LEARNING MADE EASY

Statistical Analysis with R

for dummies®

A Wiley Brand

Leverage R as a powerful statistical tool

Test your hypotheses and draw conclusions

Use R to give meaning to your data

Joseph Schmuller, PhD

Statistical Analysis with R

by Joseph Schmuller, PhD

for dummies®
A Wiley Brand

Statistical Analysis with R For Dummies®

Published by: **John Wiley & Sons, Inc.,** 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Published simultaneously in Canada

Library of Congress Control Number: 2017932881

ISBN: 978-1-119-33706-5; 978-1-119-33726-3 (ebk); 978-1-119-33709-6 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Introduction 1

Part 1: Getting Started with Statistical Analysis with R 7

CHAPTER 1: Data, Statistics, and Decisions 9

CHAPTER 2: R: What It Does and How It Does It 17

Part 2: Describing Data 49

CHAPTER 3: Getting Graphic 51

CHAPTER 4: Finding Your Center 91

CHAPTER 5: Deviating from the Average 103

CHAPTER 6: Meeting Standards and Standings 111

CHAPTER 7: Summarizing It All 123

CHAPTER 8: What's Normal? 143

Part 3: Drawing Conclusions from Data 161

CHAPTER 9: The Confidence Game: Estimation 163

CHAPTER 10: One-Sample Hypothesis Testing 179

CHAPTER 11: Two-Sample Hypothesis Testing 205

CHAPTER 12: Testing More than Two Samples 231

CHAPTER 13: More Complicated Testing 255

CHAPTER 14: Regression: Linear, Multiple, and the General Linear Model 277

CHAPTER 15: Correlation: The Rise and Fall of Relationships 313

CHAPTER 16: Curvilinear Regression: When Relationships Get Complicated 335

Part 4: Working with Probability 359

CHAPTER 17: Introducing Probability 361

CHAPTER 18: Introducing Modeling 383

Part 5: The Part of Tens 405

CHAPTER 19: Ten Tips for Excel Emigrés 407

CHAPTER 20: Ten Valuable Online R Resources 421

Index 425

Table of Contents

INTRODUCTION 1
About This Book 1
Similarity with This Other For Dummies Book 2
What You Can Safely Skip 2
Foolish Assumptions 2
How This Book Is Organized 3
Part 1: Getting Started with Statistical Analysis with R 3
Part 2: Describing Data 3
Part 3: Drawing Conclusions from Data 3
Part 4: Working with Probability 3
Part 5: The Part of Tens 4
Online Appendix A: More on Probability 4
Online Appendix B: Non-Parametric Statistics 4
Online Appendix C: Ten Topics That Just Didn't Fit in Any Other Chapter 4
Icons Used in This Book 4
Where to Go from Here 5

PART 1: GETTING STARTED WITH STATISTICAL ANALYSIS WITH R 7

CHAPTER 1: Data, Statistics, and Decisions 9
The Statistical (and Related) Notions You Just Have to Know 10
Samples and populations 10
Variables: Dependent and independent 11
Types of data 12
A little probability 13
Inferential Statistics: Testing Hypotheses 14
Null and alternative hypotheses 14
Two types of error 15

CHAPTER 2: R: What It Does and How It Does It 17
Downloading R and RStudio 18
A Session with R 21
The working directory 21
So let's get started, already 22
Missing data 26
R Functions 26
User-Defined Functions 28
Comments 29

R Structures 29
Vectors 30
Numerical vectors 30
Matrices 31
Factors 33
Lists 34
Lists and statistics 35
Data frames 36
Packages 39
More Packages 42
R Formulas 43
Reading and Writing 44
Spreadsheets 44
CSV files 46
Text files 47

PART 2: DESCRIBING DATA 49

CHAPTER 3: Getting Graphic 51
Finding Patterns 51
Graphing a distribution 52
Bar-hopping 53
Slicing the pie 54
The plot of scatter 55
Of boxes and whiskers 56
Base R Graphics 57
Histograms 57
Adding graph features 59
Bar plots 60
Pie graphs 62
Dot charts 62
Bar plots revisited 64
Scatter plots 67
Box plots 71
Graduating to ggplot2 71
Histograms 72
Bar plots 74
Dot charts 75
Bar plots re-revisited 78
Scatter plots 82
Box plots 86
Wrapping Up 89

