Texts in Statistical Science

Analysis of Categorical Data with R

Christopher R. Bilder
University of Nebraska-Lincoln
Lincoln, Nebraska, USA

Thomas M. Loughin
Simon Fraser University
Surrey, British Columbia, Canada

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140710

International Standard Book Number-13: 978-1-4987-0676-6 (eBook - PDF)

Contents

Preface **xi**

1 Analyzing a binary response, part 1: introduction **1**
1.1 One binary variable . . . 1
1.1.1 Bernoulli and binomial probability distributions . . . 1
1.1.2 Inference for the probability of success . . . 8
1.1.3 True confidence levels for confidence intervals . . . 17
1.2 Two binary variables . . . 23
1.2.1 Notation and model . . . 25
1.2.2 Confidence intervals for the difference of two probabilities . . . 29
1.2.3 Test for the difference of two probabilities . . . 34
1.2.4 Relative risks . . . 37
1.2.5 Odds ratios . . . 40
1.2.6 Matched pairs data . . . 43
1.2.7 Larger contingency tables . . . 47
1.3 Exercises . . . 48

2 Analyzing a binary response, part 2: regression models **61**
2.1 Linear regression models . . . 61
2.2 Logistic regression models . . . 62
2.2.1 Parameter estimation . . . 64
2.2.2 Hypothesis tests for regression parameters . . . 76
2.2.3 Odds ratios . . . 82
2.2.4 Probability of success . . . 86
2.2.5 Interactions and transformations for explanatory variables . . . 94
2.2.6 Categorical explanatory variables . . . 100
2.2.7 Convergence of parameter estimation . . . 110
2.2.8 Monte Carlo simulation . . . 116
2.3 Generalized linear models . . . 121
2.4 Exercises . . . 129

3 Analyzing a multicategory response **141**
3.1 Multinomial probability distribution . . . 141
3.2 $I \times J$ contingency tables and inference procedures . . . 143
3.2.1 One multinomial distribution . . . 143
3.2.2 I multinomial distributions . . . 144
3.2.3 Test for independence . . . 146
3.3 Nominal response regression models . . . 152
3.3.1 Odds ratios . . . 160
3.3.2 Contingency tables . . . 163
3.4 Ordinal response regression models . . . 170
3.4.1 Odds ratios . . . 177
3.4.2 Contingency tables . . . 179

3.4.3 Non-proportional odds model . . . 182
3.5 Additional regression models . . . 186
3.6 Exercises . . . 187

4 Analyzing a count response **195**
4.1 Poisson model for count data . . . 195
4.1.1 Poisson distribution . . . 195
4.1.2 Poisson likelihood and inference . . . 198
4.2 Poisson regression models for count responses . . . 201
4.2.1 Model for mean: Log link . . . 203
4.2.2 Parameter estimation and inference . . . 204
4.2.3 Categorical explanatory variables . . . 211
4.2.4 Poisson regression for contingency tables: loglinear models . . . 218
4.2.5 Large loglinear models . . . 222
4.2.6 Ordinal categorical variables . . . 231
4.3 Poisson rate regression . . . 240
4.4 Zero inflation . . . 244
4.5 Exercises . . . 253

5 Model selection and evaluation **265**
5.1 Variable selection . . . 265
5.1.1 Overview of variable selection . . . 265
5.1.2 Model comparison criteria . . . 267
5.1.3 All-subsets regression . . . 268
5.1.4 Stepwise variable selection . . . 272
5.1.5 Modern variable selection methods . . . 277
5.1.6 Model averaging . . . 281
5.2 Tools to assess model fit . . . 285
5.2.1 Residuals . . . 286
5.2.2 Goodness of fit . . . 292
5.2.3 Influence . . . 296
5.2.4 Diagnostics for multicategory response models . . . 301
5.3 Overdispersion . . . 301
5.3.1 Causes and implications . . . 302
5.3.2 Detection . . . 305
5.3.3 Solutions . . . 306
5.4 Examples . . . 318
5.4.1 Logistic regression - placekicking data set . . . 318
5.4.2 Poisson regression - alcohol consumption data set . . . 329
5.5 Exercises . . . 345

6 Additional topics **355**
6.1 Binary responses and testing error . . . 355
6.1.1 Estimating the probability of success . . . 355
6.1.2 Binary regression models . . . 358
6.1.3 Other methods . . . 361
6.2 Exact inference . . . 362
6.2.1 Fisher's exact test for independence . . . 362
6.2.2 Permutation test for independence . . . 367
6.2.3 Exact logistic regression . . . 372
6.2.4 Additional exact inference procedures . . . 380

6.3 Categorical data analysis in complex survey designs . . . 380
6.3.1 The survey sampling paradigm . . . 381
6.3.2 Overview of analysis approaches . . . 382
6.3.3 Weighted cell counts . . . 386
6.3.4 Inference on population proportions . . . 387
6.3.5 Contingency tables and loglinear models . . . 389
6.3.6 Logistic regression . . . 400
6.4 "Choose all that apply" data . . . 404
6.4.1 Item response table . . . 405
6.4.2 Testing for marginal independence . . . 405
6.4.3 Regression modeling . . . 412
6.5 Mixed models and estimating equations for correlated data . . . 419
6.5.1 Random effects . . . 420
6.5.2 Mixed-effects models . . . 422
6.5.3 Model fitting . . . 425
6.5.4 Inference . . . 432
6.5.5 Marginal modeling using generalized estimating equations . . . 438
6.6 Bayesian methods for categorical data . . . 443
6.6.1 Estimating a probability of success . . . 444
6.6.2 Regression models . . . 449
6.6.3 Alternative computational tools . . . 459
6.7 Exercises . . . 459

A An introduction to R **473**
A.1 Basics . . . 473
A.2 Functions . . . 475
A.3 Help . . . 476
A.4 Using functions on vectors . . . 477
A.5 Packages . . . 478
A.6 Program editors . . . 479
A.6.1 R editor . . . 479
A.6.2 RStudio . . . 479
A.6.3 Tinn-R . . . 480
A.6.4 Other editors . . . 482
A.7 Regression example . . . 482
A.7.1 Background . . . 482
A.7.2 Data summary . . . 483
A.7.3 Regression modeling . . . 485
A.7.4 Additional items . . . 491

B Likelihood methods **495**
B.1 Introduction . . . 495
B.1.1 Model and parameters . . . 495
B.1.2 The role of likelihoods . . . 496
B.2 Likelihood . . . 496
B.2.1 Definition . . . 496
B.2.2 Examples . . . 497
B.3 Maximum likelihood estimates . . . 498
B.3.1 Mathematical maximization of the log-likelihood function . . . 500
B.3.2 Computational maximization of the log-likelihood function . . . 501
B.3.3 Large-sample properties of the MLE . . . 502

B.3.4 Variance of the MLE 502
B.4 Functions of parameters 504
B.4.1 Invariance property of MLEs 504
B.4.2 Delta method for variances of functions 505
B.5 Inference with MLEs 506
B.5.1 Tests for parameters 506
B.5.2 Confidence intervals for parameters 510
B.5.3 Tests for models 511

Bibliography **513**

Index **525**

Preface

We live in a categorical world! From a positive or negative disease diagnosis to choosing all items that apply in a survey, outcomes are frequently organized into categories so that people can more easily make sense of them. However, analyzing data from categorical responses requires specialized techniques beyond those learned in a first or second course in Statistics. We offer this book to help students and researchers learn how to properly analyze categorical data. Unlike other texts on similar topics, our book is a modern account using the vastly popular R software. We use R not only as a data analysis method but also as a learning tool. For example, we use data simulation to help readers understand the underlying assumptions of a procedure and then to evaluate that procedure's performance. We also provide numerous graphical demonstrations of the features and properties of various analysis methods.

