a different sample result would have

been obtained; the MOE measures that amount of difference.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The next time you hear a media story
about a survey or poll that was conducted, take a closer look
to see if the margin of error is given; if it’s not, you should
ask why. Some news outlets are getting better about reporting
the margin of error for surveys, but what about other studies?
</p></blockquote>
<blockquote class="calibre5"><p id="a225" class="calibre6">
Calculating
margin of error for a sample mean</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">When a research question asks you to
estimate a parameter based on a numerical variable (for
example, “What’s the average age of teachers?”), the statistic
used to help estimate the results is the average of all the
responses provided by people in the sample. This is known as
the sample mean (or
average — see Chapter 5). And just like for sample proportions,
you need to report a MOE for sample means.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The general formula for margin of error
for the sample mean
(assuming a certain
condition is met) is <img alt="9780470911082-
EQ12015.eps" src="images/00198.jpg" class="calibre2"/>,
where σ
is the population standard
deviation, n is the sample
size, and z* is the
appropriate z*-value for your
desired level of confidence (which you can find in Table 12-1).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are the steps for calculating the
margin of error for a sample mean:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 1. Find the population standard deviation,
<img alt="9780470911082-EQ12023.eps"
src="images/00199.jpg" class="calibre2"/>, and the sample size,
n
.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The population standard deviation will
be given in the problem. </p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Divide the population standard deviation by the
square root of the sample size.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-EQ12016.eps"
src="images/00200.jpg" class="calibre2"/> gives you the
standard error.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Multiply by the appropriate
 z*-value (refer to Table 12-
1).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For example, the z*
-value is 1.96 if you want to be about 95% confident.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> The condition you need to meet in
order to use a z*-value in the
margin of error formula for a sample mean is either: 1) The
original population has a normal distribution to start with, or
2) The sample size is large enough so the normal distribution
can be used (that is, the Central Limit Theorem kicks in; see
Chapter 11). In general, the sample size, n,
should be above about 30 for the Central Limit

Theorem. Now, if it’s 29, don’t panic — 30 is not a magic number, it’s just a general rule of thumb. (The population standard deviation must be known either way.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you’re the manager of an ice cream shop, and you’re training new employees to be able to fill the large-size cones with the proper amount of ice cream (10 ounces each). You want to estimate the average weight of the cones they make over a one-day period, including a margin of error. Instead of weighing every single cone made, you ask each of your new employees to randomly spot check the weights of a random sample of the large cones they make and record those weights on a notepad. For n
= 50 cones sampled, the sample mean was found to be 10.3 ounces. Suppose the population standard deviation of

σ = 0.6 ounces is known.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What’s the margin of error? (Assume you want a 95% level of confidence.) It’s calculated this way:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">So to report these results, you say that based on the sample of 50 cones, you estimate that the average weight of all large cones made by the new employees over a one-day period is 10.3 ounces, with a margin of error of plus or minus 0.17 ounces. In other words, the range of likely values for the average weight of all large cones made for the day is estimated (with 95% confidence) to be between 10.30 – 0.17 = 10.13 ounces and 10.30 + 0.17 = 10.47 ounces. The new employees appear to be giving out too much ice cream (but I have a feeling the customers aren’t offended).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Notice in the ice-cream-cone example, the units are ounces, not percentages! When working with and reporting results about data, always remember what the units are. Also, be sure that statistics are reported with their correct units of measure, and if they’re not, ask what the units are.</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> In cases where
n is too small (in
general, less than 30) for the Central Limit Theorem to be
used, but you still think the data came from a normal
distribution, you can use a t*
-value instead of a
z*-value in
your formulas. A t*-value is one that
comes from a t-distribution with
n
– 1 degrees of freedom. (Chapter 10
gives you all the in-depth details on the t
-distribution.) In fact, many statisticians go ahead and
use t*-values instead of
z*
-values consistently, because if the
sample size is large,
t*-values and
z*
-values are approximately equal
anyway. In addition, for cases where you don’t know the
population standard deviation, σ, you
can substitute it with
s, the sample
standard deviation; from there you use a t*
-value instead of a
z*-value in
your formulas as well.</p></blockquote>
<blockquote class="calibre5"><p id="a226" class="calibre6">
Being
confident you’re right</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you want to be more
 than 95% confident about your results, you need
to add and subtract more than 1.96 standard errors (see Table
12-1). For example, to be 99% confident, you add and subtract
2.58 standard errors to obtain your margin of error. More
confidence means a larger margin of error, though (assuming the

sample size stays the same); so you have to ask yourself if it’s worth it. When going from 95% to 99% confidence, the z* -value increases by 2.58 – 1.96 = 0.62 (see Table 12-1). Most people don’t think adding and subtracting this much more of a MOE is worthwhile, just to be 4% more confident (99% versus 95%) in the results obtained. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You can never be completely certain that your sample results do reflect the population, even with the margin of error included. Even if you’re 95% confident in your results, that actually means that if you repeat the sampling process over and over, 5% of the time the sample won’t represent the population well, simply due to chance (not because of problems with the sampling process or anything else). In these cases, you would miss the mark. So all results need to be viewed with that in mind. </p>
</blockquote>
<p id="a227" class="calibre6">Determining the Impact of Sample Size
</p>
<blockquote class="calibre9"><p class="calibre14">The two most important ideas regarding sample size and margin of error are the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Sample size and margin of error have an inverse relationship.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
After a point, increasing n
 beyond what you already have gives you a diminished return.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">This section illustrates both concepts. </p></blockquote>
<blockquote class="calibre5"><p id="a228" class="calibre6">
Sample size and margin of error</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">The relationship between margin of error and sample size is simple: As the sample size increases, the margin of error decreases. This relationship is called an inverse because the two move in opposite directions. If you think about it, it makes sense that the more information you have, the more accurate your results are going to get (in other words, the smaller your margin of error will get). (That assumes, of course, that the data were collected and handled properly.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> In the previous section, you see that the impact of a larger confidence level is a larger MOE. But if you increase the sample size, you can offset the larger MOE and bring it down to a reasonable size! Find out more about this concept in Chapter 13.
</p></blockquote>
<blockquote class="calibre5"><p id="a229" class="calibre6">
Bigger isn’t always (that much) better!</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In the example of the poll involving the approval rating of the president (see the earlier section “Calculating margin of error for a sample proportion”), the results of a sample of only 1,000 people from well over 310,000,000 residents in the United States could get to within about 3% of what the whole population would have said, if they had all been asked. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Using the formula for margin of error for a sample proportion, you can look at how the margin of error changes dramatically for samples of different sizes. Suppose in the presidential approval poll that
 n
was 500 instead of 1,000. (Recall that = 0.52 for this example.) Therefore the margin of error for 95% confidence is
, which is equivalent to 4.38%.
When n = 1,000 in the same example, the margin of error (for 95% confidence)
is , which is equal to 3.10%. If n is increased to

1,500, the margin of error (with the same level of confidence) becomes , or 2.53%. Finally, when
n = 2,000, the margin of error is , or 2.19%.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Looking at these different results, you can see that larger sample sizes decrease the MOE, but after a certain point, you have a diminished return. Each time you survey one more person, the cost of your survey increases, and going from a sample size of, say, 1,500 to a sample size of 2,000 decreases your margin of error by only 0.34% (one third of one percent!) — from 0.0253 to 0.0219. The extra cost and trouble to get that small decrease in the MOE may not be worthwhile. Bigger isn’t always that much better!
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But what may really surprise you is that bigger can actually be worse! I explain this surprising fact in the following section.</p></blockquote>
<blockquote class="calibre5"><p id="a230" class="calibre6">
Keeping margin of error in perspective</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The margin of error is a measure of how close you expect your sample results to represent the entire population being studied. (Or at least it gives an upper limit for the amount of error you should have.) Because you’re basing your conclusions about the population on your one sample, you have to account for how much those sample results could vary just due to chance.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Another view of margin of error is that it represents the maximum expected distance between the sample results and the actual population results (if you’d been able to obtain them through a census). Of course if you had the absolute truth about the population, you wouldn’t be trying to do a survey, would you?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Just as important as knowing what the margin of error measures is realizing what the margin of error does not measure. The

margin of error does not measure anything other than chance variation. That is, it doesn’t measure any bias or errors that happen during the selection of the participants, the preparation or conduct of the survey, the data collection and entry process, or the analysis of the data and the drawing of the final conclusions.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A good slogan to remember when examining statistical results is “garbage in equals garbage out.” No matter how nice and scientific the margin of error may look, remember that the formula that was used to calculate it doesn’t have any idea of the quality of the data that the margin of error is based on. If the sample proportion or sample mean was based on a biased sample
 (one that favored certain people over others), a bad design, bad data-collection procedures, biased questions, or systematic errors in recording, then calculating the margin of error is pointless because it won’t mean a thing.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, 50,000 people surveyed sounds great, but if they were all visitors to a certain Web site, the margin of error for this result is bogus because the calculation is all based on biased results! In fact, many extremely large samples are the result of biased sampling procedures. Of course, some people go ahead and report them anyway, so you have to find out what went into the formula: good information or garbage? If it turns out to be garbage, you know what to do about the margin of error. Ignore it. (For more information on errors that can take place during a survey or experiment, see Chapters 16 and 17, respectively.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The Gallup Organization addresses the issue of what margin of error does and doesn’t measure in a disclaimer that it uses to report its survey results. Gallup tells you that besides sampling error, surveys can have additional errors or bias due to question wording and some of the logistical issues involved in conducting surveys (such as missing data due to phone numbers that are no longer current).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This means that even with the best of intentions and the most meticulous attention to details and process control, stuff happens. Nothing is ever perfect. But what you need to know is that the margin of error can’t measure

the extent of those other types of errors. And if a highly credible polling organization like Gallup admits to possible bias, imagine what’s really going on with other people’s studies that aren’t nearly as well designed or conducted.
</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a231"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

            </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a253">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 13</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a233" class="calibre10"><span class="calibre11"><span
class="bold"><span>Confidence Intervals: Making Your Best
Guesstimate</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Understanding confidence interval pieces, parts, and
interpretation</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Calculating with confidence</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Examining factors that influence the width of a
confidence interval</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Detecting misleading results</span></span></p>
</blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>M</span><span>ost statistics are used
to estimate some characteristic about a population of interest,
such as average household income, the percentage of people who
buy birthday gifts online, or the average amount of ice cream
```

consumed in the United States every year (and the resulting average weight gain — nah!). Such characteristics of a population are called parameters.Typically, people want to estimate (take a good guess at) the value of a parameter by taking a sample from the population and using statistics from the sample that will give them a good estimate. The question is: How do you define “good estimate”?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">As long as the process is done correctly (and in the media, it often isn’t!), an estimate can often get very close to the parameter. This chapter gives you an overview of confidence intervals (the type of estimates used and recommended by statisticians); why they should be used (as opposed to just a one-number estimate); how to set up, calculate, and interpret the most commonly used confidence intervals; and how to spot misleading estimates.</p></blockquote>
<p id="a234" class="calibre6">Not All Estimates Are Created Equal</p>
<blockquote class="calibre9"><p class="calibre14">Read any magazine or newspaper or listen to any newscast, and you hear a number of statistics, many of which are estimates of some quantity or another. You may wonder how they came up with those statistics. In some cases, the numbers are well researched; in other cases, they’re just a shot in the dark. Here are some examples of estimates that I came across in one single issue of a leading business magazine. They come from a variety of sources:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Even though some jobs are harder to get these days, some areas are really looking for recruits: Over the next eight years, 13,000 nurse anesthetists will be needed. Pay starts from $80,000 to $95,000.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The average number of bats used by a major league baseball player per season is 90.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"

class="calibre2"/>
The Lamborghini Murcielago can go from 0 to 60 mph in 3.7
seconds with a top speed of near 205 miles per hour.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Some of these estimates are easier to
obtain than others. Here are some observations I was able to
make about those estimates:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
How do you estimate how many nurse anesthetists are
needed over the next eight years? You can start by looking at
how many will be retiring in that time; but that won’t account
for growth. A prediction of the need in the next year or two
would be close, but eight years into the future is much harder
to do.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The average number of bats used per major league baseball
player in a season could be found by surveying the players
themselves, the people who take care of their equipment, or the
bat companies that supply the bats.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Determining car speed is more difficult but could be
conducted as a test with a stopwatch. And they should find the
average speed of many different cars (not just one) of the same
make and model, under the same driving conditions each time.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Not all
statistics are created equal. To determine whether a statistic
is reliable and credible, don’t just take it at face value.
Think about whether it makes sense and how you would go about
formulating an estimate. If the statistic is really important
to you, find out what process was used to come up with it.
(Chapter 16 handles all the elements involving surveys, and
Chapter 17 gives you the lowdown on experiments.)
</p></blockquote>
<p id="a235" class="calibre6">Linking a Statistic to a Parameter
</p>

<blockquote class="calibre9"><p class="calibre14">A
parameter is a
single number that describes a population, such as the median
household income for all households in the U.S. A
statistic
 is a single number that describes a
sample, such as the median household income of a sample of,
say, 1,200 households. You typically don’t know the values of
parameters of populations, so you take samples and use
statistics to give your best estimates.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you want to know the percentage
of vehicles in the U.S. that are pickup trucks (that’s the
parameter, in this case). You can’t look at every single
vehicle, so you take a random sample of 1,000 vehicles over a
range of highways at different times of the day. You find that
7% of the vehicles in your sample are pickup trucks. Now, you
don’t want to say that
exactly 7% of
all vehicles on U.S. roads are pickup trucks, because you know
this is only based on the 1,000 vehicles you sampled. Though
you hope 7% is close to the true percentage, you can’t be sure
because you based your results on a sample of vehicles, not on
all the vehicles in the U.S.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So what to do? You take your sample
result and add and subtract some number to indicate that you
are giving a range of possible values for the population
parameter, rather than just assuming the sample statistic
equals the population parameter (which would not be good,
although it’s done in the media all the time). This number that
is added to and subtracted from a statistic is called the
margin of error
(MOE). This plus or
minus (denoted by ±
) that’s added to any estimate helps put the
results into perspective. When you know the margin of error,
you have an idea of how much the sample results could change if
you took another sample.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The word
error in

margin of error doesn’t mean a mistake was made or the quality of the data was bad. It just means the results from a sample are not exactly equal to what you would have gotten if you had used the entire population. This gap measures error due to random chance, the luck of the draw — not due to bias. (That’s why minimizing bias is so important when you select your sample and collect your data; see Chapters 16 and 17.)</p></blockquote>
<p id="a236" class="calibre6">Getting with the Jargon
</p>
<blockquote class="calibre9"><p class="calibre14">A statistic plus or minus a margin of error is called a confidence interval:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
The word interval is used because your result becomes an interval. For example, say the percentage of kids who like baseball is 40%, plus or minus 3.5%. That means the percentage of kids who like baseball is somewhere between 40% – 3.5% = 36.5% and 40% + 3.5% = 43.5%. The lower end of the interval is your statistic minus the margin of error, and the upper end is your statistic plus the margin of error.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
With all confidence intervals, you have a certain amount of confidence in being correct (guessing the parameter) with your sample in the long run. Expressed as a percent, the amount of confidence is called the confidence level.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">You can find formulas and examples for the most commonly used confidence intervals later in this chapter.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Following are the general steps for estimating a parameter with a confidence interval. Details on Steps 1 and 4–6 are included throughout the remainder of this

chapter. Steps 2 and 3 involve sampling and data collection, which are detailed in Chapter 16 (sampling and survey data collection) and Chapter 17 (data collection from experiments). </p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Choose your confidence level and your sample size.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Select a random sample of individuals from the population.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Collect reliable and relevant data from the individuals in the sample.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Summarize the data into a statistic, such as a mean or proportion.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Calculate the margin of error.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Take the statistic plus or minus the margin of error to get your final estimate of the parameter.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> This step calculates the
confidence interval for that parameter.
</p></blockquote><div class="calibre31"> </div>
<p id="a237" class="calibre6">Interpreting Results with Confidence
</p>
<blockquote class="calibre9"><p class="calibre14"><span class="calibre15"><span>Suppose you, a research biologist, are trying to catch a fish using a hand net, and the size of your net represents the margin of error of a confidence interval. Now say your confidence level is 95%. What does this really mean? It means that if you scoop this particular net into the

```
water over and over again, you’ll catch a fish 95% of the time.
Catching a fish here means your confidence interval was correct
and contains the true parameter (in this case the parameter is
represented by the fish itself).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>But does this mean that on any given
try you have a 95% chance of catching a fish after the fact?
No. Is this confusing? It certainly is. Here’s the scoop (no
pun intended): On a single try, say you close your eyes before
you scoop your net into the water. At this point, your chances
of catching a fish are 95%. But then go ahead and scoop your
net through the water with your eyes still closed. </span>
<span><span class="calibre16"><span class="italic">After</span>
</span></span><span> that’s done, however, you open your eyes
and see one of only two possible outcomes; you either caught a
fish or you didn’t; probability isn’t involved anymore.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Likewise, </span><span><span
class="calibre16"><span class="italic">after</span></span>
</span><span> data have been collected, and the confidence
interval has been calculated, you either captured the true
population parameter or you didn’t. So you’re not saying you’re
95% confident that the parameter is in your particular
interval. What you are 95% confident about is the process by
which random samples are selected and confidence intervals are
created. (That is, 95% of the time in the long run, you’ll
catch a fish.)</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You know that this process will result
in intervals that capture the population mean 95% of the time.
The other 5% of the time, the data collected in the sample just
by random chance has abnormally high or low values in it and
doesn’t represent the population. This 5% measures errors due
to random chance only and doesn’t include bias.</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> The margin of
error is meaningless if the data that went into the study were
biased and/or unreliable. However, you can’t tell that by
looking at anyone’s statistical results. My best advice is to
look at how the data were collected before accepting a reported
margin of error as the truth (see Chapters 16 and 17 for
details on data collection issues). That means asking questions
before you believe a study.</span></span></p></blockquote>
<p id="a238" class="calibre6"><span class="calibre17"><span
class="bold"><span>Zooming In on Width</span></span></span></p>
```

<blockquote class="calibre9"><p class="calibre14">The width
 of your confidence interval is two times the
margin of error. For example, suppose the margin of error is
± 5%.
A confidence interval of 7%, plus or minus 5%, goes from 7% –
5% = 2%, all the way up to 7% + 5% = 12%. So the confidence
interval has a width of 12% – 2% = 10%. A simpler way to
calculate this is to say that the width of the confidence
interval is two times the margin of error. In this case, the
width of the confidence interval is 2 ∗
 5% = 10%.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The width of a
confidence interval is the distance from the lower end of the
interval (statistic minus margin of error) to the upper end of
the interval (statistic plus margin of error). You can always
calculate the width of a confidence interval quickly by taking
two times the margin of error.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The ultimate goal when making an
estimate using a confidence interval is to have a narrow width,
because that means you’re zooming in on what the parameter is.
Having to add and subtract a large margin of error only makes
your result much less accurate.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> So, if a small
margin of error is good, is smaller even better? Not always. A
narrow confidence interval is a good thing — to a point. To get
an extremely narrow confidence interval, you have to conduct a
much larger — and expensive — study, so a point comes where the
increase in price doesn’t justify the marginal difference in
accuracy. Most people are pretty comfortable with a margin of
error of 2% to 3% when the estimate itself is a percentage
(like the percentage of women, Republicans, or smokers).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How do you go about ensuring that your
confidence interval will be narrow enough? You certainly want
to think about this issue before collecting your data; after
the data are collected, the width of the confidence interval is
set.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Three factors affect the width of a

confidence interval:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Confidence level</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Sample size</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Amount of variability in the population</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Each of these three factors plays an
important role in influencing the width of a confidence
interval. In the following sections, you explore details of
each element and how they affect width.</p>
</blockquote>
<p id="a239" class="calibre6">Choosing a Confidence Level
</p>
<blockquote class="calibre9"><p class="calibre14">Every confidence interval (and every
margin of error, for that matter) has a percentage associated
with it that represents how confident you are that the results
will capture the true population parameter, depending on the
luck of the draw with your random sample. This percentage is
called a confidence level.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A confidence level helps you account
for the other possible sample results you could have gotten,
when you’re making an estimate of a parameter using the data
from only one sample. If you want to account for 95% of the
other possible results, your confidence level would be 95%.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> What level of confidence is typically
used by researchers? I’ve seen confidence levels ranging from
80% to 99%. The most common confidence level is 95%. In fact,
statisticians have a saying that goes, “Why do statisticians
like their jobs? Because they have to be correct only 95% of

the time.” (Sort of catchy, isn’t it? And let’s see weather
forecasters beat that.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Variability in sample results is
measured in terms of number of standard errors. A
standard
error is similar to the standard
deviation of a data set, only a standard error applies to
sample means or sample percentages that you could have gotten
if different samples were taken. (See
Chapter 11 for information on
standard errors.) </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/>Standard errors
are the building blocks of confidence intervals. A confidence
interval is a statistic plus or minus a margin of error, and
the margin of error is the number of standard errors you need
to get the confidence level you want.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Every confidence level has a
corresponding number of standard errors that have to be added
or subtracted. This number of standard errors is a called a
critical value. In a
situation where you use a
Z-distribution
to find the number of standard errors (as described later in
this chapter), you call the critical value the
z*-value
(pronounced z-star value
). See Table 13-1 for a list of
z
*-values for some of the most common
confidence levels.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> As the
confidence level increases, the number of standard errors
increases, so the margin of error increases.</p>
</blockquote>
<table border="1" class="calibre46">
<tbody class="calibre47">
<tr class="calibre48">
<td colspan="2" class="calibre49">

```
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre50"><span class="bold"><span>Table 13-1 </span>
<span><span class="calibre36"><span class="bold"><span
class="italic">z</span></span></span></span><span>*</span>
<span><span class="calibre36"><span class="bold"><span
class="italic">-</span></span></span></span><span>values for
Various Confidence Levels</span></span></span></p></blockquote>
</td>
</tr>
<tr class="calibre48">
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52"><p class="calibre39"><span class="bold">
<blockquote class="calibre5"><span class="calibre53"><span
class="bold"><span>Confidence Level</span></span></span>
</blockquote></span></p></div></blockquote><div
class="calibre19"> </div>
</td><div class="calibre19"> </div>
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52"><p class="calibre39"><span class="bold">
<blockquote class="calibre5"><span class="calibre53"><span
class="bold"><span>z*-value</span></span></span></blockquote>
</span></p></div></blockquote><div class="calibre19"> </div>
</td><div class="calibre19"> </div>
</tr>
<tr class="calibre48">
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>80%</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>1.28</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
</tr>
<tr class="calibre48">
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>90%</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
```

```
class="calibre53"><span>1.645 (by convention)</span></span></p>
</blockquote><div class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
</tr>
<tr class="calibre48">
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>95%</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>1.96</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
</tr>
<tr class="calibre48">
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>98%</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
<td border="1" class="calibre51"><blockquote class="calibre9">
<div class="calibre52">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre53"><span>2.33</span></span></p></blockquote><div
class="calibre19"> </div>
</div></blockquote></td><div class="calibre19"> </div>
</tr>
<tr class="calibre48">
<td class="calibre49">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre35"><span>99%</span></span></p></blockquote><div
class="calibre19"> </div>
</td>
<td class="calibre49">
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre35"><span>2.58</span></span></p></blockquote><div
class="calibre19"> </div>
</td>
</tr>
</tbody>
</table><div class="calibre33"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>If you want to be more than 95%
```

confident about your results, you need to add and subtract more
than about two standard errors. For example, to be 99%
confident, you would add and subtract about two and a half
standard errors to obtain your margin of error (2.58 to be
exact). The higher the confidence level, the larger the
z*
-value, the larger the margin of error, and
the wider the confidence interval (assuming everything else
stays the same). You have to pay a certain price for more
confidence.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Note that I said “assuming everything
else stays the same.” You can offset an increase in the margin
of error by increasing the sample size. See the following
section for more on this.</p></blockquote>
<p id="a240" class="calibre6">Factoring In the Sample Size
</p>
<blockquote class="calibre9"><p class="calibre14">The relationship between margin of
error and sample size is simple: As the sample size increases,
the margin of error decreases, and the confidence interval gets
narrower. This relationship confirms what you hope is true: The
more information (data) you have, the more accurate your
results are going to be. (That, of course, assumes that the
information is good, credible information. See Chapter 3 for
how statistics can go wrong.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> The margin of
error formulas for the confidence intervals in this chapter all
involve the sample size (
n) in the
denominator. For example, the formula for margin
of error for the sample mean, <img alt="9780470911082-
eq13001.eps" src="images/00207.jpg" class="calibre2"/>
(which you’ll see in great detail later in this
chapter), has an n in the denominator
of a fraction (this is the case for most margin of error
formulas). As n increases, the
denominator of this fraction increases, which makes the overall
fraction get smaller. That makes the margin of error smaller
and results in a narrower confidence interval.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"

src="images/00006.jpg" class="calibre2"/> When you need a high level of confidence, you have to increase the
z*
-value and, hence, margin of error, resulting in a wider confidence interval, which isn’t good. (See the previous section.) But you can offset this wider confidence interval by increasing the sample size and bringing the margin of error back down, thus narrowing the confidence interval.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The increase in sample size allows you to still have the confidence level you want, but also ensures that the width of your confidence interval will be small (which is what you ultimately want). You can even determine the sample size you need before you start a study: If you know the margin of error you want to get, you can set your sample size accordingly. (See the later section “Figuring Out What Sample Size You Need” for more.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When your statistic is going to be a percentage (such as the percentage of people who prefer to wear sandals during summer), a rough way to figure margin of error for a 95% confidence interval is to take 1 divided by the square root of n (the sample size). You can try different values of n
and you can see how the margin of error is affected. For example, a survey of 100 people from a large population will have a margin of error of about or plus or minus 10% (meaning the width of the confidence interval is 20%, which is pretty large). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, if you survey 1,000 people, your margin of error decreases dramatically, to plus or minus about 3%; the width now becomes only 6%. A survey of 2,500 people results in a margin of error of plus or minus 2% (so the width is down to 4%). That’s quite a small sample size to get so accurate, when you think about how large the population is (the U.S. population, for example, is over 310 million!).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Keep in mind, however, you don’t want to go too high with your

sample size, because a point comes where you have a diminished return. For example, moving from a sample size of 2,500 to 5,000 narrows the width of the confidence interval to about 2 ∗ 1.4 = 2.8%, down from 4%. Each time you survey one more person, the cost of your survey increases, so adding another 2,500 people to the survey just to narrow the interval by little more than 1% may not be worthwhile.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22">The first step in any data analysis problem (and when critiquing another person’s results) is to make sure you have good data. Statistical results are only as good as the data that went into them, so real accuracy depends on the quality of the data as well as on the sample size. A large sample size that has a great deal of bias (see Chapter 16) may appear to have a narrow confidence interval — but means nothing. That’s like competing in an archery match and shooting your arrows consistently, but finding out that the whole time you’re shooting at the next person’s target; that’s how far off you are. With the field of statistics, though, you can’t accurately measure bias; you can only try to minimize it by designing good samples and studies (see Chapters 16 and 17).</p></blockquote>
<p id="a241" class="calibre6">Counting On Population Variability</p>
<blockquote class="calibre9"><p class="calibre14">One of the factors influencing variability in sample results is the fact that the population itself contains variability. For example, in a population of houses in a fairly large city like Columbus, Ohio, you see a great deal of variety in not only the types of houses, but also the sizes and the prices. And the variability in prices of houses in Columbus should be more than the variability in prices of houses in a selected housing development in Columbus.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">That means if you take a sample of houses from the entire city of Columbus and find the average price, the margin of error should be larger than if you take a sample from that single housing development in Columbus, even if you have the same confidence level and the same sample size.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Why? Because the houses in the entire city have more variability in price, and your sample average

would change more from sample to sample than it would if you
took the sample only from that single housing development,
where the prices tend to be very similar because houses tend to
be comparable in a single housing development. So you need to
sample more houses if you’re sampling from the entire city of
Columbus in order to have the same amount of accuracy that you
would get from that single housing development.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> The standard
deviation of the population is denoted <img
alt="9780470911082-eq13003.eps" src="images/00209.jpg"
class="calibre2"/>. Notice that <img
alt="9780470911082-eq13004.eps" src="images/00210.jpg"
class="calibre2"/> appears in the numerator of the
standard error in the formula for margin of error for the
sample mean: <img alt="9780470911082-eq13005.eps"
src="images/00211.jpg" class="calibre2"/>.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Therefore, as the standard deviation
(the numerator) increases, the standard error (the entire
fraction) also increases. This results in a larger margin of
error and a wider confidence interval. (Refer to Chapter 11 for
more info on the standard error.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> More
variability in the original population increases the margin of
error, making the confidence interval wider. This increase can
be offset by increasing the sample size.</p>
</blockquote>
<p id="a242" class="calibre6">Calculating a Confidence Interval for a
Population Mean</p>
<blockquote class="calibre9"><p class="calibre14">When the characteristic that’s being
measured (such as income, IQ, price, height, quantity, or
weight) is numerical, most
people want to estimate the mean (average) value for the
population. You estimate the population mean, <img
alt="9780470911082-eq13006.eps" src="images/00212.jpg"
class="calibre2"/>, by using a sample mean, <img
alt="9780470911082-eq13007.eps" src="images/00213.jpg"
class="calibre2"/>, plus or minus a margin of error. The

result is called a confidence interval for the population mean, . Its formula depends on whether certain conditions are met. I split the conditions into two cases, illustrated in the following sections.</p></blockquote>
<blockquote class="calibre5"><p id="a243" class="calibre6">Case 1: Population standard deviation is known</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In Case 1, the population standard deviation is known. The formula for a confidence interval (CI) for a population mean in this case is , where is the sample mean, is the population standard deviation, n is the sample size, and z* represents the appropriate z*-value from the standard normal distribution for your desired confidence level. (Refer to Table 13-1 for values of z* for the given confidence levels.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> In this case, the data either have to come from a normal distribution, or if not, then n has to be large enough (at least 30 or so) for the Central Limit Theorem to kick in (see Chapter 11), allowing you to use z*-values in the formula.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To calculate a CI for the population mean (average), under the conditions for Case 1, do the following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span

```
class="calibre15"><span><span class="calibre16"><span
class="bold">1. Determine the confidence level and find the
appropriate </span></span></span><span><span class="calibre16">
<span class="bold"><span class="italic">z*</span></span></span>
</span><span><span class="calibre16"><span class="bold">-value.
</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> Refer to</span><span><span
class="calibre16"><span class="italic">
</span></span></span><span>Table 13-1.</span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">2. Find the sample mean (</span></span></span><img
alt="9780470911082-eq13012.eps" src="images/00218.jpg"
class="calibre2"/><span><span class="calibre16"><span
class="bold">) for the sample size (</span></span></span><span>
<span class="calibre16"><span class="bold"><span
class="italic">n</span></span></span></span><span><span
class="calibre16"><span class="bold">).</span></span></span>
</span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"><span class="italic">Note:</span></span></span>
</span><span> The population standard deviation is assumed to
be a known value, </span><img alt="9780470911082-eq13013.eps"
src="images/00219.jpg" class="calibre2"/><span>.</span></span>
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">3. Multiply </span></span></span><span><span
class="calibre16"><span class="bold"><span class="italic">z*
</span></span></span></span><span><span class="calibre16"><span
class="bold"> times </span></span></span><img
alt="9780470911082-eq13014.eps" src="images/00220.jpg"
class="calibre2"/><span><span class="calibre16"><span
class="bold"> and divide that by the square root of </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">n</span></span></span></span><span><span
class="calibre16"><span class="bold">.</span></span></span>
</span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> This calculation gives you the margin
of error.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
```

class="bold">4. Take <img
alt="9780470911082-eq13015.eps" src="images/00221.jpg"
class="calibre2"/> plus or minus the margin of error to obtain the
CI.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The lower end of the CI is <img
alt="9780470911082-eq13016.eps" src="images/00222.jpg"
class="calibre2"/> minus the margin of error, whereas the
upper end of the CI is <img alt="9780470911082-
eq13017.eps" src="images/00223.jpg" class="calibre2"/>
plus the margin of error.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you work for the
Department of Natural Resources and you want to estimate, with
95% confidence, the mean (average) length of walleye
fingerlings in a fish hatchery pond.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Because you want a 95% confidence
interval, your z*-value
is 1.96.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Suppose you take a random sample of
100 fingerlings and determine that the average length is 7.5
inches; assume the population standard deviation is 2.3 inches.
This means <img alt="9780470911082-eq13018.eps"
src="images/00224.jpg" class="calibre2"/>, <img
alt="9780470911082-eq13019.eps" src="images/00225.jpg"
class="calibre2"/>, and n
= 100.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Multiply 1.96 times 2.3 divided by
the square root of 100 (which is 10). The margin of error is,
therefore, ±
 1.96 ∗ (2.3 ÷ 10) = 1.96
∗ 0.23 = 0.45 inches.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Your 95% confidence interval for
the mean length of walleye fingerlings in this fish hatchery

pond is 7.5 inches ± 0.45 inches. (The lower
end of the interval is 7.5 – 0.45 = 7.05 inches; the upper end
is 7.5 + 0.45 = 7.95 inches.)</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> After you
calculate a confidence interval, make sure you always interpret
it in words a non-statistician would understand. That is, talk
about the results in terms of what the person in the problem is
trying to find out — statisticians call this interpreting the
results “in the context of the problem.” In this example you
can say: “With 95% confidence, the average length of walleye
fingerlings in this entire fish hatchery pond is between 7.05
and 7.95 inches, based on my sample data.” (Always be sure to
include appropriate units.)</p></blockquote>
<blockquote class="calibre5"><p id="a244" class="calibre6">
Case 2:
Population standard deviation is unknown and/or n is
small</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In many situations, you don’t know
<img alt="9780470911082-eq13020.eps"
src="images/00226.jpg" class="calibre2"/>, so you
estimate it with the sample standard deviation,
s;
and/or the sample size is small (less than 30),
and you can’t be sure your data came from a normal
distribution. (In the latter case, the Central Limit Theorem
can’t be used; see Chapter 11.) In either situation, you can’t
use a z*-value from the
standard normal (Z-) distribution as
your critical value anymore; you have to use a larger critical
value than that, because of not knowing what <img
alt="9780470911082-eq13021.eps" src="images/00227.jpg"
class="calibre2"/> is and/or having less data.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for a confidence interval
for one population mean in Case 2 is <img
alt="9780470911082-eq13022.eps" src="images/00228.jpg"
class="calibre2"/>, where t*
<sub
class="calibre42">n </sub><span

class="calibre41">_{– 1}
 is the critical t*
-value from the
t-distribution
with n – 1 degrees of
freedom (where n is the sample
size). The t*-values for common
confidence levels are found using the last row of the
t-
table (in the appendix). Chapter 10 gives
you the full details on the t-
distribution and how to use the t-
table.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The
t
-distribution has a similar shape to the
Z-distribution except
it’s flatter and more spread out. For small values of
n
 and a specific confidence level, the
critical values on the
t-distribution
are larger than on the
Z-distribution,
so when you use the critical values from the t
-distribution, the margin of error for your confidence
interval will be wider. As the values of n
 get larger, the
t*-values are
closer to z*-values. (Chapter
10 gives you the full details on the t
-distribution and its relationships to the
Z
-distribution.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

```
class="calibre15"><span>In the fish hatchery example from Case
1, suppose your sample size was 10 instead of 100, and
everything else was the same. The </span><span><span
class="calibre16"><span class="italic">t*</span></span></span>
<span>-value in this case comes from a </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span>-distribution with 10 – 1 = 9 degrees of freedom. This
</span><span><span class="calibre16"><span class="italic">t*-
</span></span></span><span>value is found by looking at the
</span><span><span class="calibre16"><span
class="italic">t</span></span></span><span>-table (in the
appendix). Look in the last row where the confidence levels are
located, and find the confidence level of 95%; this marks the
column you need. Then find the row corresponding to </span>
<span><span class="calibre16"><span class="italic">df</span>
</span></span><span> = 9. Intersect the row and column, and you
find </span><span><span class="calibre16"><span
class="italic">t*</span></span></span><span> = 2.262. This is
the </span><span><span class="calibre16"><span
class="italic">t*-</span></span></span><span>value for a 95%
confidence interval for the mean with a sample size of 10.
(Notice this is larger than the </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>*-value of 1.96 found in Table 13-1.) Calculating the
confidence </span><span>interval, you get </span><img
alt="9780470911082-eq13023.eps" src="images/00229.jpg"
class="calibre2"/><span>, or 5.86 to 9.15 inches. (Chapter 10
</span><span>gives you the full details on the </span><span>
<span class="calibre16"><span class="italic">t</span></span>
</span><span>-distribution and how to use the </span><span>
<span class="calibre16"><span class="italic">t</span></span>
</span><span>-table.)</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Notice this confidence interval is
wider than the one found when </span><span><span
class="calibre16"><span class="italic">n </span></span></span>
<span>= 100. In addition to having a larger critical value
(</span><span><span class="calibre16"><span class="italic">t*
</span></span></span><span> versus </span><span><span
class="calibre16"><span class="italic">z*</span></span></span>
<span>), the sample size is much smaller, which increases the
margin of error, because </span><span><span class="calibre16">
<span class="italic">n</span></span></span><span> is in its
denominator.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> In a case where
you need to use </span><span><span class="calibre16"><span
```

class="italic">s because you don't know , the confidence interval will be wider as well. It is also often the case that is unknown and the sample size is small, in which case the confidence interval is also wider.</p></blockquote>
<p id="a245" class="calibre6">Figuring Out What Sample Size You Need</p>
<blockquote class="calibre9"><p class="calibre14">The margin of error of a confidence interval is affected by size (see the earlier section "Factoring In the Sample Size"); as size increases, margin of error decreases. Looking at this the other way around, if you want a smaller margin of error (and doesn't everyone?), you need a larger sample size. Suppose you are getting ready to do your own survey to estimate a population mean; wouldn't it be nice to see ahead of time what sample size you need to get the margin of error you want? Thinking ahead will save you money and time and it will give you results you can live with in terms of the margin of error — you won't have any surprises later.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The formula for the sample size required to get a desired margin of error (MOE) when you are doing a confidence interval for is ; always round up the sample size no matter what decimal value you get. (For example, if your calculations give you 126.2 people, you can't just have 0.2 of a person — you need the whole person, so include him by rounding up to 127.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this formula, MOE is the number representing the margin of error you want, and z* is the z*-value corresponding to your desired confidence level (from Table 13-1; most people use 1.96 for a 95% confidence interval). If the population standard deviation, , is unknown, you can put in a worst-case scenario guess for it

or run a pilot study (a small trial study) ahead of time, find
the standard deviation of the sample data (s
), and use that number. This can be risky if the sample
size is very small because it’s less likely to reflect the
whole population; try to get the largest trial study that you
can, and/or make a conservative estimate for <img
alt="9780470911082-eq13088.eps" src="images/00235.jpg"
class="calibre2"/>.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Often a small trial study is worth the
time and effort. Not only will you get an estimate of
<img alt="9780470911082-eq13089.eps" src="images/00236.jpg"
class="calibre2"/> to help you determine a good sample
size, but you may also learn about possible problems in your
data collection.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> I only include
one formula for calculating sample size in this chapter: the
one that pertains to a confidence interval for a population
mean. (You can, however, use the quick and dirty formula in the
earlier section “Factoring in the Sample Size” for handling
proportions.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s an example where you need to
calculate n to estimate a
population mean. Suppose you want to estimate the average
number of songs college students store on their portable
devices. You want the margin of error to be no more than
 plus or minus 20 songs. You want a 95%
confidence interval. How many students should you sample?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because you want a 95% CI,
z*
 is 1.96 (found in Table 13-1); you know
your desired MOE is 20. Now you need a number for the
population standard deviation, <img alt="9780470911082-
eq13029.eps" src="images/00237.jpg" class="calibre2"/>.
This number is not known, so you do a pilot study of 35
students and find the standard deviation (s
) for the sample is 148 songs — use this number as a
substitute for <img alt="9780470911082-eq13030.eps"

src="images/00238.jpg" class="calibre2"/>. Using the sample size formula, you calculate the sample size you need is , which you round up to 211 students (you always round up when calculating n). So you need to take a random sample of at least 211 college students in order to have a margin of error in the number of stored songs of no more than 20. That’s why you see a greater-than-or-equal-to sign in the formula here.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You always round up to the nearest integer when calculating sample size, no matter what the decimal value of your result is (for example, 0.37). That’s because you want the margin of error to be no more than what you stated. If you round down when the decimal value is under .50 (as you normally do in other math calculations), your MOE will be a little larger than you wanted. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If you are wondering where this formula for sample size came from, it’s actually created with just a little math gymnastics. Take the margin of error formula (which contains n), fill in the remaining variables in the formula with numbers you glean from the problem, set it equal to the desired MOE, and solve for n.</p></blockquote>
<p id="a246" class="calibre6">Determining the Confidence Interval for One Population Proportion</p>
<blockquote class="calibre9"><p class="calibre14">When a characteristic being measured is categorical — for example, opinion on an issue (support, oppose, or are neutral), gender, political party, or type of behavior (do/don’t wear a seatbelt while driving) — most people want to estimate the proportion (or percentage) of people in

the population that fall into a certain category of interest. For example, consider the percentage of people in favor of a four-day work week, the percentage of Republicans who voted in the last election, or the proportion of drivers who don’t wear seat belts. In each of these cases, the object is to estimate a population proportion,
p, using a sample proportion, , plus or minus a margin of error. The result is called a confidence interval for the population proportion, p.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for a CI for a population proportion is , where
 is the sample proportion,
 n
is the sample size, and
z*
 is the appropriate value from the standard normal distribution for your desired confidence level. Refer to Table 13-1 for values of
z* for certain confidence levels.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To calculate a CI for the population proportion:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Determine the confidence level and find the appropriate
z*
-value.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Refer to
Table 13-1 for z
*-values.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Find the sample proportion,

<img alt="9780470911082-eq13035.eps"
src="images/00243.jpg" class="calibre2"/>, by dividing the number
of people in the sample having the characteristic of interest
by the sample size (n).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">Note:
 This result should be a decimal value between 0
and 1.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Multiply <img
alt="9780470911082-eq13036.eps" src="images/00244.jpg"
class="calibre2"/> and then divide that amount by
n.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Take the square root of the result from Step 3.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">5. Multiply your answer by
z*.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> This step gives you the margin of
error.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Take <img
alt="9780470911082-eq13037.eps" src="images/00245.jpg"
class="calibre2"/> plus or minus the margin of error to obtain the
CI; the lower end of the CI is <img
alt="9780470911082-eq13038.eps" src="images/00246.jpg"

class="calibre2"/> minus the margin of error, and the upper end of
the CI is <img alt="9780470911082-
eq13039.eps" src="images/00247.jpg" class="calibre2"/>
 plus the margin of
error.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> The formula
shown in the preceding example for a CI for p
 is used under the condition that the sample size is
large enough for the Central Limit Theorem to kick in and allow
us to use a z*-value (see
Chapter 11), which happens in cases when you are estimating
proportions based on large scale surveys (see Chapter 9). For
small sample sizes, confidence intervals for the proportion are
typically beyond the scope of an intro statistics course.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you want to
estimate the percentage of the time you’re expected to get a
red light at a certain intersection.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Because you want a 95% confidence
interval, your z*-value is 1.96.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. You take a random sample of 100
different trips through this intersection and find that you hit
a red light 53 times, so <img alt="9780470911082-
eq13040.eps" src="images/00248.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Find <img
alt="9780470911082-eq13041.eps" src="images/00249.jpg"
class="calibre2"/>.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Take the square root to get 0.0499.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The margin of error is, therefore,

plus or minus 1.96 ∗ (0.0499) =
0.0978, or 9.78%.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Your 95% confidence interval for
the percentage of times you will ever hit a red light at that
particular intersection is 0.53 (or 53%), plus or minus 0.0978
(rounded to 0.10 or 10%). (The lower end of the interval is
0.53 – 0.10 = 0.43 or 43%; the upper end is 0.53 + 0.10 = 0.63
or 63%.)</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> To interpret these results within the
context of the problem, you can say that with 95% confidence
the percentage of the times you should expect to hit a red
light at this intersection is somewhere between 43% and 63%,
based on your sample.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> While performing any calculations
involving sample percentages, use the decimal form. After the
calculations are finished, convert to percentages by
multiplying by 100. To avoid round-off error, keep at least 2
decimal places throughout.</p></blockquote>
<p id="a247" class="calibre6">Creating a Confidence Interval for the
Difference of Two Means</p>
<blockquote class="calibre9"><p class="calibre14">The goal of many surveys and studies is
to compare two populations, such as men versus women, low
versus high income families, and Republicans versus Democrats.
When the characteristic being compared is numerical (for
example, height, weight, or income), the object of interest is
the amount of difference in the means (averages) for the two
populations.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, you may want to compare
the difference in average age of Republicans versus Democrats,
or the difference in average incomes of men versus women. You
estimate the difference between two population means,

<img alt="9780470911082-eq13042.eps"
src="images/00250.jpg" class="calibre2"/>, by taking a
sample from each population (say, sample 1 and sample 2) and
using the difference of the two sample means <img
alt="9780470911082-eq13043.eps" src="images/00251.jpg"

class="calibre2"/>, plus or minus a margin of error. The result is a confidence interval for the difference of two population means, . The formula for the CI is different depending on certain conditions, as seen in the following sections; I call them Case 1 and Case 2.</p></blockquote>
<blockquote class="calibre5"><p id="a248" class="calibre6">
Case 1: Population standard deviations are known
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Case 1 assumes that both of the population standard deviations are known. The formula for a CI for the difference between two population means (averages) is , where
 and n
₁

are the mean and size of the first sample, and the first population’s standard deviation, , is given (known); and
n
₂<sub class="calibre42">
</sub>are the mean and size of the second sample, and the second population’s standard deviation, , is given (known). Here z* is the appropriate value from the standard normal distribution for your desired confidence level. (Refer to Table 13-1 for values of
z*
 for certain confidence levels.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To calculate a CI for the difference between two population means, do the following:

</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Determine the confidence level and find the
appropriate
z*
-value.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Refer to
Table 13-1.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Identify <img
alt="9780470911082-eq13050.eps" src="images/00258.jpg"
class="calibre2"/>, n<sub
class="calibre42">1</sub>, and
<img alt="9780470911082-eq13051.eps"
src="images/00259.jpg" class="calibre2"/>; find
<img alt="9780470911082-eq13052.eps"
src="images/00260.jpg" class="calibre2"/>,
n<sub
class="calibre42">2</sub>, and
<img alt="9780470911082-eq13053.eps"
src="images/00261.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Find the difference, (<img
alt="9780470911082-eq13054.eps" src="images/00262.jpg"
class="calibre2"/>), between the sample means.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