CHAPTER 4: **Finding Your Center** 91
Means: The Lure of Averages 91
The Average in R: mean() 93
What's your condition? 93
Eliminate $-signs forth with() 94
Exploring the data 95
Outliers: The flaw of averages 96
Other means to an end. 97
Medians: Caught in the Middle 99
The Median in R: median() 100
Statistics à la Mode 101
The Mode in R 101

CHAPTER 5: **Deviating from the Average** 103
Measuring Variation 104
Averaging squared deviations: Variance and how to calculate it 104
Sample variance. 107
Variance in R. 107
Back to the Roots: Standard Deviation. 108
Population standard deviation 108
Sample standard deviation 109
Standard Deviation in R 109
Conditions, Conditions, Conditions 110

CHAPTER 6: **Meeting Standards and Standings** 111
Catching Some Z's 112
Characteristics of z-scores 112
Bonds versus the Bambino 113
Exam scores 114
Standard Scores in R 114
Where Do You Stand? 117
Ranking in R 117
Tied scores 117
Nth smallest, Nth largest 118
Percentiles 118
Percent ranks 120
Summarizing 121

CHAPTER 7: **Summarizing It All** 123
How Many? 123
The High and the Low 125

Living in the Moments 125
A teachable moment. 126
Back to descriptives. 126
Skewness 127
Kurtosis 130
Tuning in the Frequency. 131
Nominal variables: table() et al 131
Numerical variables: hist() 132
Numerical variables: stem() 138
Summarizing a Data Frame 139

CHAPTER 8: **What's Normal?** 143
Hitting the Curve 143
Digging deeper. 144
Parameters of a normal distribution 145
Working with Normal Distributions 147
Distributions in R 147
Normal density function. 147
Cumulative density function 152
Quantiles of normal distributions. 155
Random sampling 156
A Distinguished Member of the Family 158

PART 3: DRAWING CONCLUSIONS FROM DATA 161

CHAPTER 9: **The Confidence Game: Estimation** 163
Understanding Sampling Distributions 164
An EXTREMELY Important Idea: The Central Limit Theorem 165
(Approximately) Simulating the central limit theorem. 167
Predictions of the central limit theorem 171
Confidence: It Has Its Limits! 173
Finding confidence limits for a mean 173
Fit to a t 175

CHAPTER 10: **One-Sample Hypothesis Testing** 179
Hypotheses, Tests, and Errors 179
Hypothesis Tests and Sampling Distributions. 181
Catching Some Z's Again. 183
Z Testing in R 185
t for One 187
t Testing in R. 188
Working with t-Distributions 189

Visualizing t-Distributions. 190
Plotting t in base R graphics. 191
Plotting t in ggplot2 192
One more thing about ggplot2 197
Testing a Variance 198
Testing in R 199
Working with Chi-Square Distributions 201
Visualizing Chi-Square Distributions. 201
Plotting chi-square in base R graphics 202
Plotting chi-square in ggplot2 203

CHAPTER 11: **Two-Sample Hypothesis Testing** 205
Hypotheses Built for Two 205
Sampling Distributions Revisited 206
Applying the central limit theorem 207
Z's once more. 208
Z-testing for two samples in R 210
t for Two 212
Like Peas in a Pod: Equal Variances 212
t-Testing in R. 214
Working with two vectors. 214
Working with a data frame and a formula. 215
Visualizing the results 216
Like p's and q's: Unequal variances. 219
A Matched Set: Hypothesis Testing for Paired Samples 220
Paired Sample t-testing in R. 222
Testing Two Variances 222
F-testing in R. 224
F in conjunction with t. 225
Working with *F*-Distributions 226
Visualizing *F*-Distributions 226

CHAPTER 12: **Testing More than Two Samples** 231
Testing More Than Two 231
A thorny problem 232
A solution 233
Meaningful relationships 237
ANOVA in R. 237
Visualizing the results 239
After the ANOVA 239
Contrasts in R. 242
Unplanned comparisons 243

Another Kind of Hypothesis, Another Kind of Test. 244
Working with repeated measures ANOVA. 245
Repeated measures ANOVA in R. 247
Visualizing the results 249
Getting Trendy 250
Trend Analysis in R 254

