Analysis of Neural Data

Contents

1 Introduction **1**

1.1 Data Analysis in the Brain Sciences 1

1.1.1 Appropriate analytical strategies depend crucially on the purpose of the study and the way the data are collected. 3

1.1.2 Many investigations involve a response to a stimulus or behavior. 8

1.2 The Contribution of Statistics 11

1.2.1 Statistical models describe regularity and variability of data in terms of probability distributions. 12

1.2.2 Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning. 17

1.2.3 Statistical models may be either parametric or non-parametric 19

1.2.4 Statistical model building is an iterative process that incorporates assessment of fit and is preceded by exploratory methods. 23

1.2.5 All models are wrong, but some are useful. 23

1.2.6 Statistical theory is used to understand the behavior of statistical procedures under various probabilistic assumptions. 28

1.2.7 Measuring devices often pre-process the data. 28

1.2.8 Data analytic techniques are rarely able to compensate for deficiencies in data collection. 29

1.2.9 Simple methods are essential. 30
1.2.10 It is convenient to classify data into several broad types. 30

2 Manipulating Data **33**

2.1 Describing Central Tendency and Variation 34
2.1.1 Alternative displays and summaries provide different views of the data. 34
2.1.2 Exploratory methods can be sophisticated. 38
2.2 Data Transformations 40
2.2.1 Positive values are often transformed by logarithms. . . 40
2.2.2 Non-logarithmic transformations are sometimes applied. 46

3 Probability and Random Variables **51**

3.1 The Calculus of Probability 52
3.1.1 Probabilities are defined on sets of uncertain events. . . 52
3.1.2 The conditional probability $P(A|B)$ is the probability that A occurs given that B occurs. 55
3.1.3 Probabilities multiply when the associated events are independent. 57
3.1.4 Bayes' Theorem for events gives the conditional probability $P(A|B)$ in terms of the conditional probability $P(B|A)$. 59
3.2 Random Variables 63
3.2.1 Random variables take on values determined by events. 65
3.2.2 Distributions of random variables are defined using cumulative distribution functions and probability density functions, from which theoretical means and variances may be computed. 66
3.2.3 Continuous random variables are similar to discrete random variables. 70
3.2.4 The hazard function of a random variable X at x is its conditional probability density, given that $X \geq x$. . . . 81

3.2.5 The distribution of a function of a random variable is found by the change of variables formula. 82
3.3 The Empirical Cumulative Distribution Function 84
3.3.1 Q-Q and P-P plots provide graphical checks for gross departures from a distributional form. 86
3.3.2 Q-Q and P-P plots may be used to judge the effectiveness of transformations 90
4 Random Vectors **93**
4.1 Two or More Random Variables 93
4.1.1 The variation of several random variables is described by their joint distribution. 95
4.1.2 Random variables are independent when their joint pdf is the product of their marginal pdfs. 98
4.2 Bivariate Dependence 99
4.2.1 The linear dependence of two random variables may be quantified by their correlation. 100
4.2.2 A bivariate normal distribution is determined by a pair of means, a pair of standard deviations, and a correlation coefficient. 104
4.2.3 Conditional probabilities involving random variables are obtained from conditional densities. 107
4.2.4 The conditional expectation $E(Y|X = x)$ is called the regression of Y on X. 108
4.3 Multivariate Dependence 112
4.4 The mean of a random vector is a vector and its variance is a matrix. 112
4.4.1 The dependence of two random vectors may be quantified by mutual information. 114
4.4.2 Bayes' Theorem for random vectors is analogous to Bayes' Theorem for events. 119
5 Important Probability Distributions **121**
5.1 Bernoulli Random Variables and the Binomial Distribution . . 122

5.1.1 Bernoulli random variables take values 0 or 1. 122
5.1.2 The binomial distribution results from a sum of independent and homogeneous Bernoulli random variables. 122
5.2 The Poisson Distribution 129
5.2.1 The Poisson distribution is often used to describe counts of binary events. 129
5.2.2 For large n and small p the binomial distribution is approximately the same as Poisson. 132
5.2.3 The Poisson distribution results when the binary events are independent. 135
5.3 The Normal Distribution 136
5.3.1 Normal random variables are within 1 standard deviation of their mean with probability 2/3; they are within 2 standard deviations of their mean with probability .95. 137
5.3.2 Binomial and Poisson distributions are approximately normal, for large n or large λ. 138
5.4 Some Other Common Distributions 140
5.4.1 The multinomial distribution extends the binomial to multiple categories. 140
5.4.2 The exponential distribution is used to describe waiting times without memory. 141
5.4.3 Gamma distributions are sums of exponentials. 143
5.4.4 Chi-squared distributions are special cases of Gammas. 144
5.4.5 The Beta distribution may be used to describe variation on a finite interval. 146
5.4.6 The Inverse Gaussian distribution describes the waiting time for a threshold crossing by Brownian motion. . 146
5.4.7 The t and F distributions are defined from normal and chi-squared distributions. 149
5.5 Multivariate Normal Distributions 150
5.5.1 A random vector is multivariate normal if linear combinations of its components are univariate normal. . . . 150

5.5.2 The multivariate normal pdf has elliptical contours, with probability density declining according to a χ^2 pdf. 151

6 Sequences of Random Variables 153

6.1 Random Sequences and the Sample Mean 153

6.1.1 The standard deviation of the sample mean decreases as $1/\sqrt{n}$. 155

6.1.2 Random sequences may converge according to several distinct criteria. 156

6.2 The Law of Large Numbers 159

6.2.1 As the sample size n increases, the sample mean converges to the theoretical mean. 159

6.2.2 The empirical cdf converges to the theoretical cdf. . . . 160

6.3 The Central Limit Theorem 161

6.3.1 For large n, the sample mean is approximately normally distributed. 161

6.3.2 For large n, the multivariate sample mean is approximately multivariate normal. 163

7 Estimation and Uncertainty 165

7.1 Fitting Statistical Models 165

7.2 The Problem of Estimation 169

7.2.1 The method of moments uses the sample mean and variance to estimate the theoretical mean and variance. 170

7.2.2 The method of maximum likelihood maximizes the likelihood function, which is defined up to a multiplicative constant. 171