```
class="calibre15"><span><span class="calibre16"><span
class="bold">4. Square </span></span></span><img
alt="9780470911082-eq13055.eps" src="images/00263.jpg"
class="calibre2"/><span>
</span><span><span class="calibre16"><span class="bold">and
divide it by </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">n</span></span></span></span><span><span
class="calibre43"><span class="bold"><sub
class="calibre42">1</sub></span></span></span><span><span
class="calibre16"><span class="bold">; square </span></span>
</span><img alt="9780470911082-eq13056.eps"
src="images/00264.jpg" class="calibre2"/><span><span
class="calibre16"><span class="bold"> and divide it by </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">n</span></span></span></span><span><span
class="calibre43"><span class="bold"><sub
class="calibre42">2</sub></span></span></span><span><span
class="calibre16"><span class="bold">. Add the results together
and take the square root.</span></span></span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">5. Multiply your answer from Step 4 by </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">z*</span></span></span></span><span><span
class="calibre16"><span class="bold">.</span></span></span>
</span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> This answer is the margin of error.
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">6. Take </span></span></span><img
alt="9780470911082-eq13057.eps" src="images/00265.jpg"
class="calibre2"/><span><span class="calibre16"><span
class="bold"> plus or minus the margin of error to obtain the
CI.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> The lower end of the CI is </span><img
alt="9780470911082-eq13058.eps" src="images/00266.jpg"
class="calibre2"/><span>
</span><span><span class="calibre16"><span class="italic">minus
</span></span></span><span>the margin of error, whereas the
upper end of the CI is </span><img alt="9780470911082-
eq13059.eps" src="images/00267.jpg" class="calibre2"/><span>
</span><span><span class="calibre16"><span class="italic">plus
```

the margin of error.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Suppose you want to estimate with 95%
confidence the difference between the mean (average) length of
the cobs of two varieties of sweet corn (allowing them to grow
the same number of days under the same conditions). Call the
two varieties Corn-e-stats and Stats-o-sweet. Assume by prior
research that the population standard deviations for Corn-e-
stats and Stats-o-sweet are 0.35 inches and 0.45 inches,
respectively.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Because you want a 95% confidence
interval, your z* is 1.96.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Suppose your random sample of 100
cobs of the Corn-e-stats variety averages 8.5 inches, and your
random sample of 110 cobs of Stats-o-sweet averages 7.5 inches.
So the information you have is: <img alt="9780470911082-
eq13060.eps" src="images/00268.jpg" class="calibre2"/>,
<img alt="9780470911082-eq13061.eps"
src="images/00269.jpg" class="calibre2"/>,
n
<sub
class="calibre42">1</sub>
= 100, <img
alt="9780470911082-eq13062.eps" src="images/00270.jpg"
class="calibre2"/>, <img alt="9780470911082-
eq13063.eps" src="images/00271.jpg" class="calibre2"/>,
and n₂

= 110.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The difference between the sample
means, <img alt="9780470911082-eq13064.eps"
src="images/00272.jpg" class="calibre2"/>, from Step 3,
is 8.5 – 7.5 = +1 inch. This means the average for Corn-e-stats
minus the average for Stats-o-sweet is positive, making Corn-e-
stats the larger of the two varieties, in terms of this sample.
Is that difference enough to generalize to the entire
population, though? That’s what this confidence interval is
going to help you decide.</p></blockquote><div

class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Square <img
alt="9780470911082-eq13065.eps" src="images/00273.jpg"
class="calibre2"/> (0.35) to get 0.1225; divide by 100 to
get 0.0012. Square <img alt="9780470911082-eq13066.eps"
src="images/00274.jpg" class="calibre2"/> (0.45) and
divide by 110 to get 0.2025 ÷ 110 = 0.0018. The sum is 0.0012 +
0.0018 = 0.0030; the square root is 0.0554 inches (if no
rounding was done).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Multiply 1.96 times 0.0554 to get
0.1085 inches, the margin of error.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Your 95% confidence interval for
the difference between the average lengths for these two
varieties of sweet corn is 1 inch, plus or minus 0.1085 inches.
(The lower end of the interval is 1 – 0.1085 = 0.8915 inches;
the upper end is 1 + 0.1085 = 1.1085 inches.) Notice all the
values in this interval are positive. That means Corn-e-stats
is estimated to be longer than Stats-o-sweet, based on your
data.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> To interpret these results in the
context of the problem, you can say with 95% confidence that
the Corn-e-stats variety is longer, on average, than the Stats-
o-sweet variety, by somewhere between 0.8915 and 1.1085 inches,
based on your sample.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Notice that you
could get a negative value for <img alt="9780470911082-
eq13067.eps" src="images/00275.jpg" class="calibre2"/>.
For example, if you had switched the two varieties of corn, you
would have gotten –1 for this difference. You would say that
Stats-o-sweet averaged one inch shorter than Corn-e-stats in
the sample (the same conclusion stated differently).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> If you want to avoid negative values
for the difference in sample means, always make the group with
the larger sample mean your first group — all your differences
will be positive (that’s what I do).</p>

</blockquote>
<blockquote class="calibre5"><p id="a249" class="calibre6">
Case 2:
Population standard deviations are unknown and/or sample sizes
are small</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In many situations, you don’t know
<img alt="9780470911082-eq13068.eps"
src="images/00276.jpg" class="calibre2"/>, and you
estimate them with the sample standard deviations,
s
<sub
class="calibre42">1</sub>, and
s
<sub
class="calibre42">2</sub>; and/or the
sample sizes are small (less than 30) and you can’t be sure
whether your data came from a normal distribution.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A confidence interval for the
difference in two population means under Case 2 is
<img alt="9780470911082-eq13070.eps"
src="images/00277.jpg" class="calibre2"/>, where
t*
is the critical value from the
t-distribution with
n₁
 + n₂
 – 2 degrees of freedom; n
₁
 and n₂

are the two sample sizes,
respectively; and s₁

and s

₂

are the two sample standard
deviations. This t*-value is found on
the t-table (in the
appendix) by intersecting the row for df
= n₁
 + n₂
 – 2 with the column for the confidence level you need,
as indicated by looking at the last row of the table. (See
Chapter 10.) Here we assume the population standard deviations
are similar; if not, modify by using the standard error and
degrees of freedom. See the end of the section on comparing two
means in Chapter 15.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the corn example from Case 1,
suppose the mean cob lengths of the two brands of corn, Corn-e-
stats (group 1) and Stats-o-sweet (group 2), are the same as
they were before: <img alt="9780470911082-eq13071.eps"
src="images/00278.jpg" class="calibre2"/> inches. But
this time you don’t know the population standard deviations, so
you use the sample standard deviations instead — suppose they
turn out to be s₁
 = 0.40 and s₂
 = 0.50 inches, respectively. Suppose the sample sizes,
n₁
 and n₂
, are each only 15 in this case.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Calculating the CI, you first need to
find the <span

class="italic">t*-value on the
t-distribution with
(15 + 15 – 2) = 28 degrees of freedom. (Assume the confidence
level is still 95%.) Using the t-
table (in the appendix), look at the row for 28 degrees
of freedom and the column representing a confidence level of
95% (see the labels on the last row of the table); intersect
them and you see t*₂₈
 = 2.048. Using the rest of the information you are
given, the confidence interval for the difference in mean
cob length for the two brands is <img
alt="9780470911082-eq13072a.eps" src="images/00279.jpg"
class="calibre2"/><img alt="9780470911082-eq13072b.eps"
src="images/00280.jpg" class="calibre2"/>.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">That means a 95% CI for the difference
in the mean cob lengths of these two brands of corn in this
situation is (0.0727, 1.9273) inches, with Corn-e-stats coming
out on top. (Note:
 This CI is wider than what was found in Case 1,
as expected.)</p></blockquote>
<p id="a250" class="calibre6">Estimating the Difference of Two
Proportions</p>
<blockquote class="calibre9"><p class="calibre14">When a characteristic, such as opinion
on an issue (support/don’t support), of the two groups being
compared is categorical, people
want to report on the differences between the two population
proportions — for example, the difference between the
proportion of women who support a four-day work week and the
proportion of men who support a four-day work week. How do you
do this?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You estimate the difference between two
population proportions,
p₁
 – p₂

, by taking a sample from each population and using the
difference of the two sample proportions, <img
alt="9780470911082-eq13073.eps" src="images/00281.jpg"
class="calibre2"/>, plus or minus a margin of error. The
result is called a confidence interval for the difference of two
population proportions, p₁
 – p₂
.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for a CI for the difference
between two population proportions is <img
alt="9780470911082-eq13074.eps" src="images/00282.jpg"
class="calibre2"/>, where <img alt="9780470911082-
eq13075.eps" src="images/00283.jpg" class="calibre2"/>
and n₁
<sub
class="calibre42">
</sub>are the sample propor-
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">tion and sample size of the first
sample, and <img alt="9780470911082-eq13076.eps"
src="images/00284.jpg" class="calibre2"/> and
n
<sub
class="calibre42">2</sub><sub class="calibre42">
</sub>are the sample proportion and
sample size of the second sample. z*
 is the appropriate value from the standard normal
distribution for your desired confidence level. (Refer to Table
13-1 for z*-values.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To calculate a CI for the difference
between two population proportions, do the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Determine the confidence level and find the

appropriate
z*
-value.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Refer to
Table 13-1.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Find the sample proportion
<img alt="9780470911082-eq13077.eps"
src="images/00285.jpg" class="calibre2"/> for the first sample by
taking the total number from the first sample that are in the
category of interest and dividing by the sample size,

n<sub
class="calibre42">1</sub>. Similarly, find
<img alt="9780470911082-eq13078.eps"
src="images/00286.jpg" class="calibre2"/>
for the
second sample.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Take the difference between the sample
proportions, <img alt="9780470911082-
eq13079.eps" src="images/00287.jpg" class="calibre2"/>
.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Find <img
alt="9780470911082-eq13080.eps" src="images/00288.jpg"
class="calibre2"/> and divide that by
n<sub
class="calibre42">1</sub>. Find
<img alt="9780470911082-eq13081.eps"
src="images/00289.jpg" class="calibre2"/><span><span

```
class="calibre16"><span class="bold"> and divide that
by </span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">n</span></span></span></span>
<span><span class="calibre43"><span class="bold"><sub
class="calibre42">2</sub></span></span></span><span><span
class="calibre16"><span class="bold">. Add these two results
together and take the square root.</span></span></span></span>
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 5. Multiply </span></span></span><span><span
class="calibre16"><span class="bold"><span class="italic">z*
</span></span></span></span><span><span class="calibre16"><span
class="bold"> times the result from Step 4.</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> This step gives you the margin of
error.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold">6. Take </span></span></span><img
alt="9780470911082-eq13082.eps" src="images/00290.jpg"
class="calibre2"/><span><span class="calibre16"><span
class="bold"> plus or minus the margin of error from Step 5 to
obtain the CI.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> The lower end of the CI is </span><img
alt="9780470911082-eq13083.eps" src="images/00291.jpg"
class="calibre2"/><span> minus the margin of error, and the
upper end of the CI is </span><img alt="9780470911082-
eq13084.eps" src="images/00292.jpg" class="calibre2"/><span>
plus the margin of error.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The formula shown here for a CI for
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
<span> – </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
<span><span class="calibre16"><span class="italic">
</span></span></span><span>is used under the condition that
both of the sample sizes are large enough for the Central Limit
Theorem to kick in and allow us to use a </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
```

*-value (see Chapter 11); this is true when you are
estimating proportions using large scale surveys, for example.
For small sample sizes, confidence intervals are beyond the
scope of an intro statistics course.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you work for the Las Vegas
Chamber of Commerce, and you want to estimate with 95%
confidence the difference between the percentage of females who
have ever gone to see an Elvis impersonator and the percentage
of males who have ever gone to see an Elvis impersonator, in
order to help determine how you should market your
entertainment offerings.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Because you want a 95% confidence
interval, your z*-value is 1.96.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Suppose your random sample of 100
females includes 53 females who have seen an Elvis
impersonator, so <img alt="9780470911082-eq13085.eps"
src="images/00293.jpg" class="calibre2"/> is 53 ÷ 100 =
0.53. Suppose also that your random sample of 110 males
includes 37 males who have ever seen an Elvis impersonator, so
<img alt="9780470911082-eq13086.eps"
src="images/00294.jpg" class="calibre2"/> is 37 ÷ 110 =
0.34.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The difference between these sample
proportions (females – males) is 0.53 – 0.34 = 0.19.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Take 0.53 ∗
 (1 – 0.53) and divide that by 100 to get 0.2491 ÷ 100 =
0.0025. Then take 0.34 ∗ (1 – 0.34)
and divide that by 110 to get 0.2244 ÷ 110 = 0.0020. Add these
two results to get 0.0025 + 0.0020 = 0.0045; the square root is
0.0671.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. 1.96 ∗
0.0671 gives you 0.13, or 13%, which is the margin of error.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Your 95% confidence interval for

the difference between the percentage of females who have seen an Elvis impersonator and the percentage of males who have seen an Elvis impersonator is 0.19 or 19% (which you got in Step 3), plus or minus 13%. The lower end of the interval is 0.19 – 0.13 = 0.06 or 6%; the upper end is 0.19 + 0.13 = 0.32 or 32%.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> To interpret these results within the context of the problem, you can say with 95% confidence that a higher percentage of females than males have seen an Elvis impersonator, and the difference in these percentages is somewhere between 6% and 32%, based on your sample.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Now I’m thinking there are some guys out there that wouldn’t admit they’d ever seen an Elvis impersonator (although they’ve probably pretended to be one doing karaoke at some point). This may create some bias in the results. (The last time I was in Vegas, I believe I really saw Elvis; he was driving a van taxi to and from the airport. . . .)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Notice that you could get a negative value for . For example, if you had switched the males and females, you would have gotten –0.19 for this difference. That’s okay, but you can avoid negative differences in the sample proportions by having the group with the larger sample proportion serve as the first group (here, females).</p></blockquote>
<p id="a251" class="calibre6">Spotting Misleading Confidence Intervals</p>
<blockquote class="calibre9"><p class="calibre14">When the MOE is small, relatively speaking, you would like to say that these confidence intervals provide accurate and credible estimates of their parameters. This is not always the case, however.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Not all estimates are as accurate and reliable as the sources may want you to think. For example, a Web site survey result based on 20,000 hits may have a small MOE according to the formula, but the MOE means nothing if the survey is only given to people who happened to visit that Web site.</p></blockquote>

```
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In other words, the sample isn’t even
close to being a random sample (where every sample of equal
size selected from the population has an equal chance of being
chosen to participate). Nevertheless, such results do get
reported, along with their margins of error that make the study
seem truly scientific. Beware of these bogus results! (See
Chapter</span><span><span class="calibre16"><span
class="italic">
</span></span></span><span>12 for more on the limits of
the MOE.)</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Before making
any decisions based on someone’s estimate, do the following:
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Investigate how the statistic was created; it should be
the result of a scientific process that results in reliable,
unbiased, accurate data.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Look for a margin of error. If one isn’t reported, go to
the original source and request it.</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Remember that if the statistic isn’t reliable or contains
bias, the margin of error will be meaningless.</span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>(See Chapter 16 for evaluating survey
data and see Chapter 17 for criteria for good data in
experiments.)</span></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a252"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a252" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a248" style="min-width: 10px !important; min-height:
```

```
10px !important; border: solid 1px !important;"> </a> <a
href="#a249" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a250" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a251" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a242" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a243" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a240" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a241" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a246" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a247" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a253" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a245" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a244" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a239" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a238" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a237" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a236" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a235" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a234" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a233" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a275">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 14</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a254" class="calibre10"><span class="calibre11"><span
class="bold"><span>Claims, Tests, and Conclusions</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Testing other people’s claims</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Using hypothesis tests to weigh evidence and make
decisions</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Recognizing that your conclusions could be wrong</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Y</span><span>ou hear claims involving
statistics all the time; the media has no shortage of them:
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Twenty-five percent of all women in the United States
have varicose veins. (Wow, are some claims better left unsaid,
or what?)</span></span></p></blockquote><div
```

class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Cigarette use in the U.S. continues to drop, with the
percentage of all American smokers decreasing by about 2% per
year over the last ten years.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
A 6-month-old baby sleeps an average of 14 to 15 hours in
a 24-hour period. (Yeah, right!)</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
A name-brand ready-mix pie takes only 5 minutes to make.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In today’s age of information (and big
money), a great deal rides on being able to back up your
claims. Companies that say their products are better than the
leading brand had better be able to prove it, or they could
face lawsuits. Drugs that are approved by the FDA have to show
strong evidence that their products actually work without
producing life-threatening side effects. Manufacturers have to
make sure their products are being produced according to
specifications to avoid recalls, customer complaints, and loss
of business.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Although many claims are backed up by
solid scientific (and statistically sound) research, others are
not. In this chapter, you find out how to use statistics to
investigate whether a claim is actually valid and get the
lowdown on the process that researchers should
 be using to validate claims that they make.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/>A
hypothesis
test is a statistical procedure
that’s designed to test a claim. Before diving into details, I
want to give you the big picture of a hypothesis test by
showing the main steps involved. These steps are discussed in
the following sections:</p></blockquote>

<blockquote class="calibre9"><p class="calibre30"> 1. Set up the null and alternative hypotheses.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Collect good data using a well-designed study
(see Chapters 16 and 17).</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Calculate the test statistic based on your
data.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Find the p-
value for your test statistic.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Decide whether or not to reject H

_o based on your

p-value.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">6. Understand that your conclusion may be wrong,
just by chance.</p></blockquote>
<div class="calibre31"> </div>
<p id="a255" class="calibre6">Setting Up the Hypotheses
</p>
<blockquote class="calibre9"><p class="calibre14">Typically in a hypothesis test, the
claim being made is about a population parameter
 (one number that characterizes the entire
population). Because parameters tend to be unknown quantities,
everyone wants to make claims about what their values may be.
For example, the claim that 25% (or 0.25) of all women have

varicose veins is a claim about the proportion (that’s the
parameter) of all
women (that’s the population) who have
varicose veins (that’s the variable
 — having or not having varicose veins).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Researchers often challenge claims
about population parameters. You may hypothesize, for example,
that the actual proportion of women who have varicose veins is
lower than 0.25, based on your observations. Or you may
hypothesize that due to the popularity of high heeled shoes,
the proportion may be higher than 0.25. Or if you’re simply
questioning whether the actual proportion is 0.25, your
alternative hypothesis is: “No, it isn’t 0.25.”
</p></blockquote>
<blockquote class="calibre5"><p id="a256" class="calibre6">
Defining the
null</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Every hypothesis test contains a set of
two opposing statements, or hypotheses, about a population
parameter. The first hypothesis is called the
null hypothesis,
 denoted H_o
. The null hypothesis always states that the population
parameter is equal to the claimed
value. For example, if the claim is that the average time to
make a name-brand ready-mix pie is five minutes, the
statistical shorthand notation for the null hypothesis in this
case would be as follows: H_o
: μ = 5. (That is, the
population mean is 5 minutes.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/>All null
hypotheses include an equal sign in them; there are no
≤ or
≥ signs in
H<sub
class="calibre42">o</sub><span>. Not to cop out
or anything, but the reason it’s always equal is beyond the

```
scope of this book; let’s just say you wouldn’t pay me to
explain it to you.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a257" class="calibre6">
<span class="calibre21"><span class="bold"><span>What’s the
alternative?</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Before actually conducting a hypothesis
test, you have to put two possible hypotheses on the table —
the null hypothesis is one of them. But, if the null hypothesis
is rejected (that is, there was sufficient evidence against
it), what’s your alternative going to be? Actually, three
possibilities exist for the second (or alternative) hypothesis,
denoted H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span>. Here they are,
along with their shorthand notations in the context of the pie
example:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The population parameter is </span><span><span
class="calibre16"><span class="italic">not equal</span></span>
</span><span> to the claimed value (H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span>: </span><span>μ</span><span>
</span><span><span class="calibre40">≠</span></span><span> 5).
</span></span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The population parameter is </span><span><span
class="calibre16"><span class="italic">greater than</span>
</span></span><span> the claimed value (H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span>: </span><span>μ</span><span> &gt; 5).</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The population parameter is </span><span><span
class="calibre16"><span class="italic">less than</span></span>
</span><span> the claimed value (H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span>: </span><span>μ</span><span> &lt; 5).</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Which alternative hypothesis you choose
in setting up your hypothesis test depends on what you’re
interested in concluding, should you have enough evidence to
```

refute the null hypothesis (the claim).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, if you want to test
whether a company is correct in claiming its pie takes five
minutes to make and it doesn’t matter whether the actual
average time is more or less than that, you use the not-equal-
to alternative. Your hypotheses for that test would be H
_o
: µ = 5 versus
H<sub
class="calibre42">a</sub>:
µ
≠ 5.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you only want to see whether the
time turns out to be greater than what the company claims (that
is, whether the company is falsely advertising its quick prep
time), you use the greater-than alternative, and your two
hypotheses are H<sub
class="calibre42">o</sub>:
µ = 5 versus H_a
: µ > 5.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Finally, say you work for the company
marketing the pie, and you think the pie can be made in less
than five minutes (and could be marketed by the company as
such). The less-than alternative is the one you want, and your
two hypotheses would be H
_o:
µ = 5 versus H_a
: µ < 5.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> How do you know which hypothesis to
put in H<sub
class="calibre42">o</sub> and which one to
put in H<sub
class="calibre42">a</sub><span>? Typically, the
null hypothesis says that nothing new is happening; the
previous result is the same now as it was before, or the groups
have the same average (their difference is equal to zero). In
general, you assume that people’s claims are true until proven

```
otherwise. So the question becomes: Can you prove otherwise? In
other words, can you show sufficient evidence to reject Ho?
</span></span></p></blockquote>
<p id="a258" class="calibre6"><span class="calibre17"><span
class="bold"><span>Gathering Good Evidence (Data)</span></span>
</span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>After you’ve set up the hypotheses, the
next step is to collect your evidence and determine whether
your evidence goes against the claim made in H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span>. Remember, the claim is made about the
population, but you can’t test the whole population; the best
you can usually do is take a sample. As with any other
situation in which statistics are being collected, the quality
of the data is extremely critical. (See Chapter 3 for ways to
spot statistics that have gone wrong.)</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Collecting good data starts with
selecting a good sample. Two important issues to consider when
selecting your sample are avoiding bias and being accurate. To
avoid bias when selecting a sample, make it a random sample
(one that’s got the same chance of being selected as every
other possible sample of the same size) and choose a large
enough sample size so that the results will be accurate. (See
Chapter 11 for more information on accuracy.)</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Data is collected in many different
ways, but the methods used basically boil down to two: surveys
(observational studies) and experiments (controlled studies).
Chapter 16 gives all the information you need to design and
critique surveys, as well as information on selecting samples
properly. In Chapter 17, you examine experiments: what they can
do beyond an observational study, the criteria for a good
experiment, and when you can conclude cause and effect.</span>
</span></p></blockquote>
<p id="a259" class="calibre6"><span class="calibre17"><span
class="bold"><span>Compiling the Evidence: The Test
Statistic</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>After you select your sample, the
appropriate number-crunching takes place. Your null hypothesis
(H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>) makes a
statement about the population parameter — for example, “The
proportion of all women who have varicose veins is 0.25” (in
```

other words, H<sub
class="calibre42">o</sub>:
p
 = 0.25); or the average miles per gallon of a
U.S.-built light truck is 27 (H_o
: μ = 27). The data you collect
from the sample measures the variable of interest, and the
statistics that you calculate will help you test the claim
about the population parameter.</p></blockquote>
<blockquote class="calibre5"><p id="a260" class="calibre6">
Gathering
sample statistics</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Say you're testing a claim about the
proportion of women with varicose veins. You need to calculate
the proportion of women in your sample who have varicose veins,
and that number will be your sample statistic. If you're
testing a claim about the average miles per gallon of a U.S.-
built light truck, your statistic will be the average miles per
gallon of the light trucks in your sample. And knowing you want
to measure the variability in average miles per gallon for
various trucks, you'll want to calculate the sample standard
deviation. (See Chapter 5 for all the information you need on
calculating sample statistics.)</p></blockquote>
<blockquote class="calibre5"><p id="a261" class="calibre6">
Measuring
variability using standard errors</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">After you've calculated all the
necessary sample statistics, you may think you're done with the
analysis part and ready to make your conclusions — but you're
not. The problem is you have no way to put your results into
any kind of perspective just by looking at them in their
regular units. That's because you know that your results are
based only on a sample and that sample results are going to
vary. That variation needs to be taken into account, or your
conclusions could be completely wrong. (How much do sample
results vary? Sample variation is measured by the standard
error; see Chapter 11 for more on this.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose the claim is that the
percentage of all women with varicose veins is 25%, and your
sample of 100 women had 20% with varicose veins. The standard
error for your sample percentage is 4% (according to formulas
in Chapter 11), which means that your results are expected to

vary by about twice that, or about 8%, according to the
Empirical Rule (see Chapter 12). So a difference of 5%, for
example, between the claim and your sample result (25% – 20% =
5%) isn’t that much, in these terms, because it represents a
distance of less than 2 standard errors away from the claim.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, suppose your sample percentage
was based on a sample of 1,000 women, not 100. This decreases
the amount by which you expect your results to vary, because
you have more information. Again using formulas from Chapter
11, I calculate the standard error to be 0.013 or 1.3%. The
margin of error (MOE) is about twice that, or 2.6% on either
side. Now a difference of 5% between your sample result (20%)
and the claim in H<sub
class="calibre42">o</sub> (25%) is a more
meaningful difference; it’s way more than 2 standard errors.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Exactly how meaningful are your
results? In the next section, you get more specific about
measuring exactly how far apart your sample results are from
the claim in terms of the number of standard errors. This leads
you to a specific conclusion as to how much evidence you have
against the claim in H<sub
class="calibre42">o</sub>.
</p></blockquote>
<blockquote class="calibre5"><p id="a262" class="calibre6">
Understanding
standard scores</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The number of
standard errors that a statistic lies above or below the mean
is called a standard score (for
example, a z-value is a type of
standard score; see Chapter 9). In order to interpret your
statistic, you need to convert it from original units to a
standard score. When finding a standard score, you take your
statistic, subtract the mean, and divide the result by the
standard error.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the case of hypothesis tests, you
use the value in H<sub
class="calibre42">o</sub> as the mean.
(That’s what you go with unless/until you have enough evidence

against it.) The standardized version of your statistic is called a test statistic, and it’s the main component of a hypothesis test. (Chapter 15 contains the formulas for the most common hypothesis tests.) </p></blockquote>
<blockquote class="calibre5"><p id="a263" class="calibre6"> Calculating and interpreting the test statistic</p> </blockquote>
<blockquote class="calibre9"><p class="calibre14">The general procedure for converting a statistic to a test statistic (standard score) is as follows: </p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Take your statistic minus the claimed value (the number stated in H_o). </p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Divide by the standard error of the statistic. (Different formulas for standard error exist for different problems; see Chapter 13 for detailed formulas for standard error and Chapter 15 for formulas for various test statistics.)</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Your test statistic represents the distance between your actual sample results and the claimed population value, in terms of number of standard errors. In the case of a single population mean or proportion, you know that these standardized distances should at least have an approximate standard normal distribution if your sample size is large enough (see Chapter 11). So, to interpret your test statistic in these cases, you can see where it stands on the standard normal distribution (Z -distribution).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Using the numbers from the varicose veins example in the previous section, the test statistic is found by taking the proportion in the sample with varicose veins, 0.20, subtracting the claimed proportion of all women with varicose veins, 0.25, and then dividing the result by the

standard error, 0.04. These calculations give you a test statistic (standard score) of –0.05 ÷ 0.04 = –1.25. This tells you that your sample results and the population claim in H_o are 1.25 standard errors apart; in particular, your sample results are 1.25 standard errors below the claim. Now is this enough evidence to reject the claim? The next section addresses that issue.</p></blockquote>
<p id="a264" class="calibre6">Weighing the Evidence and Making Decisions: p-Values</p>
<blockquote class="calibre9"><p class="calibre14">After you find your test statistic, you use it to make a decision about whether to reject H_o. You make this decision by coming up with a number that measures the strength of this evidence (your test statistic) against the claim in H_o. That is, how likely is it that your test statistic could have occurred while the claim was still true? This number you calculate is called the p-value; it’s the chance that someone could have gotten results as extreme as yours while H_o was still true. Similarly in a jury trial, the jury discusses how likely it is that all the evidence came out the way it did assuming the defendant was innocent.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This section shows all the ins and outs of p-values, including how to calculate them and use them to make decisions regarding H_o.</p></blockquote>
<blockquote class="calibre5"><p id="a265" class="calibre6">Connecting test statistics and p-values</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To test whether a claim in H_o should be rejected (after all, it’s all about Ho) you look at your test statistic taken from your

sample and see whether you have enough evidence to reject the claim. If the test statistic is large (in either the positive or negative directions), your data is far from the claim; the larger the test statistic, the more evidence you have against the claim. You determine “how far is far” by looking at where your test statistic ends up on the distribution that it came from. When testing one population mean, under certain conditions the distribution of comparison is the standard normal (Z-) distribution, which has a mean of 0 and a standard deviation of 1; I use it throughout this section as an example. (See Chapter 9 to find out more about the Z-distribution.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If your test statistic is close to 0, or at least within that range where most of the results should fall, then you don’t have much evidence against the claim (H_o
) based on your data. If your test statistic is out in the tails of the standard normal distribution (see Chapter 9 for more on tails), then your evidence against the claim (H_o) is great; this result has a very small chance of happening if the claim is true. In other words, you have sufficient evidence against the claim (H_o), and you reject H_o.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">But how far is “too far” from 0? As long as you have a normal distribution or a large enough sample size, you know that your test statistic falls somewhere on a standard normal distribution (see Chapter 11). If the null hypothesis (H_o) is true, most (about 95%) of the samples will result in test statistics that lie roughly within 2 standard errors of the claim. If H
_a
 is the not-equal-to alternative, any test statistic outside this range will result in H_o
<span> being rejected. See Figure 14-1 for a picture showing

```
the locations of your test statistic and their corresponding
conclusions. In the next section, you see how to quantify the
amount of evidence you have against Ho.</span></span></p>
</blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 14-1:</span><span> Decisions for Ha: not-
equal-to.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg1401.eps"
src="images/00296.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> Note that if the alternative
hypothesis is the less-than alternative, you reject H</span>
<span><span class="calibre41"><sub class="calibre42">o</sub>
</span></span><span> only if the test statistic falls in the
left tail of the distribution (below –1.64). Similarly, if
H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span> is the greater-
than alternative, you reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span> only if the test statistic falls in the right tail
(above 1.64).</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a266" class="calibre6">
<span class="calibre21"><span class="bold"><span>Defining a p-
value</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> A </span><span>
<span class="calibre16"><span class="italic">p-</span></span>
</span><span>value is a probability associated with your test
statistic. It measures the chance of getting results at least
as strong as yours if the claim (H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>) were true. In the case of testing the population mean,
the farther out your test statistic is on the tails of the
standard normal </span><span><span class="calibre16"><span
class="italic">(Z</span></span></span><span>-) distribution,
the smaller your</span><span><span class="calibre16"><span
class="italic"> p-</span></span></span><span>value will be, the
less likely your results were to have occurred, and the more
evidence you have against the claim (H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>).</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a267" class="calibre6">
```

```
<span class="calibre21"><span class="bold"><span>Calculating a
p-value</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To find the</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value for your test statistic:</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 1. Look up your test statistic on the appropriate
distribution — in this case, on the standard normal (</span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">Z-</span></span></span></span><span><span
class="calibre16"><span class="bold">) distribution (see the
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">Z</span></span></span></span>
<span><span class="calibre16"><span class="bold">-table in the
appendix).</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Find the chance that </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">Z</span></span></span></span><span><span
class="calibre16"><span class="bold"> is beyond (more extreme
than) your test statistic:</span></span></span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> • If H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span> contains a less-than alternative, find the probability
that </span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> is less than your
test statistic (that is, look up your test statistic on the
</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span>-table and find its
corresponding probability). This is the</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> • If H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span> contains a greater-than alternative, find the
probability that </span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> is greater than
your test statistic (look up your test statistic on the </span>
<span><span class="calibre16"><span class="italic">Z</span>
```

-table, find its corresponding probability,
and subtract it from one). The result is your p-
value.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> • If H_a
 contains a non-equal-to alternative, find the
probability that Z is beyond your test
statistic and double it. There are two cases:</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> If your test statistic is negative,
first find the probability that Z
 is less than your test statistic (look up your test
statistic on the Z-table and find its
corresponding probability). Then double this probability to get
the
p-value.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">If your test statistic is positive,
first find the probability that Z
is greater than your test statistic (look up your test
statistic on the Z-table, find its
corresponding probability, and subtract it from one). Then
double this result to get the p-
value.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Why do you
double the probabilities if your H_a
 contains a non-equal-to alternative? Think of the not-
equal-to alternative as the combination of the greater-than
alternative and the less-than alternative. If you’ve got a
positive test statistic, its p-
value only accounts for the greater-than portion of the
not-equal-to alternative; double it to account for the less-

than portion. (The doubling of one p-
value is possible because the Z-distribution is
symmetric.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Similarly, if you’ve got a negative
test statistic, its p-value only
accounts for the less-than portion of the not-equal-to
alternative; double it to also account for the greater-than
portion.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When testing H_o
: p = 0.25 versus
H<sub
class="calibre42">a</sub>:
 p
< 0.25 in the varicose veins example from the
previous section, the p-value turns out to
be 0.1056. This is because the test statistic (calculated in
the previous section) was –1.25, and when you look this number
up on the Z-table (in the
appendix) you find a probability of 0.1056 of being less than
this value. If you had been testing the two-sided alternative,
H<sub
class="calibre42">a</sub>:
p

≠
0.25, the p-value would be 2
∗ 0.1056, or 0.2112.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If the results
are likely to have occurred under the claim, then you fail to
reject H<sub
class="calibre42">o</sub> (like a jury
decides not guilty). If the results are unlikely to have
occurred under the claim, then you reject H_o
 (like a jury decides guilty). The cutoff point between

rejecting H<sub
class="calibre42">o</sub> and failing to
reject H<sub
class="calibre42">o</sub> is another whole
can of worms that I dissect in the next section (no pun
intended).</p></blockquote>
<p id="a268" class="calibre6">Making Conclusions</p>
<blockquote class="calibre9"><p class="calibre14">To draw conclusions about H
_o
 (reject or fail to reject) based on
a p-
value, you need to set a
predetermined cutoff point where only those p
-values less than or equal to the cutoff will result in
rejecting Ho. This cutoff point is called the
alpha level
(α),
 or significance level
for the test. While 0.05 is a very popular cutoff value for
rejecting H<sub
class="calibre42">o</sub>, cutoff points
and resulting decisions can vary — some people use stricter
cutoffs, such as 0.01, requiring more evidence before rejecting
H<sub
class="calibre42">o</sub>, and others may
have less strict cutoffs, such as 0.10, requiring less
evidence.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If H_o
 is rejected (that is, the p
-value is less than or equal to the predetermined
significance level), the researcher can say she’s found a
statistically significant result. A result is
statistically
significant if it’s too rare to have
occurred by chance assuming H_o
 is true. If you get a statistically significant result,
you have enough evidence to reject the claim, H
_o
<span>, and conclude that something different or new is

```
in effect (that is, H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span>).</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> The
significance level can be thought of as the highest
possible</span><span><span class="calibre16"><span
class="italic"> p-</span></span></span><span>value that would
reject Ho and declare the results statistically significant.
Following are the general rules for making a decision about
H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> based on a
</span><span><span class="calibre16"><span class="italic">p-
</span></span></span><span>value:</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If the</span><span><span class="calibre16"><span
class="italic"> p-</span></span></span><span>value is less than
or equal to your significance level, then it meets your
requirements for having enough evidence against H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span>; you reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If the</span><span><span class="calibre16"><span
class="italic"> p-</span></span></span><span>value is greater
than your significance level, your data failed to show evidence
beyond a reasonable doubt; you fail to reject H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span>.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>However, if you plan to make decisions
about H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> by comparing
the</span><span><span class="calibre16"><span class="italic">
p-</span></span></span><span>value to your significance level,
you must decide on your significance level ahead of time. It
wouldn’t be fair to change your cutoff point after you’ve got a
sneak peak at what’s happening in the data.</span></span></p>
</blockquote>
```