CHAPTER 13: **More Complicated Testing** 255
Cracking the Combinations 255
Interactions. 257
The analysis 257
Two-Way ANOVA in R 259
Visualizing the two-way results 261
Two Kinds of Variables . . . at Once. 263
Mixed ANOVA in R 266
Visualizing the Mixed ANOVA results 268
After the Analysis. 269
Multivariate Analysis of Variance 270
MANOVA in R 271
Visualizing the MANOVA results 273
After the analysis 275

CHAPTER 14: **Regression: Linear, Multiple, and the General Linear Model** 277
The Plot of Scatter 277
Graphing Lines. 279
Regression: What a Line! 281
Using regression for forecasting 283
Variation around the regression line 283
Testing hypotheses about regression 285
Linear Regression in R 290
Features of the linear model 292
Making predictions 292
Visualizing the scatter plot and regression line 293
Plotting the residuals 294
Juggling Many Relationships at Once: Multiple Regression. 295
Multiple regression in R 297
Making predictions 298
Visualizing the 3D scatter plot and regression plane. 298
ANOVA: Another Look. 301
Analysis of Covariance: The Final Component of the GLM 305
But wait — there's more. 311

CHAPTER 15: **Correlation: The Rise and Fall of Relationships** 313
Scatter plots Again. 313
Understanding Correlation 314
Correlation and Regression 316
Testing Hypotheses About Correlation 319
Is a correlation coefficient greater than zero? 319
Do two correlation coefficients differ? 320
Correlation in R 322
Calculating a correlation coefficient 322
Testing a correlation coefficient 322
Testing the difference between two correlation coefficients 323
Calculating a correlation matrix 324
Visualizing correlation matrices 324
Multiple Correlation 326
Multiple correlation in R 327
Adjusting R-squared 328
Partial Correlation 329
Partial Correlation in R 330
Semipartial Correlation 331
Semipartial Correlation in R 332

CHAPTER 16: **Curvilinear Regression: When Relationships Get Complicated** 335
What Is a Logarithm? 336
What Is e? 338
Power Regression 341
Exponential Regression 346
Logarithmic Regression 350
Polynomial Regression: A Higher Power 354
Which Model Should You Use? 358

PART 4: WORKING WITH PROBABILITY 359

CHAPTER 17: **Introducing Probability** 361
What Is Probability? 361
Experiments, trials, events, and sample spaces 362
Sample spaces and probability 362
Compound Events 363
Union and intersection 363
Intersection again 364
Conditional Probability 365
Working with the probabilities 366
The foundation of hypothesis testing. 366

Large Sample Spaces 366
Permutations 367
Combinations 368
R Functions for Counting Rules 369
Random Variables: Discrete and Continuous 371
Probability Distributions and Density Functions 371
The Binomial Distribution 374
The Binomial and Negative Binomial in R 375
Binomial distribution 375
Negative binomial distribution 377
Hypothesis Testing with the Binomial Distribution 378
More on Hypothesis Testing: R versus Tradition 380

CHAPTER 18: **Introducing Modeling** 383
Modeling a Distribution 383
Plunging into the Poisson distribution 384
Modeling with the Poisson distribution 385
Testing the model's fit 388
A word about chisq.test() 391
Playing ball with a model 392
A Simulating Discussion 396
Taking a chance: The Monte Carlo method 396
Loading the dice 396
Simulating the central limit theorem 401

PART 5: THE PART OF TENS 405

CHAPTER 19: **Ten Tips for Excel Emigrés** 407
Defining a Vector in R Is Like Naming a Range in Excel 407
Operating on Vectors Is Like Operating on Named Ranges 408
Sometimes Statistical Functions Work the Same Way 412
. . . And Sometimes They Don't 412
Contrast: Excel and R Work with Different Data Formats 413
Distribution Functions Are (Somewhat) Similar 414
A Data Frame Is (Something) Like a Multicolumn Named Range 416
The sapply() Function Is Like Dragging 417
Using edit() Is (Almost) Like Editing a Spreadsheet 418
Use the Clipboard to Import a Table from Excel into R 419

CHAPTER 20: **Ten Valuable Online R Resources** 421
Websites for R Users 421
R-bloggers 421
Microsoft R Application Network 422
Quick-R 422

RStudio Online Learning 422
Stack Overflow 422
Online Books and Documentation 423
R manuals 423
R documentation 423
RDocumentation 423
YOU CANanalytics 423
The R Journal 424

INDEX 425

Introduction

So you're holding a statistics book. In my humble (and absolutely biased) opinion, it's not just another statistics book. It's also not just another R book. I say this for two reasons.