The focus of this book is on the analysis of data, rather than on the mathematical development of methods. We offer numerous examples from a wide rage of disciplines—medicine, psychology, sports, ecology, and others—and provide extensive R code and output as we work through the examples. We give detailed advice and guidelines regarding which procedures to use and why to use them. While we treat likelihood methods as a tool, they are not used blindly. For example, we write out likelihood functions and explain how they are maximized. We describe where Wald, likelihood ratio, and score procedures come from. However, except in Appendix B, where we give a general introduction to likelihood methods, we do not frequently emphasize calculus or carry out mathematical analysis in the text. The use of calculus is mostly from a conceptual focus, rather than a mathematical one.

We therefore expect that this book will appeal to all readers with a basic background in regression analysis. At times, a rudimentary understanding of derivatives, integrals, and function maximization would be helpful, as would a very basic understanding of matrices, matrix multiplication, and finding inverses of matrices. However, the important points and application advice can be easily understood without these tools. We expect that advanced undergraduates in statistics and related fields will satisfy these prerequisites. Graduate students in statistics, biostatistics, and related fields will certainly have sufficient background for the book. In addition, many students and researchers outside these disciplines who possess the basic regression background should find this book useful both for its descriptions and motivations of the analysis methods and for its worked examples with R code.

The book does not require any prior experience with R. We provide an introduction to the essential features and functions of R in Appendix A. We also provide introductory details on the use of R in the earlier chapters to help inexperienced R users. Throughout the book as new R functions are needed, their basic features are discussed in the text and their implementation shown with corresponding output. We focus on using R packages that are provided by default with the initial R installation. However, we make frequent use of other R packages when they are significantly better or contain functionality unavailable in the standard R packages. The book contains the code and output as it would appear in the R Console; we make minor modifications at times to the output only to save space within the book. Code provided in the book for plotting is often meant for color display rather than the actual black-and-white display shown in the print and some electronic editions.

The data set files and R programs that are referenced in each example are available from the book's website, `http://www.chrisbilder.com/categorical`. The programs include code used to create every plot and piece of output that we show. Many of these programs contain code to demonstrate additional features or to perform more detailed and complete analyses than what is presented in the book. We strongly recommend that the book and the website be used in tandem, both for teaching and for individual learning. The website also contains many "extras" that can help readers learn the material. Most importantly, we post videos from one of us teaching a course on the subject. These videos include live, in-class recordings that are synchronized with recordings of a tablet computer screen. Instructors may find these videos useful (as we have) for a blended or flipped classroom setting. Readers outside of a classroom setting may also find these videos especially useful as a substitute for a short-course on the subject.

The first four chapters of the book are organized by type of categorical response variable. Within each of these chapters, we first introduce the measurement type, followed by the basic distributional model that is most commonly used for that type of measurement. We slowly generalize to simple regression structures, followed by multiple regressions including transformations, interactions, and categorical explanatory variables. We conclude each of these chapters with some important special cases. Chapter 5 follows with model building and assessment methods for the response variables in the first four chapters. A final chapter discusses additional topics presented as extensions to the previous chapters. These topics include solutions to problems that are frequently mishandled in practice, such as how to incorporate diagnostic testing error into an analysis, the analysis of data from "choose all that apply" questions, and methods for analyzing data arising under a complex survey sampling design. Many of these topics are broad enough that entire books have been written about them, so our treatment in Chapter 6 is meant to be introductory.

For instructors teaching a one-semester course with the book, we recommend covering most of Chapters 1–5. The topics in Chapter 6 provide supplemental material for readers to learn on their own or to provide an instructor a means to go beyond the basics. In particular, topics from Chapter 6 can make good class projects. This helps students gain experience in teaching themselves extensions to familiar topics, which they will face later in industry or in research.

An extensive set of exercises is provided at the end of each chapter (over 65 pages in all!). The exercises are deliberately variable in scope and subject matter, so that instructors can choose those that meet the particular needs of their own students. For example, some carry out an analysis step by step, while others present a problem and leave the reader to choose and implement a solution. An answer key to the exercises is available for instructors using the book for a course. Details on how to obtain the answer key are available through the book's website.

We could not have written this book without the help and support of many people. First and foremost, we thank our families, and especially our wives, Kimberly and Marie, who put in extra effort on our behalf so that we could reserve time to work on the book. We owe them a huge debt for their support and tolerance, but we are hoping that they will settle for a nice dinner. We thank Rob Calver and his staff at CRC Press for their continued support and encouragement during the writing process. We also thank the hundreds of students who have taken categorical courses from us over the last seventeen years. Their feedback helped us to hone the course material and its presentation to what they are today. We especially thank one of our students, Natalie Koziol, who wrote the MRCV package used in Section 6.4 and made the implementation of those methods available to R users. This book was written in LaTeX through LyX, and we are grateful to the many contributors to these open-source projects. Finally, we need to thank our past and present colleagues and mentors at Iowa State, Kansas State, Oklahoma State, Nebraska, and Simon Fraser Universities who have

both supported our development and brought us interesting and challenging problems to work on.

Christopher R. Bilder and Thomas M. Loughin
Lincoln, NE and Surrey, BC

Appendix B

Likelihood methods

B.1 Introduction

Likelihood methods are commonly used in statistical analyses. In particular, they provide the basis for most of the categorical data analysis techniques that are covered in this book. The main body of this book is written assuming that readers are already familiar with likelihood-based methods, at least enough to be able to apply them with some confidence. However, we recognize that readers may come from a broad range of disciplines and may not have had any previous exposure to likelihood methods. We therefore offer a brief primer on this important class of procedures. We suggest that readers who have no past exposure to likelihood methods at least look over this appendix before reading the book's main body. We cross-reference this appendix with the sections where likelihood methods are used, so that readers and instructors can also refer to this material as needed.

The scope of this appendix is deliberately limited. An understanding of the theory and asymptotics of likelihood methods is not at all required for this book, although a heuristic appreciation of them is helpful. Accordingly, the descriptions in this appendix are light in mathematics. For a more complete treatment of likelihood methods, please see texts such as Casella and Berger (2002) and Severini (2000).

Statistical Inference. Duxbury Press, 2nd edition.

B.1.1 Model and parameters

The goal of a statistical analysis is to learn something about a population (what we might call the "truth"). To achieve this goal, data are gathered, but data contain variability (or "noise") that prevents us from seeing the truth clearly. In order to extract truth from data, it helps if we start with some idea of what the truth looks like and if we know something about the origin of the noise. For example, rather than just looking at a plot of a response against an explanatory variable, it may help to speculate that the relationship should be a straight line and that the deviations around the straight line should be independent and approximately normally distributed. A *statistical model* is an assumed structure for the truth and the noise (i.e., it is an educated guess). The features of the model are combined into a probability distribution, such as a normal, Bernoulli, or Poisson, that is intended to serve as a useful approximation to reality.

Models generally contain *parameters*. Model parameters relate to the structure of the truth and/or noise and hold the place of unknown or flexible features of the model. They represent population quantities that are often of direct interest to the researcher—such as the mean of a normal distribution, or the probability of success in a binomial distribution—but not always (we often do not really care about the intercept parameter in many linear regression problems). The goal of a statistical analysis is to learn about the model parameters or some functions of the model parameters (like predictions in a regression, which are a function of the slope and intercept parameters). Functions of model parameters are also

parameters, since they are also unknown population quantities, so our discussion will not distinguish between parameters in the model and other parameters.

B.1.2 The role of likelihoods

A statistical model serves to relate the data to the parameters. We still need to find values for the parameters and use them to learn about the population. The first step—finding values for the parameters—is called *estimation*. The second step—using them to draw conclusions—is called *inference*.

There are many ways to estimate parameters from a statistical model, but the one that has been adopted almost universally because of its quality and flexibility is maximum likelihood (ML) estimation. This is because the procedure is adaptable to nearly any statistical model, leading to an automatic process for estimation. It also has associated with it a variety of tools that can be used for inference. The procedure and its tools have strengths and weaknesses. We discuss these as they arise in different settings throughout the text.