7.3 Confidence Intervals 176

7.3.1 For scientific inference, estimates are useless without some notion of precision. 176

7.3.2 Estimation of a normal mean is a paradigm case. . . . 179

7.3.3 For non-normal observations the Central Limit Theorem may be invoked. 180

7.3.4 A large-sample confidence interval for μ is obtained using the standard error $s/\sqrt{n}$. 180
7.3.5 Standard errors lead immediately to confidence intervals. 183
7.3.6 Estimates and standard errors should be reported to two digits in the standard error. 186
7.3.7 Appropriate sample sizes may be determined from desired size of standard error. 187
7.3.8 Confidence assigns probability indirectly, making its interpretation subtle. 188
7.3.9 Bayes' Theorem may be used to assess uncertainty. . . 190
7.3.10 In small samples it is customary to use the t distribution instead of the normal. 193

8 Estimation in Theory and Practice 197

8.1 Mean Squared Error 199
8.1.1 Mean squared error is bias squared plus variance. . . . 200
8.1.2 Mean squared error may be evaluated by computer simulation of pseudo-data. 206
8.1.3 In estimating a theoretical mean from observations having differing variances a weighted mean should be used, with weights inversely proportional to the variances. . . 215
8.1.4 Decision theory uses mean squared error to represent risk. 219
8.2 Estimation in Large Samples 219
8.2.1 In large samples, an estimator should be very likely to be close to its estimand. 219
8.2.2 In large samples, the precision with which a parameter may be estimated is bounded by Fisher information. . 220
8.2.3 Estimators that minimize large-sample variance are called efficient. 221
8.3 Properties of ML Estimators 222
8.3.1 In large samples, ML estimation is optimal. 222

8.3.2 The standard error of the MLE is obtained from the second derivative of the loglikelihood function. 222
8.3.3 MLEs transform along with parameters. 226
8.3.4 In large samples, ML estimation is approximately Bayesian. 227
8.3.5 Under normality, ML produces the weighted mean. . . 229
8.4 Multiparameter Maximum Likelihood 230
8.4.1 The MLE solves a set of partial differential equations. . 230
8.4.2 Least squares may be viewed as a special case of ML estimation. 232
8.4.3 The observed information is the negative of the matrix of second partial derivatives of the loglikelihood function, evaluated at $\hat{\theta}$. 234
8.4.4 When using numerical methods to implement ML estimation, some care is needed. 236
8.4.5 Maximum likelihood may produce bad estimates. . . . 238

9 Propagation of Uncertainty and the Bootstrap 239

9.1 Propagation of Uncertainty 241
9.1.1 Functions of approximately normal random vectors are approximately normal. 242
9.1.2 Random draws from the distribution of the random variable X produce random draws from the distribution of the random variable $Y = f(X)$. 251
9.2 The Bootstrap 260
9.2.1 The bootstrap is a general method of assessing uncertainty. 260
9.2.2 The parametric bootstrap draws pseudo-data from an estimated parametric distribution. 261
9.2.3 The nonparametric bootstrap draws pseudo-data from the empirical cdf. 266

10 Models, Hypotheses, and Statistical Significance 271

10.1 Chi-Squared Statistics 273

10.1.1 The chi-squared statistic compares model-fitted values to observed values. 274
10.1.2 For multinomial data, the chi-squared statistic follows, approximately, a χ^2 distribution. 276
10.1.3 The rarity of a large chi-squared is judged by its p-value. 278
10.1.4 Chi-squared may be used to test independence of two traits 280
10.2 Null Hypotheses 284
10.2.1 Statistical models are often considered null hypotheses. 284
10.2.2 Null hypotheses sometimes specify a particular value of a parameter within a statistical model. 284
10.2.3 Null hypotheses may also specify a constraint on two or more parameters. 285
10.3 Testing Null Hypotheses 288
10.3.1 The hypothesis $H_0 : \mu = 0$ for a normal random variable is a paradigm case. 288
10.3.2 In large samples the hypothesis $H_0 : \theta = \theta_0$ may be tested using the ratio $(\hat{\theta} - \theta_0)/SE(\hat{\theta})$. 289
10.3.3 For two independent samples, the hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using the t-ratio. 291
10.3.4 Computer simulation may be used to find p-values. . . 293
10.4 Interpretation and Properties of Tests 296
10.4.1 When the null hypothesis holds, the p-value should follow a uniform distribution. 296
10.4.2 Statistical tests should have the correct probability of falsely rejecting H_0, at least approximately. 298
10.4.3 A confidence interval for θ may be used to test $H_0 : \theta = \theta_0$. 299
10.4.4 Statistical tests are evaluated in terms of their probability of correctly rejecting H_0. 301
10.4.5 The performance of a statistical test may be displayed by the ROC curve. 304

10.4.6 The p-value is *not* the probability that H_0 is true. . . . 306
10.4.7 The p-value is conceptually distinct from type one error. 307

11 General Methods for Testing Hypotheses 309

11.1 Likelihood Ratio Tests 309
11.1.1 The likelihood ratio may be used to test $H_0 : \theta = \theta_0$. . 310
11.1.2 P-values for the likelihood ratio test of $H_0 : \theta = \theta_0$ may be obtained from the χ^2 distribution or by simulation. 311
11.1.3 The likelihood ratio test of $H_0 : (\omega, \theta) = (\omega, \theta_0)$ plugs in the MLE of ω, obtained under H_0. 313
11.1.4 The likelihood ratio test reproduces, exactly or approximately, many commonly-used significance tests. 315
11.2 Permutation and Bootstrap Tests 315
11.2.1 Permutation tests consider all possible permutations of the data that would be consistent with the null hypothesis. 315
11.2.2 The Bootstrap samples with replacement. 318
11.3 Multiple Tests 320
11.3.1 When multiple independent data sets are used to test the same hypothesis, the p-values are easily combined. 320
11.3.2 When multiple hypotheses are considered, statistical significance should be adjusted. 320

12 Linear Regression 321

12.1 The Linear Regression Model 323
12.1.1 Linear regression assumes linearity of $f(x)$ and independence of the noise contributions at the various observed x values. 323
12.1.2 The relative contribution of the linear signal to the total response variation is summarized by R^2. 324
12.1.3 For large samples, if the model is correct, the least-squares estimate is likely to be accurate. 326
12.2 Checking Assumptions 326

