
Springer Texts in Statistics

Series editors

R. DeVeaux
S.E. Fienberg
I. Olkin

Springer Texts in Statistics (STS) includes advanced textbooks from 3rd- to 4th-year undergraduate courses to 1st- to 2nd-year graduate courses. Exercise sets should be included. The series editors are currently Stephen Fienberg and Richard D. De Veaux. George Casella and Ingram Olkin were editors of the series for many years.

More information about this series at <http://www.springer.com/series/417>

Douglas A. Wolfe • Grant Schneider

Intuitive Introductory Statistics

 Springer

Douglas A. Wolfe
Department of Statistics
The Ohio State University
Columbus, OH, USA

Grant Schneider
Upstart Network
San Carlos, CA, USA

ISSN 1431-875X

ISSN 2197-4136 (eBook)

Springer Texts in Statistics

ISBN 978-3-319-56070-0

ISBN 978-3-319-56072-4 (eBook)

DOI 10.1007/978-3-319-56072-4

Library of Congress Control Number: 2017950163

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To our wives, Marilyn and Jingjing, for their patience and support
through the lengthy preparation of this text*

Preface

Understanding and interpreting data, both experimental data and the vast quantity of observational data that is now routinely available in the public domain, need not be linked solely to comprehension of a prescribed set of mathematical expressions. Introducing students to the beauty of statistics can be both motivational and interesting without being encumbered by equations. In fact, it is our view that the goal of an introductory statistics course should be to foster an appreciation for the role of statistics and associated data analysis approaches in our everyday lives rather than to prepare the students to be statistical analysts—if necessary, there will be time enough for that secondary emphasis as they begin to concentrate on their chosen fields of study. An introductory course should foster appreciation for the relevance and importance of using statistical methods for summarizing and interpreting data, but not at the expense of enjoying the process. While it is, of course, necessary to include common statistics such as the sample mean, standard deviation, correlation, etc. in an introductory statistics course, they do little to motivate and “grab” the students and can be introduced *after* a student has come to appreciate the nature of the basic information that is provided in collections of data. This desire leads us in this book to emphasize counting

and ranking approaches as initial tools for eliciting information from data collections rather than as a fallback only when sample means and standard deviations are not effective. Not only can students easily understand such counting and ranking techniques, but they also find them to be quite valuable as they explain their analyses to others.

The second point of emphasis in this text is that examples should not be chosen simply to illustrate how a statistical procedure can be applied. Dull, contrived examples can make even interesting statistical analyses uninteresting. On the other hand, an example to which students easily relate can go a long way in both piquing and sustaining their interest in the associated statistical analysis. While they learn a statistical technique, they also unearth some information of interest to them in its own right. We all learn best when we see the relevance of a topic in our own lives. We have worked hard to motivate the statistical discussions in our text through experimental settings and data collections that we ourselves find interesting and that raise questions that we can use statistical tools to address. While we are the first to confess that our view of the world may not match up completely with that of high school juniors and seniors or college freshmen and sophomores, we have tried to include data sets in both the examples and the exercises that are both real (not just realistic) and of general relevance for the students.

A third new feature for this text is the way in which we have chosen to present the chapter exercises. In addition to the usual set of exercises at the end of each section (and there are many), we have included a substantial number of comprehensive exercises at the end of each chapter that include conceptual exercises, data analysis/computational exercises, exercises involving hands-on student activities, and exercises associated with a student's use of the Internet to access interesting and relevant data sets and statistical analyses. This provides instructors with a wide variety of exercises to challenge the individual interests of students in their classes.

The fourth and final unique feature of this introductory text is the inclusion of the necessary **R** functions to enable instructors and students to analyze data sets without the tedium often accompanying many statistical computations. Examples throughout the text still provide all of the details of the associated statistical calculations so that students are fully aware of how the various statistics elicit information from a data set(s). However, we also provide the appropriate **R** procedures that can be used to make the same calculations. It is easy enough for instructors to bypass the **R** procedures if they choose, but including them as part of the basic course would permit their students to apply the associated statistical procedures to large data sets where the direct numerical calculations by hand would be prohibitive. The use of these **R** programs also eliminates the need to include the normal, t , and χ^2 tables, as well as the relevant nonparametric null distribution tables as part of the text. We have also organized all of the **R** programs used in the text into a documented collection that is formally registered as an **R** package IIS specifically linked to this text.

The text is specifically designed for use in an AP statistics course for high school juniors or seniors or in a one-semester or two-quarter introductory precalculus statistics course for college freshmen or sophomores. How well we have succeeded in reaching these groups of students will clearly determine the impact of our not-just-one-more introductory statistics textbook!