B.2 Likelihood

B.2.1 Definition

Statistical models are generally described in terms of a *probability mass function* or a *probability density function.* A probability mass function (PMF) for a discrete random variable provides the probabilities for each possible outcome. For a continuous random variable, the corresponding probability density function (PDF) is a little more complicated to interpret, because it assumes that measurements are made to an infinite number of decimal places. Loosely speaking, a PDF describes the relative chances of observing values from different areas of the stated probability distribution. The familiar "normal curve" is an example of a PDF. In both the discrete and continuous cases, higher values of the PMF or PDF are produced by ranges of values that are "more likely" to occur. The *joint* PMF or PDF for a sample of observations can be interpreted as how likely we are to observe the entire sample, given the distribution and its parameters.

In practice, we do not know the values of the parameters, but we do know the data. We therefore cannot know the joint mass or density for our sample exactly. We *can* calculate this quantity if we assume certain values for the parameters. If we change the values of the parameters, we get a different value for the PDF/PMF of the sample, because the data are more or less likely to occur under different parameters. For example, it is very unlikely that we would observe 5 successes in 10 Bernoulli trials when the true probability of success is 0.01. This outcome would be somewhat more likely to happen if the probability of success is 0.30, and even more likely if it is 0.50.

This is the nature of the *likelihood function*: we consider the PMF or PDF in reverse. We observe how the function changes for different values of parameters, while holding the data fixed. We can then make use of this to judge which values of the parameters lead to greater relative chances for the sample to occur. Formally, if we define the joint PMF or PDF of a sample to be $f(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y} = (y_1, \ldots, y_n)$ represents a vector[1] containing the

[1] A vector can be thought of as a simple way to express a group of values using one symbol.

n sampled values and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ represents a vector for p different parameters[2], then the likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}).$$

Larger values of the likelihood correspond to values of the parameters that are relatively better supported by the data.

A likelihood is not a probability, because the only random part, $\mathbf{y}$, is considered fixed in its construction. In particular, the likelihood is not expected to add to 1 across all values of $\boldsymbol{\theta}$. The actual numerical values of a likelihood are unimportant. The *relative* sizes of the likelihoods for different parameter values are all that matters.

When the observations are drawn independently, the likelihood function is simply the product of the PDFs or PMFs,

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}),$$

where we use the $\prod$ symbol to denote multiplying indexed terms together. Thus, likelihoods are very easy to construct for simple random samples, which is the setting for most problems in this book. Also, notice that the value of the likelihood depends on the sample, so that likelihoods—and any features calculated from them—are statistics. This means that such features are random and have probability distributions.

B.2.2 Examples

Below are some simple examples of likelihoods that are commonly used in categorical data analysis.

Example: Bernoulli

Suppose that the random variable Y takes on only two possible values. The Bernoulli PMF for Y is $f(y|\pi) = \pi^y(1-\pi)^{(1-y)}$ with probability-of-success parameter π ($0 < \pi < 1$) and $y = 1$ or 0 denoting a "success" or a "failure," respectively. Let $y_1, \ldots, y_n$ be observations of independent Bernoulli random variables from this PMF. The likelihood for the parameter π is

$$L(\pi|\mathbf{y}) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{(1-y_i)} = \pi^w(1-\pi)^{n-w}, \tag{B.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ and $w = \sum_{i=1}^{n} y_i$. This likelihood is used in Section 1.1.1.

Example: Binomial

An alternative form of the Bernoulli case occurs when the total number of successes, w, is observed instead of the individual trial results. In this case, the joint PMF of $y_1, \ldots, y_n$ cannot be found because we do not know which y_i should be 1 and which should be 0. However, we know that there are $\binom{n}{w} = n!/[w!(n-w)!]$ ways for the w successes to be observed among the n observations, so the PMF of w given π is

$$f(w|\pi) = \frac{n!}{w!(n-w)!}\pi^w(1-\pi)^{n-w}.$$

[2]The actual symbols used for the parameters in a given problem vary depending on the context of the problem; see examples later.

If only one set of n trials is run, so that only one total number of successes w is observed, then $L(\pi|w) = f(w|\pi)$. Notice that this is very similar to the Bernoulli likelihood.

Example: Poisson

The Poisson PMF for a random variable Y is $f(y|\mu) = e^{-\mu}\mu^y/y!$ with parameter $\mu > 0$ and y taking on integer values 0, 1, 2, ..., such as for a count of something. Let $y_1, \ldots, y_n$ be observations of independent Poisson random variables. The likelihood for the parameter μ is

$$L(\mu|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}, \tag{B.2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$. This likelihood is used in Section 4.1.2.

Example: Multinomial

Consider a random variable Y with responses consisting of one of c categories, labeled 1, 2, ..., c, with respective probabilities of success π_1, π_2, ..., π_c ($\sum_{k=1}^{c} \pi_k = 1$). Let $y_1, \ldots, y_n$ be observations of Y measured on independent trials of this type; that is, the possible values for each y_i are the categories $1, 2, \ldots, c$. Likelihoods can be constructed as for the Bernoulli and binomial cases above, depending on whether individual trial results $y_1, \ldots, y_n$ are observed or just the summary counts for each category, $w_1, \ldots, w_c$. The multinomial distribution is based on the summary counts, for which the PMF is

$$f(w_1, \ldots, w_c|\pi_1, \ldots, \pi_c) = \frac{n!}{w_1!w_2!\ldots w_c!}\pi_1^{w_1}\pi_2^{w_2}\ldots\pi_c^{w_c}. \tag{B.3}$$

If only one set of n trials is run, so that only one set of category counts $w_1, \ldots, w_c$ is observed, then we have $L(\pi_1, \ldots, \pi_c|w_1, \ldots, w_c) = f(w_1, \ldots, w_c|\pi_1, \ldots, \pi_c)$. This likelihood is used in Section 3.1.

B.3 Maximum likelihood estimates

We define the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}$, of a parameter as the value of the parameter at which the likelihood function from the given sample is maximized: $L(\hat{\boldsymbol{\theta}}|\mathbf{y}) \geq L(\tilde{\boldsymbol{\theta}}|\mathbf{y})$ for any possible value $\tilde{\boldsymbol{\theta}}$ of the parameter. The next example illustrates how to find this MLE through simple evaluation of a likelihood function.

Example: MLE for a sample of Bernoulli random variables (LikelihoodFunction.R)

Suppose $\sum_{i=1}^{n} y_i = 4$ successes are observed out of $n = 10$ trials. Given this information, we want to determine the most plausible value for π. The likelihood function

Table B.1: $L(\pi|\mathbf{y})$ evaluated at different values of π.

π	$L(\pi\|\mathbf{y})$
0.20	0.000419
0.30	0.000953
0.35	0.001132
0.39	0.001192
0.40	0.001194
0.41	0.001192
0.50	0.000977

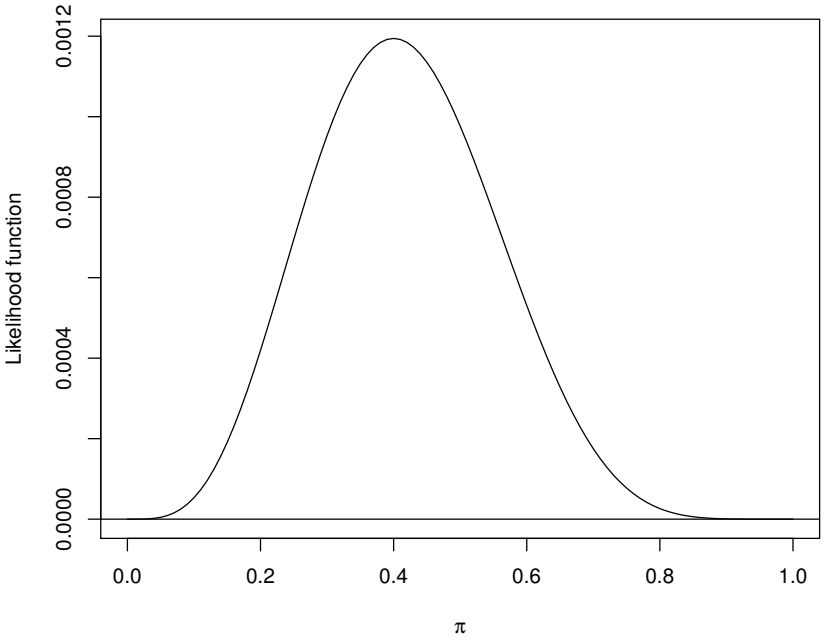

Figure B.1: Bernoulli likelihood function evaluated at $\sum_{i=1}^{n} y_i = 4$ and $n = 10$.

