

OpenIntro Statistics

First Edition

David M Diez

Postdoctoral Fellow

Department of Biostatistics

Harvard School of Public Health

david.m.diez@gmail.com

Christopher D Barr

Assistant Professor

Department of Biostatistics

Harvard School of Public Health

cdbarr@gmail.com

Mine Çetinkaya-Rundel

Assistant Professor of the Practice

Department of Statistics

Duke University

cetinkaya.mine@gmail.com

Copyright © 2011. First Edition: July, 2011.

A PDF of this textbook (OpenIntro Statistics) is also available online for free and is released by OpenIntro under Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) United States license at openintro.org. The editable source of this book is also available under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) license. Please see creativecommons.org for details on these licenses.

ISBN: 9-781461-062615

Contents

1	Introduction to data	1
1.1	Case study	1
1.2	Data basics	3
1.2.1	Observations, variables, and cars	3
1.2.2	Data Matrices	4
1.2.3	Types of variables	4
1.2.4	Relationships among variables	5
1.2.5	Associated and independent variables	8
1.3	Examining numerical data	9
1.3.1	Scatterplots for paired data	9
1.3.2	Dot plots and the mean	10
1.3.3	Histograms and shape	11
1.3.4	Variance and standard deviation	13
1.3.5	Box plots, quartiles, and the median	15
1.3.6	Robust statistics	17
1.3.7	Transforming data (special topic)	18
1.4	Considering categorical data	19
1.4.1	Contingency tables	20
1.4.2	Bar plots and proportions	20
1.4.3	Segmented bar and mosaic plots	22
1.4.4	The only pie chart you will see in this book	24
1.4.5	Comparing numerical data across groups	25
1.5	Overview of data collection principles	26
1.5.1	Populations and samples	26
1.5.2	Anecdotal evidence	27
1.5.3	Sampling from a population	28
1.5.4	Explanatory and response variables	29
1.5.5	Introducing observational studies and experiments	30
1.6	Observational studies and sampling strategies	30
1.6.1	Observational studies	30
1.6.2	Three sampling methods (special topic)	31
1.7	Experiments	34
1.7.1	Principles of experimental design	34
1.7.2	Reducing bias in human experiments	36
1.8	Case study: efficacy of sulphinpyrazone (special topic)	36
1.8.1	Variability within data	37
1.8.2	Simulating the study	38
1.8.3	Checking for independence	39

1.9	Exercises	40
1.9.1	Case study	40
1.9.2	Data basics	40
1.9.3	Examining numerical data	42
1.9.4	Considering categorical data	48
1.9.5	Overview of data collection principles	49
1.9.6	Observational studies and sampling strategies	50
1.9.7	Experiments	51
1.9.8	Case study: efficacy of sulphinpyrazone	52
2	Probability (special topic)	55
2.1	Defining probability (special topic)	55
2.1.1	Probability	56
2.1.2	Disjoint or mutually exclusive outcomes	57
2.1.3	Probabilities when events are not disjoint	59
2.1.4	Probability distributions	61
2.1.5	Complement of an event	62
2.1.6	Independence	64
2.2	Continuous distributions (special topic)	66
2.2.1	From histograms to continuous distributions	68
2.2.2	Probabilities from continuous distributions	68
2.3	Conditional probability (special topic)	69
2.3.1	Marginal and joint probabilities	70
2.3.2	Defining conditional probability	71
2.3.3	Smallpox in Boston, 1721	73
2.3.4	General multiplication rule	74
2.3.5	Independence considerations in conditional probability	75
2.3.6	Tree diagrams	76
2.3.7	Bayes' Theorem	78
2.4	Sampling from a small population (special topic)	81
2.5	Random variables (special topic)	83
2.5.1	Expectation	83
2.5.2	Variability in random variables	85
2.5.3	Linear combinations of random variables	87
2.5.4	Variability in linear combinations of random variables	89
2.6	Exercises	92
2.6.1	Defining probability	92
2.6.2	Continuous distributions	95
2.6.3	Conditional probability	96
2.6.4	Sampling from a small population	100
2.6.5	Random variables	101
3	Distributions of random variables	104
3.1	Normal distribution	104
3.1.1	Normal distribution model	105
3.1.2	Standardizing with Z scores	106
3.1.3	Normal probability table	107
3.1.4	Normal probability examples	108
3.1.5	68-95-99.7 rule	112
3.2	Evaluating the normal approximation	113