12.2.1 Residual analysis is helpful because residuals should represent unstructured noise. 326
12.2.2 Graphical examination of (x, y) data can yield crucial information. 328
12.3 Evidence of a Linear Trend 329
12.3.1 Confidence intervals for slopes are based on SE, according to the general formula. 329
12.3.2 Evidence in favor of a linear trend can be obtained from a t-test concerning the slope. 330
12.3.3 The fitted relationship may not be accurate outside the range of the observed data. 331
12.4 Correlation and Regression 332
12.4.1 The correlation coefficient is determined by the regression coefficient and the standard devitations of x and y. 333
12.4.2 Association is not causation. 333
12.4.3 When noise is added to two variables, their correlation diminishes. 334
12.5 Multiple Linear Regression 336
12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables. 338
12.5.2 Response variation may be decomposed into signal and noise sums of squares. 339
12.5.3 Multiple regression may be formulated concisely using matrices. 343
12.5.4 The linear regression model applies to polynomial regression and cosine regression. 346
12.5.5 Effects of correlated explanatory variables can not be interpreed separately. 351
12.5.6 In multiple linear regression interaction effects are often important. 352

12.5.7 Regression models with many explanatory variables can often be simplified. 353

13 Analysis of Variance **355**

13.1 One-Way and Two-Way ANOVA 355

13.1.1 ANOVA is based on a linear model. 358

13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis. 359

13.1.3 When there are only two groups, the ANOVA F-test reduces to a t-test. 361

13.1.4 Two-way ANOVA assesses the effects of one factor while adjusting for the other factor. 363

13.1.5 Additional analyses, involving multiple comparisons, may require adjustments to p-values. 365

13.2 ANOVA as Regression 368

13.2.1 The general linear model includes both regression and ANOVA models. 368

13.2.2 In multi-way ANOVA, interactions are often of interest. 371

13.3 Nonparametric Methods 373

13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks. 373

13.3.2 Permutation and bootstrap tests may be used to test ANOVA hypotheses. 378

13.3.3 When there are many levels of a factor it may be useful to model its effects as random. 379

13.4 Randomization, Causation, and Observational Studies 379

13.4.1 Randomization eliminates effects of confounding factors. 381

13.4.2 Observational studies often produce substantial evidence. 382

14 Generalized Linear and Nonlinear Regression **385**

14.1 Logistic Regression 386

14.1.1 Logistic regression may be used to analyze responses that are binomial. 386
14.1.2 The logistic regression model assumes binomial responses and converts the binomial probability p to log odds $\log \frac{p}{1-p}$. 388
14.1.3 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x. 389
14.1.4 The logit transformation is one among many that may be used for binomial responses, but it is the most commonly applied. 392
14.2 Poisson Regression 395
14.2.1 The usual Poisson regression model transforms the mean λ to $\log \lambda$. 395
14.2.2 In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends. 396
14.3 Nonlinear Regression 398
14.3.1 Nonlinear regression models may be fitted by least squares. 398
14.3.2 In solving nonlinear least-squares problems, good starting values are important, and it can be helpful to reparameterize. 400
14.3.3 Nonlinear models may also be used with binomial or Poisson responses. 403
14.4 Generalized Linear Models 406
14.4.1 Generalized linear models extend regression methods to response distributions from exponential families. . . 406
14.4.2 Generalized linear models often involve multiple explanatory variables. 409
16 Nonparametric Regression 413
16.1 Smoothers 414
16.1.1 Linear smoothers are fast. 416

16.1.2 For linear smoothers, the fitted function values are obtained via a "hat matrix," and it is easy to apply propagation of uncertainty. 416
16.2 Splines 417
16.2.1 Splines may be used to represent complicated functions. 417
16.2.2 Splines may be fit to data using linear models. 420
16.2.3 Splines are also easy to use in binomial or Poisson regression models. 421
16.2.4 With regression splines, the number and location of knots controls the smoothness of the fit. 422
16.2.5 Smoothing splines are splines with knots at each x_i, but with reduced coefficients obtained by penalized ML. 424
16.2.6 A method called BARS chooses knot sets automatically, according to a Bayesian criterion. 425
16.2.7 Cross-validation may be used to estimate MSE and control smoothness. 426
16.3 Local Fitting 426
16.3.1 Functions may be estimated by local fitting procedures. 426
16.3.2 Kernel regression estimates $f(x)$ with a weighted mean defined by a pdf. 428
16.3.3 Local polynomial regression solves a weighted least squares problem with weights defined by a kernel. 431
16.3.4 Simple theoretical considerations lead to bandwidth recommendations for linear smoothers. 432
16.4 Density Estimation 433
16.4.1 Kernels may be used to estimate a pdf. 433
16.4.2 Other nonparametric regression methods may be used to estimate a pdf. 434

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 1

Introduction

1.1 Data Analysis in the Brain Sciences

The brain sciences seek to discover mechanisms by which neural activity is generated, thoughts are created, and behavior is produced. What makes us see, hear, and understand the world around us? How can we learn intricate movements, which require continual corrections for minor variations in path? What is the basis of memory, and how is consciousness created? Answering such questions is the grand ambition of this broad enterprise and, while the workings of the nervous system are immensely complicated, several lines of now-classical research have made enormous progress: essential features of the nature of the action potential, of synaptic transmission, of visual processing, of the biochemical basis of memory, and of motor control have been discovered. These advances have formed conceptual underpinnings for modern neurophysiology, and have had a substantial impact on clinical practice. The method that produced this knowledge, the scientific method, involves both observation and experiment, but always a careful consideration of the

data. Sometimes results from an investigation have been largely qualitative, as in Milner's documentation of implicit memory retention, together with explicit memory loss, as a result of hippocampal lesioning in patient H.M. In other cases quantitative analysis has been essential, as in Hodgkin and Huxley's modeling of ion channels to describe the production of action potentials. Today's brain research builds on earlier results using a wide variety of modern techniques, including molecular methods, patch clamp recording, calcium imaging, two-photon imaging, single and multiple electrode studies, optical imaging, EEGs, and functional imaging (PET, fMRI, MEG), as well as psychophysical and behavioral studies. All of these rely, in varying ways, on vast improvements in data storage, manipulation, and display technologies. As a result, data sets from current investigations are often much larger, and more complicated, than those of earlier days. For a modern student of neural science, a working knowledge of basic methods of data analysis seems indispensible.