is $L(\pi|\mathbf{y}) = \pi^4(1-\pi)^6$. Table B.1 shows the likelihood function evaluated at a few different values of π, and Figure B.1 plots the function (R code for the table and plot is included in the corresponding program for this example). We can see that the likelihood function reaches its maximum value when $\pi = 0.4$. Therefore, the most plausible value of π, given the observed data, is 0.4, which makes sense because it is the observed proportion of successes. Formally, we say 0.4 is the MLE of π, and we denote it as $\hat{\pi} = 0.4$.

For various mathematical reasons it turns out to be easier to work with the natural log of the likelihood, $\log[L(\hat{\boldsymbol{\theta}}|\mathbf{y})]$, when we attempt to find MLEs. This causes no problem because the log transformation does not change the ordering of likelihood values across different values of $\boldsymbol{\theta}$, so the MLE also maximizes $\log[L(\boldsymbol{\theta}|\mathbf{y})]$.

B.3.1 Mathematical maximization of the log-likelihood function

For simple models with a single parameter, finding the value of $\boldsymbol{\theta}$ that maximizes the log-likelihood function is easily done using calculus. The standard technique is to differentiate the log-likelihood function with respect to the parameter, set the result equal to 0, and solve for the parameter. This process is demonstrated on some of the examples from earlier.

Example: Bernoulli

From Equation B.1, we obtain

$$\log[L(\pi|\mathbf{y})] = w\log\pi + (n-w)\log(1-\pi).$$

Differentiating, we find

$$\frac{d}{d\pi}\log[L(\pi|\mathbf{y})] = \frac{w}{\pi} - \frac{n-w}{1-\pi}.$$

Setting this equal to 0 and solving for π leads to $\hat{\pi} = w/n = \sum_{i=1}^{n} y_i/n$. Thus, the MLE for π is the sample proportion of successes, which was shown in the previous example for $\sum_{i=1}^{n} y_i = 4$ and $n = 10$.

Example: Poisson

From Equation B.2, we obtain

$$\log[L(\mu|\mathbf{y})] = -n\mu + \sum_{i=1}^{n} y_i \log\mu - \sum_{i=1}^{n}\log(y_i!).$$

Differentiating, we find

$$\frac{d}{d\mu}l(\mu|\mathbf{y}) = -n + \frac{1}{\mu}\sum_{i=1}^{n} y_i.$$

Setting this equal to 0 and solving for μ leads to $\hat{\mu} = \sum_{i=1}^{n} y_i/n$. Thus, the MLE for μ is the sample mean.

Models with several parameters Many models rely on more than one parameter. The multinomial model in Equation B.3 is an example, because there is a different probability parameter for each category. Other models use regression functions to describe the relationship between population means or probabilities and one or more explanatory variables. In these models, the maximization is carried out exactly as described above, with a separate derivative taken with respect to each parameter. Setting each derivative equal to zero results in a system of equations that needs to be solved simultaneously in order to find the MLE. Theoretically, this presents no difficulty. Practically, the process can be challenging to carry out manually, and the equations may not have closed-form solutions. For example, see the logistic regression likelihood from Section 2.2.

B.3.2 Computational maximization of the log-likelihood function

When the equations have no closed-form solution, they are solved using an educated version of trial-and-error. The general idea is to start with some initial guess at the parameter value, calculate the log-likelihood for that value, and then iteratively find parameter values with larger and larger log-likelihoods until no further improvement can be achieved. Improving the log-likelihood can be done, for example, by calculating the slope of the log-likelihood at the current guess and then moving the next guess some distance in the direction leading to larger log-likelihood values. Alternatively, one can work with the first derivative of the log-likelihood function and seek values of the parameters that cause it to be zero. There are a variety of fast, reliable computational algorithms for carrying out these procedures. One of the most widely implemented is the Newton-Raphson algorithm.

Example: MLE for a sample of Bernoulli random variables (NewtonRaphson.R)

We demonstrate the Newton-Raphson algorithm in a simple setting to find the MLE of π for the previous Bernoulli example where $\sum_{i=1}^{n} y_i = 4$ in $n = 10$ trials. We found earlier that $\hat{\pi} = \sum_{i=1}^{n} y_i/n = 0.4$.

In this context the algorithm uses the following equation to obtain an estimate $\pi^{(i+1)}$ using a previous estimate $\pi^{(i)}$:

$$\begin{aligned}\pi^{(i+1)} &= \pi^{(i)} - \frac{\left.\frac{d}{d\pi}\log[L(\pi|\mathbf{y})]\right|_{\pi=\pi^{(i)}}}{\left.\frac{d^2}{d\pi^2}\log[L(\pi|\mathbf{y})]\right|_{\pi=\pi^{(i)}}} \\ &= \pi^{(i)} - \frac{\frac{w}{\pi^{(i)}} - \frac{n-w}{1-\pi^{(i)}}}{-\frac{w}{\left(\pi^{(i)}\right)^2} - \frac{n-w}{\left(1-\pi^{(i)}\right)^2}} \end{aligned} \tag{B.4}$$

Equation B.4 comes from a first-order Taylor series expansion about $\pi^{(i)}$.[3] To begin using the algorithm, we choose a starting value $\pi^{(0)}$ that we think may be close to $\hat{\pi}$ and substitute this into Equation B.4 for $\pi^{(i)}$ to obtain $\pi^{(1)}$. If $\pi^{(1)}$ is "close enough" to $\pi^{(0)}$, we stop and use $\pi^{(1)}$ as $\hat{\pi}$; otherwise, we substitute $\pi^{(1)}$ into Equation B.4 for $\pi^{(i)}$ to find $\pi^{(2)}$. This process continues until $\left|\pi^{(i+1)} - \pi^{(i)}\right| < \varepsilon$ for some small number $\epsilon > 0$, and we say that *convergence* has been reached at iteration $i + 1$.

Using the observed data, suppose we guess $\pi^{(0)} = 0.3$, and we feel that $\varepsilon = 0.0001$ represents "close enough." Table B.2 shows the iteration history where convergence is obtained after 5 iterations. The R code that produced this table is available in the corresponding program for this example. We have included in this program the code to create plots illustrating the Newton-Raphson algorithm.

Generally, readers of this book will not need to implement a Newton-Raphson procedure like this directly. Instead, we will use R functions that take care of these details. Also, note that there are other algorithms besides Newton-Raphson that are used to find maximum likelihood estimates. Many of these are implemented in R's `optim()` function.

[3] Suppose we would like to approximate a function $f(x)$ at a point x_0. The first-order Taylor series expansion approximates $f(x)$ with $f(x_0) + (x - x_0)f'(x_0)$, where $f'(\cdot)$ is the first derivative of $f(\cdot)$ with respect to x.

Table B.2: Iterations for the Newton-Raphson algorithm.