```
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> You may be
wondering whether it’s okay to say “Accept H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>” instead of “Fail to reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>.” The answer is a big no. In a hypothesis test, you are
</span><span><span class="calibre16"><span
class="italic">not</span></span></span><span> trying to show
whether or not H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> is true (which
</span><span><span class="calibre16"><span
class="italic">accept</span></span></span><span><span
class="calibre36"><span class="italic">
</span></span></span><span>implies) — indeed, if you knew
whether H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> was true, you
wouldn’t be doing the hypothesis test in the first place.
You’re trying to show whether you have enough evidence to say
H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> is false, based
on your data. Either you have enough evidence to say it’s false
(in which case you reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>) or you don’t have enough evidence to say it’s false (in
which case you fail to reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>).</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a269" class="calibre6">
<span class="calibre21"><span class="bold"><span>Setting
boundaries for rejecting H</span><span><span class="calibre54">
<sub class="calibre42">o</sub></span></span></span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>These guidelines help you make a
decision (reject or fail to reject Ho) based on a</span><span>
<span class="calibre16"><span class="italic"> p-</span></span>
</span><span>value when your significance level is 0.05:</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>If the</span><span><span class="calibre16"><span
class="italic"> p-</span></span></span><span>value is less than
0.01 (very small), the results are considered highly
statistically significant — reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
```

.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the p-value is between
0.05 and 0.01 (but not super-close to 0.05), the results are
considered statistically significant — reject H
_o
.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the p-value is really
close to 0.05 (like 0.051 or 0.049), the results should be
considered marginally significant — the decision could go
either way.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the p-value is greater
than (but not super-close to) 0.05, the results are considered
non-significant — you fail to reject H_o
.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> When you hear a
researcher say her results are found to be statistically
significant, look for the
 p-value and
make your own decision; the researcher’s predetermined
significance level may be different from yours. If the
p
-value isn’t stated, ask for it.
</p></blockquote>
<blockquote class="calibre5"><p id="a270" class="calibre6">
Testing
varicose veins</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the varicose veins example in the
last section, the <span><span class="calibre16"><span

```
class="italic">p</span></span></span><span>-value was found to
be 0.1056. This </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>-value is fairly
large and indicates very weak evidence against H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span> by almost anyone’s standards because it’s greater
than 0.05 and even slightly greater than 0.10 (considered to be
a very large significance level). In this case you fail to
reject H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>. You didn’t have
enough evidence to say the proportion of women with varicose
veins is less than 0.25 (your alternative hypothesis). This
isn’t declared to be a statistically significant result.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>But say your</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value had been something like 0.026. A reader with a
personal cutoff point of 0.05 would reject H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span> in this case because the</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value (of 0.026) is less than 0.05. His conclusion would
be that the proportion of women with varicose veins isn’t equal
to 0.25; according to Ha in this case, you conclude it’s less
than 0.25, and the results are statistically significant.
However, a reader whose significance level is 0.01 wouldn’t
have enough evidence (based on your sample) to reject H</span>
<span><span class="calibre41"><sub class="calibre42">o</sub>
</span></span><span> because the</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value of 0.026 is greater than 0.01. These results
wouldn’t be statistically significant.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Finally, if the</span><span><span
class="calibre16"><span class="italic"> p-</span></span></span>
<span>value turned out to be 0.049 and your significance level
is 0.05, you can go by the book and say because it’s less than
0.05 you reject H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>, but you really
should say your results are marginal, and let the reader
decide. (Maybe they can flip a coin or something — “Heads we
reject H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>, tails,
we don’t!”)</span></span></p></blockquote>
<p id="a271" class="calibre6"><span class="calibre17"><span
class="bold"><span>Assessing the Chance of a Wrong
```

Decision</p>
<blockquote class="calibre9"><p class="calibre14">After you make a decision to either
reject H<sub
class="calibre42">o</sub> or fail to reject
H<sub
class="calibre42">o</sub>, the next step is
living with the consequences, in terms of how people respond to
your decision.</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If you conclude that a claim isn’t true but it actually
is, will that result
in a lawsuit, a fine, unnecessary changes in the product, or
consumer boycotts that shouldn’t have happened? It’s possible.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If you can’t disprove a claim that’s wrong, what happens
then? Will products continue to be made in the same way as they
are now? Will no new law be made, no new action taken, because
you showed that nothing was wrong? Missed opportunities to blow
the whistle have been known to occur.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Whatever
decision you make with a hypothesis test, you know there is a
chance of being wrong; that’s life in the statistics world.
Knowing the kinds of errors that can happen and finding out how
to curb the chance of them occurring are key.</p>
</blockquote>
<blockquote class="calibre5"><p id="a272" class="calibre6">
Making a false
alarm: Type-1 errors</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose a company claims that its
average package delivery time is 2 days, and a consumer group
tests this hypothesis, gets a p
-value of 0.04, and concludes that the claim is false:
They believe that the average delivery time is actually more
than 2 days. This is a big deal. If the group can stand by its
statistics, it has done well to inform the public about the
false advertising issue. But what if the group is wrong?

</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Even if the
group bases their study on a good design, collects good data,
and makes the right analysis, it can still be wrong. Why?
Because its conclusions were based on a sample of packages, not
on the entire population. And as Chapter 11 tells you, sample
results vary from sample to sample.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Just because the results from a sample
are unusual doesn’t mean they’re impossible. A
 p-
value of 0.04 means that the chance of getting
your particular test statistic, even if the claim is true, is
4% (less than 5%). You reject H_o
 in this case because that chance is small. But even a
small chance is still a chance!</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Perhaps your sample, though collected
randomly, just happens to be one of those atypical samples
whose result ended up far from what was expected. So, H
_o
 could be true, but your results lead you
to a different conclusion. How often does that happen? Five
percent of the time (or whatever your given cutoff probability
is for rejecting H<sub
class="calibre42">o</sub>).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Rejecting
H<sub
class="calibre42">o</sub> when you
shouldn’t is called a
type-1 error. I
don’t really like this name, because it seems so nondescript. I
prefer to call a type-1 error a false alarm.
 In the case of the packages, if the
consumer group made a type-1 error when it rejected the
company’s claim, they created a false alarm. What’s the result?
A very angry delivery company, I guarantee that!
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"

class="calibre2"/> To reduce the chance of false alarms, set a low cutoff probability (significance level) for rejecting H_o. Setting it to 5% or 1% will keep the chance of a type-1 error in check.</p></blockquote>
<blockquote class="calibre5"><p id="a273" class="calibre6">Missing out on a detection: Type-2 errors</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">On the other hand, suppose the company really wasn’t delivering on its claim. Who’s to say that the consumer group’s sample will detect it? If the actual delivery time is 2.1 days instead of 2 days, the difference would be pretty hard to detect. If the actual delivery time is 3 days, even a fairly small sample would probably show that something’s up. The issue lies with those in-between values, like 2.5 days.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If H_o is indeed false, you want to find out about it and reject H_o. Not rejecting H_o when you should have is called a type-2 error. I like to call it a missed detection.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Sample size is the key to being able to detect situations where H_o is false and, thus, avoiding type-2 errors. The more information you have, the less variable your results will be (see Chapter 11) and the more ability you have to zoom in on detecting problems that exist with a claim made by Ho.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This ability to detect when H_o is truly false is called the power of a test. Power is a pretty complicated

issue, but what’s important for you to know is that the higher
the sample size, the more powerful a test is. A powerful test
has a small chance for a type-2 error.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> As a preventative measure to minimize
the chances of a type-2 error, statisticians recommend that you
select a large sample size to ensure that any differences or
departures that really exist won’t be missed.</p>
</blockquote>
</div>
</div> <div class="mbppagebreak" id="a274"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a259" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a258" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a255" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a254" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a257" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a256" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a260" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a261" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a262" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a263" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a264" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a265" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a266" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a267" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a268" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a269" style="min-width: 10px !important; min-height:

10px !important; border: solid 1px !important;"> <a
href="#a273" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a272" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a271" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a270" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a275" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a274" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a288">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 15</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a276" class="calibre10"><span class="calibre11"><span
class="bold"><span>Commonly Used Hypothesis Tests: Formulas and
Examples</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Breaking down commonly used hypothesis tests</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Calculating their test statistics</span></span></p>
</blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Using the results to make informed decisions</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>F</span><span>rom product
advertisements to media blitzes on recent medical
breakthroughs, you often run across claims made about one or
more populations. For example, “We promise to deliver our
packages in two days or less” or “Two recent studies show that
a high-fiber diet may reduce your risk of colon cancer by 20%.”
Whenever someone makes a claim (also called a </span><span>
<span class="calibre16"><span class="italic">null
hypothesis</span></span></span><span>) about a population (such
```

as all packages, or all adults) you can test the claim by doing what statisticians call a
hypothesis test.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A hypothesis test involves setting up your hypotheses (a claim and its alternative), selecting a sample (or samples), collecting data, calculating the relevant statistics, and using those statistics to decide whether the claim is true.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, I outline the formulas used for some of the most common hypothesis tests, explain the necessary calculations, and walk you through some examples.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If you need more background information on hypothesis testing (such as setting up hypotheses, understanding test statistics, p-values, significance levels, and type-1 and type-2 errors), just flip to Chapter 14. All the general concepts of hypothesis testing are developed there. This chapter focuses on their application.
</p></blockquote>
<p id="a277" class="calibre6">Testing One Population Mean
</p>
<blockquote class="calibre9"><p class="calibre14">When the variable is numerical (for example, age, income, time, and so on) and only one population or group (such as all U.S. households or all college students) is being studied, you use the hypothesis test in this section to examine or challenge a claim about the population mean. For example, a child psychologist says that the average time that working mothers spend talking to their children is 11 minutes per day, on average. (For dads, the claim is 8 minutes.) The variable — time — is numerical, and the population is all working mothers. Using statistical notation,
µ represents the average number of minutes per day that all working mothers spend talking to their children, on average.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The null hypothesis is that the population mean, µ, is equal to a certain claimed value, µ_o

```
<span>. The notation for the null hypothesis is H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span>: </span><span>μ</span><span> = </span>
<span>μ</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>. So the null
hypothesis in our example is H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>: </span><span>μ</span><span> = 11 minutes, and </span>
<span>μ</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> is 11. The three
possibilities for the alternative hypothesis, H</span><span>
<span class="calibre41"><sub class="calibre42">a</sub></span>
</span><span>, are </span><span>μ</span><span>
</span><span><span class="calibre40">≠</span></span><span> 11,
</span><span>μ</span><span> &lt; 11, or </span><span>μ</span>
<span> &gt; 11, depending on what you are trying to show. (See
Chapter 14 for more on alternative hypotheses.) If you suspect
that the average time working mothers spend talking with their
kids is more than 11 minutes, your alternative hypothesis would
be H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span>: </span>
<span>μ</span><span> &gt; 11.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To test the claim, you compare the mean
you got from your sample (</span><img alt="9780470911082-
eq15047.eps" src="images/00297.jpg" class="calibre2"/><span>)
with the mean shown in H</span><span><span class="calibre41">
<sub class="calibre42">o</sub></span></span><span> (</span>
<span>μ</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>). To make a
proper comparison, you look at the difference between them, and
divide by the standard error to take into account the fact that
your sample results will vary. (See Chapter 12 for all the info
you need on standard error.) This result is your </span><span>
<span class="calibre16"><span class="italic">test statistic.
</span></span></span><span> In the case of a hypothesis test
for the population mean, the test statistic turns out (under
certain conditions) to be a </span><span><span
class="calibre16"><span class="italic">z-</span></span></span>
<span>value (a value from the </span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>-distribution; see Chapter 9 ).</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Then you can look up your test
statistic on the appropriate table (in this case, you look it
up on the </span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span>-table in the
```

appendix), and find the chance that this difference between
your sample mean and the claimed population mean really could
have occurred if the claim were true.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The test statistic for testing one
population mean (under certain conditions) is</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15001.eps"
src="images/00298.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">where <img alt="9780470911082-
eq15002.eps" src="images/00299.jpg" class="calibre2"/> is
the sample mean, σ is the population
standard deviation (assume for this case that this number is
known), and z is a value on the
Z-distribution. To
calculate the test statistic, do the following:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"> 1. Calculate the sample mean,
<img alt="9780470911082-eq15003.eps"
src="images/00300.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> 2. Find <img
alt="9780470911082-eq15004.eps" src="images/00301.jpg"
class="calibre2"/>.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> 3. Calculate the standard error:
<img alt="9780470911082-eq15005.eps"
src="images/00302.jpg" class="calibre2"/>
.
</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> 4.<span

class="calibre16">
Divide your result from Step 2 by the standard
error found in Step 3.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The conditions
for using this test statistic are that the population standard
deviation, σ, is known, and either
the population has a normal distribution or the sample size is
large enough to use the CLT (n > 30); see Chapter 11.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For our example, suppose a random
sample of 100 working mothers spend an average of 11.5 minutes
per day talking with their children. (Assume prior research
suggests the population standard deviation is 2.3 minutes.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. We are given that <img
alt="9780470911082-eq15006.eps" src="images/00303.jpg"
class="calibre2"/> is 11.5, n
= 100, and σ is 2.3.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Take 11.5 – 11 = +0.5.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Take 2.3 divided by the square root
of 100 (which is 10) to get 0.23 for the standard error.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Divide +0.5 by 0.23 to get 2.17.
That’s your test statistic, which means your sample mean is
2.17 standard errors above the claimed population mean.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The big idea of
a hypothesis test is to challenge the claim that’s being made
about the population (in this case, the population mean); that
claim is shown in the null hypothesis, H_o
. If you have enough evidence from your sample against
the claim, H<sub
class="calibre42">o</sub><span> is rejected.

```html
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To decide whether you have enough
evidence to reject H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>, calculate the
</span><span><span class="calibre16"><span class="italic">p-
</span></span></span><span>value by looking up your test
statistic (in this case 2.17) on the standard normal (</span>
<span><span class="calibre16"><span class="italic">Z-</span>
</span></span><span>) distribution — see the </span><span><span
class="calibre16"><span class="italic">Z-</span></span></span>
<span>table in the appendix — and take 1 minus the probability
shown. (You subtract from 1 because your H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span> is a greater-than hypothesis and the table shows less-
than probabilities.)</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For this example you look up the test
statistic (2.17) on the </span><span><span class="calibre16">
<span class="italic">Z</span></span></span><span>-table and
find the (less-than) probability is 0.9850, so the</span><span>
<span class="calibre16"><span class="italic"> p-</span></span>
</span><span>value is 1 – 0.9850 = 0.015. It’s quite a bit less
than your (typical) significance level 0.05, which means your
sample results would be considered unusual if the claim (of 11
minutes) was true. So reject the claim (H</span><span><span
class="calibre41"><sub class="calibre42">o</sub></span></span>
<span>: </span><span>µ</span><span> = 11 minutes). Your results
support the alternative hypothesis H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span>: </span><span>µ</span><span> &gt; 11. According to your
data, the child psychologist’s claim of 11 minutes per day is
too low; the actual average is greater than that.</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For information on how to calculate
</span><span><span class="calibre16"><span class="italic">p-
</span></span></span><span>values for the less-than or not-
equal-to alternatives, also see Chapter 14.</span></span></p>
</blockquote>
<p id="a278" class="calibre6"><span class="calibre17"><span
class="bold"><span>Handling Small Samples and Unknown Standard
Deviations: The t-Test</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In two cases, you can’t use the </span>
<span><span class="calibre16"><span class="italic">Z</span>
</span></span><span>-distribution for a test statistic for one
population mean. The first case is where the sample size is
```

small (and by small, I mean dropping below 30 or so) the second
case is when the population standard deviation,
σ, is not known, and you have to estimate it
using the sample standard deviation, s.
 In both cases, you have less reliable information on
which to base your conclusions, so you have to pay a penalty
for this by using a distribution with more variability in the
tails than a Z-distribution has.
Enter the t-distribution. (See
Chapter 10 for all things
t-distribution,
including its relationship with the Z
.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A hypothesis test for a population mean
that involves the t -distribution is
called a t-test. The formula
for the test statistic in this case is:</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15007.eps"
src="images/00304.jpg" class="calibre2"/>, where
t
<sub
class="calibre42">n-1</sub> is a value from the
t-
distribution with n
–1 degrees of freedom.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Note it is just like the test statistic
for the large sample and/or normal distribution case (see the
section “Testing One Population Mean”), except
σ is not known, so you substitute the sample
standard deviation, s, instead, and use
a
t-value rather than a
z
-value.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> Because the
t-distribution has
fatter tails than the
Z-distribution,
you get a larger p-value from the
t-distribution than
one that the standard normal (Z-
) distribution would have given you for the same test
statistic. A bigger p-value means less
chance of rejecting H<sub
class="calibre42">o.</sub> Having less data
and/or not knowing the population standard deviation should
create a higher burden of proof.</p></blockquote>
<blockquote class="calibre5"><p id="a279" class="calibre6">
Putting the t-
test to work</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose a delivery company claims they
deliver their packages in 2 days on average, and you suspect
it’s longer than that. The hypotheses are H_o
: μ = 2 versus H
_a
: μ > 2. To test this
claim, you take a random sample of 10 packages and record their
delivery times. You find the sample mean is <img
alt="9780470911082-eq15008.eps" src="images/00305.jpg"
class="calibre2"/> days, and the sample standard
deviation is 0.35 days. (Because the population standard
deviation, σ, is unknown, you
estimate it with s, the sample
standard deviation.) This is a job for the t
-test.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Because the
sample size is small (
n =10 is much
less than 30) and the population standard deviation is not
known, your test statistic has a <span><span

```
class="calibre16"><span class="italic">t-</span></span></span>
<span>distribution. Its degrees of freedom is 10 – 1 = 9. The
formula for the test statistic (referred to as the </span>
<span><span class="calibre16"><span class="italic">t-
value</span></span></span><span>) is:</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq15009.eps"
src="images/00306.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To calculate the </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>value, you look in the row in the </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>table (in the appendix) for </span><span><span
class="calibre16"><span class="italic">df</span></span></span>
<span> = 9. Your test statistic (2.71) falls between two values
in the row for </span><span><span class="calibre16"><span
class="italic">df </span></span></span><span>= 9 in the </span>
<span><span class="calibre16"><span class="italic">t</span>
</span></span><span>-table: 2.26 and 2.82 (rounding to two
decimal places). To calculate the </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>value for your test statistic, find which columns
correspond to these two numbers. The number 2.26 appears in the
0.025 column and the number 2.82 appears in the 0.010 column;
you now know the </span><span><span class="calibre16"><span
class="italic">p-</span></span></span><span>value for your test
statistic lies between 0.025 and 0.010 (that is, 0.010 &lt;
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>-value &lt; 0.025).
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Using the </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span>-table you don’t know the exact number for the </span>
<span><span class="calibre16"><span class="italic">p-</span>
</span></span><span>value, but because 0.010 and 0.025 are both
less than your significance level of 0.05, you reject H</span>
<span><span class="calibre41"><sub class="calibre42">o</sub>
</span></span><span>; you have enough evidence in your sample
to say the packages are not being delivered in 2 days, and in
fact the average delivery time is more than 2 days.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> The </span>
```

t-
table (in the appendix) doesn’t include
every possible t-value; just find
the two values closest to yours on either side, look at the
columns they’re in, and report your p-
value in relation to theirs. (If your test statistic is
greater than all the t-values in the
corresponding row of the
t-table, just
use the last one; your
p-value will be
less than its probability.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Of course you can use statistical
software, if available, to calculate exact p
-values for any test statistic; using software you get
0.012 for the exact p-value.
</p></blockquote>
<blockquote class="calibre5"><p id="a280" class="calibre6">
Relating t to
Z</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The next-to-the-last line of the
t-table shows the
corresponding values from the standard normal (
Z-
) distribution for the probabilities listed on the
top of each column. Now choose a column in the table and move
down the column looking at the t
-values. As the degrees of freedom of the
t-
distribution increase, the t
-values get closer and closer to that row of the table
where the z-values are.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>This confirms a result found in Chapter
10: As the sample size (hence degrees of freedom) increases,

```
the </span><span><span class="calibre16"><span
class="italic">t-</span></span></span><span>distribution
becomes more and more like the </span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>-distribution, so the </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>values from their hypothesis tests are virtually equal
for large sample sizes. And those sample sizes don’t even have
to be that large to see this relationship; for </span><span>
<span class="calibre16"><span class="italic">df </span></span>
</span><span>= 30 the </span><span><span class="calibre16">
<span class="italic">t</span></span></span><span>-values are
already very similar to the </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>-values shown in the bottom of the table. These results
make sense; the more data you have, the less of a penalty you
have to pay. (And of course, you can use computer technology to
calculate more exact </span><span><span class="calibre16"><span
class="italic">p-</span></span></span><span>values for any
</span><span><span class="calibre16"><span
class="italic">t</span></span></span><span>-value you like.)
</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a281" class="calibre6">
<span class="calibre21"><span class="bold"><span>Handling
negative t-values</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For a less-than alternative hypothesis
(H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span>: xx &lt; xx),
your test statistic would be a negative number (to the left of
0 on the </span><span><span class="calibre16"><span
class="italic">t-</span></span></span><span>distribution). In
this case, you want to find the percentage below, or to the
left of, your test statistic to get your </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>value. Yet negative test statistics don’t appear on the
</span><span><span class="calibre16"><span class="italic">t-
</span></span></span><span>table (in the appendix).</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Not to worry! The percentage to the
left (below) a negative </span><span><span class="calibre16">
<span class="italic">t-</span></span></span><span>value is the
same as the percentage to the right (above) the positive
</span><span><span class="calibre16"><span class="italic">t-
</span></span></span><span>value, due to symmetry. So to find
the </span><span><span class="calibre16"><span
class="italic">p-</span></span></span><span>value for your
```

negative test statistic, look up the positive version of your
test statistic on the
t-table, find
the corresponding right tail (greater-than) probability, and
use that.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose your test
statistic is –2.7105 with 9 degrees of freedom and H
_a
 is the less-than alternative. To find your
p-value, first look
up +2.7105 on the t-table; by the work
in the previous section, you know its p
-value falls between the column headings 0.025 and 0.010.
Because the t-distribution is
symmetric, the p-value for –2.7105
also falls somewhere between 0.025 and 0.010. Again you reject
H<sub
class="calibre42">o</sub> because these
values are both less than or equal to 0.05.</p>
</blockquote>
<blockquote class="calibre5"><p id="a282" class="calibre6">
Examining the
not-equal-to alternative</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> To find the
p-
value when your alternative
hypothesis (H<sub
class="calibre42">a</sub>) is not
-
equal-to, simply double the probability
that you get from the
t-table when
you look up your test statistic. Why double it? Because the
t-table shows only
greater-than probabilities, which are only half the story. To
find the p-value when you have
a not-equal-to alternative, you must add the p-

values from the less-than and greater-than alternatives.
Because the t-distribution is
symmetric, the less-than and greater-than probabilities are the
same, so just double the one you looked up on the
t
-table and you’ll have the p-
value for the not-equal-to alternative.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, if your test statistic is
2.7171 and H<sub
class="calibre42">a</sub> is a not-equal-to
alternative, look up 2.7171 on the t
-table (df = 9 again), and
you find the p-value lies between
0.025 and 0.010, as shown previously. These are the
p-
values for the greater-than alternative.
Now double these values to include the less-than alternative
and you find the p-value for your test
statistic lies somewhere between 0.025 ∗
 2 = 0.05 and 0.010 ∗ 2 =
0.020.</p></blockquote>
<p id="a283" class="calibre6">Testing One Population Proportion
</p>
<blockquote class="calibre9"><p class="calibre14">When the variable is categorical (for
example, gender or support/oppose) and only one population or
group is being studied (for example, all registered voters),
you use the hypothesis test in this section to test a claim
about the population proportion. The test looks at the
proportion (p) of individuals in
the population who have a certain characteristic — for example,
the proportion of people who carry cellphones. The null
hypothesis is H<sub
class="calibre42">o</sub>:
p
 = p<span

class="calibre41">_o
, where p_o
 is a certain claimed value of the population proportion,
p. For example, if
the claim is that 70% of people carry cellphones,
p
<sub
class="calibre42">o</sub> is 0.70. The
alternative hypothesis is one of the following:
p
> p_o
, p <
p
<sub
class="calibre42">o</sub>, or
p
 ≠ p_o
. (See Chapter 14 for more on alternative hypotheses.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the test statistic for
a single proportion (under certain conditions) is:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15010.eps"
src="images/00307.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">where <img alt="9780470911082-
eq15011.eps" src="images/00308.jpg" class="calibre2"/> is
the proportion of individuals in the sample who have that
characteristic and z is a value on the
Z-distribution (see
Chapter 9). To calculate the test statistic, do the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Calculate the sample proportion,

<img alt="9780470911082-eq15012.eps"
src="images/00309.jpg" class="calibre2"/>, by taking the number of
people in the sample who have the characteristic of interest
(for example, the number of people in the sample carrying
cellphones) and dividing that by
n, the sample size.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find <img
alt="9780470911082-eq15013.eps" src="images/00310.jpg"
class="calibre2"/>, where p<sub
class="calibre42">o</sub> is the value in H

_o.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Calculate the standard error,
<img alt="9780470911082-eq15014.eps"
src="images/00311.jpg" class="calibre2"/>.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Divide your result from Step 2 by your result
from Step 3.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">To interpret the test statistic, look
up your test statistic on the standard normal (
Z-
) distribution (in the appendix) and calculate the
p-
value (see Chapter 14 for more on
p-
value calculations).
</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The conditions
for using this test statistic are that <img
alt="9780470911082-eq15051.eps" src="images/00312.jpg"
class="calibre2"/> (see Chapter 9 for details).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose Cavifree claims
that four out of five dentists recommend Cavifree toothpaste to
their patients. In this case, the population is all dentists,
and p
is the proportion of all dentists
who recommended Cavifree. The claim is that p
 is equal to “four out of five,” or p
_o
 is 4 ÷ 5 = 0.80. You suspect that the
proportion is actually less than 0.80. Your hypotheses are
H<sub
class="calibre42">o</sub>:
p
 = 0.80 versus H_a
: p < 0.80.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose that 151 out of your sample of
200 dental patients reported receiving a recommendation for
Cavifree from their dentist. To find the test statistic for
these results, follow these steps:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Start with <img
alt="9780470911082-eq15015.eps" src="images/00313.jpg"
class="calibre2"/> and n
= 200.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Because p
_o
 = 0.80, take 0.755 – 0.80 = –0.045 (the
numerator of the test statistic).</p>
</blockquote><div class="calibre31"> </div>

```
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> 3. Next, the standard error equals
</span><img alt="9780470911082-eq15016.eps"
src="images/00314.jpg" class="calibre2"/><span> (the
denominator of the test statistic).</span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> 4. The test statistic is </span><img
alt="9780470911082-eq15017.eps" src="images/00315.jpg"
class="calibre2"/><span>.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Because the
resulting test statistic is negative, it means your sample
results are –1.61 standard errors below (less than) the claimed
value for the population. How often would you expect to get
results like this if H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span> were true? The
chance of being at or beyond (in this case less than) –1.61 is
0.0537. (Keep the negative with the number and look up –1.61 in
the </span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span>-table in the
appendix.) This result is your </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>value because H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span> is a less-than
hypothesis. (See Chapter 14 for more on this.)</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Because the </span><span><span
class="calibre16"><span class="italic">p-</span></span></span>
<span>value is greater than 0.05 (albeit not by much), you
don’t have quite enough evidence for rejecting H</span><span>
<span class="calibre41"><sub class="calibre42">o</sub></span>
</span><span>. You conclude that the claim that 80% of dentists
recommend Cavifree can’t be rejected, according to your data.
However, it’s important to report the actual </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>-value too, so others can make their own decisions.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> The letter </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span> is used two different ways in this chapter: </span>
<span><span class="calibre16"><span class="italic">p</span>
</span></span><span>-value and </span><span><span
```

class="calibre16">p
. The letter p by itself indicates
the population proportion, not the p
-value. Don’t get confused. Whenever you report a
p
-value, be sure you add –value
so it’s not confused with p,
the population proportion.
</p></blockquote>
<p id="a284" class="calibre6">Comparing Two (Independent) Population
Averages</p>
<blockquote class="calibre9"><p class="calibre14">When the variable is numerical (for
example, income, cholesterol level, or miles per gallon) and
two populations or groups are being compared (for example, men
versus women), you use the steps in this section to test a
claim about the difference in their averages. (For example, is
the difference in the population means equal to zero,
indicating their means are equal?) Two independent (totally
separate) random samples need to be selected, one from each
population, in order to collect the data needed for this test.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The null hypothesis is that the two
population means are the same; in other words, that their
difference is equal to 0. The notation for the null hypothesis
is H<sub
class="calibre42">o</sub>:
µ<sub
class="calibre42">1</sub> =
µ<sub
class="calibre42">2</sub>, where
µ<sub
class="calibre42">1</sub> represents the
mean of the first population and µ
₂
 represents the mean of the second population.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> You can also write the null hypothesis
as H<sub

class="calibre42">o</sub>:
μ<sub
class="calibre42">1</sub> –
μ<sub
class="calibre42">2 </sub>= 0, emphasizing
the idea that their difference is equal to zero if the means
are the same.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the test statistic
comparing two means (under certain conditions) is:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15018.eps"
src="images/00316.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">To calculate it, do the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Calculate the sample means
<img alt="9780470911082-eq15019.eps"
src="images/00317.jpg" class="calibre2"/>. (Assume the population
standard deviations, <img
alt="9780470911082-eq15049.eps" src="images/00318.jpg"
class="calibre2"/> and <img alt="9780470911082-
eq15050.eps" src="images/00319.jpg" class="calibre2"/>
 are given.) Let
n
<sub
class="calibre42">1</sub> and
n<sub
class="calibre42">2</sub> represent the two sample
sizes (they need not be equal).</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> See Chapter 5 for these calculations.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">2. Find the difference between the two sample

means: <img alt="9780470911082-
eq15020.eps" src="images/00320.jpg" class="calibre2"/>
.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Because
μ<sub
class="calibre42">1</sub>
– μ₂

is equal to 0 if H_o
 is true, it doesn’t need to be included in the numerator
of the test statistic. However, if the difference they are
testing is any value other than 0, you subtract that value in
the numerator of the test statistic.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Calculate the standard error using the
following equation:</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15021.eps"
src="images/00321.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Divide your result from Step 2 by your result
from Step 3.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> To interpret
the test statistic, add the following two steps to the list:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 5. Look up your test statistic on the standard
normal (
Z-
)
distribution (see the <span

class="calibre16">Z-table in the appendix)
and calculate the p-
value.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> (See Chapter 14 for more on
p-
value calculations.)</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Compare the p-
value to your significance level, such as 0.05. If
it’s less than or equal to 0.05, reject H
<sub
class="calibre42">o</sub>. Otherwise, fail to
reject H
_o
.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> (See Chapter 14 for the details on
significance levels.)</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The conditions
for using this test are that the two population standard
deviations are known and either both populations have a normal
distribution or both sample sizes are large enough for the
Central Limit Theorem (see Chapter 11).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you want to
compare the absorbency of two brands of paper towels (call the
brands Stats-absorbent and Sponge-o-matic). You can make this
comparison by looking at the average number of ounces each
brand can absorb before being saturated. H_o
 says the difference between the average absorbencies is

0 (nonexistent), and H<sub
class="calibre42">a</sub> says the
difference is not 0. In other words, one brand is more
absorbent than the other. Using statistical notation, you have
H<sub
class="calibre42">o</sub> =
µ<sub
class="calibre42">1</sub> –
µ<sub
class="calibre42">2</sub> = 0 versus
H<sub
class="calibre42">a</sub> =
µ<sub
class="calibre42">1</sub> –
µ<sub
class="calibre42">2</sub>
≠ 0.
Here, you have no indication of which paper towel may be more
absorbent, so the not-equal-to alternative is the one to use
(see Chapter 14).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you select a random sample of
50 paper towels from each brand and measure the absorbency of
each paper towel. Suppose the average absorbency of Stats-
absorbent (x₁
) for your sample is 3 ounces, and assume the population
standard deviation is 0.9 ounces. For Sponge-o-matic (
x
<sub
class="calibre42">2</sub>), the average
absorbency is 3.5 ounces according to your sample; assume the
population standard deviation is 1.2 ounces. Carry out this
hypothesis test by following the 6 steps listed above:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Given the above information, you
know <img alt="9780470911082-eq15022.eps"
src="images/00322.jpg" class="calibre2"/>,
σ<sub
class="calibre42">1</sub> = 0.9,
<img alt="9780470911082-eq15023.eps" src="images/00323.jpg"
class="calibre2"/>, σ₂
 = 1.2, n₁

 = 50, and n₂
 = 50.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. The difference between the sample
means for (Stats-absorbent – Sponge-o-matic) is <img
alt="9780470911082-eq15024.eps" src="images/00324.jpg"
class="calibre2"/>. (A negative difference simply means
that the second sample mean was larger than the first.)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The standard error is <img
alt="9780470911082-eq15025.eps" src="images/00325.jpg"
class="calibre2"/>.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Divide the difference, –0.5, by the
standard error, 0.2121, which gives you –2.36. This is your
test statistic.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. To find the p-
value, look up –2.36 on the standard normal (
Z-
) distribution — see the Z
-table in the appendix. The chance of being beyond, in
this case to the left of, –2.36 is equal to 0.0091. Because
H<sub
class="calibre42">a</sub> is a not-equal-to
alternative, you double this percentage to get 2
∗ 0.0091 = 0.0182, your p-
value. (See Chapter 14 for more on the not-equal-to
alternative.)</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. This p-
value is quite a bit less than 0.05. That means you have
fairly strong evidence to reject H_o
.</p></blockquote><div
class="calibre31"> </div>

```
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Your conclusion is that a statistically
significant difference exists between the absorbency levels of
these two brands of paper towels, based on your samples. And
Sponge-o-matic comes out on top, because it has a higher
average. (Stats-absorbent minus Sponge-o-matic being negative
means Sponge-o-matic had the higher value.)</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/><span> If one or both
of your samples happen to be under 30 in size, you use the
</span><span><span class="calibre16"><span class="italic">t-
</span></span></span><span>distribution (with degrees of
freedom equal to </span><span><span class="calibre16"><span
class="italic">n</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
<span> – 1 or </span><span><span class="calibre16"><span
class="italic">n</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
<span> – 1, whichever is smaller) to look up the </span><span>
<span class="calibre16"><span class="italic">p-</span></span>
</span><span>value. If the population standard deviations,
</span><span>σ</span><span><span class="calibre41"><sub
class="calibre42">1</sub></span></span><span> and </span>
<span>σ</span><span><span class="calibre41"><sub
class="calibre42">2</sub></span></span><span>, are unknown, you
use the sample standard deviations </span><span><span
class="calibre16"><span class="italic">s</span></span></span>
<span><span class="calibre41"><sub class="calibre42">1</sub>
</span></span><span> and </span><span><span class="calibre16">
<span class="italic">s</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
<span> instead, and you use the </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span>-distribution with the abovementioned degrees of freedom.
(See Chapter 10 for more on the </span><span><span
class="calibre16"><span class="italic">t-</span></span></span>
<span>distribution.)</span></span></p></blockquote>
<p id="a285" class="calibre6"><span class="calibre17"><span
class="bold"><span>Testing for an Average Difference (The
Paired t-Test)</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You can test for an average difference
using the test in this section when the variable is numerical
(for example, income, cholesterol level, or miles per gallon)
and the individuals in the sample are either paired up in some
way according to relevant variables such as age or perhaps
```

weight, or the same people are used twice (for example, using a pre-test and post-test). Paired tests are typically used for studies in which someone is testing to see whether a new treatment, technique, or method works better than an existing method, without having to worry about other factors about the subjects that may influence the results (see Chapter 17 for details).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The average difference (tested in this section) isn’t the same as the difference in the averages (tested in the previous section): </p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
With the difference in averages, you compare the difference in the means of two separate samples to test the difference in the means of two different populations. </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
With the average difference, you match up the subjects so they are thought of as coming from a single population, and the set of differences measured for each subject (for example, pre-test versus post-test) are thought of as one sample. The hypothesis test then boils down to a test for one population mean (as I explain earlier in this chapter).</p> </blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose a researcher wants to see whether teaching students to read using a computer game gives better results than teaching with a tried-and-true phonics method. She randomly selects 20 students and puts them into 10 pairs according to their reading readiness level, age, IQ, and so on. She randomly selects one student from each pair to learn to read via the computer game method (abbreviated CM), and the other learns to read using the phonics method (abbreviated PM). At the end of the study, each student takes the same reading test. The data are shown in Table 15-1. </p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The original data are in pairs, but you’re really interested only in the difference in reading

scores (computer reading score minus phonics reading score) for each pair, not the reading scores themselves. So the
paired differences (the differences in the pairs of scores) are your new data set. See their values in the last column of Table 15-1.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">By examining the differences in the pairs of observations, you really only have a single data set, and you only have a hypothesis test for one population mean. In this case the null hypothesis is that the mean (of the paired differences) is 0, and the alternative hypothesis is that the mean (of the paired differences) is > 0.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">If the two reading methods are the same, the average of the paired differences should be 0. If the computer method is better, the average of the paired differences should be positive; the computer reading score is larger than the phonics score.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The notation for the null hypothesis is H_o
: μ
_d
 = 0, where μ
_d is the mean of the paired differences for the population. (The
d
 in the subscript just reminds you that you’re working with the paired differences.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the test statistic for paired differences is , where is the average of all the paired differences found in the sample, and
t
_n
_{–1} is a value on the
t

```
</span><span>-distribution with </span><span><span
class="calibre16"><span class="italic">n<span><span
class="calibre41"><sub class="calibre42">d</sub></span></span>
</span></span></span><span>–1 degrees of freedom (see Chapter
10). </span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> You use a t-
distribution here because in most matched-pairs experiments the
sample size is small and/or the population standard deviation
</span><span>σ</span><span>d is unknown, so it’s estimated by
sd. (See Chapter 10 for more on the t-distribution.)</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>To calculate the test statistic for
paired differences, do the following:</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 1. For each pair of data, take the first value in
the pair minus the second value in the pair to find the paired
difference.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> Think of the differences as your new
data set.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Calculate the mean, </span></span></span><img
alt="9780470911082-eq15028.eps" src="images/00329.jpg"
class="calibre2"/><span><span class="calibre16"><span
class="bold">, and the standard deviation, </span></span>
</span><span><span class="calibre16"><span class="bold"><span
class="italic">s</span></span></span></span><span><span
class="calibre43"><span class="bold"><span class="italic"><sub
class="calibre42">d</sub></span></span></span></span><span>
<span class="calibre16"><span class="bold"><span
class="italic">,</span></span></span></span><span><span
class="calibre16"><span class="bold"> of all the differences.
</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 3. Letting </span></span></span><span><span
class="calibre16"><span class="bold"><span class="italic">nd
</span></span></span></span><span><span class="calibre16"><span
class="bold">represent the number of paired differences that
```

you have, calculate the standard error:
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15029.eps"
src="images/00330.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Divide <img
alt="9780470911082-eq15030.eps" src="images/00331.jpg"
class="calibre2"/> by the standard error from Step 3.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Because µ
<sub
class="calibre42">d</sub> is equal
to 0 if H<sub
class="calibre42">o</sub> is true, it
doesn’t really need to be included in the formula for the test
statistic. As a result, you sometimes see the test statistic
written like this:</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15031.eps"
src="images/00332.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> For the reading
scores example, you can use the preceding steps to see whether
the computer method is better in terms of teaching students to
read.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To find the statistic, follow these
steps:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Calculate the differences for each pair
(they’re shown in column 4 of Table 15-1).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Notice that the sign on each of the
differences is important; it indicates which method performed
better for that particular pair.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 2. Calculate the mean and standard deviation of
the differences from Step 1.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> My calculations found the mean of the
differences, <img alt="9780470911082-eq15032.eps"
src="images/00333.jpg" class="calibre2"/>, and the
standard deviation is
s<sub
class="calibre42">d</sub> = 4.64.
Note that n<sub
class="calibre42">d</sub>
= 10 here.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The standard error is
<img alt="9780470911082-eq15033.eps" src="images/00334.jpg"
class="calibre2"/>.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> (Remember that here, n
<sub
class="calibre42">d</sub>
is the number of pairs, which is
10.)</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">4. Take the mean of the differences (Step 2)
divided by the standard error of 1.47 (Step 3) to get 1.36, the
test statistic.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Is the result of Step 4 enough to say
that the difference in reading scores found in this experiment
applies to the whole population in general? Because the
population standard deviation, σ, is
unknown and you estimated it with the sample standard deviation
(s), you need to use

the t-distribution rather
than the Z-distribution to
find your p-value (see the
section “Handling Small Samples and Unknown Standard
Deviations: The t-Test,” earlier in
this chapter). Using the
t-table (in the
appendix) you look up 1.36 on the t-
distribution with 10 – 1 = 9 degrees of freedom to
calculate the p-value.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The p-
value in this case is greater than 0.05 because 1.36 is
smaller than (or to the left of) the value of 1.38 on the
table, and therefore its
p-value is more
than 0.10 (the p-value

for the column heading corresponding
to 1.38).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because the p-
value is greater than 0.05, you fail to reject H
_o
; you don’t have enough evidence that the
mean difference in the scores between the computer method and
the phonics method is significantly greater than 0. However,
that doesn’t necessarily mean a real difference isn’t present
in the population of all students. But the researcher can’t say
the computer game is a better reading method based on this
sample of 10 students. (See Chapter 14 for information on the
power of a hypothesis test and its relationship to sample
size.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> In many paired
experiments, the data sets are small due to costs and time
associated with doing these kinds of studies. That means the

t-
distribution (see the
t
-table in the appendix) is often used instead of
the standard normal (Z-) distribution (the
Z-table in the
appendix) when figuring out the p-
value.</p></blockquote>
<p id="a286" class="calibre6">Comparing Two Population Proportions
</p>
<blockquote class="calibre9"><p class="calibre14">This test is used when the variable is
categorical (for example, smoker/nonsmoker,
Democrat/Republican, support/oppose an opinion, and so on)
and you’re interested in the proportion of individuals with a
certain characteristic — for example, the proportion of
smokers. In this case, two populations or groups are being
compared (such as the proportion of female smokers versus male
smokers).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In order to conduct this test, two
independent (separate) random samples need to be selected, one
from each population. The null hypothesis is that the two
population proportions are the same; in other words, that their
difference is equal to 0. The notation for the null hypothesis
is H<sub
class="calibre42">o</sub>:
p
<sub
class="calibre42">1</sub> =
p
<sub
class="calibre42">2</sub>, where
p
<sub
class="calibre42">1</sub> is the proportion
from the first population, and p
₂
 is the proportion from the second
population.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Stating in