The variety of experimental paradigms across widely ranging investigative levels in the brain sciences may seem intimidating. It would take a multi-volume encyclopedia to document the details of the myriad analytical methods out there. Yet, for all the diversity of measurement and purpose, there are commonalities that make analysis of neural data a single, circumscribed and integrated subject. A relatively small number of principles, together with a handful of ubiquitous techniques—some quite old, some much newer—lay a solid foundation. One of our chief aims in writing this book has been to provide a coherent framework to serve as a starting point in understanding all types of neural data.

In addition to providing a unified treatment of analytical methods that are crucial to progress in the brain sciences, we have a secondary goal. Over many years of collaboration with neuroscientists we have observed in them a desire to learn all the data have to offer. Data collection is demanding, and time-consuming, so it is natural to want to use the most efficient and effective methods of data analysis. But we have also observed something else. Many neuroscientists take great pleasure in displaying their results not only because of the science involved but also because of the *manner in which* particular data summaries and displays are able to shed light on, and explain, neuroscientific phenomenon; in other words, they have developed a refined appreciation for the data-analytic process itself. The often-ingenious ways in-

vestigators present their data have been instructive to us, and have reinforced our own aesthetic sensibilities for this endeavor. There is deep satisfaction in comprehending a method that is at once elegant and powerful, that uses mathematics to describe the world of observation and experimentation, and that tames uncertainty by capturing it and using it to advantage. We hope to pass on to readers some of these feelings about the role of analytical techniques in illuminating and articulating fundamental concepts.

A third goal for this book comes from our exposure to numerous articles that report data analyzed largely by people who lack training in statistics. Many researchers have excellent quantitative skills and intuitions, and in most published work statistical procedures appear to be used correctly. Yet, in examining these papers we have been struck repeatedly by the absence of what we might call statistical thinking, or application of *the statistical paradigm*, and a resulting loss of opportunity to make full and effective use of the data. These cases typically do not involve an incorrect application of a statistical method (though that sometimes does happen). Rather, the lost opportunity is a failure to follow the *general approach* to the analysis of the data that we label "the statistical paradigm." Our final pedagogical goal, therefore, is to lay out the key features of this paradigm, and to illustrate its application in diverse contexts, so that readers may absorb its main tenets.

To begin, we will review several essential points that will permeate the book. Some of these concern the nature of neural data, others the process of statistical reasoning. As we go over the basic issues, we will introduce some data that will be used repeatedly.

1.1.1 Appropriate analytical strategies depend crucially on the purpose of the study and the way the data are collected.

The answer to the question, "How should I analyze my data?" always depends on what you want to know. Convenient summaries of the data are used to convey apparent tendencies. Particular summaries highlight particular aspects of the data—but they ignore other aspects. At first, the purpose of an investigation may be stated rather vaguely, as in "I would like to know how

the responses differ under these two experimental conditions." This by itself, however, is rarely enough to proceed: usually there are choices to be made, and figuring out what analysis should be performed requires a sharpening of purpose. Let us consider an example.

Example 1.1.1 SEF neural activity under two conditions Olson *et al.* (2000) examined the behavior of neurons in the Supplementary Eye Field (SEF), which is a frontal lobe region anterior to, and projecting to, the eye area in motor cortex. The general issue was whether the SEF merely relays the message to move the eyes, or whether it is involved in some higher-level processing. To address this issue, an experiment was devised in which monkeys moved their eyes in response to either an explicit external cue (the point to which the eyes were to move was illuminated) or an internally-generated translation of a complex cue (a particular pattern at fixation point determined the location to which the monkey was to move his eyes). If the SEF simply transmits the movement message to motor cortex and other downstream areas, one would expect SEF neurons to behave very similarly under the two experimental conditions. On the other hand, distinctions between the neural responses in the two conditions would indicate effects of the distinct cognitive processing tasks. While an individual neuron's activity was recorded from the SEF of an alert macaque monkey, one of the two conditions was chosen at random and applied. This was repeated many times, and the results for one neuron are given in Figure 1.1. The figure displays a pair of raster plots and peri-stimulus time histograms (PSTHs).

Visual comparison of the two raster plots and two PSTHs indicates that this neuron tends to respond somewhat more strongly under the pattern condition than under the spatial condition, at least toward the end of the trial. But such qualitative impressions are often insufficiently convincing even for a single neuron; furthermore, results for many dozens of neurons need to be reported. How should they be summarized? Should the firing rates be averaged over a suitable time interval, and then compared? If so, which interval should be used? Might it be useful to display the firing-rate histograms on top of each other somehow, for better comparison, and might the distinctions between them be quantified and then summarized across all neurons? Might it be useful to compare the peak firing rates for the two neurons, or the time at which the peaks occurred? All of these

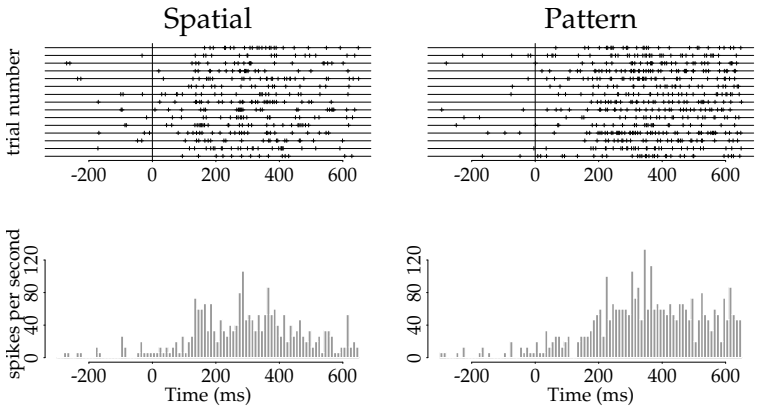


Figure 1.1: *Raster plot (TOP) and PSTH (BOTTOM) for an SEF neuron under both the external-cue or "spatial" condition (LEFT) and the complex cue or "pattern" condition (RIGHT). Each line in each raster plot contains data from a single trial, that is, a single instance in which the condition was applied. (There are 15 trials for each condition.) The tick marks represent spike times, i.e., times at which the neuron fired. The PSTH contains normalized spike counts within 10 millisecond time bins: for each time bin the number of spikes occurring in that bin is counted across all trials; this count is then divided by the number of trials, and the width of the time bin in seconds, which results in firing rate in units of spikes per second. Time is measured relative to presentation of a visual cue, which is considered time* $t = 0$. *This neuron is somewhat more active under the pattern condition, several hundred milliseconds post cue. The increase in activity may be seen from the raster plots, but is more apparent from comparison of the PSTHs.*

variations involve different ways to look at the data, and each effectively defines somewhat differently the purpose of the study.