Iteration	$\pi^{(i)}$
1	0.3000000
2	0.3840000
3	0.3997528
4	0.3999999
5	0.4000000

B.3.3 Large-sample properties of the MLE

In order to use an MLE in confidence intervals and tests, we need to know its probability distribution (the probability distribution of a statistic is also known as its "sampling distribution"). It can be shown that *all* of the MLEs we will use share certain properties relating to their sampling distributions that make them very appealing bases for inference procedures. These properties generally hold for large samples; in other words, these properties generally hold *asymptotically*, which means, as the sample size grows toward ∞. Below is a list of the properties:

1. MLEs are asymptotically normally distributed – This result is analogous to the central limit theorem for sample means. The fact that normality holds asymptotically means that in any given sample, the normal distribution is typically an *approximation* to the correct sampling distribution of $\hat{\boldsymbol{\theta}}$, and the approximation gets better with larger sample sizes.

2. MLEs are consistent – This means essentially that if you sample the whole population (or sample infinitely), the MLE will be exactly the same as the population parameter. In particular, any bias in the estimate (the difference between the expected value of the MLE and the true value of the parameter) vanishes as the sample size grows, and the variance shrinks to 0.

3. MLEs are asymptotically efficient – This means that as the sample size grows toward ∞ they achieve the smallest variance possible for estimates of their type (e.g., among all asymptotically normal estimates). An important implication of this result is that confidence intervals based on MLEs have the potential to be shorter and tests more powerful than those based on other forms of estimates.

These properties are not guaranteed to hold in samples that are not "large." The normal distribution approximation is generally very good in "large" samples, but may be very poor in "small" samples. Other estimates may have smaller variance than the MLE in finite samples. Unfortunately, there is no uniform way to define "large" or "small." However, it is often not too difficult to simulate data from the chosen model and check whether the MLE has a distribution that appears roughly normal and has an acceptably small bias.

B.3.4 Variance of the MLE

The variance of any MLE is related to the curvature of the log-likelihood function in the neighborhood of the MLE. If the log-likelihood is very flat near the maximum, then there is much uncertainty in the data regarding the location of the parameter (many different values of $\boldsymbol{\theta}$ lead to similarly large likelihoods). Conversely, if the log-likelihood has a sharp peak, then the data show little doubt about the region in which the parameter must lie.

Formally, the variance of the MLE reduces in large samples to a quantity that is estimated directly from the second derivative of the log-likelihood at its peak,

$$\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log[L(\boldsymbol{\theta}|\mathbf{y})],$$

which is referred to as the *Hessian* matrix. From the Hessian, the estimated asymptotic variance (we will refer to this more simply as the "estimated variance") is computed as

$$\widehat{Var}(\hat{\boldsymbol{\theta}}) = -E\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log[L(\boldsymbol{\theta}|\mathbf{Y})]\right)^{-1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{B.5}$$

The estimated variance can instead be approximated by

$$\widehat{Var}(\hat{\boldsymbol{\theta}}) \approx -\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log[L(\boldsymbol{\theta}|\mathbf{y})]\right)^{-1}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{B.6}$$

which is often easier to compute. Equation B.6 is *asymptotically equivalent* to Equation B.5, meaning that these estimated variances will essentially be the same in very large samples. Other asymptotically equivalent variants of these formulas are sometimes used as well. The estimated standard deviation of the statistic (i.e., the standard error) is computed by finding the square root of $\widehat{Var}(\hat{\boldsymbol{\theta}})$.

When $p > 1$,

$$\widehat{Var}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \widehat{Var}(\hat{\theta}_1) & \widehat{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \cdots & \widehat{Cov}(\hat{\theta}_1, \hat{\theta}_p) \\ \widehat{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \widehat{Var}(\hat{\theta}_2) & \cdots & \widehat{Cov}(\hat{\theta}_2, \hat{\theta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{Cov}(\hat{\theta}_1, \hat{\theta}_p) & \widehat{Cov}(\hat{\theta}_2, \hat{\theta}_p) & \cdots & \widehat{Var}(\hat{\theta}_p) \end{bmatrix}.$$

This is known as an estimated *variance-covariance matrix* (sometimes shortened to *covariance matrix* or *variance matrix*). The diagonal elements of the matrix (where the row and column numbers are the same) are the estimated variances of the MLEs. The off-diagonal elements are the covariances between pairs of MLEs. These estimated covariances measure the dependence between the MLEs, and they can be useful for finding the variances of functions of MLEs (see Appendix B.4). Because $\widehat{Cov}(\hat{\theta}_i, \hat{\theta}_j) = \widehat{Cov}(\hat{\theta}_j, \hat{\theta}_i)$, the matrix is symmetric, so that the element in row i, column j equals the element in row j, column i for any $i \neq j$.

Example: Poisson (PoissonLikelihood.R)

The purpose of this example is to find the variance for $\hat{\mu}$ in the Poisson example. Figure B.2 shows the curvature of the log-likelihood function for two different samples, where sample 1 has $\mathbf{y} = (3, 5, 6, 6, 7, 10, 13, 15, 18, 22)$ and sample 2 has $\mathbf{y} = (9, 12)$. For both samples, $\hat{\mu} = 10.5$. We can see that sample 2's log-likelihood function is relatively flat due to its small sample size, while sample 1's log-likelihood function has much more curvature due to its larger sample size. Because the variance for $\hat{\mu}$ is based on this curvature, we would expect the variance for sample 1 to be much less than the variance for sample 2.

Formally, we calculate the variance as

$$\widehat{Var}(\hat{\mu}) = -\left(\frac{\partial^2}{\partial \mu^2} \log[L(\mu|\mathbf{y})]\right)^{-1}\Bigg|_{\mu=\hat{\mu}} = \frac{\hat{\mu}^2}{\sum_{i=1}^{n} y_i} = \frac{\hat{\mu}}{n}.$$

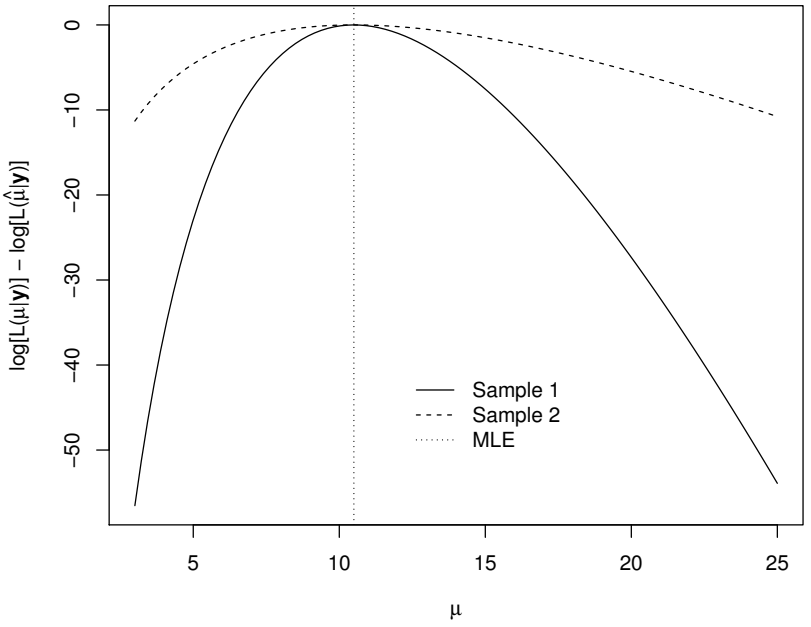

Figure B.2: Poisson log-likelihoods for samples of size $n = 10$ (sample 1) and $n = 2$ (sample 2) with common MLE $\hat{\mu} = 10.5$. Note that the two curves have been shifted vertically so that the log-likelihoods are both 0 at the MLE.

For sample 1, $\widehat{Var}(\hat{\mu}) = 10.5/10 = 1.05$, while for sample 2, $\widehat{Var}(\hat{\mu}) = 10.5/2 = 5.25$. As expected, the variance for $\hat{\mu}$ is larger for sample 2 than for sample 1. The R code for these calculations is included in the corresponding R program for this example.