H<sub
class="calibre42">o</sub> that the two
proportions are equal is the same as saying their difference is
zero. If you start with the equation p
₁
 =
p₂
 and subtract p₂
 from each side, you get p
₁
 –
p₂
 = 0. So you can write the null hypothesis either way.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the test statistic
comparing two proportions (under certain conditions) is
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15034.eps"
src="images/00335.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">where <img alt="9780470911082-
eq15035.eps" src="images/00336.jpg" class="calibre2"/> is
the proportion in the first sample with the characteristic of
interest, <img alt="9780470911082-eq15036.eps"
src="images/00337.jpg" class="calibre2"/> is the
proportion in the second sample with the characteristic of
interest, <img alt="9780470911082-eq15037.eps"
src="images/00338.jpg" class="calibre2"/> is the
proportion in the combined sample (all the individuals in the
first and second samples together) with the characteristic of
interest, and z is a value on the
Z-distribution (see
Chapter 9). To calculate the test statistic, do the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Calculate the sample proportions

<img alt="9780470911082-eq15038.eps"
src="images/00339.jpg" class="calibre2"/> and
<img alt="9780470911082-eq15039.eps" src="images/00340.jpg"
class="calibre2"/> for each sample. Let
n<sub
class="calibre42">1</sub> and
n<sub
class="calibre42">2</sub> represent the two sample
sizes (they don’t need to be equal).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the difference between the two sample
proportions, <img alt="9780470911082-
eq15040.eps" src="images/00341.jpg" class="calibre2"/>
.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Calculate the overall sample proportion
<img alt="9780470911082-eq15041.eps"
src="images/00342.jpg" class="calibre2"/>, the total number of
individuals from both samples who have the characteristic of
interest (for example, the total number of smokers, male or
female, in the sample), divided by the total number of
individuals from both samples (n<sub
class="calibre42">1</sub> +
n<sub
class="calibre42">2</sub>).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 4. Calculate the standard error:
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq15042.eps"
src="images/00343.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Divide your result from Step 2 by your result
from Step 4. This answer is your test statistic.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">To interpret the test statistic, look
up your test statistic on the standard normal (
Z-
) distribution (the Z
-table in the appendix) and calculate the
p-
value, then make decisions as usual (see Chapter
14 for more on p-values).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Consider those drug ads that
pharmaceutical companies put in magazines. The front page of an
ad shows a serene picture of the sun shining, flowers blooming,
people smiling — their lives changed by the drug. The company
claims that its drugs can reduce allergy symptoms, help people
sleep better, lower blood pressure, or fix whichever other
ailment it’s targeted to help. The claims may sound too good to
be true, but when you turn the page to the back of the ad, you
see all the fine print where the drug company justifies how
it’s able to make its claims. (This is typically where
statistics are buried!) Somewhere in the tiny print, you’ll
likely find a table that shows adverse effects of the drug when
compared to a control group
(subjects who take a fake drug), for fair comparison to those
who actually took the real drug (the treatment group;
 see Chapter 17 for more on this).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For example, Adderall, a drug for
attention deficit hyperactivity disorder (ADHD), reported that
26 of the 374 subjects (7%) who took the drug experienced
vomiting as a side effect, compared to 8 of the 210 subjects

```
(4%) who were on a </span><span><span class="calibre16"><span
class="italic">placebo </span></span></span><span>(fake drug).
Note that patients didn’t know which treatment they were given.
In the sample, more people on the drug experienced vomiting,
but is this percentage enough to say that the entire population
on the drug would experience more vomiting? You can test it to
see.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In this example, you have H</span>
<span><span class="calibre41"><sub class="calibre42">o</sub>
</span></span><span>: </span><span><span class="calibre16">
<span class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
<span> – </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
<span> = 0 versus H</span><span><span class="calibre41"><sub
class="calibre42">o</sub></span></span><span>: </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span><span class="calibre41"><sub
class="calibre42">1</sub></span></span><span> – </span><span>
<span class="calibre16"><span class="italic">p</span></span>
</span><span><span class="calibre41"><sub
class="calibre42">2</sub></span></span><span> &gt; 0, where
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">1</sub></span></span>
<span> represents the proportion of subjects who vomited using
Adderall, and </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span><span
class="calibre41"><sub class="calibre42">2</sub></span></span>
<span> represents the proportion of subjects who vomited using
the placebo.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/><span> Why does
H</span><span><span class="calibre41"><sub
class="calibre42">a</sub></span></span><span> contain a “&gt;”
sign and not a “&lt;” sign? H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
<span> represents the scenario in which those taking Adderall
experience more vomiting than those on the placebo — that’s
something the FDA (and any candidate for the drug) would want
to know about. But the order of the groups is important, too.
You want to set it up so the Adderall group is first, so that
when you take the Adderall proportion minus the placebo
proportion, you get a positive number if H</span><span><span
class="calibre41"><sub class="calibre42">a</sub></span></span>
```

 is true. If you switch the groups, the sign would have been negative.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now calculate the test statistic:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. First, determine that
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre39"></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"> The sample sizes are
n
₁ = 374 and
n
₂ = 210, respectively.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Take the difference between these sample proportions to get .
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Calculate the overall sample proportion to get .
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. The standard error is .</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Finally, the test statistic is 0.032 ÷ 0.020 = 1.60. Whew!</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">The p-
value is the percentage chance of being at or beyond (in this case to the right of) 1.60, which is 1 – 0.9452 = 0.0548. This p-value is just

slightly greater than 0.05, so, technically, you don’t have quite enough evidence to reject H_o. That means that according to your data, vomiting is not experienced any more by those taking this drug when compared to a placebo.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A p-value that’s very close to that magical but somewhat arbitrary significance level of 0.05 is what statisticians call a marginal result. In the preceding example, because the p-value of 0.0548 is close to the borderline between accepting and rejecting H_o, it’s generally viewed as a marginal result and should be reported as such.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The beauty of reporting a p-value is that you can look at it and decide for yourself what you should conclude. The smaller the p-value, the more evidence you have against H_o, but how much evidence is enough evidence? Each person is different. If you come across a report from a study in which someone found a statistically significant result, and that result is important to you, ask for the p-value so that you can make your own decision. (See Chapter 14 for more.)</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a287"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

  <a href="#a277" style="min-width: 10px !important; min-height:

10px !important; border: solid 1px !important;"> <a
href="#a276" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a288" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a286" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a287" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a279" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a278" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a282" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a283" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a280" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a281" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a291">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part V</span></span>
</span></span></p>
<p id="a289" class="calibre10"><span class="calibre11"><span
class="bold"><span>Statistical Studies and the Hunt for a
Meaningful Relationship</span></span></span></p>
<div class="calibre13">
<img alt="9780470911082 pp0501.eps" src="images/00348.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">M<span>any statistics you hear and see each
day are based on the results of surveys, experiments, and
observational studies. Unfortunately, you can’t believe
everything you read or hear.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In this part, you look at what actually
happens behind the scenes of these studies — how they are
designed and conducted and how the data is (supposed to be)
collected — so that you’ll be able to spot misleading results.
You also see what’s needed to conduct your own study correctly
and effectively.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You also analyze data from good studies
to look for relationships between two variables, where both
variables are categorical (using two-way tables) or both are
numerical (using correlation and regression). In addition, you
see how to make proper conclusions and spot problems.</span>
</span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a290"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
```

!important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

      </div></body></html>

<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<title>Statistics For Dummies, 2nd Edition</title>

<link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
<body class="calibre" id="a303">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10">Chapter 16
</p><div class="calibre19"> </div>
<p id="a292" class="calibre10">Polls, Polls, and More Polls
</p>
<blockquote class="calibre9"><p class="calibre24">
In This Chapter</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
Realizing the impact of polls and surveys
</p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
Going behind the scenes of polls and surveys
</p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
Detecting biased and inaccurate survey results
</p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14">Surveys are all the rage
amid today’s information explosion. Everyone wants to know how
the public feels about issues from prescription drug prices and
methods of disciplining children to approval ratings of the
president and ratings of reality TV shows. Polls and surveys
are a big part of American life; they’re a vehicle for quickly
getting information about how you feel, what you think, and how
you live your life, and they’re a means of quickly
disseminating information about important issues. Surveys

highlight controversial topics, raise awareness, make political
points, stress the importance of an issue, and educate or
persuade the public.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Survey results
can be powerful, because when many people hear that “such and
such percentage of the American people do this or that,” they
accept these results as the truth, and then make decisions and
form opinions based on that information. But in fact, many
surveys don’t provide
correct, complete, or even fair or balanced information.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, I discuss the impact
of surveys and how they’re used, and I take you behind the
scenes of how surveys are designed and conducted so you know
what to watch for when examining survey results and how to run
your own surveys right. I also talk about how to interpret
survey results and how to spot biased and inaccurate
information, so that you can determine for yourself which
results to believe and which to ignore.</p>
</blockquote>
<p id="a293" class="calibre6">Recognizing the Impact of Polls
</p>
<blockquote class="calibre9"><p class="calibre14">A
survey is an
instrument that collects data through questions and answers. It
is used to gather information about the opinions, behaviors,
demographics, lifestyles, and other reportable characteristics
of the population of interest. What’s the difference between a
poll and a survey? Statisticians don’t make a clear distinction
between the two, but I’ve noticed that what people call a
poll is typically a
short survey containing only a few questions (maybe that’s how
researchers get more people to respond — they call it a poll
rather than a survey!). But for all intents and purposes,
surveys and polls are the same thing.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You come into contact with surveys and
their results on a daily basis. Compared to other types of
studies, such as medical experiments, some surveys can be
relatively easy to conduct. They provide quick results that can

often make interesting headlines in newspapers or eye-catching
stories in magazines. People connect with surveys because they
feel that survey results represent the opinions of people just
like themselves (even though they may never have been asked to
participate in a survey). And many people enjoy seeing how
other people feel, what they do, where they go, and what they
care about. Looking at survey results makes people feel linked
with a bigger group, somehow. That’s what pollsters
(the people who conduct surveys) bank on, and
that’s why they spend so much time doing surveys and polls and
reporting the results of this research.</p>
</blockquote>
<blockquote class="calibre5"><p id="a294" class="calibre6">
Getting to the
source</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Who conducts surveys these days? Pretty
much anyone and everyone who has a question to ask. Some of the
groups that conduct polls and report the results include the
following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
News organizations</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Political parties and candidates running for
office</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Professional polling organizations (such as the Gallup
Organization, the Harris Poll, Zogby International, and the
National Opinion Research Center [NORC])</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Representatives of magazines, TV shows, and radio
programs</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>

Professional research organizations (like the American Medical Association, Smithsonian Institution, and Pew Research Center for the People and the Press)</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Special-interest groups (such as the National Rifle Association, Greenpeace, and American Civil Liberties Union)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Academic researchers</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The United States government</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Joe Six-Pack (who can easily conduct his own survey on the Internet)</p></blockquote><div class="calibre19"> </div>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">Ranking the worst cars of the millennium</p></blockquote>
<div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">You may be familiar with a radio show called Car Talk that’s typically aired Saturday mornings on National Public Radio and is hosted by “Click and Clack,” two brothers in Cambridge, Massachusetts, who offer wise and wacky advice to callers with strange car problems. The show’s Web site regularly offers “just for fun” surveys on a wide range of car-related topics, such as, “Who has bumper stickers on their cars, and what do they say?” One of their surveys asked the question, “What do you think was the worst car of the millennium?” Thousands upon thousands of folks responded with their votes — but, of course, these folks don’t represent all car owners. They represent only those who listen to the radio show, logged on to the Web site, and answered the survey question.</p>

```
</blockquote>
<blockquote class="calibre5"><p class="calibre34"><span
class="calibre35"><span>Just so you won’t be left hanging (and
I know you’re dying to find out!), the results of the survey
are shown in the following table. Although you may not be old
enough to remember some of these vehicles, it is certainly an
easy exercise to search the Internet for pictures and stories
about them galore. (Remember, though, that these results
represent only the opinions of </span><span><span
class="calibre36"><span class="italic">Car Talk</span></span>
</span><span> fans who took the time to get to the Web site and
take the survey.) Notice that the percentages won’t add up to
100% because the results in the table represent only the top
ten vote-getters.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/mt1601" src="images/00349.jpg"
class="calibre2"/></span></p></blockquote>
</blockquote></div><div class="calibre37"> </div>
<img alt="SB-End" src="images/00012.jpg" class="calibre2"/>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Some surveys
are just for fun, and others are more serious. Be sure to check
the source of any serious survey in which you’re asked to
participate and for which you’re given results. Groups that
have a special interest in the results should either hire an
independent organization to conduct (or at least to review) the
survey, or they should offer copies of the survey questions to
the public. Groups should also disclose in detail how the
survey was designed and conducted, so that the public can make
an informed decision about the credibility of the results.
</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a295" class="calibre6">
<span class="calibre21"><span class="bold"><span>Surveying
what’s hot</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The topics of many surveys are driven
by current events, issues, and areas of interest; after all,
timeliness and relevance to the public are two of the most
attractive qualities of any survey. Here are just a few
examples of some of the subjects being brought to the surface
by today’s surveys, along with some of the results being
reported:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>Does celebrity activism influence the political opinions
of the American public? (Over 90% of the American public says
```

no, according to CBS News.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
What percentage of Americans have dated a co-worker? (A whopping 40% have, according to a career networking Web site.) </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
How many patients surf the Web to find health-related information? (55% do, according to a national medical journal.) </p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">When you read the preceding survey results, do you find yourself thinking about what the results mean to you, rather than first asking yourself whether the results are valid? Some of the preceding survey results are more valid and accurate than others, and you should think about whether to believe the results first, before accepting them without question. Nationally known polling and research organizations such as those mentioned in the previous section are credible sources, as well as journals that are
peer-reviewed (meaning all papers published in the journal have been reviewed by others in the field and passed a certain set of standards). And the U.S. government does a good job with their data collection as well. If you are not familiar with a group conducting a survey and the results are important to you, check out the source.
</p></blockquote>
<blockquote class="calibre5"><p id="a296" class="calibre6">
Impacting lives</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Whereas some surveys are just fun to look at and think about, other surveys can have a direct impact on your life or your workplace. These life-decision surveys need to be closely scrutinized before action is taken or important decisions are made. Surveys at this level can cause politicians to change or create new laws, motivate researchers to work on the latest problems, encourage manufacturers to invent new products or change business policies and practices, and influence people’s behavior and ways of thinking. The following are some examples of survey results that can impact you:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span

class="calibre15">
Children’s healthcare suffers: A survey of 400 pediatricians by the Children’s National Medical Center in Washington, D.C., reported that pediatricians spend, on average, only 8 to 12 minutes with each patient.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Teens drink more: According to the 2009 Partnership Attitude Tracking Study, conducted by the Partnership for a Drug-Free America, the number of teens in grades 9 through 12 that use alcohol has grown by 4% (from 35% in 2008 to 39% in 2009), reversing the downward trend experienced in the ten years prior to the survey.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Always look at how researchers define the terms they’re using to collect their data. In the above example, how did they define “alcohol use”? Does it count if the teenager tried alcohol once? Does it mean they drink alcohol on a consistent basis? Results can be misleading if the range of what or who gets counted is too wide. Find out what questions were actually asked when the data was collected.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Crimes go unreported: The U.S. Bureau of Justice Crime Victimization Survey concludes that only 49.4% of violent crimes were reported to police. The reasons victims gave for not reporting crimes to the police are listed in Table 16-1.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The most frequently given reason for not reporting a violent crime to the police was that the victim considered it to be a personal matter (19.2%). Note that almost 12% of the reasons relate to perception of the reporting

process itself (for example, that it would take too much time or that the police would be bothered, biased, or ineffective).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> By the way, did you notice how large the “Other reasons” category is? This large, unexplained percentage indicates that the survey can be more specific and/or more research can be done regarding why crime victims don’t report crimes. Maybe the victims themselves aren’t even sure.</p></blockquote>
<p id="a297" class="calibre6">Behind the Scenes: The Ins and Outs of Surveys</p>
<blockquote class="calibre9"><p class="calibre14">Surveys and their results are a part of your daily experience, and you use these results to make decisions that affect your life. (Some decisions may even be life changing.) Looking at surveys with a critical eye is important. Before taking action or making decisions based on survey results, you must determine whether those results are credible, reliable, and believable. A good way to begin developing these detective skills is to go behind the scenes and see how surveys are designed, developed, implemented, and analyzed.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The survey process can be broken down into a series of ten steps:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Clarify the purpose of your survey.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Define the target population.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Choose the type and timing of the survey.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Design the introduction with ethics in mind.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="calibre15"> 5. Formulate the questions.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Select the sample.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 7. Carry out the survey.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 8. Follow up, follow up, and follow up.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 9. Organize and analyze the data.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 10. Draw conclusions.</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">Each step presents its own set of special issues and challenges, but each step is critical in terms of producing survey results that are fair and accurate. This sequence of steps helps you design, plan, and implement a survey, but it can also be used to critique someone else's survey, if those results are important to you.</p></blockquote>
<blockquote class="calibre5"><p id="a298" class="calibre6">Planning and designing a survey</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The purpose of a survey is to answer questions about a target population. The target population is the entire group of individuals that you're interested in drawing conclusions about. In most situations, surveying the entire target population (that is, conducting a full-blown census) is impossible because researchers would have to spend too much time or money to do so. Usually, the best you can do is to select a sample of individuals from the target population, survey those individuals, then draw conclusions about the

target population based on the data from that sample.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Sounds easy, right? Wrong. Many
potential problems arise after you realize that you can’t
survey everyone in the entire target population. Then, after a
sample is selected, many researchers aren’t sure what to do to
get the data they need. Unfortunately, many surveys are
conducted without taking the time needed to think through these
issues, resulting in errors, misleading results, and wrong
conclusions. In the following sections, I give specifics for
the first five steps in the survey process.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Clarifying the
purpose of your survey</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This sounds like it should just be
common sense, but in reality, many surveys have been designed
and carried out that never met their purpose, or that met only
some of the objectives, but not all of them. Getting lost in
the questions and forgetting what you’re really trying to find
out is easy to do. In stating the purpose of a survey, be as
specific as possible. Think about the types of conclusions you
would want to make if you were to write a report, and let that
help you determine your goals for the survey.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Lots of researchers can’t see the
forest for the trees. If a restaurant manager wants to
determine and compare satisfaction rates for her customers, she
needs to think ahead about what kinds of comparisons she wants
to make and what information she wants to be able to report on.
Questions that pinpoint when the customers came into the
restaurant (date and time), or even what table they were at,
are relevant. And if she wants to compare satisfaction rates
for, say, adults versus families, she needs to ask how many
people were in the party and how many were children. But if she
simply asks a couple of questions on satisfaction or throws in
every question she can think of, without considering in advance
why she needs the information, she may end up with more
questions than answers.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> The more specific you can be about the
purpose of the survey, the more easily you can design questions
that meet your objectives, and the better off you’ll be when
you need to write your report.</p></blockquote>

<blockquote class="calibre5"><p class="calibre6">Defining the target
population</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose, for example, that you want to
conduct a survey to determine the extent to which people send
and receive personal e-mail in the workplace. You may think
that the target population is e-mail users in the workplace.
However, you want to determine the extent
 to which personal e-mail is used in the
workplace, so you can’t just ask e-mail users, or your results
would be biased against those who don’t use e-mail in the
workplace. But should you also include those who don’t even
have access to a computer during their workday? (See how fast
surveys can get tricky?)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The target population that probably
makes the most sense here is all the people who use Internet-
connected computers in the workplace. Everyone in this group at
least has access to e-mail, though only some of those with
access to e-mail in the workplace actually use it, and of those
who use it, only some use it for personal e-mail. (And that’s
what you want to find out — how much they use e-mail for that
purpose.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> You need to be
clear in your definition of the target population. Your
definition is what helps you select the proper sample, and it
also guides you in your conclusions, so that you don’t
overgeneralize your results. If the researcher didn’t clearly
define the target population, this can be a sign of other
problems with the survey.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Choosing the type and
timing of the survey</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The next step in designing your survey
is to choose what type of survey is most appropriate for the
situation at hand. Surveys can be done over the phone, through
the mail, with door-to-door interviews, or over the Internet.
However, not every type of survey is appropriate for every
situation. For example, suppose you want to determine some of
the factors that relate to illiteracy in the United States. You
wouldn’t want to send a survey through the mail, because people
who can’t read won’t be able to take the survey. In that case,
a telephone interview is more appropriate.</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Choose the type of survey that's most appropriate for the target population, in terms of getting the most truthful and informative data possible. You also have to keep in mind the budget you have to work with; door-to-door interviews are more expensive than phone surveys, for example. When examining the results of a survey, be sure to look at whether the type of survey used is most appropriate for the situation, keeping budget considerations in mind.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Next you need to decide when to conduct the survey. In life, timing is everything, and the same goes for surveys. Current events shape people's opinions all the time, and although some pollsters try to determine how people feel about those events, others take advantage of events, especially negative ones, and use them as political platforms or as fodder for headlines and controversy. For example, surveys about gun control often come up after a shooting takes place. Also take note of other events that were going on at the time of the survey; for example, people may not want to answer their phones during the Super Bowl, on election night, during the Olympics, or around holidays. Improper timing can lead to bias.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In addition to the date, the time of day is also important. If you conduct a telephone survey to get people's opinions on stress in the workplace and you call them at home between the hours of 9 a.m. and 5 p.m., you're going to have bias in your results; those are the hours when the majority of people are at work (busy being stressed out!).
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Designing the introduction with ethics in mind</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">While this rule doesn't apply to little polls that you see on the Internet and in magazines, serious surveys need to provide information pertaining to important ethical issues. First, they should include what pollsters call a cover letter — an introduction that explains the purpose of the survey, what will be done with the data, whether the information the respondent supplies will be confidential or anonymous (see the sidebar

“Anonymity versus confidentiality” later in this chapter), and that the person’s participation is appreciated but not required. The cover letter should also provide the researcher’s contact information for respondents to use if they have questions or concerns.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If the survey is done by any institution or group that is federally regulated, such as a university, research institute, or a hospital, the survey has to be approved in advance by a committee designated to review, regulate, and/or monitor the research to make sure it’s ethical, scientific, and follows regulations. Such committees are called institutional review boards (IRBs), independent ethics committees (IECs), or ethical review boards (ERBs). The survey cover letter should explain who has approved the research. If you don’t see such information, ask.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Formulating the questions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After the purpose, type, timing, and ethical issues of the survey have been addressed, the next step is to formulate the questions. The way that the questions are asked can make a huge difference in the quality of the data that will be collected. One of the single most common sources of bias in surveys is the wording of the questions. Research shows that the wording of the questions can directly affect the outcome of a survey. Leading questions, also called misleading questions, are designed to favor a certain response over another. They can greatly affect how people answer the questions, and their responses may not accurately reflect how they truly feel about an issue.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, here are two ways that I’ve seen survey questions worded about a proposed school bond issue (both of which are leading questions):</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre39">Don’t you agree that a tiny percentage increase in sales tax is a worthwhile investment in improving the quality of the education of our children?
</p></blockquote><div class="calibre19"> </div>

<blockquote class="calibre5"><p class="calibre39">Don’t you think we should stop increasing the
burden on the taxpayers and stop asking for yet another sales
tax hike to fund the wasteful school system?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">From the wording of each of these
leading questions, you can easily see how the pollsters want
you to respond. To make matters worse, neither question tells
you exactly how much of a tax increase is being proposed, which
is also misleading.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The best way to
word a question is in a neutral way, giving the reader the
necessary information required to make an informed decision.
For example, the tax issue question is better worded this way:
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39">The school district is proposing a 0.01%
increase in sales tax to provide funds for a new high school to
be built in the district. What’s your opinion on the proposed
sales tax? (Possible responses: strongly in favor, in favor,
neutral, against, strongly against.)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">If the purpose of a survey is purely to
collect information rather than influence or persuade the
respondent, the questions should be worded in a neutral and
informative way in order to minimize bias. The best way to
assess the neutrality of a question is to ask yourself whether
you can tell how the person wants you to respond. If the answer
is yes, that question is a leading question and can give
misleading results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> If the results of a survey are
important to you, ask the researcher for a copy of the
questions used on the survey so you can assess the quality of
the questions. When conducting your own survey, have others
check the questions to verify that the wording is neutral and
informative.</p></blockquote>
<blockquote class="calibre5"><p id="a299" class="calibre6">
Selecting the
sample</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">After the survey has been designed, the next step is to select people to participate in the survey. Because typically you don’t have time or money to conduct a census (a survey of the entire target population), you need to select a subset of the population, called a sample.
 How this sample is selected can make all the difference in terms of the accuracy and the quality of the results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Three criteria are important in selecting a good sample, as you find out in the following sections.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">A good sample represents the target population</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">To represent the target population, the sample must be selected from the target population, the whole target population, and nothing but the target population. Suppose you want to find out how many hours of TV Americans watch in a day, on average. Asking students in a dorm at a local university to record their TV viewing habits isn’t going to cut it. Students represent only a portion of the target population.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Unfortunately, many people who conduct surveys don’t take the time or spend the money to select a representative sample of people to participate in the study, and they end up with biased survey results. When presented with survey results, find out how the sample was selected before examining the results of the survey and see how well they match the target population.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">A good sample is selected randomly</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
random sample is one in which every possible sample (of the same size) has an equal chance of being selected from the target population. The easiest example to visualize here is that of a hat (or bucket) containing individual slips of paper, each with the name of a person written on it; if the slips are thoroughly mixed before each slip of paper is drawn out, the result will be a random

sample of the target population (in this case, the population of people whose names are in the hat). A random sample eliminates bias in the sampling process.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Reputable polling organizations, such as the Gallup Organization, use a random digit-dialing procedure to telephone the members of their sample. Of course, this excludes people without telephones, but because most American households today have at least one telephone, the bias involved in excluding people without telephones is relatively small.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Beware of surveys that have a large but not randomly selected sample. Internet surveys are the biggest culprit. Someone can say that 50,000 people logged on to a Web site to answer a survey, and that means the person posting this site has gotten a lot of data. But the information is biased; research shows that people who respond to surveys tend to have stronger opinions than those that don’t respond. And if they didn’t even select the participants randomly to start with, imagine how strong (and biased) the respondents’ opinions would be. If the survey designer sampled fewer people but did so randomly, the survey results would be more accurate.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">A good sample is large enough for the results to be accurate
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you have a large sample size, and if the sample is representative of the target population and is selected at random, you can count on that information being pretty accurate. How accurate depends on the sample size, but the bigger the sample size, the more accurate the information will be (as long as that information is good information). The accuracy of most survey questions is measured in terms of a percentage. This percentage is called the margin of error, and it represents how much the researcher expects the results to vary if she were to repeat the survey many times using different samples of the same size. Read more about this in Chapter 12.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"

class="calibre2"/> A quick and dirty formula to estimate the minimum amount of accuracy of a survey involving categorical data (such as gender or political affiliation) is to take 1 divided by the square root of the sample size. For example, a survey of 1,000 (randomly selected) people is accurate to within ±0.032, or 3.2 percentage points. (See Chapter 12 for the exact formula for calculating the accuracy of a survey.) In cases where not everyone responded, you should replace the sample size with the number of respondents (see the “Following up, following up, and following up” section later in this chapter). Remember, these quick-and-dirty estimates of accuracy are conservative; using the precise formulas gives you accuracy rates that are often much better than these. (See Chapter 13 for details.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> With large populations (in the thousands, say) it’s the size of the sample, not the size of the population, that matters. For example, if you randomly sample 1,000 individuals from a large population, your accuracy level is estimated to be within 3.2 percentage points, no matter whether you sample from a small town of 10,000 people, a state of 1,000,000 people, or all of the United States. That fact was one of the things that blew my mind about statistics when I first learned it, and it still does today — it’s amazing how accurate you can get with such a comparatively small sample size.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> However, with small populations, you have to apply different methods to determine accuracy and sample size. A sample of 10 out of a population of 100 takes a much larger piece out of the pie than a sample of 10 out of 10,000 does, for example. More advanced methods involving a finite population correction handle issues that come up with small populations.</p>
</blockquote>
<blockquote class="calibre5"><p id="a300" class="calibre6">
Carrying out a survey</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The survey has been designed, and the participants have been selected. Now you have to go about the process of carrying out the survey, which is another important step — one where lots of mistakes and biases can occur.
</p></blockquote>

<blockquote class="calibre5"><p class="calibre6">Collecting the data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">During the survey itself, the participants can have problems understanding the questions, they may give answers that aren’t among the choices (in the case of a multiple choice question), or they may decide to give answers that are inaccurate or blatantly false; the latter is called response bias. (As an example of response bias, think about the difficulties involved in getting people to tell the truth about whether they’ve cheated on their income-tax forms.) </p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Some of the potential problems with the data-collection process can be minimized or avoided with careful training of the personnel who carry out the survey. With proper training, any issues that arise during the survey are resolved in a consistent and clear way, and fewer errors are made in recording the data. Problems with confusing questions or incomplete choices for answers can be resolved by conducting a pilot study on a few participants prior to the actual survey and then, based on their feedback, fixing any problems with the questions.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Personnel can also be trained to create an environment in which each respondent feels safe enough to tell the truth; ensuring that privacy will be protected also helps encourage more people to respond. To minimize interviewer bias, the interviewers must follow a script that’s the same for each subject.</p></blockquote>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">Anonymity versus confidentiality</p></blockquote><div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">If you were to conduct a survey to determine the extent of personal e-mail use at work, the response rate would probably be an issue, because many people are reluctant to discuss their use of personal e-mail in the workplace, or at least to do so truthfully. You could try to encourage people to respond by letting them know that their privacy would be protected during and after the survey.
</p></blockquote>

<blockquote class="calibre5"><p class="calibre34">When you report the results of a
survey, you generally don’t tie the information collected to
the names of the respondents, because doing so would violate
the privacy of the respondents. You’ve probably heard the terms
anonymous and
confidential before,
but what you may not realize is that these two words are
completely different in terms of privacy issues. Keeping
results confidential means
that I could tie your information to your name in my report,
but I promise that I won’t do that. Keeping results
anonymous means that
I have no way of tying your information to your name in my
report, even if I wanted to.</p></blockquote>
<blockquote class="calibre5"><p class="calibre34">If you’re asked to participate in a
survey, be sure you’re clear about what the researchers plan to
do with your responses and whether or not your name can be tied
to the survey. (Good surveys always make this issue very clear
for you.) Then make a decision as to whether you still want to
participate.</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Beware of
conflicts of interest that come up with misleading surveys. For
example, if you are being asked about the quality of your
service by the person who gave you the service, you may not
want to respond truthfully. Or, if your physical therapist
gives you an “anonymous” feedback survey on your last day and
tells you to give it to her when you’re done, the survey may
have issues of bias.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Following up,
following up, and following up</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Anyone who has ever thrown away a
survey or refused to “answer a few questions” over the phone
knows that getting people to participate in a survey isn’t
easy. If the researcher wants to minimize bias, the best way to
handle it is to get as many folks to respond as possible by

following up, one, two, or even three times. Offer dollar bills, coupons, self-addressed stamped return envelopes, chances to win prizes, and so on. Every little bit helps. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If only those folks who feel very strongly respond to a survey, that means that only their opinions will count, because the other people who didn't really care about the issue didn't respond, and their "I don't care" vote didn't get counted. Or maybe they did care, but they just didn't take the time to tell anyone. Either way, their vote doesn't count.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose 1,000 people are given a survey about whether the park rules should be changed to allow dogs without leashes. Most likely, the respondents would be those who strongly agree or disagree with the proposed rules. Suppose only 200 people responded — 100 against and 100 for the issue. That would mean that 800 opinions weren't counted. Suppose none of those 800 people really cared about the issue either way. If you could count their opinions, the results would be 800 ÷ 1,000 = 80% "no opinion," 100 ÷ 1,000 = 10% in favor of the new rules, and 100 ÷ 1,000 = 10% against the new rules. But without the votes of the 800 non-respondents, the researchers would report, "Of the people who responded, 50% were in favor of the new rules and 50% were against them." This gives the impression of a very different (and a very biased) result from the one you would've gotten if all 1,000 people had responded.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The response rate of a survey is a ratio found by taking the number of respondents divided by the number of people who were originally asked to participate. You of course want to have the highest response rate you can get with your survey; but how high is high enough to be minimizing bias? The purest of the pure statisticians feel that a good response rate is anything over 70%, but I think we need to be a little more realistic. Today's fast-paced society is saturated with surveys; many if not most response rates fall far short of 70%. In fact, response rates for today's surveys are more likely to be in the 20% to 30% range, unless the survey is conducted by a professional polling organization such as Gallup or you are being offered a new car just for filling one out. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"

src="images/00007.jpg" class="calibre2"/> Look for the response rate when examining survey results. If the response rate is too low (much less than 50%) the results are likely to be biased and should be taken with a grain of salt, or even ignored.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Don’t be fooled by a survey that claims to have a large number of respondents but actually has a low response rate; in this case, many people may have responded, but many more were asked and didn’t respond.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Note that statistical formulas at this level (including the formulas in this book) assume that your sample size is equal to the number of respondents, so statisticians want you to know how important it is to follow up with people and not end up with biased data due to non-response. However, in reality, statisticians know that you can’t always get everyone to respond, no matter how hard you try; indeed, even the U.S. Census doesn’t have a 100% response rate. One way statisticians combat the non-response problem after the data have been collected is to break down the data to see how well it matches the target population. If it’s a fairly good match, they can rest easier on the bias issue.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So which number do you put in for n in all those statistical formulas you use so often (such as the sample mean in Chapter 5)? You can’t use the intended sample size (the number of people contacted). You have to use the final sample size (the number of people who responded). In the media you most often see only the number of respondents reported, but you also need the response rate (or the total number of respondents) to be able to critically evaluate the results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Regarding the quality of results, selecting a smaller initial sample size and following them up more aggressively is a much better approach than selecting a larger group of potential respondents and having a low response rate, because of the bias introduced by non-response.</p></blockquote>
<blockquote class="calibre5"><p id="a301" class="calibre6">
Interpreting results and finding problems</p>
</blockquote>

<blockquote class="calibre9"><p class="calibre14">The purpose of a survey is to gain
information about your target population; this information can
include opinions, demographic information, or lifestyles and
behaviors. If the survey has been designed and conducted in a
fair and accurate manner with the goals of the survey in mind,
the data should provide good information as to what’s happening
with the target population (within the stated margin of error;
see Chapter 12). The next steps are to organize the data to get
a clear picture of what’s happening; to analyze the data to
look for links, differences, or other relationships of
interest; and then to draw conclusions based on the results.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Organizing and
analyzing</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After a survey has been completed, the
next step is to organize and analyze the data (in other words,
crunch some numbers and make some graphs). Many different types
of data displays and summary statistics can be created and
calculated from survey data, depending on the type of
information that was collected. (Numerical data, such as
income, have different characteristics and are usually
presented differently than categorical data, such as gender.)
For more information on how data can be organized and
summarized, see Chapters 5 through 7. Depending on the research
question, different types of analyses can be performed on the
data, including coming up with population estimates, testing a
hypothesis about the population, or looking for relationships,
to name a few. See Chapters 13, 14, 15, 18, and 19 for more on
each of these analyses, respectively.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Watch for
misleading graphs and statistics. Not all survey data are
organized and analyzed fairly and correctly. See Chapter 3 for
more about how statistics can go wrong.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">Drawing
conclusions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The conclusions are the best part of
any survey — they’re why the researchers do all of the work in
the first place. If the survey was designed and carried out
properly — the sample was selected carefully and the data were

organized and summarized correctly — the results should fairly and accurately represent the reality of the target population. But, of course, not all surveys are done right. And even if a survey is done correctly, researchers can misinterpret or overinterpret results so that they say more than they really should.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You know the saying “Seeing is believing”? Some researchers are guilty of the converse, which is “Believing is seeing.” In other words, they claim to see what they want to believe about the results. All the more reason for you to know where the line is drawn between reasonable conclusions and misleading results, and to realize when others have crossed that line.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are some common errors made in drawing conclusions from surveys:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18">
Making projections to a larger population than the study actually represents</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Claiming a difference exists between two groups when a difference isn’t really there (see Chapter 15)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Saying, “these results aren’t scientific, but . . . ,” and then going on to present the results as if they are scientific</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> To avoid common errors made when drawing conclusions, do the following:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Check whether the sample was selected properly and that the conclusions don’t go beyond the population

presented by that sample.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Look for any disclaimers about the survey
before

reading the results.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> That way, if the results aren't based
on a scientific survey (an accurate and unbiased survey),
you'll be less likely to be influenced by the results you're
reading. You can judge for yourself whether the survey results
are credible.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Be on the lookout for statistically incorrect
conclusions.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> If someone reports a difference
between two groups in terms of survey results, be sure that the
difference is larger than the reported margin of error. If the
difference is within the margin of error, you should expect the
sample results to vary by that much just by chance, and the so-
called "difference" can't really be generalized to the entire
population. (See Chapter 14 for more on this.)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Know the
limitations of any survey and be wary of any information coming
from surveys in which those limitations aren't respected. A bad
survey is cheap and easy to do, but you get what you pay for.
But don't let big expensive surveys fool you either — they can
be riddled with bias as well! Before looking at the results of
any survey, investigate how it was designed and conducted,
using the criteria and tips in this chapter, so you can judge
the quality of the results and express yourself confidently and
correctly about what is wrong.</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a302"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

```
<a href="#a293" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a292" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a295" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a294" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a297" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a296" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a299" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a298" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a303" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a302" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a301" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a300" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a321">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 17</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a304" class="calibre10"><span class="calibre11"><span
class="bold"><span>Experiments: Medical Breakthroughs or
Misleading Results?</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Distinguishing experiments from observational
studies</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Dissecting the criteria for a good experiment</span>
</span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Watching for misleading results</span></span></p>
</blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>M</span><span>edical breakthroughs seem
to come and go quickly. One day you hear about a promising new
treatment for a disease, only to find out later that the drug
didn’t live up to expectations in the last stage of testing.
Pharmaceutical companies bombard TV viewers with commercials
for pills, sending millions of people to their doctors
clamoring for the latest and greatest cures for their ills,
sometimes without even knowing what the drugs are for. Anyone
can search the Internet for details about any type of ailment,
```

disease, or symptom and come up with tons of information and
advice. But how much can you really believe? And how do you
decide which options are best for you if you get sick, need
surgery, or have an emergency?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, you go behind the
scenes of experiments, the driving force of medical studies and
other investigations in which comparisons are made —
comparisons that test, for example, which building materials
are best, which soft drink teens prefer, and so on. You find
out the difference between experiments and observational
studies and discover what experiments can do for you, how
they’re supposed to be done, how they can go wrong, and how you
can spot misleading results. With so many headlines, sound
bites, and pieces of “expert advice” coming at you from all
directions, you need to use all your critical thinking skills
to evaluate the sometimes-conflicting information you’re
presented with on a regular basis.</p>
</blockquote>
<p id="a305" class="calibre6">Boiling Down the Basics of Studies
</p>
<blockquote class="calibre9"><p class="calibre14">Although many different types of
studies exist, you can basically boil them down to two types:
experiments and observational studies. This section examines
what exactly makes experiments different from other studies.
But before I dive in to the details, I need to lay some jargon
on you.</p></blockquote>
<blockquote class="calibre5"><p id="a306" class="calibre6">
Looking at the
lingo of studies</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To understand studies, you need to find
out what their commonly used terms mean:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Subjects:
Individuals participating in the
study.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Observational
study: A study in which the

researcher merely observes the subjects and records the information. No intervention takes place, no changes are introduced, and no restrictions or controls are imposed.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Experiment:
 This study doesn't simply observe subjects in their natural state; it deliberately applies treatments to them in a controlled situation and studies their effects on the outcome.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Response:
 The response is the variable whose outcome is the million dollar question; it's the variable whose outcome is of interest. For example, if researchers want to know what happens to your blood pressure when you take a large amount of Ibuprofen each day, the response variable is blood pressure.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Factor:
 A factor is the variable whose effect on the response is being studied. For example, if you want to know whether a particular drug increases blood pressure, your factor is the amount of the drug taken. If you want to know which weight loss program is most effective, your factor would be the type of weight loss program used.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> You can have more than one factor in a study; however, in this book I stick with discussing one factor only. For the analysis of two-factor studies, including the use of Analysis of Variance (ANOVA) and multiple comparisons to compare treatment combinations, you can check out my book
Statistics II For Dummies,
 also published by Wiley.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">

Level:
 A level is one possible outcome of a factor. Each factor has a certain number of levels. In the weight loss example, the factor is the type of weight loss program and the levels would be the specific programs studied (for example Weight Watchers, South Beach, or the famous Potato Diet). Levels need not be ascending in any way; however, in a study like the drug example, the levels would be the various dosages taken each day, in increasing amounts.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Treatment:
 A treatment is a combination of the levels of the factors being studied. If you only have one factor, the levels and the treatments are the same thing. If you have more than one factor, each combination of levels of the factors is called a treatment.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> For example, if you want to study the effects of the type of weight loss program and the amount of water consumed daily, you have two factors: 1) the type of program, with 3 levels (Weight Watchers, South Beach, Potato Diet); and 2) the amount of water consumed, with, say, 3 levels (24, 48, and 64 ounces per day). In this case, there are 3
∗ 3 = 9 treatments: Weight Watchers and 24 ounces of water per day; Weight Watchers and 48 ounces of water per day, . . . all the way up to the famous Potato Diet and 64 ounces of water per day. Each subject is assigned to one treatment. (With my luck, I’d get that last treatment.)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">

Cause and effect: A factor and a response have a cause-and-effect relationship if a change in the factor results in a direct change in the response (for example, increasing calorie intake causes weight gain).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In the following sections, you see the differences between observational studies and experiments, when each is used, and what their strengths and/or weaknesses may

be.</p></blockquote>
<blockquote class="calibre5"><p id="a307" class="calibre6">
Observing
observational studies</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Just like with tools, you want to find
the right type of study for the right job. In certain
situations, observational studies are the optimal way to go.
The most common observational studies are polls
 and surveys (see Chapter
16). When the goal is simply to find out what people think and
to collect some demographic information (such as gender, age,
income, and so on), surveys and polls can’t be beat, as long as
they’re designed and conducted correctly.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In other situations, especially those
looking for cause-and-effect relationships, observational
studies aren’t optimal. For example, suppose you took a couple
of vitamin C pills last week; is that what helped you avoid
getting that cold that’s going around the office? Maybe the
extra sleep you got recently or the extra hand-washing you’ve
been doing helped you ward off the cold. Or maybe you just got
lucky this time. With so many variables in the mix, how can you
tell which one had an influence on the outcome of your not
getting a cold? An experiment that takes these other variables
into account is the way to go.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> When looking at the results of any
study, first determine what the purpose of the study was and
whether the type of study fits the purpose. For example, if an
observational study was done instead of an experiment to
establish a cause-and-effect relationship, any conclusions that
are drawn should be carefully scrutinized.</p>
</blockquote>
<blockquote class="calibre5"><p id="a308" class="calibre6">
Examining
experiments</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The object of an experiment is to see
if the response changes as a result of the factor you are
studying; that is, you are looking for cause and effect. For
example, does taking Ibuprofen cause blood pressure to
increase? If so, by how much? But because results will vary
with any experiment, you want to know that your results have a

high chance of being repeatable if you found something interesting happening. That is, you want to know that your results were unlikely to be due to chance; statisticians call such results statistically significant
. That’s the objective of any study, observational, or experimental.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A good experiment is conducted by creating a very controlled environment — so controlled that the researcher can pinpoint whether a certain factor or combination of factors causes a change in the response variable, and if so, the extent to which that factor (or combination of factors) influences the response. For example, to gain government approval for a proposed blood pressure drug, pharmaceutical researchers set up experiments to determine whether that drug helps lower blood pressure, what dosage level is most appropriate for each different population of patients, what side effects (if any) occur, and to what extent those side effects occur in each population.</p></blockquote>
<p id="a309" class="calibre6">Designing a Good Experiment
</p>
<blockquote class="calibre9"><p class="calibre14">How an experiment is designed can mean the difference between good results and garbage. Because most researchers are going to write the most glowing press releases that they can about their experiments, you have to be able to sort through the hype to determine whether to believe the results you’re being told. To decide whether an experiment is credible, check to see if it meets all
 the following criteria for a good experiment. A good experiment:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Makes comparisons</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Includes a large enough sample size so that the results are accurate</p></blockquote><div

