

LEARNING MADE EASY



Statistics Essentials

for
dummies[®]
A Wiley Brand



The “must-know”
formulas and calculations

What you need to know
about statistical techniques

Core topics in quick,
focused lessons

Deborah J. Rumsey, PhD

Professor of Statistics,
The Ohio State University



Statistics Essentials

by Deborah J. Rumsey, PhD

for
dummies[®]
A Wiley Brand

Statistics Essentials For Dummies®

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

Copyright © 2019 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

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 the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc., and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, 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. For technical support, please visit www.wiley.com/techsupport.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2019937507

ISBN 978-1-119-59030-9 (pbk); ISBN 978-1-119-59024-8 (ebk); ISBN 978-1-119-59023-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Introduction	1
CHAPTER 1: Statistics in a Nutshell	5
CHAPTER 2: Descriptive Statistics	13
CHAPTER 3: Charts and Graphs	23
CHAPTER 4: The Binomial Distribution	35
CHAPTER 5: The Normal Distribution	45
CHAPTER 6: Sampling Distributions and the Central Limit Theorem	55
CHAPTER 7: Confidence Intervals	69
CHAPTER 8: Hypothesis Tests	87
CHAPTER 9: The t-Distribution	107
CHAPTER 10: Correlation and Regression	113
CHAPTER 11: Two-Way Tables	127
CHAPTER 12: A Checklist for Samples and Surveys	137
CHAPTER 13: A Checklist for Judging Experiments	147
CHAPTER 14: Ten Common Statistical Mistakes	155
Appendix: Tables for Reference	163
Index	171

Table of Contents

INTRODUCTION	1
About This Book	1
Conventions Used in This Book	2
Foolish Assumptions	2
Icons Used in This Book	2
Where to Go from Here	3
CHAPTER 1: Statistics in a Nutshell	5
Designing Studies	5
Surveys	5
Experiments	6
Collecting Data	7
Selecting a good sample	7
Avoiding bias in your data	8
Describing Data	8
Descriptive statistics	8
Charts and graphs	9
Analyzing Data	10
Making Conclusions	10
CHAPTER 2: Descriptive Statistics	13
Types of Data	13
Counts and Percents	14
Measures of Center	15
Measures of Variability	17
Percentiles	18
Finding a percentile	19
Interpreting percentiles	20
The Five-Number Summary	21
CHAPTER 3: Charts and Graphs	23
Pie Charts	23
Bar Graphs	24
Time Charts	26
Histograms	27
Making a histogram	27
Interpreting a histogram	29
Evaluating a histogram	30

	Boxplots.....	31
	Making a boxplot	31
	Interpreting a boxplot	32
CHAPTER 4:	The Binomial Distribution	35
	Characteristics of a Binomial	35
	Checking the binomial conditions step by step	36
	Non-binomial examples.....	36
	Finding Binomial Probabilities Using the Formula.....	38
	Finding Probabilities Using the Binomial Table.....	40
	Finding probabilities when $p \leq 0.50$	40
	Finding probabilities when $p > 0.50$	41
	Finding probabilities for X greater-than, less-than, or between two values	42
	The Expected Value and Variance of the Binomial	43
CHAPTER 5:	The Normal Distribution	45
	Basics of the Normal Distribution	45
	The Standard Normal (Z) Distribution	46
	Finding Probabilities for X.....	48
	Finding X for a Given Probability	51
	Normal Approximation to the Binomial.....	53
CHAPTER 6:	Sampling Distributions and the Central Limit Theorem	55
	Sampling Distributions	55
	The mean of a sampling distribution	56
	Standard error of a sampling distribution.....	57
	Sample size and standard error.....	58
	Population standard deviation and standard error	60
	The shape.....	60
	Finding Probabilities for \bar{X}	62
	The Sampling Distribution of the Sample Proportion	63
	What proportion of students need math help?	64
	Finding Probabilities for \hat{p}	66
CHAPTER 7:	Confidence Intervals.....	69
	Making Your Best Guesstimate	69
	The Goal: Small Margin of Error	71
	Choosing a Confidence Level.....	72
	Factoring in the Sample Size	73
	Counting on Population Variability	75