B.4 Functions of parameters

B.4.1 Invariance property of MLEs

Often our interest is not limited to the model parameters that are directly estimated using the procedures in Appendix B.3.1 or B.3.2. Instead, we may want to estimate parameters that are *functions* of model parameters. As examples, (1) the odds ratio (e.g., Sections 1.2.5, 2.2.3, 3.3.1, and 4.2.3) can be written as a function of probabilities from two binomials, or from multinomial probabilities, or as a function of Poisson means; and (2) predicted values in any regression model can be written as functions of regression coefficients. The *invariance property of MLEs* states that if $\hat{\boldsymbol{\theta}}$ is the MLE for $\boldsymbol{\theta}$ and $g(\boldsymbol{\theta})$ is any function of $\boldsymbol{\theta}$ then $g(\hat{\boldsymbol{\theta}})$ is the MLE for $g(\boldsymbol{\theta})$. In words, the MLE of any function of parameters is just the same function of the parameters' MLE. The important implication of this result is that the large sample properties listed in Appendix B.3.3 hold for functions of MLEs, and all of the

inference procedures in Appendix B.5 can be applied to these functions.[4]

B.4.2 Delta method for variances of functions

The delta method is a very general and useful procedure for estimating the variance of a function of a random variable. We present it here to estimate the variance of a function of a MLE, which is not simply the same function of the MLE's variance.

Suppose that $\boldsymbol{\theta}$ consists of p parameters, $\theta_1, \theta_2, \ldots, \theta_p$ and define

$$g_1'(\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial\theta_1} g(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},\ g_2'(\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial\theta_2} g(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \ldots,\ g_p'(\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial\theta_p} g(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Then from a Taylor series approximation to $g(\boldsymbol{\theta})$ centered at $\hat{\boldsymbol{\theta}}$, one can show that

$$\widehat{Var}(g(\hat{\boldsymbol{\theta}})) \approx \sum_{i=1}^{p} [g_i'(\hat{\boldsymbol{\theta}})]^2 \widehat{Var}(\hat{\theta}_i) + 2\sum\sum_{i>j} [g_i'(\hat{\boldsymbol{\theta}})][g_j'(\hat{\boldsymbol{\theta}})]\widehat{Cov}(\hat{\theta}_i, \hat{\theta}_j), \tag{B.7}$$

where $\widehat{Var}(\hat{\theta}_i)$, $i = 1, \ldots, p$ are the estimated variances of the MLEs and $\widehat{Cov}(\hat{\theta}_i, \hat{\theta}_j)$, $i, j = 1, \ldots, p$, $i \neq j$ the estimated covariances described in Appendix B.3.4.

Example: Variance of an odds ratio

Suppose that there are two independent binomial variables: W_1 with n_1 trials, probability of success π_1 and w_1 observed successes, and W_2 with n_2 trials, probability of success π_2 and w_2 observed successes. As described in Section 1.2.5, inference on the odds ratio $OR = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$ is done based on the sampling distribution of $\log(\widehat{OR})$. The variance of $\log(\widehat{OR})$ is found using the delta method as follows.

First, note that we have $p = 2$ parameters here, so that $\boldsymbol{\theta} = (\pi_1, \pi_2)'$, and $g(\boldsymbol{\theta}) = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$. Then

$$g_1'(\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial\pi_1}[\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]\Big|_{(\pi_1,\pi_2)=(\hat{\pi}_1,\hat{\pi}_2)} = \frac{1}{\hat{\pi}_1(1-\hat{\pi}_1)},$$

and similarly $g_2'(\hat{\boldsymbol{\theta}}) = 1/[\hat{\pi}_2(1-\hat{\pi}_2)]$, where $\hat{\pi}_i = w_i/n_i$. Also, we have that $\widehat{Var}(\hat{\pi}_i) = \hat{\pi}_i(1-\hat{\pi}_i)/n_i$, and because W_1 and W_2 are independent, $\widehat{Cov}(\hat{\pi}_1, \hat{\pi}_2) = 0$. Thus,

$$\begin{aligned}
\widehat{Var}(g(\hat{\boldsymbol{\theta}})) &\approx \textstyle\sum_{i=1}^{p} [g_i'(\hat{\boldsymbol{\theta}})]^2 \widehat{Var}(\hat{\theta}_i) + 2\sum\sum_{i>j} [g_i'(\hat{\boldsymbol{\theta}})][g_j'(\hat{\boldsymbol{\theta}})]\widehat{Cov}(\hat{\theta}_i, \hat{\theta}_j) \\
&= \left[\frac{1}{\hat{\pi}_1(1-\hat{\pi}_1)}\right]^2 \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \left[\frac{1}{\hat{\pi}_2(1-\hat{\pi}_2)}\right]^2 \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2} \\
&= \frac{1}{n_1\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{1}{n_2\hat{\pi}_2(1-\hat{\pi}_2)} \\
&= \frac{1}{w_1} + \frac{1}{n_1 - w_1} + \frac{1}{w_2} + \frac{1}{n_2 - w_2}.
\end{aligned}$$

This is the formula given in Section 1.2.5.

[4]These properties are again approximate in finite samples. The sample sizes required to make them hold satisfactorily for a given function of parameters may be similar to or quite different from those required for the parameters themselves. This can be checked by simulation.

B.5 Inference with MLEs

Throughout this section we consider the problem of doing inference on a single parameter θ, which may be a model parameter or some function of model parameters. Where appropriate, we mention briefly extensions to multiple parameters.

B.5.1 Tests for parameters

Consider the hypotheses

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \neq \theta_0,$$

where θ_0 is some special value of interest. Several different procedures based on likelihood principles can be used to perform this test. All of them are approximate procedures in the sense that they may not achieve the stated type I error rate α exactly. These approximations are very accurate in large samples, but may provide poor results in small samples.

Wald tests

The Wald test (Wald, 1943) for a single parameter is the most familiar likelihood-based test procedure, because it uses the same ideas as the standard normal test that is part of every introductory statistics course. Because the MLE is asymptotically normally distributed with variance estimated as given in Equation B.6 or Equation B.7, we have that $Z_0 = (\hat{\theta} - \theta_0)/\sqrt{\widehat{Var}(\hat{\theta})} \dot{\sim} N(0,1)$ for a large sample when the null hypothesis is true.[5] Thus, we reject H_0 if

$$|Z_0| = \frac{|(\hat{\theta} - \theta_0)|}{\sqrt{\widehat{Var}(\hat{\theta})}} > Z_{1-\alpha/2}, \tag{B.8}$$

where the critical value $Z_{1-\alpha/2}$ is a $1 - \alpha/2$ quantile from a standard normal distribution. Alternatively, a p-value is calculated as $2P(Z > |Z_0|)$, where Z has a standard normal distribution.[6]

A Wald test is usually simple to conduct, but it is not always very effective. In particular, the $Z_{1-\alpha/2}$ critical value is guaranteed to be close to the correct critical value only in very large samples. In small samples, the normal approximation may be poor and $Z_{1-\alpha/2}$ may not be an appropriate approximation to the true critical value of the test. Thus, this test is recommended only when the sample size is large (in the context of the problem) or when no other test is possible.

A version of the Wald test is available to test hypotheses involving more than one parameter. This can be useful in regression problems with a categorical explanatory variable, which is represented in the regression by several indicator variables, each with a separate parameter. A null hypothesis of no association between the explanatory variable and the

[5] $\dot{\sim}$ means "approximately distributed as."

[6] Note that the critical value is from the normal distribution and not a t-distribution, even though we are estimating the variance in the denominator of Z_0. This is because the t-distribution arises specifically when the variance in the denominator of the test statistic is based on a sum-of-squares calculation on data from a normal distribution. The variance in Z_0 above is based on Equation B.6, which most often is not a sum-of-squares calculation.

response implies that these parameters must all be zero together. Let $\boldsymbol{\theta}_0$ be the hypothesized value of a p-dimensional parameter $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ be the parameter estimate, and $\widehat{Var}(\hat{\boldsymbol{\theta}})$ be its estimated variance. Then $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is tested using the Wald statistic

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'[\widehat{Var}(\hat{\boldsymbol{\theta}})]^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

which has an approximate χ^2_p distribution in large samples.