```
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Chooses
subjects that most accurately represent the target
population</span></span></span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Assigns
subjects randomly to the treatment group(s) and the control
group</span></span></span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Controls for
possible confounding variables</span></span></span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Is
ethical</span></span></span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Collects good
data</span></span></span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Applies the
proper data analysis</span></span></span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Makes
appropriate conclusions</span></span></span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>In this section, each of these criteria
is explained and illustrated with examples.</span></span></p>
```

</blockquote>
<blockquote class="calibre5"><p id="a310" class="calibre6">
Designing the experiment to make comparisons</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Every experiment has to make bonafide comparisons to be credible. This seems to go without saying, but researchers often are so gung-ho to prove their results that they forget (or just don’t bother) to show that their factor, and not some other factor(s), including random chance, was the actual cause for any differences found in the response.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose a researcher is convinced that taking vitamin C prevents colds, and she assigns subjects to take one vitamin C pill per day and follows them for 6 months. Suppose the subjects get very few colds during that time. Can she attribute these results to the vitamin C and nothing else? No; there’s no way of knowing whether the subjects would have been just as healthy without the vitamin C, due to some other factor(s), or just by chance. There’s nothing to compare the results to.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To tease out the real effect (if any) that your factor has on the response, you need a baseline to compare the results to. This baseline is called the control. Different methods exist for creating a control in an experiment; depending on the situation, one method typically rises to the top as being the most appropriate. Three common methods for including control are to administer: 1) a fake treatment; 2) a standard treatment; or 3) no treatment. The following sections describe each method.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When examining the results of an experiment, make sure the researchers established a baseline by creating a control group. Without a control group, you have nothing to compare the results to, and you never know whether the treatment being applied was the real cause of any differences found in the response.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Fake treatments — the placebo effect</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">A fake treatment (also called a
placebo) is not
distinguishable from a “real” treatment by the subject. For
example, when drugs are administered, a subject assigned to the
placebo will receive a fake pill that looks and tastes exactly
like a real pill; it’s just filled with an inert substance like
sugar instead of the actual drug. A placebo establishes a
baseline measure for what responses would have taken place
anyway, in lieu of any treatment (this would have helped the
vitamin C study mentioned under “Designing the experiment to
make comparisons”). But a fake treatment also takes into
account what researchers call the placebo effect,
 a response that people have (or think
they’re having) because they know they’re getting some type of
“treatment” (even if that treatment is a fake treatment, such
as sugar pills).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Pharmaceutical companies are required
to account for the placebo effect when examining both the
positive and negative effects of a drug. When you see an ad for
a drug in a magazine, you see the positive results of the drug
standing out in big, bright, happy, colorful visuals. Then look
at the back of the page and you see it’s entirely filled in
black with words written in 3-point font. Embedded somewhere on
that page, you can find one or more tiny tables that show the
number and nature of side effects reported by each
treatment
group (subjects who received an
actual treatment) as well as the control group
 (subjects who were administered a
placebo).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If the control
group is on a placebo, you may expect the subjects not to
report any side effects, but you would be wrong. If you are
taking a pill, you know it could be an actual drug, and you are
being asked whether or not you’re experiencing side effects,
you might be surprised at what your response would be.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you don’t take the placebo effect
into account, you have to believe that any side effects (or
positive results) reported are actually due to the drug. This

gives an artificially high number of reported side effects because at least some of those reports are likely due to the placebo effect and not to the drug itself. If you have a control group to compare with, you can subtract the percentage of people in the control group who reported the side effects from the percentage of people in the treatment group that reported the side effects, and examine the magnitude of the numbers that remain. You’re in essence looking at the net number of reported side effects due to the drug, rather than the gross number of side effects, some of which are due to the placebo effect.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The placebo effect has been shown to be real. If you want to be fair about examining the reported side effects (or positive reactions) of a treatment, you have to also take into account the side effects (or positive reactions) that the control group reports — those reactions that are due to the placebo effect only.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Standard treatments</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> In some situations, such as when the subjects have very serious diseases, offering a fake treatment as an option may be unethical. One famous example of a breech in ethics occurred in 1997. The U.S. government was harshly criticized for financing an HIV study that examined new dosage levels of AZT, a drug known at that time to cut the risk of HIV transmission from pregnant mothers to their babies by two-thirds. This particular study, in which 12,000 pregnant women with HIV in Africa, Thailand, and the Dominican Republic participated, had a deadly design. Researchers gave half of the women various dosages of AZT, but the other half of the women received sugar pills. Of course, had the U.S. government realized that a placebo was being given to half of the subjects, it wouldn’t have supported the HIV study. It’s not ethical to give a fake treatment to anyone with a deadly disease for which a standard treatment is available (in this case, the standard dosage of AZT).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When ethical reasons bar the use of fake treatments, the new treatment is compared to at least one existing or standard treatment that is known to be an effective treatment. After researchers have enough data to see that one

of the treatments is working better than the other, they
generally stop the experiment and put everyone on the better
treatment; again, for ethical reasons.</p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6">No treatment
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">“No treatment” means the researcher
can’t help but tell which group the subject is in, due to the
nature of the experiment. The subjects in this case aren’t
receiving any type of intervention in terms of their behavior,
but they still serve as a control, establishing a baseline of
data to compare their results with those in the treatment
group(s). For example, if you want to determine whether speed
walking around the block ten times a day lowers a person’s
resting heart rate after six months, the subjects in your
control group know they aren’t going to be speed walking —
obviously you can’t do fake speed walking (although faking
exercising and still reaping the benefits would be great,
wouldn’t it?).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> In situations where the control group
receives no treatment, you still make sure the groups of
subjects (speed walkers versus non–speed walkers) are similar
in as many ways as possible, and that the other criteria for a
good experiment are being met. (See “Designing a Good
Experiment” for the list of criteria.)</p>
</blockquote>
<blockquote class="calibre5"><p id="a311" class="calibre6">
Selecting the
sample size</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The size of a (good) sample greatly
affects the accuracy of the results. The larger the sample
size, the more accurate the results, and the more powerful the
statistical tests (in terms of being able to detect real
results when they exist). In this section, I hit the
highlights; Chapter 14 has the details.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> The word sample
 is often attributed to surveys where a random
sample is selected from the target population (see Chapter 16).
However, in the setting of experiments, a sample means the

group of subjects who have volunteered to participate.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Limiting small
samples to small conclusions</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You may be surprised at the number of
research headlines that have been made regarding large
populations that were based on very small samples. Such
headlines can be of concern to statisticians, who know that
detecting true statistically significant results in a large
population using a small sample is difficult because small data
sets have more variability from sample to sample (see Chapter
12). When sample sizes are small and big conclusions have been
made by the researcher, either the researchers didn’t use the
right hypothesis test to analyze their data (for example, using
the Z-distribution rather
than the t-distribution; see
Chapter 10) or the difference was so large that it would be
very difficult to miss. The latter isn’t always the case,
however.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Be wary of
research conclusions that find significant results based on
small sample sizes (especially for experiments involving many
treatments but only a few subjects assigned to each treatment).
Statisticians want to see at least five subjects per treatment,
but (much) more is (much) better. You do need to be aware of
some of the limitations of experiments such as cost, time, as
well as ethical issues, and realize that the number of subjects
for experiments is often smaller than the number of
participants in a survey.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If the results are important to you,
ask for a copy of the research report and look to see what type
of analysis was done on the data. Also look at the sample of
subjects to see whether this sample truly represents the
population about which the researchers are drawing conclusions.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Defining sample
size</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When asking questions about

sample size, be specific about what you mean by the term. For example, you can ask how many subjects were selected to participate and also ask for the number who actually completed the experiment; these two numbers can be very different. Make sure the researchers can explain any situations in which the research subjects decided to drop out or were unable (for some reason) to finish the experiment. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, an article in The
New York Times titled “Marijuana Is Called an Effective Relief in Cancer Therapy” says in the opening paragraph that marijuana is “far more effective” than any other drug in relieving the side effects of chemotherapy. When you get into the details, you find out that the results are based on only 29 patients (15 on the treatment, 14 on a placebo). Then you find out that only 12 of the 15 patients in the treatment group actually completed the study. What happened to the other three subjects?</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Sometimes researchers draw their conclusions based on only those subjects who completed the study. This can be misleading, because the data don’t include information about those who dropped out (and why), which may be leading to biased data. For a discussion of the sample size you need to achieve a certain level of accuracy, see Chapter 13.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Accuracy isn’t the only issue in terms of having “good” data. You still need to worry about eliminating bias by selecting a random sample (see Chapter 16 for more on how random samples are taken). </p></blockquote>
<blockquote class="calibre5"><p id="a312" class="calibre6"> Choosing the subjects</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The first step in carrying out an experiment is selecting the subjects (participants). Although researchers would like their subjects to be selected randomly from their respective populations, in most cases, this just isn’t appropriate. For example, suppose a group of eye

researchers wants to test out a new laser surgery on nearsighted people. They need a random sample of subjects, so they randomly select various eye doctors from across the country and randomly select nearsighted patients from these doctors’ files. They call up each person selected and say, “We’re experimenting with a new laser surgery technique for nearsightedness, and you’ve been selected at random to participate in our study. When can you come in for the surgery?” Something tells me that this approach wouldn’t go over very well with many people receiving the call (although some would probably jump at the chance, especially if they didn’t have to pay for the procedure).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> The point is that getting a truly random sample of people to participate in an experiment is generally more difficult than getting a random sample of folks to participate in a survey. However, statisticians can build techniques into the design of an experiment to help minimize the potential bias that can occur.
</p></blockquote>
<blockquote class="calibre5"><p id="a313" class="calibre6">
Making random assignments</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One way to minimize bias in an experiment is to introduce some randomness. After the sample has been decided on, the subjects are randomly divided into treatment and control groups. The treatment groups receive the various treatments being studied, and the control group receives the current (or standard) treatment, no treatment, or a placebo. (See the section “Designing the experiment to make comparisons” earlier in this chapter.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Making random assignments of subjects to treatments is an extremely critical step toward minimizing bias in an experiment. Suppose a researcher wants to determine the effects of exercise on heart rate. The subjects in his treatment group run 5 miles and have their heart rates measured before and after the run. The subjects in his control group sit on the couch the whole time and watch reruns of old TV shows. Which group would you rather be in? Some health nuts out there would no doubt volunteer for the treatment group. If you’re not crazy about the idea of running five miles, you may opt for the easy way out and volunteer to be a couch potato. (Or maybe you hate to watch old reruns

so much that you’d run five miles to avoid that.)</p></blockquote>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">Finding volunteers</p></blockquote><div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">To find subjects for their experiments, researchers often advertise for volunteers and offer them incentives such as money, free treatments, or follow-up care for their participation. Medical research on humans is complicated and difficult, but it’s necessary in order to really know whether a treatment works, how well it works, what the dosage should be, and what the side effects are. In order to prescribe the right treatments in the right amounts in real-life situations, doctors and patients depend on these studies being representative of the general population. In order to recruit such representative subjects, researchers have to do a broad advertisement campaign and select enough participants with enough different characteristics to represent a cross section of the populations of folks who will be prescribed these treatments in the future.</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<blockquote class="calibre9"><p class="calibre14">What impact would this selective volunteering have on the results of the study? If only the health nuts (who probably already have excellent heart rates) volunteer to be in the treatment group, the researcher will be looking only at the effect of the treatment (running five miles) on very healthy and active people. He won’t see the effect that running five miles has on the heart rates of couch potatoes. This non-random assignment of subjects to the treatment and control groups could have a huge impact on the conclusions he draws from this study.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> To avoid major bias in the results of an experiment, subjects must be randomly assigned to treatments by a third party and not be allowed to choose which group they will be in. The goal of random assignment is to create homogenous groups; any unusual characteristics or biases have an equal chance of appearing in any of the groups. Keep this in mind when you evaluate the

results of an experiment.</p></blockquote>
<blockquote class="calibre5"><p id="a314" class="calibre6">
Controlling
for confounding variables</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you’re participating in a
research study that looks at factors influencing whether you
catch a cold. If a researcher records only whether you got a
cold after a certain period of time and asks questions about
your behavior (how many times per day you washed your hands,
how many hours of sleep you get each night, and so on), the
researcher is conducting an observational study. The problem
with this type of observational study is that without
controlling for other factors that may have had an influence
and without regulating which action you were taking when, the
researcher won’t be able to single out exactly which of your
actions (if any) actually impacted the outcome.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The biggest
limitation of observational studies is that they can’t really
show true cause-and-effect relationships, due to what
statisticians call confounding variables. A confounding
variable is a variable or factor
that was not controlled for in the study but can have an
influence on the results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, one news headline boasted,
“Study links older mothers, long life.” The opening paragraph
said that women who have a first baby after age 40 have a much
better chance of living to be 100, compared to women who have a
first baby at an earlier age. When you get into the details of
the study (done in 1996) you find out, first of all, that it
was based on 78 women in suburban Boston who were born in 1896
and had lived to be at least 100, compared to 54 women who were
also born in 1896 but died in 1969 (the earliest year the
researchers could get computerized death records). This so-
called “control group” lived to be exactly 73, no more and no
less. Of the women who lived to be at least 100 years of age,
19% had given birth after age 40, whereas only 5.5% of the
women who died at age 73 had given birth after age 40.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">I have a real problem with these
conclusions. What about the fact that the “control group” was
based only on mothers who died in 1969 at age 73? What about

all the other mothers who died before age 73, or who died between the ages of 73 and 100? What about other variables that may affect both mothers' ages at the births of their children and longer life spans — variables such as financial status, marital stability, or other socioeconomic factors? The women in this study were in their thirties during the Depression; this may have influenced both their life span and if or when they had children.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> How do researchers handle confounding variables? They control for them as best they can, for as many of them as they can anticipate, trying to minimize their possible effect on the response. In experiments involving human subjects, researchers have to battle many confounding variables.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, in a study trying to determine the effect of different types and volumes of music on the amount of time grocery shoppers spend in the store (yes, they do think about that), researchers have to anticipate as many possible confounding variables ahead of time and then control for them. What other factors besides volume and type of music may influence the amount of time you spend in a grocery store? I can think of several factors: gender, age, time of day, whether you have children with you, how much money you have, the day of the week, how clean and inviting the store is, how nice the employees are, and (most importantly) what your motive is — are you shopping for the whole week, or are you just running in to grab a candy bar?</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How can researchers begin to control for so many possible confounding factors? Some of them can be controlled for in the design of the study, such as the time of the day, day of the week, and reason for shopping. But other factors (such as the perception of the store environment) depend totally on the individual in the study. The ultimate form of control for those person-specific confounding variables is to use pairs of people that are matched according to important variables, or to just use the same person twice: once with the treatment and once without. This type of experiment is called a matched-pairs design. (See Chapter 15 for more on this.)</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Before
believing any medical headlines (or any headlines with
statistics, for that matter), look to see how the study was
conducted. Observational studies can’t control for confounding
variables, so their results are not as statistically meaningful
(no matter what the statistics say) as the results of a well-
designed experiment are. In cases where an experiment can’t be
done (after all, no one can force you to have a baby after or
before age 40), make sure the observational study is based on a
large enough sample that represents a cross-section of the
population. And think about possible confounding variables that
may affect the conclusions being drawn.</p>
</blockquote>
<blockquote class="calibre5"><p id="a315" class="calibre6">
Respecting
ethical issues</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The trouble with experiments is that
some experimental designs are not ethical. You can’t force
research subjects to smoke in order to see whether they get
lung cancer, for example — you can only look at people who have
lung cancer and work backward to see what factors
(variables being studied) may have caused the
disease. But because you can’t control for the various factors
you’re interested in — or for any other variables, for that
matter — singling out any one particular cause becomes
difficult with observational studies. That’s why so much
evidence was needed to show that smoking causes lung cancer,
and why the tobacco companies only recently had to pay huge
penalties to victims.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Although the causes of cancer and other
diseases can’t be determined ethically by conducting
experiments on humans, new treatments for cancer can be (and
are) tested using experiments. Medical studies that involve
experiments are called
clinical trials.
The U.S. government has a registry of federally and
privately supported clinical trials conducted in the United
States and around the world; it also has information available
on who may participate in various clinical trials
.
 Check out <a
href="http://www.clinicaltrials.gov">www.clinicaltrials.gov

 for more information.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Serious experiments (such as those
funded by and/or regulated by the U.S. government) must pass a
huge series of tests that can take years to carry out. The
approval of a new drug, for example, goes through a very
lengthy, comprehensive, and detailed process regulated and
monitored by the FDA (Federal Drug Administration). One reason
the cost of prescription drugs is so high is the massive amount
of time and money needed to conduct research and development of
new drugs, most of which fail to pass the tests and have to be
scrapped.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Any experiments involving human
subjects are also regulated by the federal government and have
to gain approval by a committee created for the purpose of
protecting “the rights and welfare of the participants.” The
committees set up for different organizations have different
names (such as Institutional Review Board [IRB], Independent
Ethics Committee [IEC], or Ethical Review Board [ERB], to name
a few) but they all serve the same purpose. Research conducted
on animals is more nebulous in terms of regulations and
continues to be a topic of much debate and controversy in the
U.S. and around the world.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Surveys, polls,
and other observational studies are fine if you want to know
people’s opinions, examine their lifestyles without
intervention, or examine some demographic variables. If you
want to try to determine the cause of a certain outcome or
behavior (that is, a reason why something happened), an
experiment is a much better way to go. If an experiment isn’t
possible because of ethics concerns (or because of expense or
other reasons), a large body of observational studies examining
many different factors and coming up with similar conclusions
is the next best thing. (See Chapter 18 for more about cause-
and-effect relationships.)</p></blockquote>
<blockquote class="calibre5"><p id="a316" class="calibre6">
Collecting
good data</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What constitutes “good” data?
Statisticians use three criteria for evaluating data quality;
each of the criteria really relates most strongly to the
quality of the measurement instrument that’s used in the
process of collecting the data. To decide whether you’re
looking at good data from a study, look for these

characteristics:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
The data are reliable — you can get repeatable results with subsequent measurements. Many bathroom scales give unreliable data. You get on the scale, and it gives you one number. You don’t believe the number, so you get off, get back on, and get a different number. (If the second number is lower, you’ll most likely quit at this point; if not, you may continue getting on and off until you see a number you like.) Or you can do what some researchers do: Take three measurements, find the average, and use that; at least this will improve the reliability a bit.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Unreliable data come from unreliable measurement instruments or unreliable data collection methods. Errors can go beyond the actual scales to more intangible measurement instruments, like survey questions, which can give unreliable results if they’re written in an ambiguous way (see Chapter 16).</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> Find out how the data were collected when examining the results of a study. If the measurements are unreliable, the data could be inaccurate.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
The data are valid — they measure what they’re supposed to measure. Checking the validity of data requires you to step back and look at the big picture. You have to ask the question: Do these data measure what they should be measuring? Or should the researchers have been collecting altogether different data? The appropriateness of the measurement instrument used is important. For example, many educators say that a student’s transcript is not a valid measure of their ability to perform well in college. Alternatives include a more holistic approach, taking into account not only grades, but adding weight to elements such as service, creativity, social involvement, extracurricular activities, and the like.</p></blockquote><div class="calibre19"> </div>

<blockquote class="calibre5"><p class="calibre39"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Before
accepting the results of an experiment, find out what data were
measured and how they were measured. Be sure the researchers
are collecting valid data that are appropriate for the goals of
the study.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The data are
unbiased — they contain no systematic errors that either add to
or subtract from the true values.
Biased data are data that systematically overmeasure or
undermeasure the true result. Bias can occur almost anywhere
during the design or implementation of a study. Bias can be
caused by a bad measurement instrument (like that bathroom
scale that’s “always” 5 pounds over), by survey questions that
lead participants in a certain way, or by researchers who know
what treatment each subject received and who have preconceived
expectations.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Bias is
probably the number-one problem in collecting good data.
However, you can minimize bias with methods similar to those
discussed in Chapter 16 for surveys and in the “Making random
assignments” section earlier in this chapter, and by making
your experiments double-blind whenever possible.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Double-blind means
neither the subjects nor the researchers know who got what
treatment or who is in the control group. The subjects need to
be oblivious to which treatment they’re getting so that the
researchers can measure the placebo effect. And researchers
should be kept in the dark so they don’t treat subjects
differently by either expecting or not expecting certain
responses from certain groups. For example, if a researcher
knows you’re in the treatment group to study the side effects
of a new drug, she may expect you to get sick and therefore may
pay more attention to you than if she knew you were in the
control group. This can result in biased data and misleading
results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">If the researcher knows who got what
treatment but the subjects don’t know, the study is called a
blind
study (rather than a double-blind
study). Blind studies are better than nothing, but double-blind
studies are best. In case you’re wondering: In a double-blind
study, does anyone know which
treatment was given to which subjects? Relax; typically a third
party, such as a lab assistant, does that part.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In some cases the subjects know which
group they’re in because it’s unconcealable — for example, when
comparing the benefits of doing yoga versus jogging. However,
bias can be reduced by not telling the subjects the precise
purpose of the study. This irregular type of plan would have to
be reviewed by an institutional review board to make sure it
isn’t unethical to do; see the earlier section “Respecting
ethical issues.”</p></blockquote>
<blockquote class="calibre5"><p id="a317" class="calibre6">
Analyzing the
data properly</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After the data have been collected,
they’re put into that mysterious box called the
statistical
analysis for number crunching. The
choice of analysis is just as important (in terms of the
quality of the results) as any other aspect of a study. A
proper analysis should be planned in advance, during the design
phase of the experiment. That way, after the data are
collected, you won’t run into any major problems during the
analysis.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s the bottom line when selecting
the proper analysis: Ask yourself the question, “After the data
are analyzed, will I be able to legitimately and correctly
answer the question that I set out to answer?” If the answer is
“no,” then that analysis isn’t appropriate.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Some basic types of statistical
analyses include confidence intervals
(used when you’re trying to estimate a population value, or the
difference between two population values); hypothesis tests

 (used when you want to test a claim about
one or two populations, such as the claim that one drug is more
effective than another); and correlation and
regression analyses (used when you
want to show if and/or how one quantitative variable can
predict or cause changes in another quantitative variable). See
Chapters 13, 15, and 18, respectively, for more on each of
these types of analyses.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> When choosing
how you’re going to analyze your data, you have to make sure
that the data and your analysis will be compatible. For
example, if you want to compare a treatment group to a control
group in terms of the amount of weight lost on a new (versus an
existing) diet program, you need to collect data on how much
weight each person lost — not just each person’s weight at the
end of the study.</p></blockquote>
<blockquote class="calibre5"><p id="a318" class="calibre6">
Making
appropriate conclusions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In my opinion, the biggest mistakes
researchers make when drawing conclusions about their studies
are the following (discussed in the following sections):
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Overstating their results</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Making connections or giving explanations that aren’t
backed up by the statistics</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Going beyond the scope of the study in terms of whom the
results apply to</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre6">Overstating the
results</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Many times, the headlines in the media overstate actual research results. When you read a headline or otherwise hear about a study, be sure to look further to find out the details of how the study was done and exactly what the conclusions were.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Press releases often overstate results, too. For example, in a recent press release by the National Institute for Drug Abuse, the researchers claimed that use of the street drug Ecstasy was down from the previous year. However, when you look at the actual statistical results in the report, you find that the percentage of teens
from the sample who said they’d used Ecstasy was lower than those from the previous year, but this difference was not found to be statistically significant when they tried to project it onto the population of
all
teens. This discrepancy means that although fewer teens in the sample used Ecstasy that year, the difference wasn’t enough to account for more than chance variability from sample to sample. (See Chapter 14 for more about statistical significance.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Headlines and leading paragraphs in press releases and news articles often overstate the actual results of a study. Big results, spectacular findings, and major breakthroughs make the news these days, and reporters and others in the media constantly push the envelope in terms of what is and isn’t newsworthy. How can you sort out the truth from exaggeration? The best thing to do is to read the fine print.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Taking the results one step beyond the actual data</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A study that links having children later in life to longer life spans illustrates another point about research results. Do the results of this observational study mean that having a baby later in life can make you live longer? “No,” said the researchers. Their explanation of the results was that having a baby later in life may be due to women having a “slower” biological clock, which presumably would then result in the aging process being slowed down.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">My question to these researchers is,
“Then why didn’t you study that,
 instead of just looking at their ages?” The study
didn’t include any information that would lead me to conclude
that women who had children after age 40 aged at a slower rate
than other women, so in my view, the researchers shouldn’t make
that conclusion. Or the researchers should state clearly that
this view is only a theory and requires further study. Based on
the data in this study, the researchers’ theory seems like a
leap of faith (although since I became a new mom at age 41,
I’ll hope for the best!).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Frequently in a press release or news
article, the researcher will give an explanation about
why
he thinks the results of the study turned
out the way they did and what implications these results have
for society as a whole when the “why” hasn’t been studied yet.
These explanations may have been in response to a reporter’s
questions about the research — questions that were later edited
out of the story, leaving only the juicy quotes from the
researcher. Many of these after-the-fact explanations are no
more than theories that have yet to be tested. In such cases,
you should be wary of conclusions, explanations, or links drawn
by researchers that aren’t backed up by their studies.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Be aware that
the media wants to make you read the article (they get paid
to do that), so they will have strong headlines, or will make
unconfirmed “cause-effect” statements because it is their job
to sell the story. It is
your job to
be wary.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Generalizing results
to people beyond the scope of the study
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You can make conclusions only about the
population that’s represented by your sample. If you sample men
only, you can’t make conclusions about women. If you sample
healthy young people, you can’t make your conclusions about
everyone. But many researchers try to do just that, and it can
give misleading results.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">Here’s how you can determine whether a researcher’s conclusions measure up (Chapter 16 has more on samples and populations):</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Find out what the target population is (that is, the group that the researcher wants to make conclusions about).</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find out how the sample was selected and see whether the sample is representative of that target population (and not some more narrowly defined population).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Check the conclusions made by the researchers and make sure they’re not trying to apply their results to a broader population than they actually studied.
</p></blockquote><div class="calibre31"> </div>
<p id="a319" class="calibre6">Making Informed Decisions
</p>
<blockquote class="calibre9"><p class="calibre14">Just because someone says they conducted a “scientific study” or a “scientific experiment” doesn’t mean it was done right or that the results are credible (not that I’m saying you should discount everything that you see and hear). Unfortunately, I’ve come across a lot of bad experiments in my days as a statistical consultant. The worst part is that if an experiment was done poorly, you can’t do anything about it after the fact except ignore the results — and that’s exactly what you need to do.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Here are some tips that help you make an informed decision about whether to believe the results of an experiment, especially one whose results are very important to you:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
When you first hear or see the result, grab a pencil and write down as much as you can about what you heard or read, where you heard or read it, who did the research, and what the main results were.

 (I keep pencil and paper in my TV
room and in my purse just for this purpose.)</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Follow up on
your sources until you find the person who did the original
research and then ask them for a copy of the report or paper.
</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Go through the
report and evaluate the experiment according to the eight steps
for a good experiment described in the “Designing a Good
Experiment” section of this chapter.
(You really don’t have to understand everything written in a
report in order to do that.)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Carefully
scrutinize the conclusions that the researcher makes regarding
his or her findings. Many
researchers tend to overstate results, make conclusions beyond
the statistical evidence, or try to apply their results to a
broader population than the one they studied.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Never be
afraid to ask questions of the media, the researchers, and even
your own experts. For example, if
you have a question about a medical study, ask your doctor. He
or she will be glad that you’re an empowered and well-informed
patient!</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
And finally,
don’t get overly skeptical, just because you’re now a lot more
aware of all the bad practices going on out there.
Not everything is bad. There are many more

good researchers, credible results, and well-informed reporters than not. You have to maintain a sense of being cautious and ready to spot problems without discounting everything.
</p></blockquote><div class="calibre19"> </div>
</div>
</div> <div class="mbppagebreak" id="a320"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

  <a href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px

!important; min-height: 10px !important; border: solid 1px !important;"> </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a339">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 18</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a322" class="calibre10"><span class="calibre11"><span
class="bold"><span>Looking for Links: Correlation and
Regression</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Exploring statistical relationships between numerical
variables</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Looking at correlation and linear regression</span>
</span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Making predictions based on known relationships</span>
</span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Considering correlation versus causation</span></span>
</p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">T<span>oday’s media provide a steady stream
of information, including reports on all the latest links that
have been found by researchers. </span><span>Just today I heard
that increased video game use can negatively affect a child’s
```

attention span, the amount of a certain hormone in a woman’s body can predict when she will enter menopause, and the more depressed you get, the more chocolate you eat, and the more chocolate you eat, the more depressed you get (how depressing!).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Some studies are truly legitimate and help improve the quality and longevity of our lives. Other studies are not so clear. For example, one study says that exercising 20 minutes three times a week is better than exercising 60 minutes one time a week, another study says the opposite, and yet another study says there is no difference. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you are a confused consumer when it comes to links and correlations, take heart; this chapter can help. You’ll gain the skills to dissect and evaluate research claims and make your own decisions about those headlines and sound bites that you hear each day alerting you to the latest correlation. You’ll discover what it truly means for two variables to be correlated, when a cause-and-effect relationship can be concluded, and when and how to predict one variable based on another.</p></blockquote>
<p id="a323" class="calibre6">Picturing a Relationship with a Scatterplot</p>
<blockquote class="calibre9"><p class="calibre14">An article in Garden Gate magazine caught my eye: “Count Cricket Chirps to Gauge Temperature.” According to the article, all you have to do is find a cricket, count the number of times it chirps in 15 seconds, add 40, and voilà! You’ve just estimated the temperature in Fahrenheit.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The National Weather Service Forecast Office even puts out its own “Cricket Chirp Converter.” You enter the number of cricket chirps recorded in 15 seconds, and the converter gives you the estimated temperature in four different units, including Fahrenheit and Celsius. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A fair amount of research does support the claim that frequency of cricket chirps is related to temperature. For the purpose of illustration I’ve taken only a subset of some of the data (see Table 18-1).</p> </blockquote>
<blockquote class="calibre9"><p class="calibre30"><span

class="calibre15"><img alt="/Table 18-1" src="images/00351.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Notice that each observation is
composed of two variables that are tied together: the number of
times the cricket chirped in 15 seconds (the X
-variable) and the temperature at the time the data was
collected (the Y-variable).
Statisticians call this type of two-dimensional data
bivariate
data. Each observation contains one
pair of data collected simultaneously. For example, row one of
Table 18-1 depicts a pair of data (18, 57).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Bivariate data is typically organized
in a graph that statisticians call a scatterplot.
 A scatterplot has two dimensions, a
horizontal dimension (the
X-axis) and a
vertical dimension (the
Y-axis). Both
axes are numerical; each one contains a number line. In the
following sections, I explain how to make and interpret a
scatterplot.</p></blockquote>
<blockquote class="calibre5"><p id="a324" class="calibre6">
Making a
scatterplot</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Placing
observations (or points) on a scatterplot is similar to playing
the game Battleship. Each observation has two coordinates; the
first corresponds to the first piece of data in the pair
(that’s the X coordinate; the
amount that you go left or right). The second coordinate
corresponds to the second piece of data in the pair (that’s the
Y-coordinate; the
amount that you go up or down). You place the point
representing that observation at the intersection of the two
coordinates.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 18-1 shows a scatterplot for the

cricket chirps and temperature data listed in Table 18-1.
Because I ordered the data according to their
X
-values, the points on the scatterplot correspond
from left to right to the observations given in Table 18-1, in
the order listed.</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 18-1: Scatterplot of cricket
chirps in relation to outdoor temperature.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1801.eps"
src="images/00352.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre5"><p id="a325" class="calibre6">
Interpreting a
scatterplot</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> You interpret a
scatterplot by looking for trends in the data as you go from
left to right:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the data show an uphill pattern as you move from left
to right, this indicates a positive relationship
between X and Y. As the
 X-
values increase (move right), the
Y
-values increase (move up) a certain amount.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the data show a downhill pattern as you move from left
to right, this indicates a negative relationship
between X and Y. As the
X
-values
increase (move right) the
Y
-values decrease (move down) by a certain

amount.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If the data don’t seem to resemble any kind of pattern
(even a vague one), then no relationship exists between
 X
and
Y.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">One pattern of special interest is a
linear pattern, where
the data has a general look of a line going uphill or downhill.
Looking at Figure 18-1, you can see that a positive linear
relationship does appear between number of cricket chirps and
the temperature. That is, as the cricket chirps increase, the
temperature increases as well.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> In this chapter
I explore linear relationships only. A linear relationship
between X and Y exists when the
pattern of X- and
Y
-values resembles a line, either uphill (with a
positive slope) or downhill (with a negative slope). Other
types of trends may exist in addition to the uphill/downhill
linear trends (for example, curves or exponential functions);
however, these trends are beyond the scope of this book. The
good news is that many relationships do fall under the
uphill/downhill linear scenario.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Scatterplots
show possible associations or relationships between two
variables. However, just because your graph or chart shows
something is going on, it doesn’t mean that a cause-and-effect
relationship exists.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, a doctor observes that
people who take vitamin C each day seem to have fewer colds.
Does this mean vitamin C prevents colds? Not necessarily. It
could be that people who are more health conscious take vitamin

C each day, but they also eat healthier, are not overweight, exercise every day, and wash their hands more often. If this doctor really wants to know if it’s the vitamin C that’s doing it, she needs a well-designed experiment that rules out these other factors. (See the later section “Explaining the Relationship: Correlation versus Cause and Effect” for more information.)</p></blockquote>
<p id="a326" class="calibre6">Quantifying Linear Relationships Using the Correlation</p>
<blockquote class="calibre9"><p class="calibre14">After the bivariate data have been organized graphically with a scatterplot (see the preceding section), and you see some type of linear pattern, the next step is to do some statistics that can quantify or measure the extent and nature of the relationship. In the following sections, I discuss correlation, a statistic measuring the strength and direction of a linear relationship between two variables; in particular, how to calculate and interpret correlation and understand its most important properties.</p></blockquote>
<blockquote class="calibre5"><p id="a327" class="calibre6">Calculating the correlation</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the earlier section “Interpreting a scatterplot,” I say data that resembles an uphill line has a positive linear relationship and data that resembles a downhill line has a negative linear relationship. However, I didn’t address the issue of whether or not the linear relationship was strong or weak. The strength of a linear relationship depends on how closely the data resembles a line, and of course varying levels of “closeness to a line” exist.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Can one statistic measure both the strength and direction of a linear relationship between two variables? Sure! Statisticians use the correlation coefficient to measure the strength and direction of the linear relationship between two numerical variables X and Y. The correlation coefficient for a sample of data is denoted by r.</p></blockquote>

```
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Although the
street definition of </span><span><span class="calibre16"><span
class="italic">correlation</span></span></span><span> applies
to any two items that are related (such as gender and political
affiliation), statisticians use this term only in the context
of two numerical variables. The formal term for correlation is
the </span><span><span class="calibre16"><span
class="italic">correlation coefficient.</span></span></span>
<span> Many different correlation measures have been created;
the one used in this case is called the </span><span><span
class="calibre16"><span class="italic">Pearson correlation
coefficient</span></span></span><span> (but from now on I'll
just call it the correlation).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>The formula for the correlation
(</span><span><span class="calibre16"><span
class="italic">r</span></span></span><span>) is</span></span>
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq18001.eps"
src="images/00353.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>where </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span> is the number of pairs of data; </span><img
alt="9780470911082-eq18002.eps" src="images/00354.jpg"
class="calibre2"/><span> and </span><img alt="9780470911082-
eq18003.eps" src="images/00355.jpg" class="calibre2"/><span>
are the sample means of all the </span><span><span
class="calibre16"><span class="italic">x</span></span></span>
<span>-values and all the </span><span><span class="calibre16">
<span class="italic">y</span></span></span><span>-values,
respectively; and </span><span><span class="calibre16"><span
class="italic">s</span></span></span><span><span
class="calibre43"><span class="italic"><sub
class="calibre42">x</sub></span></span></span><span> and
</span><span><span class="calibre16"><span
class="italic">s</span></span></span><span><span
class="calibre43"><span class="italic"><sub
class="calibre42">y</sub></span></span></span><span> are the
sample standard deviations of all the</span><span><span
class="calibre16"><span class="italic"> x- </span></span>
</span><span>and</span><span><span class="calibre16"><span
class="italic"> y-</span></span></span><span>values,
respectively.</span></span></p></blockquote>
```

```html
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Use the
following steps to calculate the correlation, </span><span>
<span class="calibre16"><span class="italic">r,</span></span>
</span><span> from a data set:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 1. Find the mean of all the </span></span></span>
<span><span class="calibre16"><span class="bold"><span
class="italic">x</span></span></span></span><span><span
class="calibre16"><span class="bold">-values (</span></span>
</span><img alt="9780470911082-eq18004.eps"
src="images/00356.jpg" class="calibre2"/><span><span
class="calibre16"><span class="bold">) and the mean of all the
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">y</span></span></span></span>
<span><span class="calibre16"><span class="bold">-values
(</span></span></span><img alt="9780470911082-eq18005.eps"
src="images/00357.jpg" class="calibre2"/><span><span
class="calibre16"><span class="bold">).</span></span></span>
</span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> See Chapter 5 for more on calculating
the mean.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 2. Find the standard deviation of all the </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">x</span></span></span></span><span><span
class="calibre16"><span class="bold">-values (call it </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">s</span></span></span></span><span><span
class="calibre43"><span class="bold"><span class="italic"><sub
class="calibre42">x</sub></span></span></span></span><span>
<span class="calibre16"><span class="bold">) and the standard
deviation of all the </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">y</span></span></span></span><span><span
class="calibre16"><span class="bold">-values (call it </span>
</span></span><span><span class="calibre16"><span class="bold">
<span class="italic">s</span></span></span></span><span><span
class="calibre43"><span class="bold"><span class="italic"><sub
class="calibre42">y</sub></span></span></span></span><span>
<span class="calibre16"><span class="bold">).</span></span>
</span></span></p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
```

class="calibre15"> See Chapter 5 to find out how to
calculate the standard deviation.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. For each (x,
y) pair in the data set,
take x
minus
<img alt="9780470911082-eq18006.eps"
src="images/00358.jpg" class="calibre2"/> and
 y minus
<img alt="9780470911082-eq18007.eps"
src="images/00359.jpg" class="calibre2"/>, and multiply them
together to get <img alt="9780470911082-
eq18008.eps" src="images/00360.jpg" class="calibre2"/>
.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Add up all the results from Step 3.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. Divide the sum by
s<sub
class="calibre42">x</sub>