The several possible ways of examining firing rate, just mentioned, have in common the aggregation of data across trials. A quite different idea would be to examine the relationship of neural spiking activity and reaction time, on a trial-by-trial basis, and then to see how that changes across conditions.

This intriguing possibility, however, would require a different experiment: in the experiment of Olson *et al.* the eye movement occurred long after the cue,[1] so there was no observed behavior corresponding to reaction time. This is an extreme case of the way analytical alternatives depend on the purpose of the experiment. □

Example 1.1.1 illustrates the way a particular purpose shaped the design of the experiment, and thus the data that were collected, which constrained the possible analytic strategies. In thinking about the way the data are collected, one particular distinction is especially important: that of *steady-state* versus systematically evolving conditions. In many studies, an experimental manipulation leads to a measured response that evolves in a somewhat predictable way over time. In Example 1.1.1 the neuronal firing rate, as represented by the PSTH, evolves over time, with the firing rate increasing roughly 200 milliseconds after the cue. This may be contrasted with observation of a phenomenon that has no predictable time trend, experimentally-induced or otherwise. Typically, such situations arise when one is making baseline measurements, in which some neural signal is observed while the organism or isolated tissue is at rest. Sometimes a key piece of laboratory apparatus must be observed in steady state to establish background conditions. Here is an important example.

Example 1.1.2 MEG background noise Magnetoencephalography (MEG) is an imaging technique used to measure the magnetic fields produced by electrical activity in the brain. MEG recordings are used clinically to localize a brain tumor, for example, or characterize an epileptic focus; they are used by neuroscientists to study such things as language production, memory formation, and the neurological basis of diseases such as schizophrenia.

The MEG signals are generated from the net effect of ionic currents flowing in the dendrites of cortical neurons during synaptic transmission. According to Maxwell's equations, any electrical current produces a magnetic field oriented orthogonally (perpendicularly) to the current flow, according to the right-hand rule. MEG measures this magnetic field. Magnetic fields

[1] They used a random delay followed by a separate cue to move; this helped ensure that movement and anticipatory effects would not contaminate the processing effects of interest.

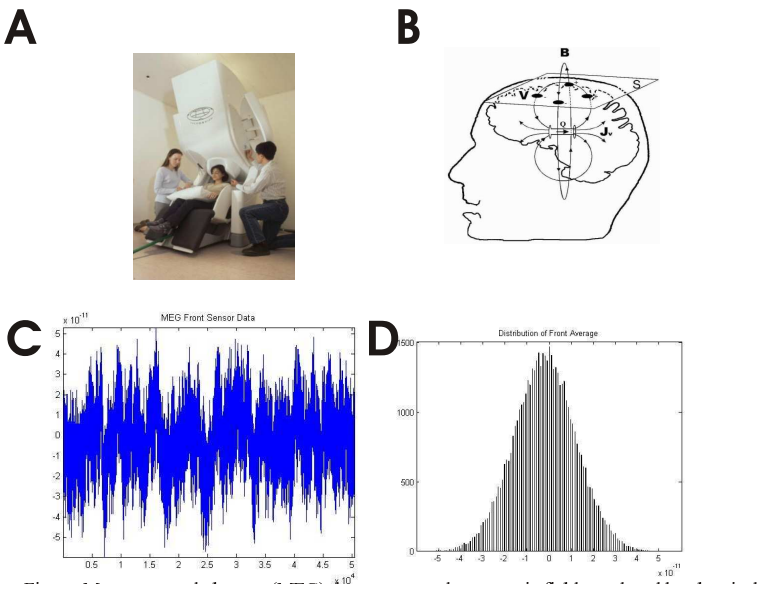


Figure 1.2: *MEG imaging. Panel C displays a MEG signal at a single SQUID detector.*
TO BE RE-DONE

are relatively unaffected by the tissue through which the signal passes on the way to a detector, but the signals are very weak. Two things make detection possible. One is that MEG uses highly sensitive detectors called superconducting quantum interference devices (SQUIDs). The second is that currents from many neighboring neuronal dendrites have similar orientations, so that their magnetic fields reinforce each other. The layer of pyramidal cells in the cortex are generally perpendicular to its surface, and their generated fields tend to be oriented outward, toward the detectors sitting outside the head.

A detectable MEG signal is produced by the net effects of currents from approximately 50,000 active neurons. See Figure 1.2.

Because the signals are weak, and the detectors extremely sensitive, it is important to assess MEG activity prior to imaging patients. Great pains are taken to remove sources of magnetic fields from the room in which the detector is located. Nonetheless, there remains a background signal that must be identified under steady-state conditions. □

Many analytical methods assume a steady state exists. The mathematical formulation of "steady state," based on *stationarity*, will be discussed in Chapter **??**.

1.1.2 Many investigations involve a response to a stimulus or behavior.

In contrast to the steady state conditions in Example 1.1.2, many experiments involve perturbation or stimulation of a sytem, producing a temporally evolving response. This does *not* correspond to a steady state. The SEF experiment was a stimulus-response study. Functional imaging also furnishes good examples.

Example 1.1.3 fMRI in a visuomotor experiment Functional magnetic resonance imaging (fMRI) uses change in magnetic resonance (MR) to infer change in neural activity, within small patches (voxels) of brain tissue. When neurons are active they consume oxygen from the blood, which produces a local increase in blood flow after a delay of several seconds. Oxygen in the blood is bound to hemoglobin, and the magnetic resonance of hemoglobin changes when it is oxygenated. By using an appropriate MR pulse sequence, the change in oxygenation can be detected as the blood-oxygen-level dependent (BOLD) signal, which follows a few seconds after the increase in neural activity. The relationship between neural activity and BOLD is not known in detail, but since the 1990s fMRI has been used to track changes in BOLD in relation to the execution of a task, giving at least a rough guide to the location of sustained functional neural activity.