First, many statistics books teach you the concepts but don't give you an easy way to apply them. That often leads to a lack of understanding. Because R is ready-made for statistics, it's a tool for applying (and learning) statistics concepts.

Second, let's look at it from the opposite direction: Before I tell you about one of R's features, I give you the statistical foundation it's based on. That way, you understand that feature when you use it — and you use it more effectively.

I didn't want to write a book that only covers the details of R and introduces some clever coding techniques. Some of that is necessary, of course, in any book that shows you how to use a software tool like R. My goal was to go way beyond that.

Neither did I want to write a statistics "cookbook": when-faced-with-problem-category-#152-use-statistical-procedure-#346. My goal was to go way beyond that, too.

Bottom line: This book isn't just about statistics or just about R — it's firmly at the intersection of the two. In the proper context, R can be a great tool for teaching and learning statistics, and I've tried to supply the proper context.

About This Book

Although the field of statistics proceeds in a logical way, I've organized this book so that you can open it up in any chapter and start reading. The idea is for you to find the information you're looking for in a hurry and use it immediately — whether it's a statistical concept or an R-related one.

On the other hand, reading from cover to cover is okay if you're so inclined. If you're a statistics newbie and you have to use R to analyze data, I recommend that you begin at the beginning.

Similarity with This Other For Dummies Book

You might be aware that I've written another book: *Statistical Analysis with Excel For Dummies* (Wiley). This is not a shameless plug for that book. (I do that elsewhere.)

I'm just letting you know that the sections in this book that explain statistical concepts are much like the corresponding sections in that one. I use (mostly) the same examples and, in many cases, the same words. I've developed that material during decades of teaching statistics and found it to be very effective. (Reviewers seem to like it, too.) Also, if you happen to have read the other book and you're transitioning to R, the common material might just help you make the switch.

And, you know: If it ain't broke. . . .

What You Can Safely Skip

Any reference book throws a lot of information at you, and this one is no exception. I intended for it all to be useful, but I didn't aim it all at the same level. So if you're not deeply into the subject matter, you can avoid paragraphs marked with the Technical Stuff icon.

As you read, you'll run into sidebars. They provide information that elaborates on a topic, but they're not part of the main path. If you're in a hurry, you can breeze past them.

Foolish Assumptions

I'm assuming this much about you:

- You know how to work with Windows or the Mac. I don't describe the details of pointing, clicking, selecting, and other actions.
- You're able to install R and RStudio (I show you how in Chapter 2) and follow along with the examples. I use the Windows version of RStudio, but you should have no problem if you're working on a Mac.

How This Book Is Organized

I've organized this book into five parts and three appendixes (which you can find on this book's companion website at www.dummies.com/go/statisticalanalysiswithr).

Part 1: Getting Started with Statistical Analysis with R

In Part 1, I provide a general introduction to statistics and to R. I discuss important statistical concepts and describe useful R techniques. If it's been a long time since your last course in statistics or if you've never even had a statistics course, start with Part 1. If you have never worked with R, *definitely* start with Part 1.

Part 2: Describing Data

Part of working with statistics is to summarize data in meaningful ways. In Part 2, you find out how to do that. Most people know about averages and how to compute them. But that's not the whole story. In Part 2, I tell you about additional statistics that fill in the gaps, and I show you how to use R to work with those statistics. I also introduce R graphics in this part.

Part 3: Drawing Conclusions from Data

Part 3 addresses the fundamental aim of statistical analysis: to go beyond the data and help you make decisions. Usually, the data are measurements of a sample taken from a large population. The goal is to use these data to figure out what's going on in the population.

This opens a wide range of questions: What does an average mean? What does the difference between two averages mean? Are two things associated? These are only a few of the questions I address in Part 3, and I discuss the R functions that help you answer them.

Part 4: Working with Probability

Probability is the basis for statistical analysis and decision-making. In Part 4, I tell you all about it. I show you how to apply probability, particularly in the area of modeling. R provides a rich set of capabilities that deal with probability. Here's where you find them.

Part 5: The Part of Tens

Part V has two chapters. In the first, I give Excel users ten tips for moving to R. In the second, I cover ten statistical- and R-related topics that wouldn't fit in any other chapter.