∗
s
_y

.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 6. Divide the result by
n
– 1, where
n is the number of (

x,
y) pairs. (It’s the same as
multiplying by 1 over n
– 1.)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> This gives you the correlation,
r.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre14">For example, suppose you have the data
set (3, 2), (3, 3), and (6, 4). You calculate the correlation
coefficient r via the following
steps. (Note for this data the x
-values are 3, 3, 6, and the y
-values are 2, 3, 4.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. <img alt="9780470911082-
eq18009.eps" src="images/00361.jpg" class="calibre2"/>
 is 12 ÷ 3 = 4, and
<img alt="9780470911082-eq18010.eps"

src="images/00362.jpg" class="calibre2"/> is 9 ÷ 3 = 3.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. The standard deviations are
s<sub
class="calibre42">x</sub>
 = 1.73 and

s<sub
class="calibre42">y</sub>
 = 1.00.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> See Chapter 5 for step-by-step
calculations.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The differences found in Step 3 multiplied
together are: (3 – 4)(2
– 3) = (–
 1)(
 – 1) = +1; (3 –
 4)(3
– 3) = (–
 1)(0) = 0; (6
– 4)(4
– 3) = (2)(1) = +2.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Adding the Step 3 results, you get 1 + 0 + 2 =
3.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

```
class="calibre15"><span><span class="calibre16"><span
class="bold"> 5. Dividing by </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">s</span></span></span></span><span><span
class="calibre43"><span class="bold"><span class="italic"><sub
class="calibre42">x</sub></span></span></span></span><span>
<span class="calibre16"><span class="bold">
</span></span></span><span>∗</span><span><span
class="calibre16"><span class="bold">
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">s</span></span></span></span>
<span><span class="calibre43"><span class="bold"><span
class="italic"><sub class="calibre42">y</sub></span></span>
</span></span><span><span class="calibre16"><span class="bold">
gives you 3 ÷ (1.73 </span></span></span><span>∗</span><span>
<span class="calibre16"><span class="bold"> 1.00) = 3 ÷ 1.73 =
1.73.</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 6. Now divide the Step 5 result by 3 </span>
</span></span><span>–</span><span><span class="calibre16"><span
class="bold"> 1 (which is 2), and you get the correlation
</span></span></span><span><span class="calibre16"><span
class="bold"><span class="italic">r </span></span></span>
</span><span><span class="calibre16"><span class="bold">= 0.87.
</span></span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre5"><p id="a328" class="calibre6">
<span class="calibre21"><span class="bold"><span>Interpreting
the correlation</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> The correlation
</span><span><span class="calibre16"><span
class="italic">r</span></span></span><span> is always between
+1 and –1. To interpret various values of </span><span><span
class="calibre16"><span class="italic">r</span></span></span>
<span> (no hard and fast rules here, just Rumsey’s rule of
thumb), see which of the following values your correlation is
closest to:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Exactly
</span></span></span><span>–</span><span><span
class="calibre16"><span class="bold">1:</span></span></span>
```

```
<span> A perfect downhill (negative) linear relationship</span>
</span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>–</span><span><span class="calibre16"><span
class="bold">0.70: </span></span></span><span>A strong downhill
(negative) linear relationship</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>–</span><span><span class="calibre16"><span
class="bold">0.50: </span></span></span><span>A moderate
downhill (negative) relationship</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>–</span><span><span class="calibre16"><span
class="bold">0.30: </span></span></span><span>A weak downhill
(negative) linear relationship</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">0: </span>
</span></span><span>No linear relationship</span></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">+0.30:</span>
</span></span><span> A weak uphill (positive) linear
relationship</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">+0.50:</span>
</span></span><span> A moderate uphill (positive)
relationship</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">+0.70:</span>
</span></span><span> A strong uphill (positive) linear
```

relationship</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Exactly +1:
 A perfect uphill (positive) linear
relationship</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> If the
scatterplot doesn’t indicate there’s at least somewhat of a
linear relationship, the correlation doesn’t mean much. Why
measure the amount of linear relationship if there isn’t enough
of one to speak of? However you can take the idea of no linear
relationship two ways: 1) If no relationship at all exists,
calculating the correlation doesn’t make sense because
correlation only applies to linear relationships; and 2) If a
strong relationship exists but it’s not linear, the correlation
may be misleading, because in some cases a strong curved
relationship exists yet the correlation turns out to be strong.
That’s why it’s critical to examine the scatterplot first.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 18-2 shows examples of what
various correlations look like, in terms of the strength and
direction of the relationship. Figure 18-2a shows a correlation
of +1, Figure 18-2b shows a correlation of –0.50, Figure 18-2c
shows a correlation of +0.85, and Figure 18-2d shows a
correlation of +0.15. Comparing Figures 18-2a and c, you see
Figure 18-2a is a perfect uphill straight line, and Figure 18-
2c shows a very strong uphill linear pattern. Figure 18-2b is
going downhill but the points are somewhat scattered in a wider
band, showing a linear relationship is present, but not as
strong as in Figures 18-2a and 18-2c. Figure 18-2d doesn’t show
much of anything happening (and it shouldn’t, since its
correlation is very close to 0).</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 18-2: Scatterplots with
correlations of a) +1.00; b) –0.50; c) +0.85; and d) +0.15.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1802.eps"
src="images/00363.jpg" class="calibre2"/></p>
</blockquote>

```html
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Many folks make the mistake of thinking
that a correlation of –1 is a bad thing, indicating no
relationship. Just the opposite is true! A correlation of –1
means the data are lined up in a perfect straight line, the
strongest linear relationship you can get. The “–” (minus) sign
just happens to indicate a negative relationship, a downhill
line.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> How close is close enough to –1 or +1
to indicate a strong enough linear relationship? Most
statisticians like to see correlations beyond at least +0.5 or
–0.5 before getting too excited about them. Don’t expect a
correlation to always be 0.99 however; remember, this is real
data, and real data aren’t perfect.</span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>For my subset of the cricket chirps
versus temperature data from the earlier section “Picturing a
Relationship with a Scatterplot,” I calculated a correlation of
0.98, which is almost unheard of in the real world (these
crickets are </span><span><span class="calibre16"><span
class="italic">good!</span></span></span><span>).</span></span>
</p></blockquote>
<blockquote class="calibre5"><p id="a329" class="calibre6">
<span class="calibre21"><span class="bold"><span>Examining
properties of the correlation</span></span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> Here are
several important properties of the correlation coefficient:
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The correlation is always between –1 and +1, as I explain
in the preceding section.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The correlation is a unitless measure, which means that
if you change the units of</span><span><span class="calibre16">
<span class="italic"> X </span></span></span><span>or </span>
<span><span class="calibre16"><span class="italic">Y,</span>
</span></span><span> the correlation won’t change. For example,
```

changing the temperature from Fahrenheit to Celsius won’t
affect the correlation between the frequency of chirps (
X
) and the outside temperature (
Y
).</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The variables X and
 Y
can be switched in the data set without changing
the correlation. For example, if height and weight have a
correlation of 0.53, weight and height have the same
correlation.</p></blockquote><div
class="calibre19"> </div>
<p id="a330" class="calibre6">Working with Linear Regression
</p>
<blockquote class="calibre9"><p class="calibre14">In the case of two numerical
variables X and
Y,
 when at least a moderate correlation has been
established through both the correlation and the scatterplot,
you know they have some type of linear relationship.
Researchers often use that relationship to predict the
(average) value of Y for a given value
of X
using a straight line. Statisticians
call this line the regression line.
If you know the slope and the y
-intercept of that regression line, then you can plug in
a value for X and predict the
average value for Y. In other words,
you predict (the average)
 Y from
X.
 In the following sections, I provide the
basics of understanding and using the linear regression

equation (I explain how to make predictions with linear
regression later in this chapter).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Never do a
regression analysis unless you have already found at least a
moderately strong correlation between the two variables. (My
rule of thumb is it should be at or beyond either positive or
negative 0.50, but other statisticians may have different
criteria.) I’ve seen cases where researchers go ahead and make
predictions when a correlation is as low as 0.20! By anyone’s
standards, that doesn’t make sense. If the data don’t resemble
a line to begin with, you shouldn’t try to use a line to fit
the data and make predictions (but people still try).
</p></blockquote>
<blockquote class="calibre5"><p id="a331" class="calibre6">
Figuring out
which variable is X and which is Y</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Before moving forward to find the
equation for your regression line, you have to identify which
of your two variables is
 X and which
is
Y. When doing correlations (as I
explain earlier in this chapter), the choice of which variable
is X
and which is Y
doesn’t matter, as long as you’re consistent for all the
data. But when fitting lines and making predictions, the choice
of X
and Y
does make a difference.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> So how do you
determine which variable is which? In general,
 Y
is the variable that you want to predict, and
X is the variable you
are using to make that prediction. In the earlier cricket
chirps example, you are using the number of chirps to predict
the temperature. So in this case the variable <span>

```
<span class="calibre16"><span class="italic">Y</span></span>
</span><span> is the temperature, and the variable </span>
<span><span class="calibre16"><span class="italic">X</span>
</span></span><span> is the number of chirps. Hence</span>
<span><span class="calibre16"><span class="italic"> Y </span>
</span></span><span>can be predicted by</span><span><span
class="calibre16"><span class="italic"> X </span></span></span>
<span>using the equation of a line if a strong enough linear
relationship exists.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/><span> Statisticians call the </span><span>
<span class="calibre16"><span class="italic">X</span></span>
</span><span>-variable (cricket chirps in my earlier example)
the </span><span><span class="calibre16"><span
class="italic">explanatory variable,</span></span></span><span>
because if</span><span><span class="calibre16"><span
class="italic"> X </span></span></span><span>changes, the slope
tells you (or explains) how much</span><span><span
class="calibre16"><span class="italic"> Y </span></span></span>
<span>is expected to change in response. Therefore, the </span>
<span><span class="calibre16"><span class="italic">Y</span>
</span></span><span> variable is called the </span><span><span
class="calibre16"><span class="italic">response variable.
</span></span></span><span> Other names for </span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span> and </span><span><span class="calibre16"><span
class="italic">Y</span></span></span><span> include the </span>
<span><span class="calibre16"><span
class="italic">independent</span></span></span><span> and
</span><span><span class="calibre16"><span
class="italic">dependent</span></span></span><span> variables,
respectively.</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a332" class="calibre6">
<span class="calibre21"><span class="bold"><span>Checking the
conditions</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/><span> In the case of
two numerical variables, you can come up with a line that
enables you to predict</span><span><span class="calibre16">
<span class="italic"> Y </span></span></span><span>from </span>
<span><span class="calibre16"><span class="italic">X,</span>
</span></span><span> if (and only if) the following two
conditions from the previous sections are met:</span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
```

class="calibre2"/>
The scatterplot must form a linear pattern.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The correlation,
r, is moderate
to strong (typically beyond 0.50 or –0.50).</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Some researchers actually don’t check
these conditions before making predictions. Their claims are
not valid unless the two conditions are met.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">But suppose the correlation is high; do
you still need to look at the scatterplot? Yes. In some
situations the data have a somewhat curved shape, yet the
correlation is still strong; in these cases making predictions
using a straight line is still invalid. Predictions need to be
made based on a curve. (This topic is outside the scope of this
book; if you are interested, see Statistics II For
Dummies, where I tackle nonlinear
relationships.)</p></blockquote>
<blockquote class="calibre5"><p id="a333" class="calibre6">
Calculating
the regression line</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For the crickets and temperature data,
you can see that the scatterplot in Figure 18-1 shows a linear
pattern. The correlation between cricket chirps and temperature
was found earlier in this chapter to be very strong (
r
 = 0.98). You now can find one line that
best fits the data (in terms of having the smallest overall
distance to the points). Statisticians call this technique for
finding the best-fitting line a simple linear

regression analysis using the least squares
method.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The formula for
the <span

class="italic">best-fitting line (or
regression line)
is y
= mx
 + b, where
m
 is the slope of the line and
b
 is the y-intercept. This
equation itself is the same one used to find a line in algebra;
but remember, in statistics the points don’t lie perfectly on a
line — the line is a model around which the data lie if a
strong linear pattern exists.</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The slope of a line is
the change in Y over the change in
X. For example, a
slope of 10/3 means as the x-
value increases (moves right) by 3 units, the
y
-value moves up by 10 units on average.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The y-intercept is that
place on the y-axis where the
value of x is zero. For
example, in the equation 2
x – 6, the line
crosses the y-axis at the point –
6. The coordinates of this point are (0, –6); when a line
crosses the y-axis, the

x-
value is always 0.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="technicalstuff.eps"
src="images/00008.jpg" class="calibre2"/> To come up with
the best-fitting line, you need to find values for
m
 and
b that fit the
pattern of data the best, for your given criteria. Different
criteria exist and can lead to other lines, but the criteria I
use in this book (and in all introductory level statistics
courses in general) is to find the line that minimizes what
statisticians call the
sum of squares for error (SSE)
. The SSE is the sum of all the squared
differences from the points on the proposed line to the actual
points in the data set. The line with the lowest possible SSE
wins and its equation is used as the best-fitting line. This
process is where the name
the least-squares method
 comes from.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You may be thinking that you have to
try lots and lots of different lines to see which one fits
best. Fortunately, you have a more straightforward option
(although eyeballing a line on the scatterplot does help you
think about what you’d expect the answer to be). The best-
fitting line has a distinct slope and y-
intercept that can be calculated using formulas (and, I
may add, these formulas aren’t too hard to calculate).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> To save a great deal of time
calculating the best fitting line, first find the “big five,”
five summary statistics that you’ll need in your calculations:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. The mean of the x
values (denoted <img alt="9780470911082-
eq18011.eps" src="images/00364.jpg" class="calibre2"/>)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. The mean of the<span

class="calibre16"> y
values (denoted <img alt="9780470911082-
eq18012.eps" src="images/00365.jpg" class="calibre2"/>)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. The standard deviation of
the x
values (denoted s
<sub
class="calibre42">x</sub>)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. The standard deviation of
the y
values (denoted s
<sub
class="calibre42">y</sub>)
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. The correlation between
 X
and
 Y (denoted
r)</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre5"><p class="calibre6">Finding the
slope</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the slope,
m,
 of the best-fitting line is
</p></blockquote>
<blockquote class="calibre5"><p class="calibre39"><img alt="9780470911082-eq18013.eps"
src="images/00366.jpg" class="calibre2"/></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">where r
 is the correlation between X
and Y, and
s

<sub
class="calibre42">x</sub> and
s<sub
class="calibre42">y</sub> are the
standard deviations of the x
-values and the
y-values,
respectively. You simply divide s
<sub
class="calibre42">y</sub><sub class="calibre42">
</sub>by s
<sub
class="calibre42">x</sub> and
multiply the result by
r.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Note that the slope of the best-fitting
line can be a negative number because the correlation can be a
negative number. A negative slope indicates that the line is
going downhill. For example, an increase in police officers is
related to a decrease in the number of crimes in a linear
fashion; the correlation and hence the slope of the best-
fitting line is negative in this case.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> The correlation and the slope of the
best-fitting line are not the same. The formula for slope takes
the correlation (a unitless measurement) and attaches units to
it. Think of s<sub
class="calibre42">y</sub> ÷
s
_x
 as the variation (resembling change) in
Y
over the variation in X,
 in units of<span

class="italic"> X and
Y.
 For example, variation in temperature (degrees
Fahrenheit) over the variation in number of cricket chirps (in
15 seconds).</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Finding the y-
intercept</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The formula for the y
-intercept, b, of the best-
fitting line is <img alt="9780470911082-eq18014.eps"
src="images/00367.jpg" class="calibre2"/>, where
<img alt="9780470911082-eq18015.eps" src="images/00368.jpg"
class="calibre2"/> and <img alt="9780470911082-
eq18016.eps" src="images/00369.jpg" class="calibre2"/>
are the means of the x-values and the
y-values,
respectively, and m is the slope (the
formula for which is given in the preceding section).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> So to calculate
the y-intercept,
b
, of the best-fitting line, you start by
finding the slope, m, of the best-
fitting line using the steps listed in the preceding section.
You then multiply m by <img
alt="9780470911082-eq18017.eps" src="images/00370.jpg"
class="calibre2"/> and subtract your result from
<img alt="9780470911082-eq18018.eps" src="images/00371.jpg"
class="calibre2"/>.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> Always calculate the slope before the
y-
intercept. The formula for the
y-

intercept contains the slope!
</p></blockquote>
<blockquote class="calibre5"><p id="a334" class="calibre6">
Interpreting
the regression line</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Even more important than being able to
calculate the slope and
y-intercept to
form the best-fitting regression line is the ability to
interpret their values; I explain how to do so in the following
sections.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Interpreting the
slope</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The slope is interpreted in algebra as
rise
over run. If, for example, the slope
is 2, you can write this as 2/1
 and say that as you move from point to point on the
line, as
the
value of the X
variable
increases by 1,
the value of the Y
variable
increases by 2. In a regression
context, the slope is the heart and soul of the equation
because it tells you how much you can expect Y
to change as X increases.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In general, the units for slope are the
units of the Y variable per units
of the <span

class="italic">X variable. It’s a ratio of change in Y per change in X. Suppose in studying the effect of dosage level in milligrams (mg) on systolic blood pressure (mmHg), a researcher finds that the slope of the regression line is –2.5. You can write this as –2.5/1 and say that systolic blood pressure is expected to decrease by 2.5 mmHg on average per 1 mg increase in drug dosage.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Always make sure to use proper units when interpreting slope. If you don’t consider units, you won’t really see the connection between the two variables at hand. For example if Y is exam score and X = study time, and you find the slope of the equation is 5, what does this mean? Not much without any units to draw from. Including the units, you see you get an increase of 5 points (change in Y) for every 1 hour increase in studying (change in X). Also be sure to watch for variables that have more than one common unit, such as temperature being in either Fahrenheit or Celsius; know which unit is being used. </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If using a 1 in the denominator of slope is not super-meaningful to you, you can multiply the top and bottom by any number (as long as it’s the same number) and interpret it that way instead. In the systolic blood pressure example, instead of writing slope as –2.5/1 and interpreting it as a drop of 2.5 mmHg per 1 mg increase of the drug, you can multiply the top and bottom by 10 to get –25/10 and say an increase in dosage of 10 mg results in a decrease in systolic blood pressure of 25 mmHg.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Interpreting the y-intercept</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">The y-
intercept is the place where the regression line
 y = mx +
b crosses the y
-axis where x = 0, and is denoted
by b (see the earlier
section “Finding the y-intercept”).
Sometimes the y-intercept can be
interpreted in a meaningful way, and sometimes not. This
uncertainty differs from slope, which is always interpretable.
In fact, between the two elements of slope and
y-
intercept, the slope is the star of the show, with
the y-intercept serving
as the less-famous but still noticeable sidekick.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> At times the
y-
intercept makes no sense. For
example, suppose you use rain to predict bushels per acre of
corn. You know if the data set contains a point where rain is
0, the bushels per acre must be 0 as well. As a result, if the
regression line crosses the y-
axis somewhere else besides 0 (and there is no guarantee
it will cross at 0 – it depends on the data), the
y-
intercept will make no sense. Similarly, in this
context a negative value of y
 (corn production) cannot be interpreted.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Another situation where you can’t
interpret the y-intercept is when
data are not present near the point where x

= 0. For example, suppose you want to use students' scores on Midterm 1 to predict their scores on Midterm 2. The y-intercept represents a prediction for Midterm 2 when the score on Midterm 1 is 0. You don't expect scores on a midterm to be at or near 0 unless someone didn't take the exam, in which case her score wouldn't be included in the first place.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Many times, however, the y-intercept is of interest to you, it has meaning, and you have data collected in the area where x = 0. For example, if you're predicting coffee sales at football games in Green Bay, Wisconsin, using temperature, some games get cold enough to have temperatures at or even below 0 degrees Fahrenheit, so predicting coffee sales at these temperatures makes sense. (As you may guess, they sell more and more coffee as the temperature dips.)</p>
</blockquote>
<blockquote class="calibre5"><p id="a335" class="calibre6">Putting it all together with an example: The regression line for the crickets</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the earlier section "Picturing a Relationship with a Scatterplot," I introduce the example of cricket chirps related to temperature. The "big five" statistics, which I explain in "Calculating the regression line," are shown in Table 18-2 for the subset of cricket data. (Note: I'm rounding for ease of explanation only.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The slope, m, for the best-fitting line for the subset of cricket chirps versus temperature data is . So as the number of chirps increases by 1 chirp per 15 seconds, the temperature is expected to increase by 0.90 degrees Fahrenheit on average. To get a more meaningful interpretation, you can multiply the top

and bottom of the slope by 10 and say as chirps increase by 10
(per 15 seconds) temperature increases 9 degrees Fahrenheit.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now, to find the y-
intercept, b, you take
<img alt="9780470911082-eq18022.eps" src="images/00374.jpg"
class="calibre2"/>, or 67 – (0.90)(26.5) = 43.15. So the
best-fitting line for predicting temperature from cricket
chirps based on the data is y
= 0.90x + 43.15, or
temperature (in degrees Fahrenheit) = 0.90
∗ (number of chirps in 15 seconds) + 43.2.
Now can you use the y-intercept to
predict temperature when no chirping is going on at all?
Because no data was collected at or near this point, you cannot
make predictions for temperature in this area. You can’t
predict temperature using crickets if the crickets are silent.
</p></blockquote>
<p id="a336" class="calibre6">Making Proper Predictions
</p>
<blockquote class="calibre9"><p class="calibre14">After you have determined a strong
linear relationship and you find the equation of the best
fitting line
using y
= mx +
b
, you use that line to predict (the average)
 y
for a given x-
value. To make predictions, you plug the
x-
value into the equation and solve for
y.
 For example, if your equation is
 y

= 2
x + 1 and you
want to predict y for
 x
= 1, then plug 1 into the equation for
 x
to get
 y = 2(1) + 1 =
3.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Keep in mind that you choose the values
of X
(the explanatory variable) that you
plug in; what you predict is Y,
 the response variable, which totally depends on
X.
 By doing this, you are using one variable
that you can easily collect data on to predict a
 Y
variable that is difficult or not possible to
measure. This process works well as long as X
and Y are correlated.
This concept is the big idea of regression.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Using the examples from the previous
section, the best-fitting line for the crickets is
 y
= 0.90x + 43.2. Say you’re
camping outside, listening to the crickets, and remember you
can predict temperature by counting cricket chirps. You count
35 chirps in 15 seconds, put in 35 for x,
 and find that y = 0.9(35) + 43.2 =
74.7. (Yeah, you memorized the formula before you went camping
just in case you needed it.) So because the crickets chirped 35
times in 15 seconds, you figure the temperature is probably
about 75 degrees Fahrenheit.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Just because
you have a regression line doesn’t mean you can plug in

any
value for X
and do a good job of predicting Y.
 Making predictions using x-
values that fall outside the range of your data is a no-
no. Statisticians call this extrapolation;
 watch for researchers who try to make
claims beyond the range of their data.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, in the chirping data, no
data is collected for fewer than 18 chirps or more than 39
chirps per 15 seconds (refer to Table 18-1). If you try to make
predictions outside this range, you are going into uncharted
territory; the farther outside this range you go with your
x-
values, the more dubious your
predictions for y will get. Who’s to
say the line still works outside of the area where data were
collected? Do you really think that crickets will chirp faster
and faster without limit? At some point they would either pass
out or burn up! And what does a negative number of chirps
really mean? (Is this similar to asking what the sound of one
hand clapping is?)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Be aware that
not every data point will necessarily fit the regression line
well, even if the correlation is high. A point or two may fall
outside the overall pattern of the rest of the data; such
points are called outliers. One or two
outliers probably won’t affect the overall fit of the
regression line much, but in the end you can see that the line
didn’t do well at those specific points.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">The numerical difference between the
predicted value of y from the line and
the actual y-value you got from
your data is called a

residual.
Outliers have large residuals compared to the rest of the
points; they are worth investigating to see if there was an
error in the data at those points or if there is something
particularly interesting in the data to follow up on. (I give a
much more detailed look at residuals in the book
Statistics II For
Dummies.)</p>
</blockquote>
<p id="a337" class="calibre6">Explaining the Relationship: Correlation
versus Cause and Effect</p>
<blockquote class="calibre9"><p class="calibre14">Scatterplots and correlations identify
and quantify relationships between two variables. However, if a
scatterplot shows a definite pattern and the data are found to
have a strong correlation, that doesn’t necessarily mean that a
cause-and-effect relationship exists between the two variables.
A cause-and-effect relationship
 is one where a change in
one variable (in this case
X
)
causes a change in another variable
(in this case Y). (In other words,
the change in Y is not only
associated with a change in X
, but also directly caused by X
.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, suppose a well-controlled
medical experiment is conducted to determine the effects of
dosage of a certain drug on blood pressure. (See a total
breakdown of experiments in Chapter 17.) The researchers look
at their scatterplot and see a definite downhill linear
pattern; they calculate the correlation, and it’s strong. They
conclude that increasing the dosage of this drug causes a
decrease in blood pressure. This cause-and-effect conclusion is
okay because they controlled for other variables that could
affect blood pressure in their experiment, such as other drugs
taken, age, general health, and so on.</p>

</blockquote>
<blockquote class="calibre9"><p class="calibre14">However, if you made a scatterplot and
examined the correlation between ice cream consumption versus
murder rates in New York City, you would also see a strong
linear relationship (this one is uphill). Yet no one would
claim that more ice cream consumption causes more murders to
occur.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What’s going on here? In the first
case, the data were collected through a well-controlled medical
experiment, which minimizes the influence of other factors that
may affect blood pressure. In the second example, the data were
based just on observation, and no other factors were examined.
Researchers subsequently found out that this strong
relationship exists because increases in murder rates and ice
cream sales are both related to increases in temperature.
Temperature in this case is called a confounding
variable; it affects both
 X
and
 Y but was not
included in the study (see Chapter 17).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Whether two
variables are found to be causally associated depends on how
the study was conducted. I’ve seen many instances in which
people try to claim cause-and-effect relationships just by
looking at scatterplots or correlations. Why would they do
this? Because they want to believe it (in other words for them
it’s “believing is seeing,” rather than the other way around).
Beware of this tactic. In order to establish cause and effect,
you need to have a well-designed experiment or a boatload of
observational studies. If someone is trying to establish a
cause-and-effect relationship by showing a chart or graph, dig
deeper to find out how the study was designed and how the data
were collected, and evaluate the study appropriately using the
criteria outlined in Chapter 17.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The need for a well-designed experiment
in order to claim cause and effect is often ignored by some
researchers and members of the media, who give us headlines
such as “Doctors can lower malpractice lawsuits by spending
more time with patients.” In reality, it was found that doctors
who have fewer lawsuits are the type who spend a lot of time

with patients. But that doesn’t mean taking a bad doctor and having him spend more time with his patients will reduce his malpractice suits; in fact, spending more time with them may create even more problems.</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a338"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

  <a href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px

!important; min-height: 10px !important; border: solid 1px !important;"> </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a357">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 19</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a340" class="calibre10"><span class="calibre11"><span
class="bold"><span>Two-Way Tables and Independence</span>
</span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Setting up two-way tables with categorical
variables</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Delving into marginal, joint, and conditional
distributions</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Checking for independence and dependence</span></span>
</p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Having perspective on the results of two-way
tables</span></span></p></blockquote><div
class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>C</span><span><span class="calibre16">
<span class="italic">ategorical variables</span></span></span>
<span> place individuals into groups based on certain
```

characteristics, behaviors, or outcomes, such as whether you ate breakfast this morning (yes, no) or political affiliation (Democrat, Republican, Independent, “other”). Oftentimes people look for relationships between two categorical variables; hardly a day goes by that you don’t hear about another relationship that’s reported to have been found.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are just a few examples I found on the Internet recently:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Dog owners are more likely to take their animal to the vet than cat owners.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Heavy use of social-networking Web sites in teens is linked to depression.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Children who play more video games do better in science classes.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">With all this information being given to you about variables that are related, how do you decide what to believe? For example, does heavy use of social-networking Web sites cause depression, or is it the other way around? Or perhaps a third variable out there is related to both of them, such as problems in the home.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, you see how to organize and analyze data from two categorical variables. You find out how to use proportions to make comparisons and look at overall patterns and how to check for independence of two categorical variables. You see how to describe dependent relationships appropriately and to evaluate results claiming to indicate cause-and-effect relationships, making predictions, and/or projecting their results to a population.
</p></blockquote>
<p id="a341" class="calibre6">Organizing a Two-Way Table
</p>

<blockquote class="calibre9"><p class="calibre14">To explore links between two
categorical variables, you first need to organize the data
that’s been collected, and a table is a great way to do that. A
two-
way table classifies individuals
into groups based on the outcomes of two categorical variables
(for example, gender and opinion).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose your local community developers
are building a campground, and they’ve decided pets will be
allowed as long as they’re on a leash. They are now trying to
decide whether the campground should have a separate section
for pets. You have a hunch that non–pet campers in the area may
be more in favor of a separate pet area than pet campers, so
you decide to find out what the members of the camping
community think. You randomly select 100 campers from the local
area and conduct a pet camping survey, recording each person’s
opinion on having a pet section (yes, no) and if they camp with
pets (yes, no). You now have a spreadsheet with 100 rows of
data, one for each person you surveyed. Each row has two pieces
of data: one column for whether the person is a pet camper
(yes, no) and one column for that person’s opinion on having a
pet section (support, oppose). Suppose the first 10 rows of
your data set look like what’s shown in Table 19-1.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 19-1" src="images/00375.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">From this small portion of your data
set, you can start to break it down yourself. For example,
looking at column 2 results, you see that half the respondents
(5 ÷ 10 = 0.50) camp with pets and the other half do not. Of
those who camp with pets (that is, of those five people who
have a yes in column 2), three of them (60%) support having a
separate section; and the same results are true for non–pet
campers. These results from these 10 campers likely don’t apply
to all 100 campers surveyed; however, if you tried to examine
the raw data from all 100 rows of this data set by hand, you
wouldn’t make much progress in seeing patterns without a lot of
hard work.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In order to get a handle on what’s
happening in a large data set when you are examining two
categorical variables, you organize your data into a two-way
table. The following sections take you through it.

</p></blockquote>
<blockquote class="calibre5"><p id="a342" class="calibre6">
Setting up the cells</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A two-way table organizes categorical data from two variables by using rows to represent one variable (such as pet camping — yes or no) and columns to represent the other variable (such as opinion on a pet section — support or oppose). Each person appears exactly once in the table.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Continuing with the camping example I start earlier in this chapter, in Table 19-2 I summarize the results from all 100 campers surveyed.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Table 19-2 has 2 ∗
 2 = 4 numbers in it. These numbers represent the
cells
 of the two-way table; each one represents an intersection of a row and column. The cell in the upper left corner of the table represents the 20 people who are pet campers supporting a pet section. In the upper right cell 10 people are pet campers opposing a pet section. In the lower left are the 55 non–pet campers who want a pet section; the 15 people in the lower right are non–pet campers opposing a pet section.</p></blockquote>
<blockquote class="calibre5"><p id="a343" class="calibre6">
Figuring the totals</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Before getting to the nitty-gritty analysis of a two-way table in the later section “Interpreting Results from a Two-Way Table,” you calculate some totals and add them to the table for later reference. You summarize each variable separately by calculating the marginal totals,
which represent the total number in each row (for the first variable) and the total number in each column (for the second variable). The <span

class="italic">marginal row totals
form an additional column on the right side of the table, and
the marginal column totals
 form an additional row on the bottom of the table.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, in Table 19-2 in the
preceding section, the marginal row total for row 1, the number
of pet campers, is 20 + 10 = 30; the marginal row total for
non–pet campers (row 2) is 55 + 15 = 70. The marginal column
total for those wanting a pet section (column 1) is 20 + 55 =
75; and the marginal column total for those not wanting a
separate section (column 2) is 10 + 15 = 25.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The
grand
total is the total of all the cells
in the table and is equal to the sample size. (Note the
marginal totals are not included in the grand total, only the
cells.) The grand total sits in the lower right-hand corner of
the two-way table. In this example, the grand total is 20 + 10
+ 55 + 15 = 100. Table 19-3 shows the marginal row and column
totals and the grand total for the pet camping survey data.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The marginal row totals always sum to
the grand total, because everyone in the survey either camps
with a pet or they don’t. In the last column of Table 19-3 you
see that 30 + 70 = 100. Similarly the marginal column totals
always sum to the grand total; everyone in the survey either
wants a pet section or they don’t; in the last row of Table 19-
3 you see 75 + 25 = 100.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table 19-3" src="images/00377.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> When organizing
a two-way table, always include the marginal totals and the
grand total. It gets you off on the right foot when analyzing
the data.</p></blockquote>
<p id="a344" class="calibre6">Interpreting Two-Way Tables
</p>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">After the two-way table is set up (with
the help of the information in the previous section), you
calculate percents to explore the data to answer your research
questions. Here are some questions of interest from the camping
data earlier in this chapter (each question will be handled in
the following sections, respectively):</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
What percentage of the campers are in favor of a pet
section?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
What percentage of the campers are pet campers who
support a pet section?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Do more non–pet campers support a pet section, compared
to pet campers?</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">The answers to these (and any other)
questions about the data come from finding and working with the
proportions, or percentages, of individuals within certain
parts of the table. This process involves calculating and
examining what statisticians call distributions.
 A distribution in the case of a two-way
table is a list of all the possible outcomes for one variable
or a combination of variables, along with their corresponding
proportions (or percentages).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, the distribution for the
pet camping variable lists the percentages of people who do and
do not camp with pets. The distribution for the combination of
the pet camping variable (yes, no) and the opinion variable
(support, oppose) lists the percentages of: 1) pet campers who
support a pet section; 2) pet campers who oppose a pet section;
3) non–pet campers who support a pet section; and 4) the non–
pet campers who oppose a pet section.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"

src="images/00006.jpg" class="calibre2"/> For any distribution, all the percentages must sum to 100%. If you’re using proportions (decimals), they must sum to 1.00. Each individual has to be somewhere, and he can’t be in more than one place at one time.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the following sections, you see how to find three types of distributions, each one helping you to answer its corresponding question in the preceding list.</p></blockquote>
<blockquote class="calibre5"><p id="a345" class="calibre6">Singling out variables with marginal distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you want to examine one variable at a time in a two-way table, you don’t look in the cells of the table, but rather in the margins. As seen in the earlier section “Figuring the totals,” the marginal totals represent the total number in each row (or column) separately. In the two-way table for the pet camping survey (refer to Table 19-3), you see the marginal totals for the pet camping variable (yes/no) in the right-hand column, and you find the marginal totals for the opinion variable (support/oppose) in the bottom row.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If you want to make comparisons between two groups (for example, pet campers versus non–pet campers), however, the results are easier to interpret if you use proportions instead of totals. If 350 people were surveyed, visualizing a comparison is easier if you’re told that 60% are in Group A and 40% are in Group B, rather than saying 210 people are in Group A and 140 are in Group B.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To examine the results of a two-way table based on a single variable, you find what statisticians call the marginal distribution for that variable. In the following sections, I show you how to calculate and graph marginal distributions.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Calculating marginal distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To find a

marginal distribution for one variable in a two-way table, you take the marginal total for each row (or column) divided by the grand total.</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
If your variable is represented by the rows (for example, the pet camping variable in Table 19-3), use the marginal row totals in the numerators and the grand total in the denominators. Table 19-4 shows the marginal distribution for the pet camping variable (yes, no).</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
If your variable is represented by the columns (for example, opinion on the pet section policy, shown in Table 19-3), use the marginal column totals for the numerators and the grand total for the denominators. Table 19-5 shows the marginal distribution for the opinion variable (support, oppose).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> In either case, the sum of the proportions for any marginal distribution must be 1 (subject to rounding). All results in a two-way table are subject to rounding error; to reduce rounding error, keep at least 2 digits after the decimal point throughout.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Graphing marginal distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">You graph a marginal distribution using either a pie chart or a bar graph. Each graph shows the proportion of individuals within each group for a single variable. Figure 19-1a is a pie chart summarizing the pet camping variable, and Figure 19-1b is a pie chart showing the breakdown of the opinion variable. You see that the results of these two pie charts correspond with the marginal distributions in Tables 19-4 and 19-5, respectively.</p>
</blockquote>