Likelihood ratio tests

Values of θ that have likelihoods near the maximum are more plausible guesses for the true value of the parameter than those whose likelihoods are much lower than the maximum. A comparison of the value of the likelihood at its peak against the best value of the likelihood when parameters are constrained to follow the null hypothesis is therefore a measure of evidence against the null. It turns out that the best way to make this comparison is through the likelihood ratio (LR) statistic

$$\Lambda = \frac{L(\theta_0|\mathbf{y})}{L(\hat{\theta}|\mathbf{y})}. \tag{B.9}$$

Note that $\Lambda \leq 1$ because $L(\hat{\theta}|\mathbf{y})$ is the maximum value of the likelihood for the given data. The value of Λ is near 1 when $L(\theta_0|\mathbf{y})$ is close to $L(\hat{\theta}|\mathbf{y})$. Closeness is judged by the fact that $-2\log(\Lambda) = 2[\log L(\hat{\theta}|\mathbf{y}) - \log L(\theta_0|\mathbf{y})]$ has an asymptotic χ^2_1 distribution when the null hypothesis is true. Thus, the *likelihood ratio test* (LRT) for $\mathrm{H_0} : \theta = \theta_0$ vs. $\mathrm{H_a} : \theta \neq \theta_0$ rejects H_0 when $-2\log(\Lambda) > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom.

Example: Poisson (LR-Plots.R)

The left plot in Figure B.3 gives a plot of the log-likelihood function for the previous Poisson example with $n = 10$ (sample 1). Evidence against $\mu = 9$ is fairly light, because its log-likelihood is fairly close to the peak at $\hat{\mu} = 10.5$. However, there is more evidence against $\mu = 5$, whose log-likelihood is much farther from the peak.

A LRT of $\mathrm{H_0} : \mu = 9$ vs. $\mathrm{H_a} : \mu \neq 9$ leads to

$$\begin{aligned} -2\log(\Lambda) &= -2\log\left(\frac{L(\mu = 9|\mathbf{y})}{L(\mu = 10.5|\mathbf{y})}\right) \\ &= -2\log\left(L(\mu = 9|\mathbf{y})\right) + 2\log\left(L(\mu = 10.5|\mathbf{y})\right) \\ &= 2.37. \end{aligned}$$

With $\chi^2_{1,0.95} = 3.84$, we do not reject $H_0 : \mu = 9$. In a similar manner for $\mathrm{H_0} : \mu = 5$ vs. $\mathrm{H_a} : \mu \neq 5$, we calculate $-2\log(\Lambda) = 45.81$ leading to a rejection of the null hypothesis.

The right plot in Figure B.3 demonstrates the results from the two hypothesis tests differently. We represent the rejection region here by finding the set of all possible values of μ such that $-2\log(\Lambda) < \chi^2_{1,1-\alpha}$.[7] The figure again shows that $H_0 : \mu = 9$ is not rejected, but $H_0 : \mu = 5$ is rejected.

The LRT generalizes to a broad range of problems involving multiple parameters and hypotheses involving intervals or constraints on the parameters. The general approach is

[7] Solve for μ in $-2\log\left(L(\mu|\mathbf{y})\right) + 2\log\left(L(\mu = 10.5|\mathbf{y})\right) = \chi^2_{1,0.95}$ in order to find the rejection regions.

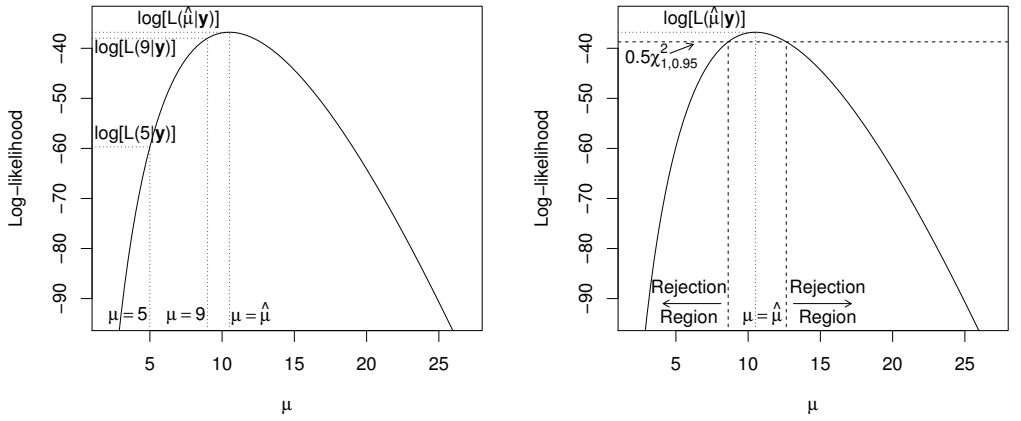

Figure B.3: Left: Poisson log-likelihood for the sample of size 10 with three values of μ indicated. Right: The rejection region for a likelihood ratio test using $\alpha = 0.05$. Any value of μ that lies in the rejection region has its null hypothesis rejected.

that the numerator of Equation B.9 is replaced by the maximum value of the likelihood function across all parameters that satisfy H_0. Additionally, the denominator is replaced by the maximum value of the likelihood function without constraints. The degrees of freedom for the χ^2 distribution are determined by the number of constraints placed on the parameters. This is explained further throughout the book where LRTs are used.

Many problems use models that contain extra parameters that are not involved in the null hypothesis (for example, hypotheses about the mean of a normal model generally do not place constraints on the variance). In these cases, the extra parameters not specified in H_0, say ϕ, are set at their MLEs under the conditions on θ. That is, in the denominator of Equation B.9, the likelihood is maximized with respect to θ and ϕ simultaneously. In the numerator, the likelihood is again maximized for both parameters simultaneously, but subject to the constraint that θ satisfies H_0. Thus, the values for ϕ that yield the best likelihoods in the numerator and denominator may be different.

The LRT is often not simple to do by hand, but there are very good computational techniques that can efficiently maximize likelihoods subject to constraints. This makes LRTs broadly applicable to a wide range of testing problems. Also, the accuracy of the LRT critical value is generally much better for a given sample size than that of the Wald test critical value, so it is generally preferred over Wald tests when both are available.

Score test

An alternative approach to testing hypotheses using likelihoods comes from examining properties of the likelihood function at the null hypothesis. As Figure B.4 shows, the slope of the log-likelihood near the peak should have a smaller magnitude than the slope far from the peak. We can therefore use this slope as a measure of the evidence in the data against H_0. This slope, also called the *score*, is just the first derivative of the log likelihood, evaluated at θ_0.

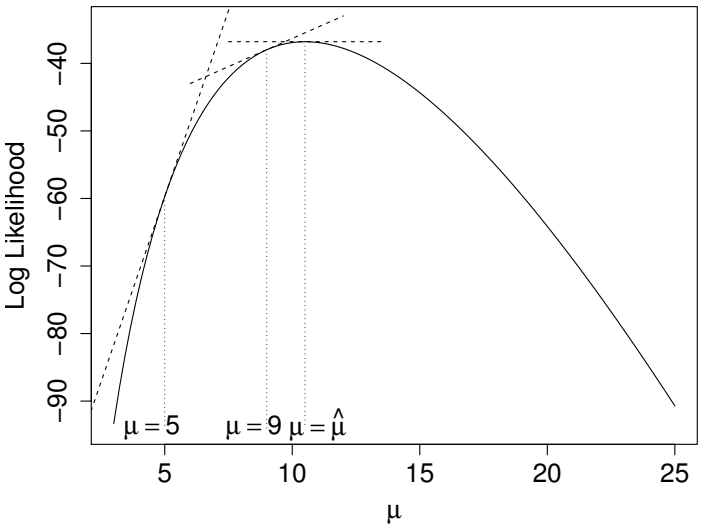

Figure B.4: Poisson log likelihood for artificial sample of size 10 showing the score function (slope) at three values of θ. The score is 0 at the MLE. The magnitude is larger for the θ farther from the MLE ($\theta = 5$) than for the one closer ($\theta = 9$). Code for the plot is in ScorePlot.R.