Many friends and colleagues have helped with both the initial development and improvement of this text over the years. We owe a particular debt of gratitude to Brad Hartlaub for his invaluable help in initiating this project in the first place and his dedicated effort to move it forward over a number of succeeding years. We also appreciated his input from teaching much of this material in statistics classes at Kenyon College. Similarly, we owe thanks to Deborah Rumsey and Elizabeth Stasny for their feedback from using early drafts of the project in introductory statistics courses at The Ohio State

University. We also owe thanks to Jungwon Byun, Ben Chang, Neha Hebbar, Cindy Smith, and all of the attendees of the so-called “data party” at Upstart for providing a fertile testing ground for the material and concepts. Finally, we owe a special thank you to the computer support group in the Department of Statistics at The Ohio State University for their patience in helping us work through the various versions of MSWord that were confronted over the many years in preparation of the text.

Our editors Michael Penn and Hannah Bracken, who originally signed us to publish with Springer, and Rebekah McClure, who assumed our project when Hannah returned to school, were dedicated from the start of the project and provided tremendous support to see it through to publication. Our production manager Christina Oliver skillfully guided the manuscript through the production process.

To everyone who helped over the many years, our heartfelt thanks.

Columbus, OH, USA
San Carlos, CA, USA

Douglas A. Wolfe
Grant Schneider

Contents

1	Exploratory Data Analysis: Observing Patterns and Departures from Patterns	1
1.1	Interpreting Graphical Displays of Data Collections	2
1.1.1	Construction of a Histogram	17
1.2	Numerically Summarizing One-Variable Data Collections	48
1.2.1	Effects of a Linear Transformation	78
1.3	Comparing One-Variable Data Collections	95
	Chapter 1 Comprehensive Exercises	124
	1.A. Conceptual	124
	1.B. Data Analysis/Computational	128
	1.C. Activities	138
	1.D. Internet Archives	140
2	Exploring Bivariate and Categorical Data	143
2.1	Exploring the Relationship Between Two Quantitative Variables	143

2.1.1	Common Types of Relationships – No Association, Positive Association, Negative Association	144
2.1.2	Scatterplot Smoothing	152
2.1.3	Including a Third Variable on Scatterplots	161
2.2	Measuring the Strength of Association	169
2.2.1	Properties of r	175
2.2.2	An Alternative Measure of Association	178
2.3	Exploring the Relationship between Two Categorical Variables (Frequency Tables)	182
Chapter 2	Comprehensive Exercises	190
2.A.	Conceptual	190
2.B.	Data Analysis/Computational	192
2.C.	Activities	195
2.D.	Internet Archives	197
3	Designing a Survey or Experiment: Deciding What and How to Measure	199
3.1	Methods of Data Collection	201
3.2	Planning and Conducting Surveys or Polls	218
3.3	Planning and Conducting Experiments	227
Chapter 3	Comprehensive Exercises	237
3.A.	Conceptual	237
3.B.	Data Analysis/Computational	238
3.C.	Activities	240
3.D.	Internet Archives	241
4	Understanding Random Events: Producing Models Using Probability and Simulation	243
4.1	Probability as Relative Frequency: Law of Large Numbers	245

4.2	Some Basic Probability Rules	253
4.2.1	Addition Rule	255
4.2.2	Conditional Probability	257
4.2.3	Multiplication Rule	258
4.3	Discrete Random Variables and Their Probability Distributions	266
4.3.1	Binomial Distribution	268
4.3.2	Geometric Distribution	272
4.4	Simulating Probability Distributions	279
4.5	Expected Values and Standard Deviations for Random Variables	285
4.6	Combining Random Variables	291
4.7	Normal Distributions	295
4.7.1	Probability Calculations for Normal Distributions	299
4.7.2	Using Normal Distributions as Models for Measurements	306
	Chapter 4 Comprehensive Exercises	315
4.A.	Conceptual	315
4.B.	Data Analysis/ Computational	317
4.C.	Activities	325
4.D.	Internet Archives	328
5	Sampling Distributions and Approximations	331
5.1	The Sampling Distribution for a Sample Average	333
5.1.1	Comparing Two Averages	339
5.2	Sampling Distributions for Proportions and Counts	346
5.2.1	Comparing Two Proportions	350