```html
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 19-1:</span><span> Pie charts showing
marginal distributions for a) pet camping variable; and b)
opinion variable.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg1901.eps"
src="images/00380.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>From the results of the two separate
marginal distributions for the pet camping and opinion
variables, you say that the majority of all the campers in this
sample are non–pet campers (70%) and the majority of all the
campers in this sample (75%) support the idea of having a pet
section.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> While marginal
distributions show us how each variable breaks down on its own,
they don’t tell us about the connection between two variables.
For the camping example, you know what percentage of all
campers support a new pet section, but you can’t distinguish
the opinions of the pet campers from the non–pet campers.
Distributions for making such comparisons are found in the
later section, “Comparing groups with conditional
distributions.”</span></span></p></blockquote>
<blockquote class="calibre5"><p id="a346" class="calibre6">
<span class="calibre21"><span class="bold"><span>Examining all
groups — a joint distribution</span></span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Story time: A certain auto manufacturer
conducted a survey to see what characteristics customers prefer
in their small pickup trucks. They found that the most popular
color for these trucks was red and the most popular option was
four-wheel drive. In response to these results, the company
started making more of their small pickup trucks red with four-
wheel drive.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Guess what? They struck out; people
weren’t buying those trucks. Turns out that the customers who
bought the red trucks were more likely to be women, and women
didn’t use four-wheel drive as often as men did. Customers who
bought the four-wheel drive trucks were more likely to be men,
and they tended to prefer black ones over red ones. So the most
popular outcome of the first variable (color) paired with the
```

most popular outcome of the second variable (options on the vehicle) doesn’t necessarily add up to the most popular combination of the two variables.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> To figure out which combination of two categorical variables contains the highest proportion, you need to compare the cell proportions (for example, the color and vehicle options together) rather than the marginal proportions (the color and vehicle option separately). The joint distribution of both variables in a two-way table is a listing of all possible row and column combinations and the proportion of individuals within each group. You use it to answer questions involving two characteristics; such as “What proportion of the voters are Democrat and female?” or, “What percentage of the campers are pet campers who support a pet section?” In the following sections, I show you how to calculate and graph joint distributions.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Calculating joint distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A joint distribution shows the proportion of the data that lies in each cell of the two-way table. For the pet camping example, the four row-column combinations are:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
All campers who camp with pets and support a pet section.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
All campers who camp with pets and oppose a pet section.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
All campers who don’t camp with pets and support a pet section.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">

All campers who don’t camp with pets and oppose a pet section.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> The key phrase in all of the proportions mentioned in the preceding list is all campers. You are taking the entire group of all campers in the survey and breaking them into four separate groups. When you see the word all, think joint distribution. Table 19-6 shows the joint distribution for all campers in the pet camping survey.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To find a joint distribution for a two-way table, you take the cell count (the number of individuals in a cell) divided by the grand total, for each cell in the table. The total of all these proportions should be 1 (subject to rounding error). </p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To get the numbers in the cells of Table 19-6, take the cells of Table 19-3 and divide by their corresponding grand total (100, in this case). Using the results listed in Table 19-6, you report the following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
20% of all campers surveyed camp with pets and support a pet section. (See the upper left-hand cell of the table.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
10% of all campers surveyed camp with pets and oppose a pet section. (See the upper right-hand cell of the table.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
55% of all campers surveyed don’t camp with pets and do

support the pet section policy. (See the lower left-hand cell
of the table.)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
15% of all campers surveyed don’t camp with pets and
oppose the pet section policy. (See the lower right-hand cell
of the table.)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Adding all the proportions shown in
Table 19-6, you get 0.20 + 0.10 + 0.55 + 0.15 = 1.00. Every
camper shows up in one and only one of the cells of the table.
</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Graphing joint
distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To graph a joint distribution from a
two-way table, you make a single pie chart with four slices,
representing each proportion of the data that falls within a
row-column combination. Groups containing more individuals get
a bigger piece of the overall pie, and hence get more weight
when all the votes are counted up. Figure 19-2 is a pie chart
showing the joint distribution for the pet camping survey data.
</p></blockquote>
<div class="calibre1">
<p class="calibre38">Figure 19-2: Pie chart showing the
joint distribution of the pet camping and opinion variables.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1902.eps"
src="images/00382.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">From the pie chart shown in Figure 19-
2, you see some results that stand out. The majority of campers
in this sample (0.55 or 55%) don’t camp with pets and support a
separate section for pets. The smallest slice of the pie
represents those campers who camp with pets and are opposed to
a separate section for pets (0.10 or 10%).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">A joint distribution gives you a
breakdown of the entire group by both variables at once and

allows you to compare the cells to each other and to the whole group. The results in Figure 19-2 show that if they were asked to vote today as to whether or not to have a pet section, when all the votes were added up, most of the weight would be placed on the opinions of non–pet campers, because they make up the majority of campers in the survey (70%, according to Table 19-4), and the pet campers would have less of a voice, because they are a smaller group (30%).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> A limitation of a joint distribution is that you can’t fairly compare two groups to each other (for example pet campers versus non–pet campers) because the joint distribution puts more weight on larger groups. The next section shows how to fairly compare the groups in a two-way table.</p></blockquote>
<blockquote class="calibre5"><p id="a347" class="calibre6">
Comparing groups with conditional distributions</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You need a different type of distribution other than a joint distribution to compare the results from two groups (for example comparing opinions of pet campers versus non–pet campers). Conditional distributions are used when looking for relationships between two categorical variables; the individuals are first split into the groups you want to compare (for example, pet campers and non–pet campers); then the groups are compared based on their opinion on a pet section (yes, no). In the following sections, I explain how to calculate and graph conditional distributions.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Calculating conditional distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To find conditional distributions for the purpose of comparison, first split the individuals into groups according to the variable you want to compare. Then for each group, take the cell count (the number of individuals in a particular cell) divided by the marginal total for that group. Do this for all the cells in that group. Now repeat for the other group, using its marginal total as the denominator and the cells within its group as the numerators. (See the earlier section “Figuring the totals” for more about marginal totals.) You now have two conditional

distributions, one for each group, and you fairly compare the results for the two groups.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">For the pet camping survey data example (earlier in this chapter), you compare the opinions of two groups: pet campers and non–pet campers; in statistical terms you want to find the conditional distributions of opinion based on the pet camping variable. That means you split the individuals into the pet camper and non–pet camper groups, and then for each group, you find the percentages of who supports and opposes the new pet section. Table 19-7 shows these two conditional distributions in table form (working off Table 19-3).</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Notice that Table 19-7 differs from Table 19-6 in the earlier section “Calculating joint distributions” in terms of how the values in the table add up. This represents the key difference between a joint distribution and a conditional distribution that allows you to make fair comparisons using the conditional distribution:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
In Table 19-6, the proportions in the cells of the entire table sum to 1 because the entire group is broken down by both variables at once in a joint distribution.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
In Table 19-7, the proportions in each row of the table sum to 1 because each group is treated separately in a conditional distribution.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre6">Graphing conditional distributions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">One effective way to graph conditional distributions is to make a pie chart for each group (for example, one for pet campers and one for non–pet campers) where each pie chart shows the results of the variable being studied (opinion: yes or no).</p></blockquote>

<blockquote class="calibre9"><p class="calibre14">Another method is to use a stacked bar
graph. A stacked bar graph is
a special bar graph where each bar has a height of 1 and
represents an entire group (one bar for pet campers and one bar
for non–pet campers). Each bar shows how that group breaks down
regarding the other variable being studied (opinion: yes or
no).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 19-3 is a stacked bar graph
showing two conditional distributions. The first bar is the
conditional distribution of opinion for the pet camping group
(row 1 of Table 19-7) and the second bar represents the
conditional distribution of opinion for the non–pet camping
group (row 2 of Table 19-7).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Using Table 19-7 and Figure 19-3, first
look at the opinions of each group. More than 50% of the pet
campers support the pet section (the exact number rounds to
67%), so you say the majority of pet campers support a pet
section. Similarly, the majority of non–pet campers (about 79%,
way more than half) support a pet section.</p>
</blockquote>
<div class="calibre1">
<p class="calibre38">Figure 19-3: Stacked bar graph
showing the conditional distributions of opinion for pet
campers and non–pet campers.</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg1903.eps"
src="images/00384.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Now you compare the opinions of the two
groups by comparing the percentage of supporters in the pet
camping group (67%) to the percentage of supporters in the non-
pet camping group (79%). While both groups have a majority of
supporters of the pet section, you see more of the non–pet
campers support the policy than pet campers (because 79% >
67%). By comparing the conditional distributions, you’ve found
that a relationship appears to exist between opinion and pet
camping, and your original hunch that non-pet campers in the
area may be more in favor of a separate pet area than pet
campers is correct, based on this data.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"> The difference in the results found in Figure 19-3 isn’t as large as you may have thought by looking at the joint distribution in Figure 19-2. The conditional distribution takes into account and adjusts for the number in each group being compared, while the joint distribution puts everyone in the same boat. That’s why you need conditional distributions to make fair comparisons.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> When making my conclusions regarding the pet-camping data, the operative words I use are “a relationship
 appears to exist.” The results of the pet camping survey are based on only your sample of 100 campers. To be able to generalize these results to the whole population of pet campers and non–pet campers in this community (which is really what you want to do), you need to take into account that these sample results will vary, and when they do vary, will they still show the same kind of difference? That’s what a hypothesis test will tell you (all the details are in Chapter 14).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> To conduct a hypothesis test for a relationship between two categorical variables (when each variable has only two categories, like yes/no or male/female), you either do a test for two proportions (see Chapter 15) or a Chi-square test (which is covered in my book Statistics II For Dummies,
 also published by Wiley). If one or more of your variables have more than two categories, such as Democrats/Republicans/Other, you must use the Chi-square test to test for independence in the population.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Be mindful that you may run across a report in which someone is trying to give the appearance of a stronger relationship than really exists, or trying to make a relationship less obvious by how the graphs are made. With pie charts, the sample size often is not reported, leading you to believe the results are based on a large sample when they may not be. With bar graphs, they stretch or shrink the scale to make differences appear larger

or smaller, respectively. (See Chapter 6 for more information on misleading graphs of categorical data.)</p>
</blockquote>
<p id="a348" class="calibre6">Checking Independence and Describing Dependence</p>
<blockquote class="calibre9"><p class="calibre14">The main reason researchers collect data on two categorical variables is to explore possible relationships or connections between the variables. For example, if a survey finds that more females than males voted for the incumbent president in the last election, then you conclude that gender and voting outcome are related. If a relationship between two categorical variables has been found (that is, the results from the two groups are different), then statisticians say they’re dependent.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">However, if you find that the percentage of females who voted for the incumbent is the same as the percentage of males who voted for the incumbent, then the two variables (gender and voting for the incumbent) have no relationship and statisticians say those two variables are independent. In this section, you find out how to check for independence and describe relationships found to be dependent.</p>
</blockquote>
<blockquote class="calibre5"><p id="a349" class="calibre6">Checking for independence</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Two categorical variables are independent if the percentages for the second variable (typically representing the results you want to compare, such as support or oppose) do not differ based on the first variable (typically representing the groups you want to compare, such as men versus women). You can check for independence with the methods that I cover in this section.</p></blockquote>
<blockquote class="calibre5"><p class="calibre6">Comparing the results of two conditional distributions</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"

src="images/00006.jpg" class="calibre2"/> Two categorical variables are independent if the conditional distributions are the same for all groups being compared. The variables are independent because breaking them down and comparing them by group doesn’t change the results. In the election example I introduced at the beginning of “Checking Independence and Describing Dependence,” independence means the conditional distribution for opinion is the same for the males and the females.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you do a survey of 200 voters to see if gender is related to whether they voted for the incumbent president, and you summarize your results in Table 19-8.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To see whether gender and voting are independent, you find the conditional distribution of voting pattern for the males and the conditional distribution of voting pattern for the females. If they’re the same, you’ve got independence; if not, you’ve got dependence. These two conditional distributions have been calculated and appear in rows 1 and 2, respectively, of Table 19-9. (See the earlier section “Comparing groups with conditional distributions” for details.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To get the numbers in Table 19-9, I started with Table 19-8 and divided the number in each cell by its marginal row total to get a proportion. Each row in Table 19-9 sums to 1 because each row represents its own conditional distribution. (If you’re male, you either voted for the incumbent or you didn’t — same for females.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Row 1 of Table 19-9 shows the conditional distribution of voting pattern for males. You see 40% voted for the incumbent and 60% not. Similarly, row 2 of the table shows the conditional distribution of voting pattern for females; again, 40% voted for the incumbent and 60% did not. Because these distributions are the same, men and women voted the same way; gender and voting pattern are independent.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p></blockquote>

```html
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Figure 19-4 shows the conditional
distributions of voting pattern for males and females using a
graph called a stacked bar chart. Because the bars look exactly
alike, you conclude that gender and voting pattern are
independent.</span></span></p></blockquote>
<div class="calibre1">
<p class="calibre38"><span class="calibre35"><span
class="bold">Figure 19-4:</span><span> Bar graph showing the
conditional distributions of voting pattern for males versus
females.</span></span></p>
</div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="9780470911082-fg1904.eps"
src="images/00387.jpg" class="calibre2"/></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> To have
independence, you don’t need the percentages within each bar to
be 50-50 (for example, 50% males in favor and 50% males
opposed). It’s not the percentages within each bar (group) that
have to be the same; it’s the percentages across the bars
(groups) that need to match (for example, 60% of males in favor
and 60% of females in favor).</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/><span> Instead of
comparing rows of a two-way table to determine independence,
you can compare the columns. In the voting example you’d be
comparing the gender breakdowns for the group who voted for the
incumbent to the gender breakdowns for the group who didn’t
vote for the incumbent. The conclusion of independence would be
the same as what you found previously, although the percentages
you calculate would be different.</span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold"><span>Comparing marginal
and conditional to check for independence</span></span></span>
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Another way to check for independence
is to see whether the marginal distribution of voting pattern
(overall) equals the conditional distribution of voting pattern
for each of the gender groups (males and females). If these
distributions are equal, then gender doesn’t matter. Again,
gender and voting pattern are independent.</span></span></p>
</blockquote>
```

<blockquote class="calibre9"><p class="calibre14">Looking at the voting pattern example, you find the conditional distribution of voting pattern for the males (first bar in Figure 19-4) is 40% yes and 60% no. To find the marginal (overall) distribution of voting pattern (males and females together), take the marginal column totals in the last row of Table 19-8 (80 yes and 120 no) and divide through by 200 (the grand total). You get 80 ÷ 200 = 0.40 or 40% yes, and 120 ÷ 200 = 0.60 or 60% no. (See the section “Calculating marginal distributions” earlier in this chapter for more explanation.) The marginal distribution of overall voting pattern matches the conditional distribution of voting pattern for males, so voting pattern is independent of gender.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Here’s where a small table with only two rows and two columns cuts you a break. You have to compare only one of the conditionals to the marginal because you have only two groups to compare. If the voting pattern for the males is the same as the overall voting pattern, then the same will be true for the females. To check for independence when you have more than two groups, you use a Chi-square test (discussed in my book Statistics II For Dummies, published by Wiley).
</p></blockquote>
<blockquote class="calibre5"><p id="a350" class="calibre6">
Describing a dependent relationship</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Two categorical variables are
dependent if the conditional distributions are different for at least two of the groups being compared. In the election example from the previous section, the groups are males and females, and the variable being compared is whether the person voted for the incumbent president.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Dependence in this case means knowing that the outcome of the first variable does affect the outcome of the second variable. In the election example, if dependence had been found, it would mean that males and females didn’t have the same voting pattern for the incumbent (for example, more males voting for the incumbent than females). (Pollsters use this kind of data to help steer their campaign strategies.)
</p></blockquote>

<blockquote class="calibre9"><p class="calibre22"> Other ways of saying two variables are dependent are to say they are related, or associated. However, statisticians don’t use the term correlation to indicate relationships between categorical variables. The word correlation in this context applies to the linear relationship between two numerical variables (such as height and weight), as seen in Chapter 18. (This mistake occurs in the media all the time, and it drives us statisticians crazy!)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s an example to help you better understand dependence: A recent press release put out by The Ohio State University Medical Center caught my attention. The headline said that aspirin can prevent polyps in colon-cancer patients. Having had a close relative who succumbed to this disease, I was heartened at the prospect that researchers are making progress in this area and decided to look into it.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The researchers studied 635 colon-cancer patients; they randomly assigned approximately half of them to an aspirin regimen (317 people) and the other half to a placebo (fake pill) regimen (318 people). They followed the patients to see which ones developed subsequent polyps and which did not. The data from the study are summarized in Table 19-10.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Comparing the results in the rows of Table 19-10 to check for independence means finding the conditional distribution of outcomes (polyps or not) for the aspirin group and comparing it to the conditional distribution of outcomes for the placebo group. Making these calculations, you find that 54 ÷ 317 = 17% of patients in the aspirin group developed polyps (the rest, 83%, did not), compared to 86 ÷ 318 = 27% of the placebo group who developed subsequent polyps (the rest, 73%, did not).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because the percentage of patients

developing polyps is much smaller for the aspirin group
compared to the placebo group (17% versus 27%), a dependent
relationship appears to exist between aspirin-taking and the
development of subsequent polyps among the colon-cancer
patients in this study. (But does it carry over to the
population? You find out in the section “Projecting from sample
to population” later in this chapter.)</p>
</blockquote>
<p id="a351" class="calibre6">Cautiously Interpreting Results
</p>
<blockquote class="calibre9"><p class="calibre14">It’s easy to get carried away when a
relationship between two variables has been found; you see this
happen all the time in the media. For example, a study reports
that eating eggs doesn’t affect your cholesterol as once
thought; in the details of the report you see the study was
conducted on a total of 20 men who were all in excellent
health, on low-fat diets, who exercised several times a week.
Ten men in good health ate two eggs a day and their cholesterol
didn’t change much, compared to ten men who didn’t eat two eggs
per day. Do these results carry over to the entire population?
Can’t tell — the subjects in the study don’t represent the rest
of us. (See Chapter 17 for the scoop on evaluating
experiments.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this section, you see how to put the
results from a two-way table into proper perspective in terms
of what you can and can’t say and why. This basic understanding
gives you the ability to critically evaluate and make decisions
about results presented to you (not all of which are correct).
</p></blockquote>
<blockquote class="calibre5"><p id="a352" class="calibre6">
Checking for
legitimate cause and effect</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Researchers studying two variables
often look for links that indicate a cause-and-effect
relationship. A cause-and-effect relationship
 between two categorical variables means as you
change the value of one variable and all else remains the same,
it causes a change in the second variable — for example, if
being on an aspirin regimen decreases the chance of developing
subsequent polyps in colon-cancer patients.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span

class="calibre15">However, just because two variables are found to be related (dependent) doesn’t mean they have a cause-and-effect relationship. For example, observing that people who live near power lines are more likely to visit the hospital in a year’s time due to illness doesn’t necessarily mean the power lines caused the illnesses.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> The most effective way to conclude a cause-and-effect relationship is by conducting a well-designed experiment (where possible). All the details are laid out in Chapter 17, but I touch on the main points here. A well-designed experiment meets the following three criteria:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It minimizes bias (systematic favoritism of subjects or outcomes).</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It repeats the experiment on enough subjects so the results are reliable and repeatable by another researcher.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
It controls for other variables that may affect the outcome that weren’t included in the study.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">In the earlier section “Describing a dependent relationship,” I discuss a study involving the use of aspirin to prevent polyps in cancer patients. Because of the way the data was collected for this study, you can be confident about the conclusions drawn by the researchers; this study was a well-designed experiment, according to the criteria established in Chapter 17. To avoid problems, the researchers in this study did the following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Randomly chose who took the aspirin and who received a fake pill</p></blockquote><div
class="calibre19"> </div>

<blockquote class="calibre9"><p class="calibre18">
Had large enough sample sizes to obtain accurate information</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Controlled for other variables by conducting the experiment on patients in similar situations with similar backgrounds</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Because their experiment was well-designed, the researchers concluded that a cause-and-effect relationship was found for the patients in this study. The next test is to see whether they can project these results to the population of all colon-cancer patients. If so, they are truly entitled to the headline “Aspirin Prevents Polyps in Colon-Cancer Patients.” The next section walks you through the test.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Whether two related variables are found to be causally associated depends on how the study was conducted. A well-designed experiment is the most convincing way to establish cause and effect. In cases where an experiment would be unethical (for example, proving that smoking causes lung cancer by forcing people to smoke), a mountain of convincing observational studies (where you collect data on people who smoke and people who don’t) would be needed to show that an association between two variables crosses over into a cause-and-effect relationship.</p>
</blockquote>
<blockquote class="calibre5"><p id="a353" class="calibre6">
Projecting from sample to population</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In the aspirin/polyps experiment discussed in the earlier section “Describing a dependent relationship,” I compare the percentage of patients developing subsequent polyps for the aspirin group versus the non-aspirin group and got the results 17% and 27%, respectively. For this sample, the difference is quite large, so I’m cautiously optimistic that these results would carry over to the population of all cancer patients. But what if the numbers were closer, such as 17% and 20%? Or 17% compared to 19%? How

different do the proportions have to be in order to signal a meaningful association between the two variables?
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Percentages compared using data from your sample reflect relationships within your sample. However, you know that results change from sample to sample. To project these conclusions to the population of all colon-cancer patients (or any population being studied), the difference in percentages found by the sample has to be statistically significant.
Statistical significance means even though you know results will vary, even taking that variation into account it’s very unlikely the differences were due to chance. That way, the same conclusion about a relationship can be made about the whole population, not just for a particular data set.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">I analyzed the data from the aspirin/polyps study using a hypothesis test for the difference of two proportions (found in Chapter 15). The proportions being compared were the proportion of patients taking aspirin who developed subsequent polyps and the proportion of patients not taking aspirin who developed subsequent polyps. Looking at these results, my p-value is less than 0.0024. (A p-value measures how likely you were to have gotten the results from your sample if the populations really had no difference; see Chapter 14 to get the scoop on p-values.)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Because this p
-value is so small, the difference in proportions between the aspirin and non-aspirin groups is declared to be statistically significant, and I conclude that a relationship exists between taking aspirin and developing fewer subsequent polyps.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> You can’t make conclusions about relationships between variables in a population based only on the sample results in a two-way table.

You must take into account the fact that results change from sample to sample. A hypothesis test gives limits for how different the sample results can be to still say the variables are independent. Beware of conclusions based only on sample data from a two-way table.</p></blockquote>
<blockquote class="calibre5"><p id="a354" class="calibre6">
Making prudent predictions</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A common goal of research (especially medical studies) is to make predictions, recommendations, and decisions after a relationship between two categorical variables is found. However, as a consumer of information, you have to be very careful when interpreting results; some studies are better designed than others.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The colon-cancer study from the previous section shows that patients who took aspirin daily had a lower chance of developing subsequent polyps (17% compared to 27% for the non-aspirin group). Because this was a well-designed experiment and the hypothesis test for generalizing to the population was significant, making predictions and recommendations for the population of colon-cancer patients based on these sample results is appropriate. They’ve indeed earned the headline of their press release: “Aspirin Prevents Polyps in Colon-Cancer Patients.”</p>
</blockquote>
<blockquote class="calibre5"><p id="a355" class="calibre6">
Resisting the urge to jump to conclusions</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Try not to jump to conclusions when you hear or see a relationship being reported regarding two categorical variables. Take a minute to figure out what’s really going on, even when the media wants to sweep you away with a dramatic result.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">For example, as I write this, a major news network reports that men are 40% more likely to die from cancer than women. If you’re a man, you may think you should panic. But when you examine the details, you find something different. Researchers found that men are much less likely to go to the doctor than women, so by the time cancer is found, it’s more advanced and difficult to treat. As a result, men were more likely to die of cancer after its diagnosis. (They

aren’t necessarily more likely to get
 cancer; that’s for a different study.) This study was
meant to promote early detection as the best protection and
encourage men to keep their annual checkups. The message would
have been clearer had the media reported it correctly (but
that’s not as exciting or dramatic).</p>
</blockquote>
</div>
</div> <div class="mbppagebreak" id="a356"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a350" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a356" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a351" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a357" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a352" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a353" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a343" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a342" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a341" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a340" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a347" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a346" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a345" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a344" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a354" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a349" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a348" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">

    </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a360">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Part VI</span>
</span></span></span></p>
<p id="a358" class="calibre10"><span class="calibre11"><span
class="bold"><span>The Part of Tens</span></span></span></p>
<div class="calibre13">
<img alt="9780470911082 pp0601.eps" src="images/00389.jpg"
class="calibre2"/>
</div>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre25"><span class="bold"><span>In this part . . .
</span></span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15">W<span>here would a statistics book be
without some statistics of its own? This part contains ten
methods for being a statistically savvy sleuth and ten tips for
boosting your score on a statistics exam. You can use this
quick, concise reference to help critique or design a survey,
detect common statistical abuses, and ace your introductory
statistics course.</span></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a359"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a358" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a359" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a360" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a377">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 20</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a361" class="calibre10"><span class="calibre11"><span
class="bold"><span>Ten Tips for the Statistically Savvy
Sleuth</span></span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Recognizing common statistical mistakes made by
researchers and the media</span></span></p></blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Avoiding mistakes when doing your statistics</span>
</span></p></blockquote><div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>his book is not only
about understanding the statistics that you come across in the
media and in your workplace; it’s even more about digging
deeper to examine whether those statistics are correct,
reasonable, and fair. You have to be vigilant — and a bit
skeptical — to deal with today’s information explosion, because
many of the statistics you find are wrong or misleading, either
by error or by design. If you don’t critique the information
you’re consuming, in terms of its correctness, completeness,
and fairness, who will? In this chapter, I outline ten tips for
detecting common statistical mistakes made by researchers and
by the media and ways to avoid making them yourself.</span>
</span></p></blockquote>
<p id="a362" class="calibre6"><span class="calibre17"><span
```

class="bold">Pinpoint Misleading Graphs
</p>
<blockquote class="calibre9"><p class="calibre14">Most graphs and charts contain great
information that makes a point clearly, concisely, and fairly.
However, many graphs give incorrect, mislabeled, and/or
misleading information; or they simply lack important
information that the reader needs to make critical decisions
about what is being presented. Some of these shortcomings occur
by mistake; others are incorporated by design in hopes you
won’t notice. If you’re able to pick out problems with a graph
before you contemplate any conclusions, you won’t be taken in
by misleading graphs.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Figure 20-1 shows examples of four
important types of data displays: pie charts, bar graphs, time
charts, and histograms. In this section I point out some of the
ways you can be misled if these types of graphs are not made
properly. (For more information on making charts and graphs
correctly and identifying misleading ones, see Chapters 6 and
7.)</p></blockquote>
<blockquote class="calibre5"><p id="a363" class="calibre6">
Pie
charts</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Pie charts are exactly what they sound
like: circular (pie-shaped) charts that are divided into slices
that represent the percentage (relative frequency) of
individuals that fall into different groups. Groups represent a
categorical variable, such as gender, political party, or
employment status. Figure 20-1a is a pie chart showing a
breakdown of voter opinions on some issue (call it Issue 1).
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here’s how to sink your teeth into a
pie chart and test it for quality:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Check to be sure the percentages add up to 100 percent,
or close to it (any round-off error should be small).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Be careful when you see a slice of the pie called
“other”; this is the catch-all category. If the slice for

“other” is too large (larger than other slices), the pie chart is too vague. On the other extreme, pie charts with many tiny slices give you information overload.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Watch for distortions that come with the three-dimensional (“exploded”) pie charts, in which the slice closest to you looks larger than it really is because of the angle at which it’s presented.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Look for a reported total number of individuals who make up the pie chart so you can determine how big the sample was before it was divided up into slices. If the size of the data set (the number of respondents) is too small, the information isn’t reliable.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p id="a364" class="calibre6">
Bar graphs</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A bar graph is similar to a pie chart, except that instead of being in the shape of a circle that’s divided up into slices, a bar graph represents each group as a bar, and the height of the bar represents the number (frequency) or percentage (relative frequency) of individuals in that group. Figure 20-1b is a relative frequency–style bar graph showing voter opinions on some issue (call it Issue 1); its results correspond with the pie chart shown in Figure 20-1a.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">When examining a bar graph:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Check for the sample size. If the bars represent frequencies, you find the sample size by summing them up; if the bars represent relative frequencies, you need the sample size to know how much data went into making the graph.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">

Consider the units being represented by the height of the bars and what the results mean in terms of those units. For example, are they showing the total number of crimes, or the crime rate (also known as total number of crimes per capita)?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Evaluate the starting point of the axis where the counts (or percents) are shown, and watch for the extremes: If the heights of the bars fluctuate from 200 to 300 but the counts axis starts at 0, the heights of the bars won’t look much different. However, if the starting point on the counts axis is 200, you are basically chopping off the bottoms of all the bars, and what differences remain (ranging from 0 to 100) will look more dramatic than they should.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Check out the range of values on the axis where the counts (or percents) are shown. If the heights of the bars range from 6 to 108 but the axis shows 0 to 500, the graph will have a great deal of white space and differences in the bars become hard to distinguish. However, if the axis goes from 5 to 110 with almost no breathing room, the bars will be stretched to the limit, making differences between groups look larger than they should.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p id="a365" class="calibre6">
Time charts</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A time chart shows how a numerical variable changes over time (for example, stock prices, car sales, or average temperature). Figure 20-1c is an example of a time chart showing the percentage of yes voters from 2002 to 2010, in 2-year increments.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are some issues to watch for with time charts:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Watch the scale on the vertical (quantity) axis as well as the horizontal (timeline) axis; results can be made to look more or less dramatic than they actually are by simply changing the scale.</p></blockquote><div

class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Take into account the units being portrayed by the chart
and be sure they are equitable for comparison over time; for
example, are dollar amounts being adjusted for inflation?
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Beware of people trying to explain why a trend is
occurring without additional statistics to back themselves up.
A time chart generally shows what
 is happening. Why
 it’s happening is another story!</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Watch for situations in which the time axis isn’t marked
with equally spaced jumps. This often happens when data are
missing. For example, the time axis may have equal spacing
between 2001, 2002, 2005, 2006, 2008 when it should actually
show empty spaces for the years in which no data are available.
</p></blockquote><div class="calibre19"> </div>
<div class="calibre1">
<p class="calibre38">Figure 20-1: Four types of graphs: a)
pie chart; b) bar graph; c) time chart; and d) histogram.
</p>
</div>
<blockquote class="calibre9"><p class="calibre30"><img alt="9780470911082-fg2001.eps"
src="images/00390.jpg" class="calibre2"/></p>
</blockquote>
<blockquote class="calibre5"><p id="a366" class="calibre6">

Histograms</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">A
histogram is a
graph that breaks the sample into groups according to a
numerical variable (such as age, height, weight, or income) and
shows either the number of individuals (frequency) or the
percentage of individuals (relative frequency) that fall into

each group. Figure 20-1d is a frequency style histogram showing the ages of voters in a certain election.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Some items to watch for regarding histograms include the following:</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre18">
Watch the scale used for the vertical (frequency/relative frequency) axis, looking especially for results that are exaggerated or played down through the use of inappropriate scales.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Check out the units on the vertical axis to see whether they report frequencies or relative frequencies; if they’re relative frequencies, you need the sample size to determine how much data you’re looking at.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Look at the scale used for the groupings of the numerical variable on the horizontal axis. If the groups are based on small intervals (for example, 0–2, 2–4, and so on), the heights of the bars may look choppy and overly volatile. If the groups are based on large intervals (for example, 0–100, 100–200, and so on), the data may give a smoother appearance than is realistic. </p></blockquote><div class="calibre19"> </div>
<p id="a367" class="calibre6">Uncover Biased Data</p>
<blockquote class="calibre9"><p class="calibre14">Bias in statistics is the result of a systematic error that either overestimates or underestimates the true value. For example, if I use a ruler to measure plants and that ruler is 1/2-inch short, all of my results are biased; they’re systematically lower than their true values.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Here are some of the most common sources of biased data:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"

class="calibre2"/>
Measurement instruments may be systematically off. For
example, a police officer’s radar gun may say you were going 76
miles per hour when you know you were only going 72 miles per
hour. Or a badly adjusted scale may always add 5 pounds to your
weight.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The way the study is designed can create bias. For
example, a survey question that asks, “Have you
ever
 disagreed with the government?” will overestimate
the percentage of people who are generally unhappy with the
government. (See Chapter 16 for ways to minimize bias in
surveys.)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The sample of individuals may not represent the
population of interest — for example, examining student study
habits by only going to the campus library. (See more in the
section, “Identify Non-Random Samples” later in this chapter.)
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Researchers aren’t always objective. Suppose in a drug
study one group of patients is given a sugar pill and the other
group is given the real drug. If the researchers know who
received the real drug, they may inadvertently pay more
attention to those patients to see if it’s working; they may
even project results onto the patients (such as saying, “I bet
you’re feeling better, aren’t you?”). This creates a bias in
favor of the drug. (See Chapter 17 for more information on
setting up good experiments.)</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> To spot biased
data, examine how the data were collected. Ask questions about
the selection of the participants, how the study was conducted,
what questions were used, what treatments (medications,
procedures, therapy, and so on) were given (if any) and who
knew about them, what measurement instruments were used and how
they were calibrated, and so on. Look for systematic errors or

favoritism, and if you see too much of it, ignore the results.
</p></blockquote>
<p id="a368" class="calibre6">Search for a Margin of Error
</p>
<blockquote class="calibre9"><p class="calibre14">The word error
 has a somewhat negative connotation, as if an
error is something that is always avoidable. In statistics,
that’s not always the case. For example, a certain amount of
what statisticians call
sampling error
will always occur whenever someone tries to estimate a
population value using anything other than the entire
population. Just the act of selecting a sample from the
population means you leave out certain individuals, and that
means you’re not going to get the precise, exact population
value. No worries, though. Remember that statistics means never
having to say you’re certain — you have to only get close. And
if the sample is large enough, the sampling error will be small
(assuming it’s good data of course).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">To evaluate a statistical result, you
need a measure of its accuracy — typically through the margin
of error. The margin of error tells you how much the researcher
expects her results to vary from sample to sample. (For more
information on margin of error, see Chapter 12.) When a
researcher or the media fail to report the margin of error,
you’re left to wonder about the accuracy of the results, or
worse, you just assume that everything is fine, when in many
cases, it’s not.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> When looking at
statistical results in which a number is being estimated (for
example, the percentage of all Americans who think the
president is doing a good job), always check for the margin of
error. If it’s not included, ask for it! (Or if given enough
other pertinent information, you can calculate the margin of
error yourself using the formulas in Chapter 13.)
</p></blockquote>
<p id="a369" class="calibre6">Identify Non-Random Samples
</p>
<blockquote class="calibre9"><p class="calibre14">If you’re trying to study a population

but you can only study a sample of individuals from it, how can you ensure that your sample represents the population? The most important criteria is to select your sample in a random fashion; that is, to take a random sample. You know a sample is random if it had the same chance of being selected as every other possible sample of the same size did. (It’s like pulling names from a hat.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Many surveys aren’t based on random samples, however. For example, TV polls asking viewers to “call us with your opinion” don’t represent random samples. In fact they don’t represent samples at all; when you take a sample, you select individuals from the population; for call-in polls, the individuals select themselves.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Experiments (particularly medical studies) typically can’t involve a random sample of individuals, for ethical reasons. You can’t call someone and say, “You were chosen at random to participate in a sleep study. You’ll need to come down to our lab tomorrow and stay there for two nights.” Such types of experiments are conducted using subjects that volunteer to participate — they’re not randomly selected first.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> But even though you can’t randomly select the subjects (participants) for your experiment, you can still get valid results if you incorporate the randomness in a different way — by randomly assigning the subjects to the treatment group and the control group. If the groups were assigned at random, they have a good chance of being very similar, except for what treatment they received. That way, if you do find a large enough difference in the outcomes of the groups, you can attribute those differences to the treatment, rather than to other factors.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Before making any decisions about statistical results from a survey, look to see how the sample of individuals was selected. If the sample wasn’t selected randomly, take the results with a grain of salt (see Chapter 16). If you’re looking at the results of an experiment, find out whether the subjects were randomly assigned to the treatment and control groups; if not, ignore

the results (see Chapter 17).</p></blockquote>
<p id="a370" class="calibre6">Sniff Out Missing Sample Sizes
</p>
<blockquote class="calibre9"><p class="calibre14">Both the quality and quantity of
information is important in assessing how accurate a statistic
will be. The more good data that goes into a statistic, the
more accurate that statistic will be. The quality issue is
tackled in the section “Uncover Biased Data” earlier in this
chapter. When the quality has been established, you need to
assess the accuracy of the information, and for that you need
to look at how much information was collected (that is, you
have to know the sample size).</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Small sample sizes make results less
accurate (unless your population was small to begin with). Many
headlines aren’t exactly what they appear to be when the
details reveal a study that was based on a small sample.
Perhaps even worse, many studies don’t even report the sample
size at all, which should lead you to be skeptical of the
results. (For example, an old chewing gum ad said, “Four out of
five dentists surveyed recommend [this gum] for their patients
who chew gum.” What if they really did ask only five dentists?)
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Don’t think
about this too much, but according to statisticians (who are
picky about precision), 4 out of 5 is much different than 4,000
out of 5,000, even though both fractions equal 80 percent. The
latter represents a much more precise (repeatable) result
because it’s based on a much higher sample size. (Assuming it’s
good data, of course.) If you ever wondered how math and
statistics are different, here’s your answer! (Chapter 12 has
more on precision.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">However, more data isn’t always better
data — it depends on how well the data were collected (see
Chapter 16). Suppose you want to gather the opinions of city
residents on a city council proposal. A small random sample
with well-collected data (such as a mail survey of a small
number of homes chosen at random from a city map) is much
better than a large non-random sample with poorly collected
data (for example, posting a Web survey on the city manager’s
Web site and asking for people to respond).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Always look for
the sample size before making decisions about statistical
information. The smaller the sample size, the less precise the
information. If the sample size is missing from the article,
get a copy of the full report of the study, contact the
researcher, or contact the journalist who wrote the article.
</p></blockquote>
<p id="a371" class="calibre6">Detect Misinterpreted Correlations
</p>
<blockquote class="calibre9"><p class="calibre14">Everyone wants to look for connections
between variables; for example, what age group is more likely
to vote Democrat? If I take even more vitamin C, am I even less
likely to get a cold? How does staring at the computer all day
affect my eyesight? When you think of connections or
associations between variables, you probably think of
correlation. Yes, correlation is one of the most commonly used
statistics — but it’s also one of the most misunderstood and
misused, especially throughout the media.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Some important points about correlation
are as follows (see Chapter 18 for all the additional
information):</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The
statistical definition of correlation
(denoted by

r) is the measure of
strength and direction of the linear relationship between two
numerical variables. A correlation
tells you whether the variables increase together or go in
opposite directions and the extent to which the pattern is
consistent across the data set.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
The
statistical term <span

class="calibre16">correlation
 is only used in the
context of two numerical variables (such as height and weight).
 It does not apply to two
categorical variables (such as political party and gender).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"> For example, voting pattern and gender
may be related, but using the word correlated
to describe their relationship isn’t “sc”
(statistically correct, get it?). You can say two categorical
variables are associated.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If a strong
correlation and scatterplot exist between two numerical
variables, you should be able to draw a straight line through
the points, and the points should lie close to the line.
 If a line doesn’t fit the data well, the
variables likely won’t have a strong correlation (
r
), and vice versa. (See Chapter 18 for information
on line-fitting, also known as linear regression.
)</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> A weak
correlation implies that a linear relationship doesn’t exist
between the two variables, but this doesn’t necessarily mean
the variables aren’t related at all. They may have some other
type of relationship besides a linear relationship. For
example, bacteria multiply at an exponential rate over time
(their numbers explode, doubling faster and faster).
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Correlation
doesn’t automatically mean cause and effect.
 For example, suppose Susan reports based on her
observations that people who drink diet soda have more acne

than people who don’t. If you’re a diet soda drinker, don’t break out just yet! This correlation may be a freak coincidence that only happened to the people she observed. At most, it means more research needs to be done (beyond observation) in order to draw any connections between diet soda and acne. (Susan can read Chapter 17 to find out how to design a good experiment.)</p></blockquote><div class="calibre19"> </div>
<p id="a372" class="calibre6">Reveal Confounding Variables</p>
<blockquote class="calibre9"><p class="calibre14">A confounding variable is a variable that isn’t included in a study but whose influence can affect the results and create confusing (confounding) conclusions. For example, suppose a researcher reports that eating seaweed helps you live longer, but when you examine the study, you find out that it was based on a sample of people who regularly eat seaweed in their diets and are over the age of 100. When you read the interviews of these people, you discover some of their other secrets to long life (besides eating seaweed): They slept an average of 8 hours a day, drank a lot of water, and exercised every day. So did the seaweed cause them to live longer? You can’t tell, because several confounding variables (exercise, water consumption, and sleeping patterns) may also have contributed.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The best way to control for confounding variables is to conduct a well-designed experiment (see Chapter 17), which involves setting up two groups that are alike in as many ways as possible, except that one group receives a specified treatment and the other group receives a control (a fake treatment, no treatment, or a standard, non-experimental treatment).You then compare the results from the two groups, attributing any significant differences to the treatment (and to nothing else, in an ideal world).</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> This seaweed study wasn’t a designed experiment; it was an observational study. In observational studies, no control for any variables exists; people are merely observed, and information is recorded. Observational studies are great for surveys and polls, but not for showing cause-and-effect relationships,

because they don’t control for confounding variables. A well-designed experiment provides much stronger evidence.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">If doing an experiment is unethical (for example, showing smoking causes lung cancer by forcing half of the subjects in the experiment to smoke ten packs a day for 20 years while the other half of the subjects smoke nothing), then you must rely on mounting evidence from many observational studies over many different situations, all leading to the same result. (See Chapter 17 for all the details on designing experiments.)</p></blockquote>
<p id="a373" class="calibre6">Inspect the Numbers</p>
<blockquote class="calibre9"><p class="calibre14">Just because a statistic appears in the media doesn’t mean it’s correct. In fact, errors appear all the time (by mistake or by design), so stay on the lookout for them. Here are some tips for spotting botched numbers:
</p></blockquote>
<blockquote class="calibre9"><p class="calibre18">
Make sure everything adds up to what it’s reported to.
 With pie charts, be sure all the percentages add up to 100 percent (subject to a small amount of rounding error).</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Double-check even the most basic of calculations. For example, a pie chart shows that about 83.33 percent of Americans are in favor of an issue, but the accompanying article reports “7 out of every 8” Americans are in favor of the issue. Are these statements saying the same thing? No; 7 divided by 8 is 87.5 percent — if you want 83.33 percent, it’s 5 out of 6.</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Look for the response rate of a survey; don’t just be happy with the number of participants. (The response rate is the number of people who responded divided by the total

number of people surveyed times 100 percent.) If the response rate is much lower than 50 percent, the results may be biased, because you don’t know what the non-respondents would have said. (See Chapter 16 for the full scoop on surveys and their response rates.)</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18">
Question the type of statistic used, to determine whether it’s appropriate. For example, suppose the number of crimes went up, but so did the population size. Instead of reporting the number of crimes, the media need to report the crime rate (number of crimes per capita).</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"> Statistics are based on formulas that take the numbers you give them and crunch out what you ask them to crunch out. The formulas don’t know whether the final answers are correct or not. The people behind the formulas should know better, of course. Those who don’t know better will make mistakes; those who do know better might fudge the numbers anyway and hope you don’t catch on. You, as a consumer of information (also known as a certified skeptic), must be the one to take action. The best policy is to ask questions.</p></blockquote>
<p id="a374" class="calibre6">Report Selective Reporting
</p>
<blockquote class="calibre9"><p class="calibre14">You cannot credit studies in which a researcher reports his one statistically significant result but fails to mention the reports of his other 25 analyses, none of which came up significant. If you had known about all the other analyses, you may have wondered whether this one statistically significant result is truly meaningful, or simply due to chance (like the idea that a monkey typing randomly on the typewriter would eventually write Shakespeare). It’s a legitimate question.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">The misleading practice of analyzing data until you find something is what statisticians call data snooping or data fishing.
 Here’s an example: Suppose Researcher Bob

wants to figure out what causes first graders to argue with each other so much in school (he must not be a parent or he wouldn’t even try to touch this one!). He sets up a study in which he observes a classroom of first graders every day for a month and records their every move. He gets back to his office, enters all his data, hits a button that asks the computer to perform every analysis known to man, and sits back in his chair eagerly awaiting the results. After all, with all this data he’s bound to find something.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">After poring through his results for several days, he hits pay dirt. He runs out of his office and tells his boss he’s got to put out a press release saying a ground-breaking study finds that first graders argue most when 1) the day of the week ends in the letter y
 or 2) when the goldfish in their classroom aquarium swims through the hole in its sunken pirate ship. Great job, Researcher Bob! I’ve got a feeling that a month of watching a group of first graders took the edge off his data analysis skills.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> The bottom line is that if you collect enough data and analyze it long enough, you’re bound to find something, but that something may be totally meaningless or just a fluke that’s not repeatable by other researchers.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">How do you protect yourself against misleading results due to data fishing? Find out more details about the study, starting with how many tests were done in total, and how many of those tests were found to be non-significant. In other words, get the whole story if you can, so that you can put the significant results into perspective.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> To avoid being reeled in by someone’s data fishing, don’t just go with the first result that you hear, especially if it makes big news and/or seems a little suspicious. Contact the researchers and ask for more information about their data, or wait to see whether other researchers can verify and replicate their results.</p></blockquote>
<p id="a375" class="calibre6"><span class="calibre17"><span

```
class="bold"><span>Expose the Anecdote</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Ah, the anecdote — one of the strongest
influences on public opinion and behavior ever created. And one
of the least valid. An </span><span><span class="calibre16">
<span class="italic">anecdote</span></span></span><span> is a
story or result based on a single person’s experience or
situation. For example:</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The waitress who won the lottery — twice.</span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The cat that learned how to ride a bicycle.</span></span>
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The woman who lost a hundred pounds in two days on the
new miracle potato diet.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span>The celebrity who claims to have used an over-the-counter
hair color for which she is a spokesperson (yeah, right).
</span></span></p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Anecdotes make great news; the more
sensational the better. But sensational stories are outliers
from the norm of life. They don’t happen to most people.</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>You may think you’re out of reach of
the influence of anecdotes. But what about those times when you
let one person’s experience influence you? Your neighbor loves
his Internet service provider, so you try it, too. Your friend
had a bad experience with a certain brand of car, so you don’t
bother to test-drive it. Your dad knows somebody who died in a
car crash because she was trapped in the car by her seat belt,
so he decides never to wear his.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>While some decisions are okay to make
based on anecdotes, some of the more important decisions you
make should be based on real statistics and real data that come
```

from well-designed studies and careful research.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> An anecdote is
really a data set with a sample size of only one. You have no
information to compare it to, no statistics to analyze, no
possible explanations or information to go on — just a single
story. Don’t let anecdotes have much influence over you.
Instead, rely on scientific studies and statistical information
based on large random samples of individuals who represent
their target populations (not just a single situation). When
someone tries to persuade you by telling you an anecdote just
say, “Show me the data!”</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a376"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a371" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a369" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a368" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a365" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a364" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a367" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a366" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a361" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a377" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;">
<a href="#a363" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a362" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a372" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a376" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a
href="#a374" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> <a

```
href="#a373" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a375" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a370" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a392">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Chapter 21</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a378" class="calibre10"><span class="calibre11"><span
class="bold"><span>Ten Surefire Exam Score Boosters</span>
</span></span></p>
<blockquote class="calibre9"><p class="calibre24"><span
class="calibre26"><span class="bold"><span class="italic">
<span>In This Chapter</span></span></span></span></p>
</blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Getting into the zone</span></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre27"><span
class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Developing savvy strategies</span></span></p>
</blockquote>
<blockquote class="calibre5"><p border="1" class="calibre27">
<span class="calibre15"><img alt="arrow" src="images/00010.jpg"
class="calibre2"/>
<span>Preventing silly mistakes</span></span></p></blockquote>
<div class="calibre28"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>I</span><span>’ve taught more than
40,000 students in my teaching career (don’t try to guess how
old I am, it’s not polite!), and each student has taken at
least three exams for me. That makes over 120,000 exams I’ve
graded or had a hand in grading, and believe me, I’ve seen it
all. I’ve seen excellent answers, disastrous answers, and
everything in between. I’ve gotten notes from students in the
margins asking me to go easy on them because their dog ran away
and they didn’t have time to study. I’ve seen some answers that
even I couldn’t figure out. I’ve laughed, I’ve cried, and I’ve
```

beamed with pride at what my students have come up with in exam situations.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">In this chapter, I’ve put together a list of ten strategies most often used by students who do well on exams. These students are not necessarily smarter than everyone else (although you do have to know your material, of course), but they are much better prepared. As a result, they are able to handle new problems and situations without getting thrown off; they make fewer little mistakes that chip away at an exam score; and they are less likely to have that deer-in-the headlights look, not being able to start a problem. They are more likely to get the right answer (or at least get partial credit) because they are good at labeling information and organizing their work. No doubt about it — preparation is the key to success on a stat exam.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">You too can be a successful statistics student — or more successful, if you’re already doing well — by following the simple strategies outlined in this chapter. Remember, every point counts, and they all add up, so let’s start boosting your exam score right away!</p></blockquote>
<p id="a379" class="calibre6">Know What You Don’t Know, and then Do Something about It</p>
<blockquote class="calibre9"><p class="calibre14">Figuring out what you know and what you don’t know can be hard when you are taking a statistics class. You read the book and can understand all the examples in your notes, but you can’t do your homework problems. You can answer all your roommate’s statistics questions, but you can’t answer your own. You walk out of an exam thinking you did well, but when you see your grade, you are shocked.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">What’s happening here? The bottom line is, you have to be aware of what you know and what you don’t know if you want to be successful. This is a very tough skill to develop, but it’s well worth it. Students often find out what they don’t know the hard way — by losing points on exam questions. Mistakes are okay, we all make them — what matters is when you make them. If you make a mistake before the exam while you still have time to figure out what you’re doing wrong, it doesn’t cost you

anything. If you make that same mistake on an exam, it’ll cost you points.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Here’s a strategy for figuring out what you know and what you don’t know. Go through your lecture notes and place stars by any items from the notes that you don’t understand. You can also “test” yourself, as I describe later in “Yeah-yeah trap #2,” and make a list of problems that stumped you. Take your notes and list to your professor and ask him to go through the problem areas with you. Your questions will be specific enough that your professor can zoom in when he’s talking with you, give you specific information and examples, and then check to make sure you understand each idea before moving on to the next item. Meeting with your professor won’t take long; sometimes getting one question answered has a ripple effect and clears up other questions farther down on your list.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Leave no stone unturned when it comes to making sure you understand all the concepts, examples, formulas, notation, and homework problems before you walk into the exam. I always tell my students that 30 minutes with me has a potential of raising your grade by 10%, because I’m awfully good at explaining things and answering questions — and I’m probably better at it than any roommate, brother-in-law, or friend who took the class four years ago with another professor. A quick office visit with your professor is well worth your time — especially if you bring a detailed list of questions with you. If for some reason your professor is not available, see if you have access to a tutor for help.</p></blockquote>