	Confidence Interval for a Population Mean	76
	Confidence Interval for a Population Proportion.....	77
	Confidence Interval for the Difference of Two Means.....	78
	Confidence Interval for the Difference of Two Proportions	80
	Interpreting Confidence Intervals	81
	Spotting Misleading Confidence Intervals.....	84
CHAPTER 8:	Hypothesis Tests	87
	Doing a Hypothesis Test.....	87
	Identifying what you're testing	88
	Setting up the hypotheses	88
	Finding sample statistics	90
	Standardizing the evidence: The test statistic	90
	Weighing the evidence and making decisions: p-values.....	91
	General steps for a hypothesis test.....	94
	Testing One Population Mean	94
	Testing One Population Proportion	96
	Comparing Two Population Means.....	97
	Testing the Mean Difference: Paired Data	99
	Testing Two Population Proportions	102
	You Could Be Wrong: Errors in Hypothesis Testing.....	104
	A false alarm: Type-1 error	104
	A missed detection: Type-2 error.....	105
CHAPTER 9:	The <i>t</i>-Distribution	107
	Basics of the <i>t</i> -Distribution	107
	Understanding the <i>t</i> -Table	108
	<i>t</i> -Distributions and Hypothesis Tests.....	109
	Finding critical values	110
	Finding <i>p</i> -values	110
	<i>t</i> -Distributions and Confidence Intervals.....	111
CHAPTER 10:	Correlation and Regression	113
	Picturing the Relationship with a Scatterplot.....	113
	Making a scatterplot.....	114
	Interpreting a scatterplot.....	114
	Measuring Relationships Using the Correlation.....	115
	Calculating the correlation.....	115
	Interpreting the correlation.....	117
	Properties of the correlation	118

Finding the Regression Line	118
Which is X and which is Y?.....	119
Checking the conditions.....	119
Understanding the equation	119
Finding the slope.....	120
Finding the y-intercept	121
Interpreting the slope and y-intercept	121
Making Predictions.....	124
Avoid Extrapolation!.....	124
Correlation Doesn't Necessarily Mean Cause-and-Effect.....	125
CHAPTER 11: Two-Way Tables	127
Organizing and Interpreting a Two-Way Table	127
Defining the outcomes.....	128
Setting up the rows and columns	128
Inserting the numbers.....	129
Finding the row, column, and grand totals.....	129
Finding Probabilities within a Two-Way Table	130
Figuring joint probabilities	130
Calculating marginal probabilities	131
Finding conditional probabilities	132
Checking for Independence	134
CHAPTER 12: A Checklist for Samples and Surveys	137
The Target Population Is Well Defined	138
The Sample Matches the Target Population.....	138
The Sample Is Randomly Selected	139
The Sample Size Is Large Enough.....	139
Nonresponse Is Minimized	140
The importance of following up.....	140
Anonymity versus confidentiality.....	141
The Survey Is of the Right Type	142
Questions Are Well Worded.....	142
The Timing Is Appropriate.....	143
Personnel Are Well Trained.....	143
Proper Conclusions Are Made	144
CHAPTER 13: A Checklist for Judging Experiments.....	147
Experiments versus Observational Studies	147
Criteria for a Good Experiment	148
Inspect the Sample Size.....	148
Small samples — small conclusions	148
Original versus final sample size.....	149

Examine the Subjects.....	149
Check for Random Assignments	150
Gauge the Placebo Effect	150
Identify Confounding Variables	151
Assess Data Quality.....	152
Check out the Analysis.....	153
Scrutinize the Conclusions	153
Overstated results	154
Ad-hoc explanations.....	154
Generalizing beyond the scope.....	154
CHAPTER 14: Ten Common Statistical Mistakes	155
Misleading Graphs	155
Pie charts	156
Bar graphs	156
Time charts	156
Histograms	157
Biased Data	157
No Margin of Error	158
Nonrandom Samples.....	158
Missing Sample Sizes	159
Misinterpreted Correlations	159
Confounding Variables	160
Botched Numbers	161
Selectively Reporting Results	161
The Almighty Anecdote.....	162
APPENDIX: TABLES FOR REFERENCE.....	163
INDEX.....	171

Introduction

This book is designed to give you the essential, nitty-gritty information typically covered in a first semester statistics course. It's bottom-line information for you to use as a refresher, a resource, a quick reference, and/or a study guide. It helps you decipher and make important decisions about statistical polls, experiments, reports, and headlines with confidence, being ever aware of the ways people can mislead you with statistics, and how to handle it.