3.2.1	Normal probability plot	113
3.2.2	Constructing a normal probability plot (special topic)	116
3.3	Geometric distribution (special topic)	117
3.3.1	Bernoulli distribution	118
3.3.2	Geometric distribution	119
3.4	Binomial distribution (special topic)	121
3.4.1	The binomial distribution	122
3.4.2	Normal approximation to the binomial distribution	125
3.5	More discrete distributions (special topic)	128
3.5.1	Negative binomial distribution	128
3.5.2	Poisson distribution	131
3.6	Exercises	133
3.6.1	Normal distribution	133
3.6.2	Evaluating the Normal approximation	136
3.6.3	Geometric distribution	137
3.6.4	Binomial distribution	138
3.6.5	More discrete models	141
4	Foundations for inference	143
4.1	Variability in estimates	144
4.1.1	Point estimates	144
4.1.2	Point estimates are not exact	145
4.1.3	Standard error of the mean	145
4.1.4	Basic properties of point estimates	148
4.2	Confidence intervals	149
4.2.1	Capturing the population parameter	149
4.2.2	An approximate 95% confidence interval	149
4.2.3	A sampling distribution for the mean	150
4.2.4	Changing the confidence level	152
4.2.5	Interpreting confidence intervals	153
4.2.6	Nearly normal with known SD (special topic)	154
4.3	Hypothesis testing	156
4.3.1	Hypothesis testing framework	156
4.3.2	Testing hypotheses using confidence intervals	157
4.3.3	Decision errors	159
4.3.4	Formal testing using p-values	161
4.3.5	Two-sided hypothesis testing with p-values	166
4.3.6	Choosing a significance level	168
4.4	Examining the Central Limit Theorem	169
4.5	Inference for other estimators	172
4.5.1	A general approach to confidence intervals	172
4.5.2	Generalizing hypothesis testing	173
4.6	Sample size and power (special topic)	176
4.6.1	Finding a sample size for a certain margin of error	176
4.6.2	Power and the Type 2 Error rate	177
4.6.3	Statistical significance versus practical significance	179
4.7	Exercises	180
4.7.1	Variability in estimates	180
4.7.2	Confidence intervals	182
4.7.3	Hypothesis testing	186

4.7.4	Examining the Central Limit Theorem	189
4.7.5	Inference for other estimators	190
4.7.6	Sample size and power	191
5	Large sample inference	192
5.1	Paired data	192
5.1.1	Paired observations and samples	192
5.1.2	Inference for paired data	193
5.2	Difference of two means	195
5.2.1	Point estimates and standard errors for differences of means	195
5.2.2	Confidence interval for the difference	197
5.2.3	Hypothesis tests based on a difference in means	197
5.2.4	Summary for inference of the difference of two means	200
5.2.5	Examining the standard error formula	201
5.3	Single population proportion	201
5.3.1	Identifying when a sample proportion is nearly normal	201
5.3.2	Confidence intervals for a proportion	202
5.3.3	Hypothesis testing for a proportion	203
5.3.4	Choosing a sample size when estimating a proportion	204
5.4	Difference of two proportions	206
5.4.1	Sampling distribution of the difference of two proportions	206
5.4.2	Intervals and tests for $p_1 - p_2$	206
5.4.3	Hypothesis testing when $H_0 : p_1 = p_2$	208
5.5	When to retreat	211
5.6	Testing for goodness of fit using chi-square (special topic)	212
5.6.1	Creating a test statistic for one-way tables	212
5.6.2	The chi-square test statistic	213
5.6.3	The chi-square distribution and finding areas	214
5.6.4	Finding a p-value for a chi-square test	216
5.6.5	Evaluating goodness of fit for a distribution	219
5.7	Testing for independence in two-way tables (special topic)	222
5.7.1	Expected counts in two-way tables	223
5.7.2	The chi-square test statistic for two-way tables	225
5.8	Exercises	227
5.8.1	Paired data	227
5.8.2	Difference of two means	228
5.8.3	Single population proportion	229
5.8.4	Difference of two proportions	235
5.8.5	When to retreat	238
5.8.6	Testing for goodness of fit using chi-square	238
5.8.7	Testing for independence in two-way tables	239
6	Small sample inference	241
6.1	Small sample inference for the mean	241
6.1.1	The normality condition	242
6.1.2	Introducing the t distribution	242
6.1.3	Working with the t distribution	243
6.1.4	The t distribution as a solution to the standard error problem	245
6.1.5	One sample confidence intervals with small n	246
6.1.6	One sample t tests with small n	248