<div border="1" class="calibre32"><blockquote class="calibre5">
<blockquote class="calibre5"><p class="calibre10">All-purpose pointers for succeeding in class</p></blockquote>
<div class="calibre33"> </div>
<blockquote class="calibre5"><p class="calibre34">Here’s some general advice my students have found helpful:</p></blockquote>
<blockquote class="calibre44"><p class="calibre45">
I know you’ve heard this before, but you really are at an advantage if you go to class every day so you have a full set of notes to review. It also ensures you didn’t miss any of the

little things that add up to big points on an exam.
</p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Don’t just write down what the professor wrote down —
that’s for amateurs. The professionals also write down anything
else he made a big deal about but didn’t write down. That’s
what separates the As from the Bs.</p>
</blockquote>
<blockquote class="calibre44"><p class="calibre45"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Do little things to stay organized while you go through
the course; you won’t get overwhelmed later when it’s crunch
time. The day I invested 5 dollars and bought a good mechanical
pencil, a good eraser, a cheap three-hole punch for my
handouts, and a tiny stapler was one of the best days of my
student life. Okay, it’ll probably cost you 10 dollars for
these items today, but trust me, it’ll be worth it!
</p></blockquote>
<blockquote class="calibre44"><p class="calibre45"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Get to your know your professor and let her get to know
you. Introducing yourself on the first day makes a big
impression; getting face time (as well as some good help) by
asking a question after class (if you have one) or stopping in
during office hours never hurts. Don’t worry about whether your
questions are silly — it’s not what level you’re at now that
counts; it’s your desire to get to the next level and do well
in the class that’s important. That’s what your professor wants
to see.</p></blockquote>
</blockquote></div><div class="calibre37"> </div>

<p id="a380" class="calibre6">Avoid “Yeah-Yeah” Traps
</p>
<blockquote class="calibre9"><p class="calibre14">What’s a “yeah-yeah” trap? It’s a term
I use when you get caught saying “Yeah-yeah, I got this; I know
this, no problem,” but then comes the exam and whoa — you
didn’t have it, you didn’t know it, and Houston, you actually
had a problem. Yeah-yeah traps are bad because they lull you
into thinking you know everything, you don’t have any
questions, and you’ll get 100% on the exam, when the truth is
you still need to resolve some issues.</p>
</blockquote>

<blockquote class="calibre9"><p class="calibre14">Although many different yeah-yeah traps
exist, I point out the two most common ones in this section and
help you avoid them. I call them (cleverly) yeah-yeah trap #1
 and
yeah-yeah trap #2
. Both of these traps are subtle, and they can sneak up
on even the most conscientious students, so if you recognize
yourself in this section, don’t feel bad. Just think how many
points you’ll be saving yourself when you get out of “yeah-
yeah” mode and into “wait a minute — here’s something I need to
get straightened out!” mode.</p></blockquote>
<blockquote class="calibre5"><p id="a381" class="calibre6">
Yeah-yeah trap
#1</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Yeah-yeah trap #1 happens when you
study by looking through your lecture notes over and over
again, saying “yeah, I get that,” “I understand that,” and
“okay, I can do that,” but you don’t actually try the problems
from scratch totally on your own. If you understand a problem
that’s already been done by someone else, it only means you
understand what that person did when they
 worked the problem. It doesn’t say anything about
whether you could have done
it on your own in an exam situation when the pressure is on and
you’re staring at a blank space where your answer is supposed
to be. Big difference!</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">I fall into yeah-yeah trap #1 too. I
read through my DVR (digital video recording) manual from
beginning to end, and it all made total sense to me. But a week
later when I went to record a movie, I had no clue how to do
it. Why not? I understood the information as I was reading
along, but I didn’t try to apply it for myself, and when the
time came I couldn’t remember how to do it.</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">Students always tell me, “If someone
sets up the problem for me, I can always figure it out.” The
problem is, almost anyone can solve a problem that’s already
been set up. In fact, the whole point is being able to set it
up, and no one is going to do that for you on an exam.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span

class="calibre15"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Avoid yeah-yeah
trap #1 by going through your notes, pulling out a set of
examples that your professor used, and writing each one on a
separate piece of paper (just the problem, not the solution).
Then mix up the papers and make an “exam” out of them. For each
problem, try to start it by writing down just the very first
step. Don’t worry about finishing the problems; just
concentrate on starting them. After you’ve done this step for
all the problems, go back into your lecture notes and see if
you started them right. (On the back of each problem, write
down where it came from in your notes so you can check your
answers faster.)</p></blockquote>
<blockquote class="calibre5"><p id="a382" class="calibre6">
Yeah-yeah trap
#2</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Yeah-yeah trap #2 is even more subtle
than yeah-yeah trap #1. A student comes into my office after
the exam and says, “Well I worked every problem in the notes, I
redid all the homework problems, I worked all the old exams you
posted, and I did great on all of them; I hardly got a single
problem wrong. But when I took the exam, I bombed it.”
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">What happened? Nine out of ten times,
students in yeah-yeah trap #2 did indeed work all those
problems, and spent hours upon hours doing so. But whenever
they got stuck and couldn’t finish a problem, they peeked at
the solutions (which they kept sitting right next to them), saw
where they went wrong, said “yeah-yeah, that was a silly
mistake — I knew that!” and continued on to finish the problem.
In the end they thought they got the problems correct all by
themselves, but on an exam they lost some (if not all) of the
points, depending on where they originally got stuck.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">So how do you avoid yeah-yeah trap #2?
By making a test run under “real” exam conditions where the
pressure is on. Here’s how:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Study as much as you need to, in whatever
manner you need to, until you are ready to test your knowledge.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 2. Sit down with a practice exam, or if one isn’t
available, make your own by choosing some problems from
homework, your notes, or the book and shuffling them up.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> Just like at a real exam, you also
need a pencil, a calculator, and any other materials you are
allowed to bring to your exam — and nothing else! Putting your
book and notes away may make you feel anxious, frustrated, or
exposed when you do a test run of an exam, but you really need
to find out what you can do on your own before you do the real
thing.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Some teachers
allow you to bring a review sheet (also
sometimes called a memory sheet or —
cringe — a cheat sheet), a sheet
of paper on which you can write any helpful information you
want, subject to limitations that your professor may give. If
your teacher allows review sheets at tests, use one for your
practice test, too.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30">3. Turn on the oven timer for however long your
exam is scheduled to last, and then get started.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Work as many problems as you can to the best
of your ability, and when you are finished (or time runs out),
put your pencil down.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 5. When your “exam” is over, get into the lotus
position and breathe in, hold it, and breathe out three times.
Then look at the solutions and grade your paper the way your
professor would.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> If you couldn’t start a problem, even

if you just forgot one little thing and you immediately recognized it when you saw the solutions — you can’t say “Yeah-yeah, I knew that; I wouldn’t make that mistake on a real exam”; you have to say “No, I couldn’t start it on my own. I would have gotten 0 points for that problem. I need to figure this out.”</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"> You don’t get a second chance on a real exam, so when you’re studying, don’t be afraid to admit when you can’t do a problem correctly on your own; just be glad you caught it, and figure out how to fix the problem so you’ll get it right next time. Go back over it in your notes, read about it in the book, ask your professor, try more problems of the same type, or ask your study buddy to quiz you on it. Also, try to see a pattern in the type of problems that you were missing points on or getting wrong altogether. Figure out why you missed what you missed. Did you read the questions too fast, which caused you to answer them incorrectly? Was it a vocabulary or a notation issue? How did your studying align with what was on the test? And so on.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Being critical of yourself is hard, and finding out you didn’t know something you thought you knew is a little scary. But if you put yourself out there and find your mistakes before they cost you points, you’ll zoom in on your weaknesses, turn them into strengths, boost your knowledge, and get a higher exam score.
</p></blockquote>
<p id="a383" class="calibre6">Make Friends with Formulas
</p>
<blockquote class="calibre9"><p class="calibre14">Many students are not comfortable with formulas (unless you are a math nerd, in which case formulas make you shout for joy). That unease is understandable — I used to be intimidated by them too (formulas, that is — not math nerds). The trouble is, you really can’t survive too long without eventually using a formula in a statistics class, so becoming comfortable with them right from the start is important. A formula tells you much more than how to calculate something. It shows the thinking process behind the calculations. For example, the big picture regarding standard deviation can be seen by analyzing its formula:
</p></blockquote>

```
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq21001.eps"
src="images/00391.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Subtracting the mean, </span><img
alt="9780470911082-eq21002.eps" src="images/00392.jpg"
class="calibre2"/><span>, from a value in the data set, </span>
<img alt="9780470911082-eq21003.eps" src="images/00393.jpg"
class="calibre2"/><span>, measures how far above or below the
mean that number is. Because you don’t want the positive and
negative differences to cancel each other out, you square them
all to make them positive (but remember that this gives you
square units). Then you add them up and divide by </span><span>
<span class="calibre16"><span class="italic">n</span></span>
</span><span> – 1, which is near to finding an average, and
take the square root to get back into original units. In a
general sense, you are finding something like the average
distance from the mean.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Stepping back even further, you can
tell from the formula that the standard deviation can’t be
negative, because everything is squared. You also know the
smallest it can be is zero, which occurs when all the data are
the same (that is, all are equal to the mean). And you see how
data that is far from the mean will contribute a larger number
to the standard deviation than data that is close to the mean.
</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>And here’s another perk. Because you
understand the formula for standard deviation now, you know
what it’s really measuring: the spread of the data around the
mean. So when you get an exam question saying “Measure the
spread around the mean,” you’ll know what to do. Bam!</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><span
class="calibre15"><span><img alt="tip.eps"
src="images/00005.jpg" class="calibre2"/></span><span> In order
to feel comfortable about formulas, follow these tips:</span>
</span></p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Get into the
right mind-set.</span></span></span><span> Think of formulas as
mathematical shorthand and nothing more. All you have to do is
be able to decipher them. Oftentimes you’re allowed to bring a
review sheet to your exam, or you’ll be given a formula sheet
```

with your exam, so you may not have to make things harder by
memorizing them.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Understand
every part of every formula. In
order for any formula to be useful, you have to understand all
its components. For example, before you can use the formula for
standard deviation, you need to know what <img
alt="9780470911082-eq21004.eps" src="images/00394.jpg"
class="calibre2"/> and <img alt="9780470911082-
eq21005.eps" src="images/00395.jpg" class="calibre2"/>
mean and what <img alt="9780470911082-eq21006.eps"
src="images/00396.jpg" class="calibre2"/> stands for.
Otherwise it’s totally useless.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Practice using
formulas from day one. Use them to
verify the calculations done in lecture or in your book. If you
get a different answer from what’s shown, figure out what you
are doing wrong. Making mistakes here is okay — you caught the
problem early, and that’s all that counts.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Whenever you
use a formula to do a problem, write it down first and then
plug in the numbers in the second step.
 The more often you write down a formula, the more
comfortable you will be using it on an exam. And if (heaven
forbid!) you copy the formula down wrong, your instructor will
be able to follow your error, which may mean some partial
credit for you!</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Chances are, if
you’ve learned some formulas in your class, you’re going to
need to use them on your exam. Don’t expect to be able to use
formulas with confidence on an exam if you haven’t practiced
with them and written them down many, many times beforehand.
Practice when the problems are easy so when they get harder you

won’t have to worry as much.</p></blockquote>
<p id="a384" class="calibre6">Make an “If-Then-How” Chart
</p>
<blockquote class="calibre9"><p class="calibre14">Quarterbacks always talk about trying
to get the game to “slow down” for them so they feel like they
have more time to think and react. You want the same thing when
you take a statistics exam. (See, you and your NFL hero really
do have something in common!) The game starts slowing down for
a quarterback when he begins to see patterns in the way the
defense lines up against him, rather than feeling like every
play brings a completely different look. Similarly for you, the
exam starts to “slow down” when the problems start falling into
categories as you read them, rather than each one appearing to
be totally different from anything you’ve ever seen before.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">To make this happen, many of my
students find help in making what I call an if-then-how chart
. An
if-then-how chart
 maps out the types of problems you are likely to run
into, strategies to solve them, and examples for quick
reference. The basic idea of the if-then-how
 chart is to say “If
the problem asks for X, then
 I solve it by doing Y, and here’s
how
.” An if-then-how chart
contains three columns:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
If: In the
if column, write down
a succinct description of what you are asked to find or do. For
example, if the problem asks you to test a claim about the
population mean (see Chapter 14 for more about claims), write
“Test a claim — population mean.” If you are asked to give your
best estimate of the population mean (Chapter 13 has the scoop

on estimates), write “Estimate population mean.”
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> Problems are
worded in different ways, because that’s how the real world
works. Pay attention to different wordings that in essence boil
down to the same problem, and add them to the appropriate place
in the if column where the
actual problem is already listed. For example, one problem may
ask you to estimate the population mean; another problem may
say, “Give a range of likely values for the population mean.”
These questions ask for the same thing, so include both in your
if column.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Then:
In your then
 column you write the exact statistical procedure,
formula, or technique you need to solve that type of problem
using the statistical lingo. For example, when your
if

column says “Test a claim —
population mean,” your
then column
should say “Hypothesis test for μ.”
When your
if statement reads
“Estimate population mean” your then
 column should read “Confidence interval for
μ.”</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><img alt="tip.eps" src="images/00005.jpg"
class="calibre2"/> To match strategies to situations,
look carefully at how the examples in your lecture notes and

your book were done and use them as your guide.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
How: In the
how column, write an
example, a formula, and/or a quick note to yourself that will
spark your mind and send you off running in the right
direction. Write whatever you need to feel comfortable (no
one’s going to see it but you, so make it your way!). For
example, suppose your
if column says
“Estimate the population mean,” and your then
column says “Confidence interval — population
mean.” In the how column, you can
write the formula.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre14">Although I just took a lot of time and
talking to walk you through it, making an if-then-how chart is
much easier done than said. Below is an example of an entry in
an if-then-how chart for the confidence interval problem I just
laid out.</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><img alt="/mt2101" src="images/00397.jpg"
class="calibre2"/></p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Using these three columns, fill in your
if-then-how chart with each different type of problem you’ve
covered in class. Don’t write down every little example; look
for patterns in the problems and boil down the number of
scenarios to a doable list.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> If-then-how
charts should be customized to your needs, so the only way it’s
going to work is if you make it yourself. No two people think
alike; what works for your friend may not work for you.
However, it might be helpful to compare your chart with a
friend’s once you are both finished, to see if you’ve left
anything out.</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"><img alt="tip.eps" src="images/00005.jpg"

class="calibre2"/> If you’re allowed to bring a review sheet to exams, I suggest putting your if-then-how chart on one side. On the other side, write down those little nuggets of information your professor gave you in lecture but didn’t write down. If you aren’t allowed to have a review sheet during the exam, call me crazy, but I’ll argue that you should still make one to study from. Making one really helps you sort out all the ideas so when you take the exam you’ll be much more clear about what to look for and how to set up and solve problems. Lots of students come out of an exam saying they didn’t even use their review sheet, and that’s when you know you’ve done a good job putting one together: When it went on the sheet, it went into your mind!</p></blockquote>
<p id="a385" class="calibre6">Figure Out What the Question Is Asking</p>
<blockquote class="calibre9"><p class="calibre14">Students often tell me that they don’t understand what a problem is asking for. That’s the million dollar question, isn’t it? And it’s not a trivial matter. Oftentimes the actual question is embedded somewhere in the language of the problem; it isn’t usually as clear as: “Find the mean of this data set.”</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> For example, a question may ask you to “interpret” a statistical result. What does “interpret” really mean? To most professors the word “interpret” means to explain in words that a nonstatistician would understand.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you are given some computer output analyzing number of crimes and number of police officers, and you are asked to interpret the correlation between them. First you pick off the number from the output that represents the correlation (say it’s –0.85); then you talk about its important features in language that is easy for others to understand. The answer I would like to see on an exam goes something like this: “The correlation between number of police officers and number of crimes is –0.85; they have a strong negative linear relationship. As the number of police officers increases, number of crimes decreases.”
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If you know what the problem is asking for, you have a better chance of actually solving it. You’ll gain confidence when you know what

you are supposed to do. On the flip side, if you don’t know
what the problem is asking, even starting it will be very hard.
Your anxiety will go up, which can affect your ability to work
other problems as well. So how do you boil down a problem to
figure out exactly what it’s asking for? Here are some tips to
follow:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Check the very
last sentence of the problem — that’s usually where the
question is located. Rather than
reading the entire problem a second (and third and fourth) time
and getting yourself all worked up, just read it once and then
focus on the end of the problem.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Practice
boiling down questions ahead of time.
 Look at all the examples from your lecture notes, your
homework problems, and problems in your textbook and try to
figure out what each problem is asking for. Eventually you’ll
start to see patterns in the way problems are worded, and
you’ll get better at figuring out what they are really asking
for.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Ask your
professor what clues you should look for, and bring example
problems with you. She will be
impressed because you are trying to figure out the big picture,
and oh, how professors love those “big picture” questions! And
after she helps you, you can add those to your if-then-how
chart (see “Make an ‘If-Then-How’ Chart”).</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Translate the
wording of the problem into a statistical statement.
 This involves labeling not only what you
are given (as discussed in the next section), but also what you
want to find.</p></blockquote><div
class="calibre19"> </div>

<blockquote class="calibre5"><p class="calibre39">For example, Professor Barb wants to give 20 percent of her students an A on her statistics exam; your job is to find the cutoff exam score for an A, and this translates to “find the score representing the 80th percentile.”</p></blockquote><div class="calibre19"> </div>
<p id="a386" class="calibre6">Label What You’re Given
</p>
<blockquote class="calibre9"><p class="calibre22"> Many students try to work problems by pushing around numbers that are given in the problem. This approach may work with easy problems, but everyone hits the wall at some point and needs more support to solve harder problems. You’ll benefit from getting into the habit of labeling everything properly — labeling is the critical connection between the if

column and the
then

column in your if-then-how chart (described earlier in this chapter). You may read a problem and know what you need to do, but without understanding how to use what you’re given in the problem, you won’t be able to solve it correctly. To really understand the numbers the problem gives you, take each one and write down what it stands for.</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Suppose you’re given the following problem to solve: “You want to use the size of a house in a certain city (in square feet) to predict its price (in thousands). You collect data on 100 randomly selected homes that have recently been sold. You find the mean price is $219,100 with standard deviation of $60,100, and you know the mean size is 1,993 square feet, with standard deviation of 349 square feet. You find the correlation between size and price for these homes is +0.90. Find the best-fitting regression line that you can use to predict house price using size.”
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Your first step is labeling everything. Knowing you use size to predict price, you figure size must be

the x variable and price
must be the y variable. You then
label the means <img alt="9780470911082-eq21008.eps"
src="images/00398.jpg" class="calibre2"/> (square feet)
and <img alt="9780470911082-eq21009.eps"
src="images/00399.jpg" class="calibre2"/> (in thousands)
respectively; the standard deviations are labeled <img
alt="9780470911082-eq21010.eps" src="images/00400.jpg"
class="calibre2"/> (square feet) and <img
alt="9780470911082-eq21011.eps" src="images/00401.jpg"
class="calibre2"/> (in thousands), respectively, and the
correlation is labeled
r = 0.90. The
sample size is n = 100. Now you can
plug your numbers into the right formulas. (See Chapter 18
regarding correlation and regression.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">When you know you have to work with a
regression line and that formulas are involved, having all the
given information organized and labeled, ready to go, is very
comforting. It’s one less thing to think about. (The problem in
this particular example is solved in the section “Make the
Connection and Solve the Problem.”) If that example doesn’t
convince you, here are six more reasons to label what you are
given in a problem:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Labeling
allows you to check your work more easily.
When you go back to check your work (as I advise
in the section “Do the Math — Twice”), you’ll quickly see what
you were thinking when you did the problem the first time.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Your professor
will be impressed. He will see your
labels and realize you at least know what the given information
stands for. That way if your calculations go haywire, you still
have a chance for partial credit.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span

```
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Labeling saves
time.</span></span></span><span> I know that writing down more
information seems like a strange way to save time, but by
labeling all the items, you can pull out the info you need in a
flash.</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><span> For example, suppose you need to do a
95% confidence interval for the population mean (using what you
know from Chapter 13) and you’re told that the sample mean is
60, the population standard deviation is 10, and the sample
size is 200. You know the formula has to involve </span><img
alt="9780470911082-eq21012.eps" src="images/00402.jpg"
class="calibre2"/><span>, </span><img alt="9780470911082-
eq21013.eps" src="images/00403.jpg" class="calibre2"/><span>,
and </span><span><span class="calibre16"><span
class="italic">n</span></span></span><span>, and you see one
that does:</span></span></p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq21014.eps"
src="images/00404.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><span> Because you’ve already labeled
everything, you just grab what you need, put it into the
formula, throw in a </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>*-value of 1.96 (the
critical value corresponding to a 95% confidence level), and
crunch it out to get the answer:</span></span></p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre5"><p class="calibre39"><span
class="calibre15"><img alt="9780470911082-eq21015.eps"
src="images/00405.jpg" class="calibre2"/></span></p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><span
class="calibre15"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
<span><span class="calibre16"><span class="bold">Labels keep
your mind organized. </span></span></span><span>You are less
likely to get buried in calculations and forget what you’re
doing if your work involves symbols and not just numbers. By
sorting out the information you are given, you’re less likely
to resort to reading the problem over and over again, raising
your anxiety level each time.</span></span></p></blockquote>
<div class="calibre19"> </div>
```

<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
You use the
labels to figure out which formula or technique you need to use
to solve a problem. For example, if
you think you need a hypothesis test but no claim is made about
the population mean, hold up. You may need a confidence
interval instead; this realization saves you precious time
because you won’t be spinning your wheels in the wrong
direction. Labels help you quickly narrow down your options.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Labeling helps
you resist the urge to just write down numbers and push them
around on the paper. More often than
not, number-pushing leads to wrong answers and less (if any)
partial credit if your answer is wrong. Your professor may not
be able to follow you, or just doesn’t want to spend all that
time trying to figure it out (sorry to say, but this happens
sometimes).</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Labeling saves
you anxiety, time, and points when you take your exam. But in
order to be successful on exam day, you need to start this
practice early on, while the problems are easy to do. Don’t
expect to suddenly be able to sort out the information on exam
day if you never did it before; it’s not gonna happen. Make it
your habit right away and you won’t freak out when you see a
new problem. You’ll at least be able to break it down into
smaller chunks, which always helps.</p>
</blockquote>
<p id="a387" class="calibre6">Draw a Picture</p>
<blockquote class="calibre9"><p class="calibre14">You’ve heard the expression “A picture
is worth a thousand words.” As a statistics professor, I say,
“A picture is worth a thousand points (or at least half the
points on a given problem).” When the given information and/or
the question being asked can be expressed in a picture form,
you should do it. Here’s why:</p></blockquote>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>

A picture can
help you see what’s going on in the problem.
 For example, if you know exam scores have a
normal distribution with mean 75 and standard deviation 5 (see
Chapter 9 for more about normal distribution), you draw a bell-
shaped curve, marking off the mean in the center and three
standard deviations on each side. You can now visualize the
scenario you’re dealing with.</p></blockquote>
<div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
You can use
the drawing to help figure out what you are trying to find.
 For example, if you need to know
the probability that Bob scored more than 70 points on the
exam, you shade in the area to the right of 70 on your drawing,
and you’re on your way.</p></blockquote><div
class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Your professor
knows that you understand the basics of the problem, increasing
your chance for partial credit. On
the other hand, someone who got the problem wrong doesn’t get
much sympathy if the professor knows drawing a simple picture
would have avoided the whole problem.</p>
</blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre18"><img alt="check.jpg" src="images/00004.jpg"
class="calibre2"/>
Students who
draw pictures tend to get more problems correct than students
who don’t. Without a picture you can
easily lose track of what’s needed, and make mistakes like
finding P(X < 70) instead of
P(X > 70), for
example. Also, checking for and spotting errors before you turn
in your exam is easier if you have a picture to look at.
</p></blockquote><div class="calibre19"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="remember.eps"
src="images/00006.jpg" class="calibre2"/> Drawing a
picture may seem like a waste of valuable time on an exam, but
it’s actually a time-saver because it gets you going in the

right direction, keeps you focused throughout the problem, and helps ensure you answer the right question. Drawing a picture can also help you analyze your final numerical answer and either confirm you've got it right, or quickly a spot and fix an error and save yourself some points. (Be sure to draw pictures while studying so they come naturally during an exam.) </p></blockquote>
<p id="a388" class="calibre6">Make the Connection and Solve the Problem</p>
<blockquote class="calibre9"><p class="calibre22"> When you've figured out what the problem is asking, you have everything labeled, and you have your pictures drawn, it's time to solve the problem. After doing the prep work, nine times out of ten you'll remember a technique you learned from class, a formula that contains the items you've labeled, and/or an example you worked through. Use or remember your if-then-how chart and you'll be on your way. (See “Make an ‘If-Then-How’ Chart” if you need more info.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Breaking down a problem means having less to think about at each step, and in a stressful exam situation where you may forget your own name, that's a real plus! (This strategy reminds me of the saying, “How do you eat an elephant? One bite at a time.”)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">In the example of using size of a home to predict its price (see the earlier section “Label What You're Given”), you know the mean and standard deviation of size, the mean and standard deviation of price, and the correlation between them; and you've labeled them all. The question asks you to find the equation of the best-fitting regression line to predict price based on size of the home; you know that means find the equation y = a + bx where x = size (square feet) and y = price (thousands of dollars), b is the slope of the

regression line, and a is the
y
-intercept. (Flip to Chapter 18 for more
about this formula.)</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Now you recognize what to do — you have
to find a and
b.
 You remember (or can find) that
those formulas are <img alt="9780470911082-eq21016.eps"
src="images/00406.jpg" class="calibre2"/> and <img
alt="9780470911082-eq21017.eps" src="images/00407.jpg"
class="calibre2"/>. Grab the numbers you’ve
labeled (<img alt="9780470911082-eq21018.eps"
src="images/00408.jpg" class="calibre2"/>), put them into
the formulas, and solve (sounds like a commercial for a frozen
dinner doesn’t it?). You find the slope is
<img alt="9780470911082-eq21019.eps" src="images/00409.jpg"
class="calibre2"/> and the y
-intercept is <img alt="9780470911082-eq21020.eps"
src="images/00410.jpg" class="calibre2"/>, so the
equation of the best-fitting regression line is
<img alt="9780470911082-eq21021.eps"
src="images/00411.jpg" class="calibre2"/>. (See Chapter
18 for the details of regression.)</p>
</blockquote>
<p id="a389" class="calibre6">Do the Math — Twice</p>
<blockquote class="calibre9"><p class="calibre14">I can still remember some of the
struggles I had way back in high school algebra. For the
longest time 3 times 2 was equal to 5 for me; this mistake (and
others like it) caused me to miss a handful of points on every
exam and homework assignment, and I just could not get past it.
One day I decided I’d had enough of losing points here and
there for silly errors, and I did something about it. From that
day on, I wrote out all of my work, step by step, and resisted
the urge to do steps in my head. When I got my final answer,
instead of moving on, I went back and checked every step, and I
did so with the mind-set that a mistake had probably slipped in
somewhere and it was my job to find it before anyone else did.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">This approach forced me to look at each
step with fresh eyes, as if I were grading someone else’s

paper. I caught more mistakes because I never skipped over a step without bothering to check it. I finally stopped thinking 3 times 2 was 5 because I caught myself in the act enough times. My exam grades went up, just because I started checking things more carefully. It reminds me of the carpenter's saying, "Measure twice, cut once." They waste a lot less wood that way. </p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> Every time you find and fix a mistake before you turn in your exam, you're getting a handful of points back for yourself. Find your errors before your professor does, and you'll be amazed how those points add up. However, remember that time is not unlimited on an exam, so try to get the problems right the first time. Labeling everything, drawing pictures, writing down formulas, and showing all your work will definitely help!
</p></blockquote>
<p id="a390" class="calibre6">Analyze Your Answers
</p>
<blockquote class="calibre9"><p class="calibre14">A very prominent statistician I know has a framed piece of paper on his office wall. It's a page of an exam he took way back when he was a student. It's got a big red circle around one of his answers, which happens to be the number 2. Why was writing the number 2 for an answer such a problem? Because the question asked him to find a probability, and probabilities are always between 0 and 1. As a result, he didn't get any points for that problem, not even partial credit. In fact, I'll bet his professor wanted to give him negative points for making such a mistake. (They really don't like it when you totally miss the boat.)</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre22"> Always take the time to check your final answer to see if it makes sense. A negative standard deviation, a probability more than 1, or a correlation of –121.23 is not going to go over well with your professor, and it will not be treated like a simple math error. It will be treated as a fundamental error in not knowing (or perhaps caring) what the result should look like.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre22"> If you know an answer you got can't possibly be right, but you cannot for the life of you figure

out where you went wrong, don’t waste any more time on it. Just write a note in the margin that says you know your answer can’t be right but you can’t figure out your error. This helps separate you from the regular Joe who found a probability of 10,524.31 (yes, I’ve seen it) and merrily moved on.
</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">By the way, you may be wondering why this world-class statistician still keeps this exam page framed on his office wall. He says it’s to keep him humble. Learn from his example and never move on to the next problem without stepping back and saying “does this answer even make sense?”</p></blockquote>
</div>
</div> <div class="mbppagebreak" id="a391"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block !important; page-break-before: always !important; break-before: always !important; white-space: pre-wrap !important">

          <a href="#a391" style="min-width: 10px !important; min-height:

10px !important; border: solid 1px !important;">   </div></body></html>

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a395">
<div class="calibre1">
<div class="calibre1">
<p class="calibre10"><span class="calibre23"><span
class="bold"><span class="underline"><span>Appendix</span>
</span></span></span></p><div class="calibre19"> </div>
<p id="a393" class="calibre10"><span class="calibre11"><span
class="bold"><span>Tables for Reference</span></span></span>
</p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>T</span><span>his appendix includes
tables for finding probabilities and/or critical values for the
three distributions used in this book: the </span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>-distribution (standard normal), the </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span>-distribution, and the binomial distribution.</span>
</span></p></blockquote>
<p class="calibre6"><span class="calibre17"><span class="bold">
<span>The Z-Table</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Table A-1 shows less-than-or-equal-to
probabilities for the </span><span><span class="calibre16">
<span class="italic">Z</span></span></span><span>-distribution;
that is, </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>
</span><span><span class="calibre40">≤</span></span><span>
</span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>) for a given
</span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>-value. (See Chapter
9 for calculating </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>-values for a normal
distribution; see Chapter 11 for calculating </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span>-values for a sampling distribution.) To use Table A-1,
```

do the following:</p></blockquote>
<blockquote class="calibre9"><p class="calibre30">1. Determine the z-value for your particular
problem.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The z-
value should have one leading digit before the decimal
point (positive, negative, or zero) and two digits after the
decimal point; for example z
 = 1.28, –2.69, or 0.13.</p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the row of the table corresponding to the
leading digit and first digit after the decimal point.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For example, if your
z
-value is 1.28, look in the “1.2” row; if
z
 = –1.28, look in the “–1.2” row.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Find the column corresponding to the second
digit after the decimal point. </p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For example, if your
z-
value is 1.28 or –1.28, look in the “.08” column.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Intersect the row and column from Steps 2 and
3.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> This number is the probability that

```
</span><span><span class="calibre16"><span
class="italic">Z</span></span></span><span> is less than or
equal to your </span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>-value. In other
words, you’ve found </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span>
</span><span><span class="calibre40">≤</span></span><span>
</span><span><span class="calibre16"><span
class="italic">z</span></span></span><span>). For example, if
</span><span><span class="calibre16"><span
class="italic">z</span></span></span><span> = 1.28, you see
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">Z</span></span></span>
<span> ≤ 1.28) = 0.8997. For </span><span><span
class="calibre16"><span class="italic">z</span></span></span>
<span> = –1.28, you see </span><span><span class="calibre16">
<span class="italic">p</span></span></span><span>(</span><span>
<span class="calibre16"><span class="italic">Z</span></span>
</span><span> ≤ –1.28) = 0.1003.</span></span></p></blockquote>
<div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-1a" src="images/00412.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-1b" src="images/00413.jpg"
class="calibre2"/></span></p></blockquote>
<p class="calibre6"><span class="calibre17"><span class="bold">
<span>The t-Table</span></span></span></p>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Table A-2 shows right-tail
probabilities for selected </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span>-distributions (see Chapter 10 for more on the </span>
<span><span class="calibre16"><span class="italic">t</span>
</span></span><span>-distribution). </span></span></p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><span>Follow these steps to use Table A-2 to
find right-tail probabilities and </span><span><span
class="calibre16"><span class="italic">p</span></span></span>
<span>-values for hypothesis tests involving </span><span><span
class="calibre16"><span class="italic">t</span></span></span>
<span> (see Chapter 15):</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
```

class="bold"> 1. Find the t-value for which you want
the right-tail probability (call it
t), and find the sample
size (for example, n)
.</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the row corresponding to the degrees of
freedom (
df
) for
your problem (for example, n
– 1). Go across that row to find the two

t-values between which your
t
 falls.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For example, if your
t
 is 1.60 and your n
 is 7, you look in the row for df
 = 7 – 1 = 6. Across that row you find your
t
 lies between t
-values 1.44 and 1.94.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span

class="bold"> 3. Go to the top of the columns containing the
two t
-values from
Step 2.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> The right-tail (greater-than)
probability for your t-value is somewhere
between the two values at the top of these columns. For
example, your t = 1.60 is between
t-values 1.44 and
1.94 (df = 6); so the right
tail probability for your
t is between
0.10 (column heading for
t = 1.44); and
0.05 (column heading for
t = 1.94).
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre22"><img alt="headsup_lewis.eps"
src="images/00007.jpg" class="calibre2"/> The row near
the bottom with Z in the
df
 column gives right-tail (greater-than)
probabilities from the
Z-distribution
(Chapter 10 shows Z’s relationship with
t).

</p></blockquote>
<blockquote class="calibre9"><p class="calibre14">Use Table A-2 to find
t*
-values (critical values) for a confidence
interval involving t (see Chapter 13):
</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span

class="calibre15"> 1. Determine the confidence level you need (as a
percentage).</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Determine the sample size (for example,
n
).
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Look at the bottom row of the table where the
percentages are shown. Find your % confidence level there.
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Intersect this column with the row
representing your degrees of freedom (
df).
 This is the t-value you need for
your confidence interval.</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> For example, a 95% confidence interval
with df=6 has
t*
=2.45. (Find 95% on the last line and go up
to row 6.)</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><img alt="/Table A-2" src="images/00414.jpg"
class="calibre2"/></p></blockquote>
<p class="calibre6">
The Binomial Table</p>
<blockquote class="calibre9"><p class="calibre14">Table A-3 shows probabilities for the
binomial distribution (see Chapter 8).</p>
</blockquote>
<blockquote class="calibre9"><p class="calibre14">To use Table A-3, do the following:

</p></blockquote>
<blockquote class="calibre9"><p class="calibre30"> 1. Find these three numbers for your particular
problem:</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> • The sample size, n
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> • The probability of success,
p
</p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> • The x-
value for which you want p
(X
=
x)</p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 2. Find the section of Table A-3 that’s devoted
to your
n.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 3. Look at the row for your
x-value and the column for
your p.
</p></blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"> 4. Intersect that row and column.
You have found p
(X<span>

```
</span><span><span class="calibre40">=</span></span><span>
</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>).</span></span></p>
</blockquote><div class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span><span class="calibre16"><span
class="bold"> 5. To get the probability of being less than,
greater than, greater than or equal to, less than or equal to,
or between two values of </span></span></span><span><span
class="calibre16"><span class="bold"><span
class="italic">X</span></span></span></span><span><span
class="calibre16"><span class="bold">, you add the appropriate
values of Table A-3 using the steps found in Chapter 8.</span>
</span></span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><span> For example, if </span><span><span
class="calibre16"><span class="italic">n</span></span></span>
<span>=10, </span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>=0.6, and you want
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>(</span><span><span
class="calibre16"><span class="italic">X</span></span></span>
<span>=9), go to the </span><span><span class="calibre16"><span
class="italic">n</span></span></span><span>=10 section, the
</span><span><span class="calibre16"><span
class="italic">x</span></span></span><span>=9 row, and the
</span><span><span class="calibre16"><span
class="italic">p</span></span></span><span>=0.6 column to find
0.04.</span></span></p></blockquote><div
class="calibre31"> </div>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-3a" src="images/00415.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-3b" src="images/00416.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-3c" src="images/00417.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre30"><span
class="calibre15"><img alt="/Table A-3d" src="images/00418.jpg"
class="calibre2"/></span></p></blockquote>
</div>
</div>  <div class="mbppagebreak" id="a394"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">
```

```
<a href="#a394" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a395" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a393" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```

```
<?xml version='1.0' encoding='utf-8'?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Statistics For Dummies, 2nd Edition</title>

  <link href="stylesheet.css" rel="stylesheet"
type="text/css"/>
<link href="page_styles.css" rel="stylesheet" type="text/css"/>
</head>
  <body class="calibre" id="a398">
<div id="a396" class="calibre1">
<div class="calibre1">
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><img alt="IFC_top" src="images/00419.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold">To access the cheat sheet
specifically for this book, go to <a
href="http://www.dummies.com/cheatsheet/statistics">www.dummies
.com/cheatsheet/statistics</a>.</span></span></p></blockquote>
<blockquote class="calibre9"><p class="calibre14"><span
class="calibre15"><img alt="IFC_bottom" src="images/00420.jpg"
class="calibre2"/></span></p></blockquote>
<blockquote class="calibre5"><p class="calibre6"><span
class="calibre7"><span class="bold">Find out "HOW" at <a
href="http://www.dummies.com">Dummies.com</a></span></span></p>
</blockquote>
</div>
</div>  <div class="mbppagebreak" id="a397"></div>
<div id="5KzRDSqba4c6hbtT4ckd63" style="display:block
!important; page-break-before: always !important; break-before:
always !important; white-space: pre-wrap !important">

<a href="#a396" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a>
<a href="#a397" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#a398" style="min-width: 10px !important; min-height:
10px !important; border: solid 1px !important;"> </a> <a
href="#5KzRDSqba4c6hbtT4ckd63" style="min-width: 10px
!important; min-height: 10px !important; border: solid 1px
!important;"> </a> </div></body></html>
```