Topics I work you through include graphs and charts, descriptive statistics, the binomial, normal, and t -distributions, two-way tables, simple linear regression, confidence intervals, hypothesis tests, surveys, experiments, and of course the most frustrating yet critical of all statistical topics: sampling distributions and the Central Limit Theorem.

About This Book

This book departs from traditional statistics texts and reference/supplement books and study guides in these ways:

- » **Clear and concise step-by-step procedures** that intuitively explain how to work through statistics problems and remember the process.
- » **Focused, intuitive explanations** empower you to know you're doing things right and whether others do it wrong.
- » **Nonlinear approach** so you can quickly zoom in on that concept or technique you need, without having to read other material first.
- » **Easy-to-follow examples** reinforce your understanding and help you immediately see how to apply the concepts in practical settings.
- » **Understandable language** helps you remember and put into practice essential statistical concepts and techniques.

Conventions Used in This Book

I refer to statistics in two different ways: as numerical results (such as means and medians) or as a field of study (for example, “Statistics is all about data.”).

The second convention refers to the word *data*. I’m going to go with the plural version of the word *data* in this book. For example “data are collected during the experiment” — not “data is collected during the experiment.”

Foolish Assumptions

I assume you’ve had some (not necessarily a lot of) previous experience with statistics somewhere in your past. For example, you can recognize some of the basic statistics such as the mean, median, standard deviation, and perhaps correlation; you can handle some graphs; and you can remember having seen the normal distribution. If it’s been a while and you are a bit rusty, that’s okay; this book is just the thing to jog your memory.

If you have very limited or no prior experience with statistics, allow me to suggest my full-version book, *Statistics For Dummies*, to build up your foundational knowledge base. But if you are someone who has not seen these ideas before and either doesn’t have time for the full version, or you like to plunge into details right away, this book can work for you.

I assume you’ve had a basic algebra background and can do some of the basic mathematical operations and understand some of the basic notation used in algebra like x , y , summation signs, taking the square root, squaring a number, and so on. (If you’d like some backup on the algebra part, I suggest you consider *Algebra I For Dummies* and *Algebra II For Dummies* [Wiley]).

Icons Used in This Book

Here are the road signs you’ll encounter on your journey through this book:



TIP

Tips refer to helpful hints or shortcuts you can use to save time.



TECHNICAL
STUFF

Read these to get the inside track on why a certain concept is important, what its impact will be on the results, and highlights to keep on your radar.



WARNING

These alert you to common errors that can cause problems, so you can steer around them.



REMEMBER

These point out things in the text that you should, if possible, stash away somewhere in your brain for future use.

Where to Go from Here

This book is written in a nonlinear way, so you can start anywhere and still be able to understand what's happening. However, I can make some recommendations for those who are interested in knowing where to start.

For a quick overview of the topics to refresh your memory, check out Chapter 1. For basic number crunching and graphs, see Chapters 2 and 3. If you're most interested in common distributions, see Chapters 4 (binomial); 5 (normal); and 9 (t -distribution). Confidence intervals and hypothesis testing are found in Chapters 7 and 8. Correlation and regression are found in Chapter 10, and two-way tables and independence are tackled in Chapter 11. If you are interested in evaluating and making sense of the results of medical studies, polls, surveys, and experiments, you'll find all the info in Chapters 12 and 13. Common mistakes to avoid or watch for are seen in Chapter 14.

IN THIS CHAPTER

- » Getting the big picture of the field of statistics
- » Overviewing the steps of the scientific method
- » Seeing the role of statistics at each step

Chapter 1

Statistics in a Nutshell

The most common description of statistics is that it's the process of analyzing data — number crunching, in a sense. But statistics is not just about analyzing the data. It's about the whole process of using the scientific method to answer questions and make decisions. That process involves designing studies, collecting good data, describing the data with numbers and graphs, analyzing the data, and then making conclusions. In this chapter, I review each of these steps and show where statistics plays the all-important role.