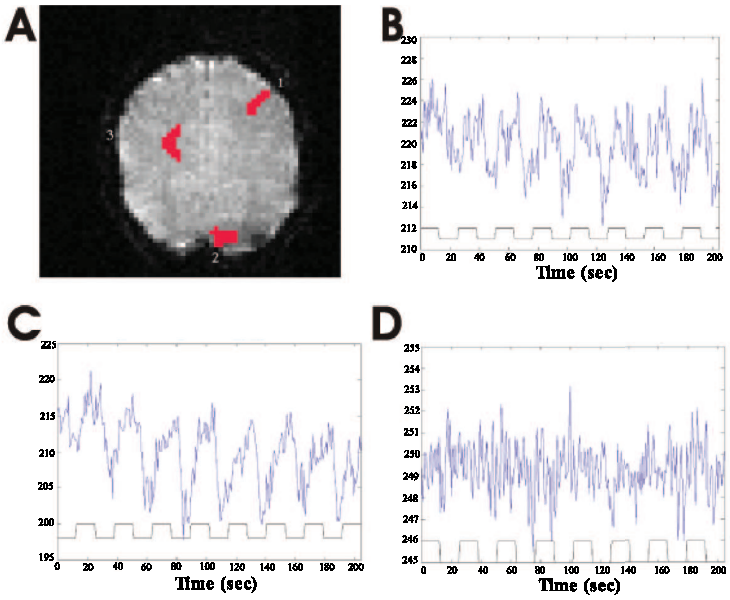


Figure 1.3: *An fMRI image with several traces of the signal across time. Panel A displays an image indicating three locations from which voxel signals were examined. Panels B-D display the signals themselves, averaged across the voxels. They correspond, respectively, to motor cortex, primary visual cortex, and white matter.* **TO BE RE-DONE**

Figure 1.3 displays images from one subject in a combined visual and motor fMRI experiment. The subject was presented with a full-field flickering checkerboard, in a repeating pattern of 12.8 seconds OFF followed by 12.8 seconds ON. This was repeated 8 times. Alternating out of phase with the flickering checkerboard pattern the subject also executed a finger tapping task (12.8 seconds ON followed by 12.8 seconds OFF). The brain was imaged once every 800 milliseconds for the duration of the experiment. The slice shown

was chosen to transect both the visual and motor cortices. Three regions of interest have been selected, corresponding to (1) motor (2) visual cortex, and (3) white matter. Parts B through D of the figure illustrate the raw time series taken from each of these regions, along with timing diagrams of the input stimuli. As expected, the motor region is more active during finger tapping (but the BOLD signal responds several seconds after the neural activity) while the visual region is more active during the flickering visual image (again with several seconds lag). The response within white matter serves as a control. □

The focus of stimulus-response experiments is usually the relationship between stimulus and response. This may suggest strategies for analysis of the data. If we let X denote the stimulus and Y the response, we might write the relationship as follows:

$$X \longrightarrow Y \tag{1.1}$$

where the arrow indicates that X leads to Y. Chapters 13, **??**, **??**, and **??** are devoted to regression methods, which are designed for situations in which X might lead to Y.

In Example 1.1.1, Y could be the average firing rate in a specified window of time, such as 200 to 600 milliseconds following the cue, and X could represent the experimental condition. In other words, the particular experimental condition leads to a corresponding average firing rate. In Example **??**, Y could be the value of the BOLD response, and X could represent whether the checkerboard was on or off 5 seconds prior to the response Y.

The tools of Chapters 13, **??**, **??**, and **??** are somewhat broader than the relation (1.1) may imply. While the arrow suggests a mechanistic relationship (the stimulus occurred, and that made Y occur), it is often preferable to step back and remain agnostic about a causal connection. A more general notion is that the variables X and Y are *associated*, meaning that they tend to vary together. A wide variety of neuroscientific studies seek to establish associations among variables. Such studies might relate a pair of behavioral measures, for example, or they might involve spike trains from a pair of neurons recorded simultaneously, EEGs from a pair of electrodes on the scalp, or MEG signals from a pair of SQUID detectors. Measures of association are

discussed in Chapters **??** and 13. Chapter **??** also contains a brief discussion of the distinction between association and causation, and some issues to consider when one wishes to infer causation from an observed association.

1.2 The Contribution of Statistics

Many people think of statistics as a collection of particular data-analytic techniques, such as analysis of variance, chi-squared goodness-of-fit, linear regression, etc. And so it is. But the field of statistics, as an academically specialized discipline, strives for something much deeper, namely, the development and characterization of data collection and analysis methods according to well-defined principles, as a means of quantifying knowledge about underlying phenomena and rationalizing the learning and decision-making process. As we said above, one of the main pedagogical goals of this book is to impart to the reader some sense of the way data analytic issues are framed within the discipline of statistics. In trying to achieve this goal, we find it helpful to articulate the nature of the statistical paradigm as concisely as possible. After numerous conversations with colleagues, we have arrived at the conclusion that among many components of the statistical paradigm, summarized below, two are the most fundamental.

Two Fundamental Tenets of the Statistical Paradigm:

1. Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.

2. Statistical methods may be analyzed to determine how well they are likely to perform.

In the remainder of this section we will elaborate, adding a variety of comments and clarifications.

1.2.1 Statistical models describe regularity and variability of data in terms of probability distributions.

When data are collected, repeatedly, under conditions that are as nearly identical as an investigator can make them, the measured responses nevertheless exhibit variation. The spike trains generated by the SEF neuron in Example 1.1.1 were collected under experimental conditions that were essentially identical; yet, the spike times, and the number of spikes, varied from trial to trial. The most fundamental principle of the statistical paradigm, its starting point, is that this variation may be described by probability. Chapters 3 and 5 are devoted to spelling out the details, so that it will become clear what we mean when we say that probability describes variation. But the idea is simple enough: probability describes familiar games of chance, such as rolling dice, so when we use probability also to describe variation, we are making an analogy; we do not know all the reasons why one measurement is different than another, so it is *as if* the variation in the data were generated by a gambling device. Let us consider a simple but interesting example.