Online Appendix A: More on Probability

This online appendix continues what I start in Part 4. The material is a bit on the esoteric side, so I've stashed it in an appendix.

Online Appendix B: Non-Parametric Statistics

Non-parametric statistics are based on concepts that differ somewhat from most of the rest of the book. In this appendix, you learn these concepts and see how to use R to apply them.

Online Appendix C: Ten Topics That Just Didn't Fit in Any Other Chapter

This is the Grab Bag appendix, where I cover ten statistical- and R-related topics that wouldn't fit in any other chapter.

Icons Used in This Book

Icons appear all over *For Dummies* books, and this one is no exception. Each one is a little picture in the margin that lets you know something special about the paragraph it sits next to.


TIP

This icon points out a hint or a shortcut that can help you in your work (and perhaps make you a finer, kinder, and more insightful human being).


REMEMBER

This one points out timeless wisdom to take with you on your continuing quest for statistics knowledge.

WARNING

Pay attention to the information accompanied by this icon. It's a reminder to avoid something that might gum up the works for you.

TECHNICAL STUFF

As I mention in the earlier section "What You Can Safely Skip," this icon indicates material you can blow past if it's just too technical. (I've kept this to a minimum.)

Where to Go from Here

You can start reading this book anywhere, but here are a couple of hints. Want to learn the foundations of statistics? Turn the page. Introduce yourself to R? That's Chapter 2. Want to start with graphics? Hit Chapter 3. For anything else, find it in the table of contents or the index and go for it.

In addition to what you're reading right now, this product comes with a free access-anywhere Cheat Sheet that presents a selected list of R functions and describes what they do. To get this Cheat Sheet, visit `www.dummies.com` and type **Statistical Analysis with R For Dummies Cheat Sheet** in the search box.

1

Getting Started with Statistical Analysis with R

IN THIS PART . . .

Find out about R's statistical capabilities

Explore how to work with populations and samples

Test your hypotheses

Understand errors in decision-making

Determine independent and dependent variables

IN THIS CHAPTER

- Introducing statistical concepts
- Generalizing from samples to populations
- Getting into probability
- Testing hypotheses
- Two types of error

Chapter 1
Data, Statistics, and Decisions

Statistics? That's all about crunching numbers into arcane-looking formulas, right? Not really. Statistics, first and foremost, is about *decision-making*. Some number-crunching is involved, of course, but the primary goal is to use numbers to make decisions. Statisticians look at data and wonder what the numbers are saying. What kinds of trends are in the data? What kinds of predictions are possible? What conclusions can we make?

To make sense of data and answer these questions, statisticians have developed a wide variety of analytical tools.

About the number-crunching part: If you had to do it via pencil-and-paper (or with the aid of a pocket calculator), you'd soon get discouraged with the amount of computation involved and the errors that might creep in. Software like R helps you crunch the data and compute the numbers. As a bonus, R can also help you comprehend statistical concepts.

Developed specifically for statistical analysis, R is a computer language that implements many of the analytical tools statisticians have developed for decision-making. I wrote this book to show how to use these tools in your work.

The Statistical (and Related) Notions You Just Have to Know

The analytical tools that that R provides are based on statistical concepts I help you explore in the remainder of this chapter. As you'll see, these concepts are based on common sense.

Samples and populations

If you watch TV on election night, you know that one of the main events is the prediction of the outcome immediately after the polls close (and before all the votes are counted). How is it that pundits almost always get it right?

The idea is to talk to a *sample* of voters right after they vote. If they're truthful about how they marked their ballots, and if the sample is representative of the *population* of voters, analysts can use the sample data to draw conclusions about the population.

That, in a nutshell, is what statistics is all about — using the data from samples to draw conclusions about populations.

Here's another example. Imagine that your job is to find the average height of 10-year-old children in the United States. Because you probably wouldn't have the time or the resources to measure every child, you'd measure the heights of a representative sample. Then you'd average those heights and use that average as the estimate of the population average.

Estimating the population average is one kind of *inference* that statisticians make from sample data. I discuss inference in more detail in the upcoming section "Inferential Statistics: Testing Hypotheses."


REMEMBER

Here's some important terminology: Properties of a population (like the population average) are called *parameters*, and properties of a sample (like the sample average) are called *statistics*. If your only concern is the sample properties (like the heights of the children in your sample), the statistics you calculate are *descriptive*. If you're concerned about estimating the population properties, your statistics are *inferential*.