In particular, for independent random samples, the score is

$$U_0 = \frac{\partial}{\partial\theta} \log[L(\theta|\mathbf{y})]\Big|_{\theta=\theta_0}.$$

The central limit theorem ensures that U_0 is asymptotically normally distributed. Under the null hypothesis, the average slope across all possible data sets is zero: $E(U_0) = 0$. It can be shown that the asymptotic variance of the score—measuring how variable the slopes would be at H_0 from likelihood functions calculated based on different data sets—turns out to be estimated analogously to Equation B.5. Specifically,

$$\widehat{Var}(U_0) = -E\left(\frac{\partial^2}{\partial\theta^2} \log[L(\theta|\mathbf{y})]\right)\Big|_{\theta=\theta_0}.$$

A score test is then carried out much like the Wald test, comparing $U_0/\sqrt{\widehat{Var}(U_0)}$ to $Z_{1-\alpha/2}$. Multiparameter extensions are carried out just as in the Wald test.

The score test is also based on asymptotics, so that the critical value is approximate for any finite sample. It generally performs better than the Wald test, but not necessarily better than the LRT. Its main advantage is that it uses the likelihood only at the null hypotheses. In some complicated problems, the null hypothesis represents a considerable simplification to the general model—for example, by setting certain parameters to 0—so that calculations are much easier to carry out at θ_0 than anywhere else.

B.5.2 Confidence intervals for parameters

Wald

Like the Wald test, the Wald confidence interval is based on familiar relationships using the normal distribution. We can write $(\hat{\theta} - \theta)/\sqrt{\widehat{Var}(\hat{\theta})} \dot{\sim} N(0,1)$, where $\hat{\theta}$ and $\widehat{Var}(\hat{\theta})$ are the same as defined in Appendix B.3. Thus,

$$P\left(Z_{\alpha/2} < (\hat{\theta} - \theta)/\sqrt{\widehat{Var}(\hat{\theta})} < Z_{1-\alpha/2}\right) \approx 1 - \alpha,$$

where $Z_{1-\alpha/2}$ is a $1 - \alpha/2$ quantile from a standard normal distribution. After rearranging terms, we obtain

$$P\left(\hat{\theta} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\theta})} < \theta < \hat{\theta} - Z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta})}\right) \approx 1 - \alpha.$$

Recognizing that $-Z_{\alpha/2} = Z_{1-\alpha/2}$, this leads to the familiar form of a $(1-\alpha)100\%$ confidence interval for θ as

$$\hat{\theta} \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\theta})}. \tag{B.10}$$

An alternative approach to finding a confidence interval is to "invert" a test. That is, we seek the set of values for θ_0 for which $H_0 : \theta = \theta_0$ is *not* rejected. This procedure is easily carried out for the Wald test starting from Equation B.8 and leads to the same interval as Equation B.10.

Likelihood ratio

log

Likelihood ratio confidence intervals are found by inverting the LRT. A $(1 - \alpha)100\%$ confidence interval for θ is the set of all possible values of θ such that

$$-2[L(\theta|\mathbf{y})/L(\hat{\theta}|\mathbf{y})] \leq \chi^2_{1,1-\alpha}. \tag{B.11}$$

In Figure B.3 this is the interval between the two areas labeled "Rejection Region." As was the case with tests, likelihood ratio confidence intervals tend to be more accurate than Wald in the sense of having a true confidence level closer to the stated $1 - \alpha$ level for a given sample size. However, there is rarely a closed form for the solution, so iterative numerical procedures are needed to locate the endpoints of the interval where equality holds in Equation B.11. These computations can be difficult to carry out in some more complex problems, so not all software packages compute them.

In situations with multiple parameters, such as regression models, we frequently find confidence intervals for individual parameters. For example, suppose that a model contains two parameters, θ_1 and θ_2, and that we want to find a $(1 - \alpha)100\%$ LR confidence interval for θ_1. Because $L(\theta_1, \theta_2|\mathbf{y})$ changes as a function of both parameters, we need to account somehow for θ_2 in the process of computing Equation B.11. One approach is to fix θ_2 at some value, say d, and then find the values of θ_1 that satisfy $-2[L(\theta_1, d|\mathbf{y})/L(\tilde{\theta}_1(d), d|\mathbf{y})] \leq \chi^2_{1,1-\alpha}$, where $\tilde{\theta}_1(d)$ is the MLE of θ_1 when $\theta_2 = d$. However, this may lead to a different confidence interval for each value of d. As an alternative, we can let the value of θ_2 be its MLE for each value of θ_1 that we consider. That is, we fix the denominator of the LR statistic to be the overall maximum, $L(\hat{\theta}_1, \hat{\theta}_2|\mathbf{y})$, and for each different θ_1 we try in the numerator, we set $L(\theta_1, \theta_2|\mathbf{y})$ at the maximum that it achieves across all values θ_2. If we let $\tilde{\theta}_2(c)$ be the MLE of θ_2 when we fix $\theta_1 = c$, then the profile LR confidence interval is the set of values for c that satisfy $-2[L(c, \tilde{\theta}_2(c)|\mathbf{y})/L(\hat{\theta}_1, \hat{\theta}_2|\mathbf{y})] \leq \chi^2_{1,1-\alpha}$.

Score

Score confidence intervals are also found by inverting the score test. This is not as easy to do as it is for the Wald test, however. In the latter case, the standard error in the denominator of the test statistic does not change as one examines different values of θ_0, so that the rearrangement of Equation B.8 is easy. For the score test, the denominator changes with θ_0, and so the rearrangement can result in complicated mathematics unless the form of $\widehat{Var}(U_0)$ is fairly simple. Instead, the interval typically needs to be found iterative numerical procedures similar to those described in Appendix B.3.2. An example where the mathematics *can* be worked out with relative ease is the Wilson score interval in Section 1.1.2.

B.5.3 Tests for models

Many forms of regression are used in categorical analysis. Their use in practice often requires comparing several models involving different subsets of explanatory variables or comparing models with and without groups of explanatory variables (e.g., groups of dummy variables representing a categorical explanatory variable). When model parameters are estimated using maximum likelihood estimation, standard model comparison techniques are available based on likelihood ratio tests.

Consider two models: a *full model*, M_1, consisting of a set of p_1 explanatory variables, and the *reduced model*, M_0, containing a proper subset of p_0 explanatory variables. It is important that the reduced model does not contain any variables that are not also in the full model. Comparing M_0 and M_1 is equivalent to a hypothesis test specifying M_0 as the null hypothesis and M_1 as the alternative. Fit both M_0 and M_1 to the data and let L_{M_1} and L_{M_0} be their respective maximized likelihoods. If we use $\Lambda(M_0,M_1)$ to denote the likelihood ratio L_{M_0}/L_{M_1}, then the LRT for $H_0 : M_0$ vs. $H_a :$ M_1 rejects the null hypothesis if

$$-2\log\left(\Lambda(M_0,M_1)\right) > \chi^2_{(p_1-p_0),1-\alpha}. \tag{B.12}$$

A rejection of H_0 means that at least one of the parameters (thus, one of the variables) that make up the difference between M_0 and M_1 is important to include in a model already containing M_0. Failure to reject H_0 suggests that this simpler model may suffice and that the extra variables do not contribute significantly to the explanatory power of the model. Of course, we can never conclude that the null hypothesis is true for any hypothesis test, so it is not possible to say that M_0 is a "significantly better model" than M_1.