Example 1.2.1 Blindsight in patient P.S. Marshall and Halligan (1988, *Nature*, 336: 766–767) reported an interesting neuropsychological finding from a particular patient, identified as P.S. This patient was a 49 year-old woman who had suffered damage to her right parietal cortex that reduced her capacity to process visual information coming from the left side of her visual space. For example, she would frequently read words incorrectly by omitting left-most letters ("smile" became "mile") and when asked to copy simple line drawings, she accurately drew the right-hand side of the figures but omitted the left-hand side without any conscious awareness of her error. To show that she could actually see what was on the left but was simply neglecting it—a phenomenon known as *blindsight*—the examiners presented P.S. with a pair of cards showing identical green line drawings of a house, except that on one of the cards bright red flames were depicted on the left side of the house. They presented to P.S. both cards, one above the other (the one placed above being selected at random), and asked her to choose which house she would prefer to live in. She thought this was silly "because they're the same" but when forced to make a response chose the non-burning

house on 14 out of 17 trials. This would seem to indicate that she did, in fact, see the left side of the drawings but was unable to fully process the information. But how convincing is it that she chose the non-burning house on 14 out of 17 trials? Might she have been guessing?

If, instead, P.S. had chosen 17 out of 17 trials there would have been very strong evidence that her processing of the visual information affected her decision-making, while, on the other hand, a choice of 9 out of 17 clearly would have been consistent with guessing. The intermediate outcome 14 out of 17 is of interest as a problem in data analysis precisely because it feels fairly convincing, but leaves us unsure: a thorough, quantitative analysis of the uncertainty would be very helpful.

The standard way to begin is to recognize the variability in the data, namely, that P.S. did not make the same choice on every trial; we then say that the choice made by P.S. on each trial was a random event, that the probability of her choosing the non-burning house on each trial was a value p, and that the responses on the different trials were independent of each other. These three assumptions use probability to describe the variability in the data. Once these three assumptions are made it becomes possible to quantify the uncertainty about p and the extent to which the data are inconsistent with the value $p = .5$, which would correspond to guessing. In other words, it becomes possible to make statistical inferences. □

The key step in this example is the introduction of probability to describe variation. Once that first step is taken, the second step of making inferences about the phenomenon becomes possible. Because the inferences are statistical in nature, and they require the introduction of probability, we usually refer to the probability framework—with its accompanying assumptions—as a *statistical model.* Statistical models provide a simple formalism for describing the way the repeatable, regular features of the data are combined with the variable features. In Example 1.2.1 we may think of p as the propensity for P.S. to choose the non-burning house. According to this statistical model, p is a kind of regularity in the data in the sense that it is unchanging across trials. The variation in the data comes from the probabilistic nature of the choice: what P.S. will choose is somewhat unpredictable, so we attribute a degree of uncertainty to unknown causes and describe it as if predicting her choice were a game of chance.

Probability is also often introduced to describe small fluctuations around a specified formula or "law." We typically consider such fluctuations "noise," in contrast to the systematic part of the variation in some data, which we call the "signal." For instance, as we explain in Chapter 13, when the underlying, systematic mathematical specification (the signal) would have the form

$$y = f(x)$$

we will replace it with a statistical model having the form

$$Y = f(x) + \epsilon \tag{1.2}$$

where ϵ represents noise and the variable Y is capitalized to indicate its now-random nature: it becomes "signal plus noise." The simplest case occurs when $f(x)$ is a line, having the form $f(x) = \beta_0 + \beta_1 x$, where we use coefficients β_0 and β_1 (instead of writing $f(x) = a + bx$) to conform to statistical convention. Here is an example.

Example 1.2.2 Neural conduction velocity Hursh (1939, *Amer. J. Physiology*) presented data on the relationship between a neuron's conduction velocity and its axonal diameter, in adult cats. (Data from kittens were consistent with the adult cat data.) Hursh measured maximal velocity among fibers in several nerve bundles, and then also measured the diameter of the largest fiber in the bundle. The resulting data, together with a fitted line, are shown in Figure 1.4. In this case the line $y = \beta_0 + \beta_1 x$ represents the approximate linear relationship between maximal velocity y and diameter x, and the data follow the line reasonably closely. This implies, for example, that if fiber A has twice the diameter of fiber B, A will be able to propagate an action potential roughly twice as fast as B. □

The method used to fit the line to the data in Figure 1.4 is called *least squares regression*. It is very simple. Suppose we have a line $y = \beta_0^* + \beta_1^* x$ that is fitted by some method, possibly least-squares or possibly another method. Suppose there are n data pairs of the form (x, y) and let us label them with a subscript so that they take the form (x_i, y_i) with $i = 1, 2, \ldots, n$. That is, (x_1, y_1) would be the first data pair, (x_2, y_2) the second, and so forth. The y-coordinate on the line $y = \beta_0^* + \beta_1^* x$ corresponding to x_i is