Designing Studies

Once a research question is defined, the next step is designing a study in order to answer that question. This amounts to figuring out what process you'll use to get the data you need. In this section, I overview the two major types of studies: observational studies and experiments.

Surveys

An *observational study* is one in which data are 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 on many different forms: paper surveys sent through the mail; websites; call-in polls conducted by TV networks; and phone surveys. 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, people who were selected to participate but do not respond, or an entire group in the population who had no chance of even being selected. These potential problems mean a survey has to be well thought-out before it's given.

A downside of surveys is that they can only report relationships between variables that are found; they cannot claim cause and effect. For example, if in a survey researchers notice that the people who drink more than one Diet Coke per day tend to sleep fewer hours each night than those who drink at most one per day, they cannot conclude that Diet Coke is causing the lack of sleep. Other variables might explain the relationship, such as number of hours worked per week. See all the information about surveys, their design, and potential problems in Chapter 12.

Experiments

An *experiment* imposes one or more treatments on the participants in such a way that clear comparisons can be made. Once the treatments are applied, the response is recorded. For example, to study the effect of drug dosage on blood pressure, one group might take 10 mg of the drug, and another group might take 20 mg. Typically, a control group is also involved, where subjects each receive a fake treatment (a sugar pill, for example).

Experiments take place in a controlled setting, and are designed to minimize biases that might occur. Some potential problems include: researchers knowing who got what treatment; a certain condition or characteristic wasn't accounted for that can affect the results (such as weight of the subject when studying drug dosage); or lack of a control group. But when designed correctly, if a difference in the responses is found when the groups are compared, the researchers can conclude a cause and effect relationship. See coverage of experiments in Chapter 13.

It is perhaps most important to note that no matter what the study, it has to be designed so that the original questions can be answered in a credible way.

Collecting Data

Once a study has been designed, be it a survey or an experiment, the subjects are chosen and the data are ready to be collected. This phase of the process is also critical to producing good data.

Selecting a good sample

First, a few words about selecting individuals to participate in a study (much, much more is said about this topic in Chapter 12). In statistics, we have a saying: “Garbage in equals garbage out.” If you select your subjects in a way that is *biased* — that is, favoring certain individuals or groups of individuals — then your results will also be biased.

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. That’s too restrictive — those folks don’t represent a cross-section of the city. Similarly, Bob could put up a website survey and ask people to use it to vote. However, only those who know about the site, have Internet access, and want to respond will give him data. Typically, only those with strong opinions will go to such trouble. So, again, these individuals don’t represent all the folks in the city.

In order to minimize bias, you need to select your sample of individuals *randomly* — that is, using some type of “draw names out of a hat” process. Scientists use a variety of methods to select individuals at random (more in Chapter 12), but getting a random sample is well worth the extra time and effort to get results that are legitimate.

Avoiding bias in your data

Say you're conducting a phone survey on job satisfaction of Americans. If you call them at home during the day between 9 a.m. and 5 p.m., you'll miss out on all those who work during the day; it could be that day workers are more satisfied than night workers, for example. 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 an answer that isn't on your list of possible answers? A host of problems can occur when collecting survey data; Chapter 12 gives you tips on avoiding and spotting them.

Experiments are sometimes even more challenging when it comes to collecting data. Suppose you want to test blood pressure; what if the instrument you are 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 are just some of the problems in data collection that can arise with experiments; Chapter 13 helps you find and minimize them.

Describing Data

Once data are collected, the next step is to summarize it all to get a handle on the big picture. Statisticians describe data in two major ways: with pictures (that is, charts and graphs) and with numbers, called *descriptive statistics*.

Descriptive statistics

Data are also summarized (most often in conjunction with charts and/or graphs) by using what statisticians call descriptive statistics. *Descriptive statistics* are numbers that describe a data set in terms of its important features.

If the data are categorical (where individuals are placed into groups, such as gender or political affiliation), they are typically