REMEMBER

Now for an important convention about notation: Statisticians use Greek letters (μ, σ, ρ) to stand for parameters, and English letters ($\bar{X}$, s, r) to stand for statistics. Figure 1-1 summarizes the relationship between populations and samples, and between parameters and statistics.

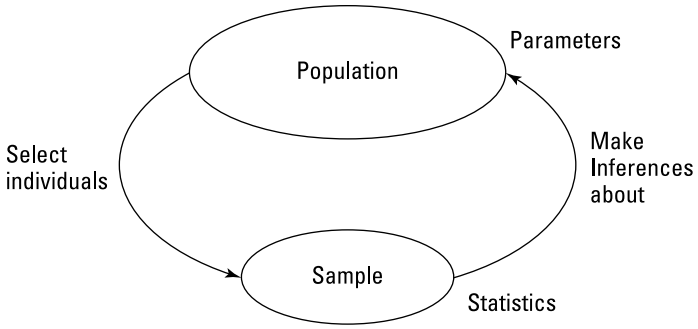

FIGURE 1-1: The relationship between populations, samples, parameters, and statistics.

Variables: Dependent and independent

A *variable* is something that can take on more than one value — like your age, the value of the dollar against other currencies, or the number of games your favorite sports team wins. Something that can have only one value is a *constant*. Scientists tell us that the speed of light is a constant, and we use the constant π to calculate the area of a circle.

Statisticians work with *independent* variables and *dependent* variables. In any study or experiment, you'll find both kinds. Statisticians assess the relationship between them.

For example, imagine a computerized training method designed to increase a person's IQ. How would a researcher find out if this method does what it's supposed to do? First, he would randomly assign a sample of people to one of two groups. One group would receive the training method, and the other would complete another kind of computer-based activity — like reading text on a website. Before and after each group completes its activities, the researcher measures each person's IQ. What happens next? I discuss that topic in the upcoming section "Inferential Statistics: Testing Hypotheses."

For now, understand that the independent variable here is Type of Activity. The two possible values of this variable are IQ Training and Reading Text. The dependent variable is the change in IQ from Before to After.

REMEMBER

A dependent variable is what a researcher *measures*. In an experiment, an independent variable is what a researcher *manipulates*. In other contexts, a researcher can't manipulate an independent variable. Instead, he notes naturally occurring values of the independent variable and how they affect a dependent variable.

REMEMBER

In general, the objective is to find out whether changes in an independent variable are associated with changes in a dependent variable.


REMEMBER

In the examples that appear throughout this book, I show you how to use R to calculate characteristics of groups of scores, or to compare groups of scores. Whenever I show you a group of scores, I'm talking about the values of a dependent variable.

Types of data

When you do statistical work, you can run into four kinds of data. And when you work with a variable, the way you work with it depends on what kind of data it is. The first kind is *nominal* data. If a set of numbers happens to be nominal data, the numbers are labels – their values don't signify anything. On a sports team, the jersey numbers are nominal. They just identify the players.

The next kind is *ordinal* data. In this data-type, the numbers are more than just labels. As the name "ordinal" might tell you, the order of the numbers is important. If I ask you to rank ten foods from the one you like best (one), to the one you like least (ten), we'd have a set of ordinal data.

But the difference between your third-favorite food and your fourth-favorite food might not be the same as the difference between your ninth-favorite and your tenth-favorite. So this type of data lacks equal intervals and equal differences.

Interval data gives us equal differences. The Fahrenheit scale of temperature is a good example. The difference between 30° and 40° is the same as the difference between 90° and 100°. So each degree is an interval.

People are sometimes surprised to find out that on the Fahrenheit scale, a temperature of 80° is not twice as hot as 40°. For ratio statements ("twice as much as", "half as much as") to make sense, "zero" has to mean the complete absence of the thing you're measuring. A temperature of 0° F doesn't mean the complete absence of heat – it's just an arbitrary point on the Fahrenheit scale. (The same holds true for Celsius.)

The fourth kind of data, *ratio*, provides a meaningful zero point. On the Kelvin Scale of temperature, zero means "absolute zero," where all molecular motion (the basis of heat) stops. So 200° Kelvin is twice as hot as 100° Kelvin. Another example is length. Eight inches is twice as long as four inches. "Zero inches" means "a complete absence of length."