6.2	The t distribution for the difference of two means	249
6.2.1	Sampling distributions for the difference in two means	250
6.2.2	Two sample t test	250
6.2.3	Two sample t confidence interval	254
6.2.4	Pooled standard deviation estimate (special topic)	254
6.3	Small sample hypothesis testing for a proportion (special topic)	255
6.3.1	When the success-failure condition is not met	255
6.3.2	Generating the null distribution and p-value by simulation	256
6.3.3	Generating the exact null distribution and p-value	258
6.4	Hypothesis testing for two proportions (special topic)	258
6.4.1	Large sample framework for a difference in two proportions	259
6.4.2	Simulating a difference under the null distribution	260
6.4.3	Null distribution for the difference in two proportions	262
6.5	Exercises	263
6.5.1	Small sample inference for the mean	263
6.5.2	The t distribution for the difference of two means	266
6.5.3	Small sample hypothesis testing for a proportion	270
6.5.4	Hypothesis testing for two proportions	272
7	Introduction to linear regression	274
7.1	Line fitting, residuals, and correlation	276
7.1.1	Beginning with straight lines	276
7.1.2	Fitting a line by eye	277
7.1.3	Residuals	277
7.1.4	Describing linear relationships with correlation	280
7.2	Fitting a line by least squares regression	282
7.2.1	An objective measure for finding the best line	282
7.2.2	Conditions for the least squares line	283
7.2.3	Finding the least squares line	283
7.2.4	Interpreting regression line parameter estimates	286
7.2.5	Extrapolation is treacherous	286
7.2.6	Using R^2 to describe the strength of a fit	288
7.3	Types of outliers in linear regression	288
7.4	Inference for linear regression	290
7.4.1	Midterm elections and unemployment	290
7.4.2	Understanding regression output from software	291
7.4.3	An alternative test statistic	294
7.5	Exercises	295
7.5.1	Line fitting, residuals, and correlation	295
7.5.2	Fitting a line by least squares regression	302
7.5.3	Types of outliers in linear regression	305
7.5.4	Inference for linear regression	306
8	Multiple regression and ANOVA	309
8.1	Introduction to multiple regression	309
8.1.1	Using categorical variables with two levels as predictors	309
8.1.2	Including and assessing many variables in a model	311
8.1.3	Adjusted R^2 as a better estimate of explained variance	313
8.2	Model selection	314
8.2.1	Identifying variables that may not be helpful in the model	314

8.2.2	Two model selection strategies	315
8.3	Checking model assumptions using graphs	317
8.4	ANOVA and regression with categorical variables	321
8.4.1	Is batting performance related to player position in MLB?	323
8.4.2	Analysis of variance (ANOVA) and the F test	325
8.4.3	Reading regression and ANOVA output from software	327
8.4.4	Graphical diagnostics for an ANOVA analysis	328
8.4.5	Multiple comparisons and controlling Type 1 Error rate	329
8.4.6	Using ANOVA for multiple regression	332
8.5	Exercises	334
8.5.1	Introduction to multiple regression	334
8.5.2	Model selection	337
8.5.3	Checking model assumptions using graphs	338
8.5.4	ANOVA and regression with categorical variables	340
A	Bibliography	341
B	End of chapter exercise solutions	345
C	Distribution tables	361
C.1	Normal Probability Table	361
C.2	t Distribution Table	364
C.3	Chi-Square Probability Table	366