5.2.2	Comparing Several Proportions	352
5.2.3	Using Ranks and Counts to Compare Two Samples	354
5.3	Approximating Sampling Distributions	366
5.4	Simulating Sampling Distributions	380
Chapter 5	Comprehensive Exercises	397
5.A.	Conceptual	397
5.B.	Data Analysis/Computational	400
5.C.	Activities	413
5.D.	Internet Archives	414
6	Statistical Inference: Estimating Probabilities and Testing and Confirming Models	417
6.1	Point Estimation	420
6.2	Interval Estimation	432
6.3	Hypothesis Testing	466
Chapter 6	Comprehensive Exercises	510
6.A.	Conceptual	510
6.B.	Data Analysis/Computational	517
6.C.	Activities	530
6.D.	Internet Archives	533
7	Statistical Inference for the Center of a Population	537
7.1	Exact Inference for the Center of a Population under a Minimal Assumption	539
7.2	Exact Inference for the Center of a Continuous Population Under the Assumption of Population Symmetry	554
7.3	Inference for the Center of a Normal Distribution– Procedures Associated with the Sample Mean and Sample Standard Deviation	574

7.4	Discussion of Methods of Inference for the Center of a Population	593
7.5	Approximate Inference for the Center of a Population when the Number of Sample Observations is Large	604
7.6	Approximate Inference for the Median of an Arbitrary Distribution – Bootstrapping the Sample Median	616
	Chapter 7 Comprehensive Exercises	621
	7.A. Conceptual	621
	7.B. Data Analysis/Computational	623
	7.C. Activities	628
	7.D. Internet Archives	630
8	Statistical Inference for Matched Pairs or Paired Replicates Data	633
8.1	Inference for Continuous Paired Replicates or Matched Pairs Data	636
8.2	Inference for Qualitative Differences—Data from Paired Replicates or Matched Pairs Experiments	649
	Chapter 8 Comprehensive Exercises	655
	8.A. Conceptual	655
	8.B. Data Analysis/Computational	657
	8.C. Activities	666
	8.D. Internet Archives	667
9	Statistical Inference for Two Populations—Independent Samples	669
9.1	Approximate Inference for the Difference in Proportions for Two Populations	671

9.2	Inference for the Difference in Medians for Any Two Continuous Populations	687
9.3	Approximate Inference for the Difference in Means for Two Populations–Procedures Based on the Two Sample Averages and Sample Standard Deviations	708
9.4	Inference for the Difference in Means for Two Normal Populations with Equal Variances--Procedures Based on the Two Sample Averages and a Pooled Sample Standard Deviation	734
9.5	Discussion of the Methods of Inference for the Difference Between the Centers of Two Populations with Independent Samples	748
	Chapter 9 Comprehensive Exercises	749
	9.A. Conceptual	749
	9.B. Data Analysis/Computational	750
	9.C. Activities	766
	9.D. Internet Archives	769
10	Statistical Inference for Two-Way Tables of Count Data	773
10.1	General Test for Differences in Population Proportions	776
10.2	Test for Association (Independence) between Two Categorical Attributes	787
10.3	Exact Procedure for Testing Equality of Two Population Proportions	801
10.4	Goodness-of-fit Test for Probabilities in a Multinomial Distribution with $I > 2$ Categories	808

Chapter 10 Comprehensive Exercises	817
10.A. Conceptual	817
10.B. Data Analysis/Computational	820
10.C. Activities	832
10.D. Internet Archives	834
11 Statistical Inference for Bivariate Populations	839
11.1 Correlation Procedures for Bivariate Normal Populations	840
11.2 Rank-Based Correlation Procedures	851
11.3 Fitting a Least Squares Line to Bivariate Data	860
11.4 Linear Regression Inference for Normal Populations	867
11.5 Rank-Based Linear Regression Inference	874
Chapter 11 Comprehensive Exercises	883
11.A. Conceptual	883
11.B. Data Analysis/Computational	889
11.C. Activities	901
11.D. Internet Archives	902
12 Statistical Inference for More Than Two Populations	907
12.1 One-way Rank-Based General Alternatives ANOVA for More Than Two Populations	909
12.2 One-way General Alternatives ANOVA for More Than Two Normal Populations	916
12.3 One-way Rank-Based Ordered Alternatives ANOVA for More Than Two Populations	926
Chapter 12 Comprehensive Exercises	935
12.A. Conceptual	935
12.B. Data Analysis/Computational	936
12.C. Activities	942
12.D. Internet Archives	944

Appendix A: Listing of Datasets Usage Locations	
Throughout <i>IIS</i>	947
Appendix B: Listing of <i>R</i> Functions Usage Locations	
Throughout <i>IIS</i>	951
Bibliography	955
Index	967