REMEMBER

An independent variable or a dependent variable can be either nominal, ordinal, interval, or ratio. The analytical tools you use depend on the type of data you work with.

A little probability

When statisticians make decisions, they use probability to express their confidence about those decisions. They can never be absolutely certain about what they decide. They can only tell you how probable their conclusions are.

What do we mean by probability? Mathematicians and philosophers might give you complex definitions. In my experience, however, the best way to understand probability is in terms of examples.

Here's a simple example: If you toss a coin, what's the probability that it turns up heads? If the coin is fair, you might figure that you have a 50-50 chance of heads and a 50-50 chance of tails. And you'd be right. In terms of the kinds of numbers associated with probability, that's $\frac{1}{2}$.

Think about rolling a fair die (one member of a pair of dice). What's the probability that you roll a 4? Well, a die has six faces and one of them is 4, so that's $\frac{1}{6}$. Still another example: Select one card at random from a standard deck of 52 cards. What's the probability that it's a diamond? A deck of cards has four suits, so that's $\frac{1}{4}$.

These examples tell you that if you want to know the probability that an event occurs, count how many ways that event can happen and divide by the total number of events that can happen. In the first two examples (heads, 4), the event you're interested in happens only one way. For the coin, we divide one by two. For the die, we divide one by six. In the third example (diamond), the event can happen 13 ways (Ace through King), so we divide 13 by 52 (to get $\frac{1}{4}$).

Now for a slightly more complicated example. Toss a coin and roll a die at the same time. What's the probability of tails and a 4? Think about all the possible events that can happen when you toss a coin and roll a die at the same time. You could have tails and 1 through 6, or heads and 1 through 6. That adds up to 12 possibilities. The tails-and-4 combination can happen only one way. So the probability is $\frac{1}{12}$.

In general, the formula for the probability that a particular event occurs is

$$\Pr(\text{event}) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible events}}$$

At the beginning of this section, I say that statisticians express their confidence about their conclusions in terms of probability, which is why I brought all this up in the first place. This line of thinking leads to *conditional* probability — the probability that an event occurs given that some other event occurs. Suppose that I roll a die, look at it (so that you don't see it), and tell you that I rolled an odd number. What's the probability that I've rolled a 5? Ordinarily, the probability of a 5 is $\frac{1}{6}$,

but "I rolled an odd number" narrows it down. That piece of information eliminates the three even numbers (2, 4, 6) as possibilities. Only the three odd numbers (1,3, 5) are possible, so the probability is ⅓.

What's the big deal about conditional probability? What role does it play in statistical analysis? Read on.

Inferential Statistics: Testing Hypotheses

Before a statistician does a study, he draws up a tentative explanation — a *hypothesis* that tells why the data might come out a certain way. After gathering all the data, the statistician has to decide whether or not to reject the hypothesis.

That decision is the answer to a conditional probability question — what's the probability of obtaining the data, given that this hypothesis is correct? Statisticians have tools that calculate the probability. If the probability turns out to be low, the statistician rejects the hypothesis.

Back to coin-tossing for an example: Imagine that you're interested in whether a particular coin is fair — whether it has an equal chance of heads or tails on any toss. Let's start with "The coin is fair" as the hypothesis.

To test the hypothesis, you'd toss the coin a number of times — let's say, a hundred. These 100 tosses are the sample data. If the coin is fair (as per the hypothesis), you'd expect 50 heads and 50 tails.

If it's 99 heads and 1 tail, you'd surely reject the fair-coin hypothesis: The conditional probability of 99 heads and 1 tail given a fair coin is very low. Of course, the coin could still be fair and you could, quite by chance, get a 99-1 split, right? Sure. You never really know. You have to gather the sample data (the 100 toss-results) and then decide. Your decision might be right, or it might not.

Juries make these types of decisions. In the United States, the starting hypothesis is that the defendant is not guilty ("innocent until proven guilty"). Think of the evidence as "data." Jury-members consider the evidence and answer a conditional probability question: What's the probability of the evidence, given that the defendant is not guilty? Their answer determines the verdict.

Null and alternative hypotheses

Think again about that coin-tossing study I just mentioned. The sample data are the results from the 100 tosses. I said that we can start with the hypothesis that