Preface

OpenIntro is an organization focused on developing free and affordable education materials. *OpenIntro Statistics*, our first project, is intended for introductory statistics courses at the high school through university levels. The chapters are as follows:

- 1. Introduction to data.** Data structures, variables, summaries, graphics, and basic data collection techniques.
- 2. Probability (special topic).** The basic principles of probability. An understanding of this chapter is not required for the main content in Chapters 3-8.
- 3. Distributions of random variables.** Introduction to the normal model and other key distributions.
- 4. Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean.
- 5. Large sample inference.** Inferential methods for one or two sample means and proportions using the normal model, and also contingency tables via chi-square.
- 6. Small sample inference.** Inference for means using the t distribution, as well as simulation and randomization techniques for proportions.
- 7. Introduction to regression.** An introduction to linear regression. Most of this chapter could be covered after Chapter 1.
- 8. Multiple regression and ANOVA.** An introduction to multiple regression and one-way ANOVA for an accelerated course.

This textbook was written to allow flexibility in choosing and ordering course topics. The material is divided into two pieces: main text and special topics. The goal of the main text is to allow people to move towards statistical inference and modeling sooner rather than later. Special topics, labeled in the table of contents and in section titles, may be added to a course as they arise naturally in the curriculum.

OpenIntro Statistics would also serve as a helpful supplement in a course preparing students for the Advanced Placement Statistics exam, either through the textbook or use of the online resources outlined below.

Examples, exercises, and appendices

Examples and within-chapter exercises have been clearly labeled throughout the textbook and may be identified by their distinctive bullets:

- **Example 0.1** Large filled bullets signal the start of an example.

Full solutions to examples are provided and often include an accompanying table or figure.

- ⊙ **Exercise 0.2** Large empty bullets signal readers that an exercise has been inserted into the text for additional practice and guidance. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for the vast majority of within-chapter exercises in footnotes¹.

There are a large selection of exercises at the end of each chapter useful for practice or homework assignments. Many of these questions have multiple parts, and odd-numbered questions include brief solutions in Appendix B. These end-of-chapter exercises are also available online in a public question bank at **openintro.org**, and the available selection is constantly growing based on teacher contributions. Numbered citations in end-of-chapter exercises may be found in Appendix B.

Probability tables for the normal, t , and chi-square distributions are in Appendix C, and PDF copies of these tables are also available from **openintro.org** for anyone to use, share, or modify.

Online resources and getting involved

We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved by using the many online resources, which are all free and rapidly expanding, or by creating new material. Students can test their knowledge with practice quizzes for each chapter, or try an application of concepts learned using real data. A companion *R* package has also been released, which includes most data sets introduced in this book². Teachers can download the source files for labs, slides, data sets, textbook figures, or create their own custom quizzes and problem sets for students to take at **openintro.org**. Anyone can download a PDF or the source files of this textbook for modifying and sharing. All of these products are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

We value your feedback. If there is a particular component of the project you especially like or think needs improvement, we want to hear from you. You may find our contact information in the About section of the website.

Acknowledgements

This project would not be possible without the dedication and volunteer hours of all those involved. No one has received any monetary compensation from this project, and we hope you will join us in extending a *thank you* to all those volunteers below.

The authors would like to thank Filipp Brunshteyn, Rob Gould, and Chris Pope for their involvement and contributions to OpenIntro. A special thank you to Andrew Bray, who has developed the R labs and diligently worked to translate end-of-chapter exercises to the project website. We also deeply appreciate the contribution of Meenal Patel, who has helped raise the professional profile of OpenIntro by designing a business system and website for the project. We are also very grateful to Dalene Stangl, Dave Harrington, Jan de Leeuw, Kevin Rader, and Philippe Rigollet for providing us with valuable feedback on this textbook.

¹Solutions are usually located down here!

²Diez DM, Barr CD, Çetinkaya-Rundel M. 2011. **openintro**: OpenIntro data sets and supplement functions. <http://cran.r-project.org/web/packages/openintro>