$$\hat{y}_i = \beta_0^* + \beta_1^* x_i.$$

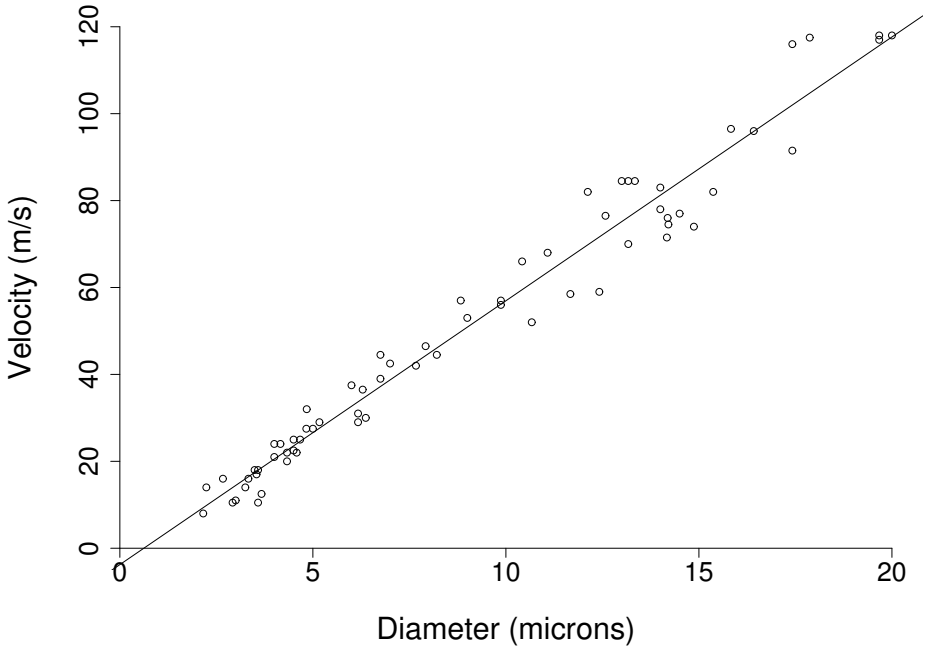


Figure 1.4: *Conduction velocity of action potentials, as a function of diameter. The x-axis is diameter in microns; the y-axis is velocity in meters per second. Based on Hursh (1939, Figure 2). Also shown is the least-squares regression line.*

The number $\hat{y}_i$ is called the *fitted value* at x_i and we may think of it as predicting y_i. We then define the i-th *residual* as

$$e_i = y_i - \hat{y}_i.$$

The value e_i is the error at x_i in fitting, or the error of prediction, i.e., it is the vertical distance between the observation (x_i, y_i) and the line at x_i. See Figure 1.5. To judge the quality of the fit of the line, we examine the e_i's. An overall assessment of the fit must somehow combine the magnitudes of the e_i's, making them as small as possible in aggregate. It is reasonable to have positive and negative values of the residuals be equally important, which is another way of saying that we should look at their magnitudes (the absolute values $|e_i|$, ignoring sign). We could sum the magnitudes $\sum |e_i|$ and

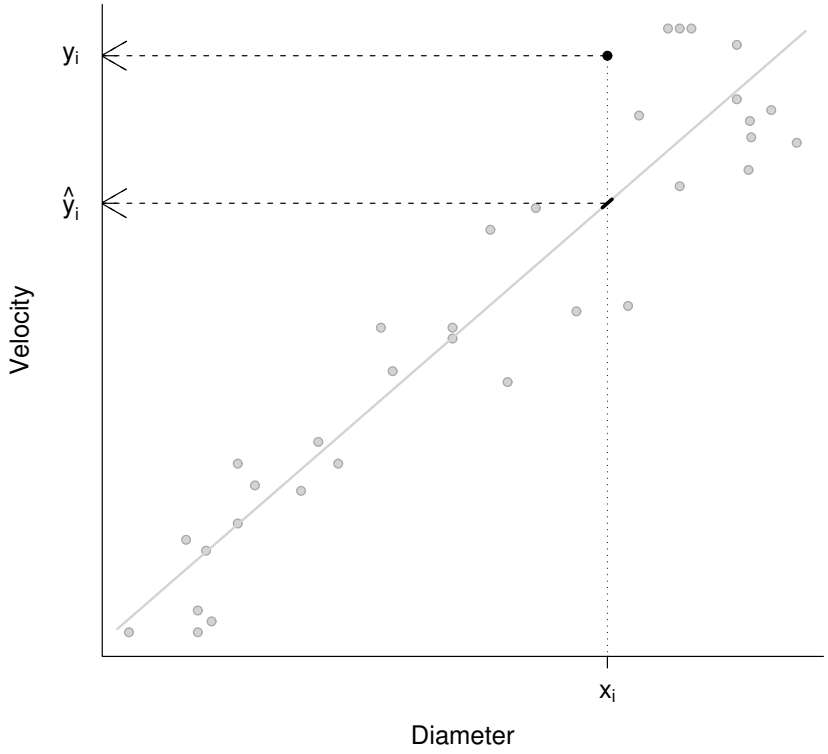


Figure 1.5: *Plot of the Hursh data set, with data points in gray except for a particular point* (x_i, y_i) *which is shown in black to identify the corresponding fitted value* $\hat{y}_i$.

choose the line that makes this sum as small as possible. That is sometimes done, but the solution can not be obtained in closed form, and it is harder to analyze mathematically. Instead, the method of least squares uses the more tractable criterion $\sum e_i^2$. Specifically, for each possible choice of β_0^* and β_1^*, we may compute the sum of squares $\sum e_i^2$ then choose the values of β_0^* and β_1^* that minimizes this sum of squares. A relatively easy way to minimize the sum of squares is to apply calculus: we differentiate $\sum e_i^2$ with respect to each of β_0^* and β_1^*, set the derivatives equal to 0, and solve the resulting pair of equations. We omit the details.

The least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of β_0^* and β_1^* that minimize $\sum e_i^2$. The least-squares line is then

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Equation 1.2 is not yet a statistical model. If we write

$$Y_i = f(x_i) + \epsilon_i, \tag{1.3}$$

take

$$f(x) = \beta_0 + \beta_1 x$$

and let the noise term ϵ_i be a *random variable* we obtain a *linear regression model*. Random variables are introduced in Chapter 3. The key point here is that linear regression describes the regularity of the data by a straight line and the variability (the deviations from the line) by a probability distribution (the distribution of the noise random variable ϵ_i). Regression methods are discussed in detail in Chapter 13.

1.2.2 Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.

The introduction of a statistical model not only provides guidance in determining fits to data, as in Example 1.2.2, but also assessments of uncertainty.

Example 1.2.1 (continued) Let us return to the question of whether the responses of P.S. were consistent with guessing. In this framework, guessing would correspond to $p = .5$ and the problem then becomes one of assessing what these data tell us about the value of p. As we will see in Chapter 7, standard statistical methods give an approximate 95% confidence interval for p of $(.64, 1.0)$. This is usually interpreted by saying we are 95% confident the value of p lies in the interval $(.64, 1.0)$, which is a satisfying result: while this interval contains a range of values, indicating considerable uncertainty, we are nonetheless highly confident that the value of p is inconsistent with guessing. □

This confidence illustrates the expression of "knowledge and uncertainty." It is an example of "inductive reasoning" in the sense that we reason from the data back to the quantity p assumed in the model. As described in Chapter 7, statistical theory uses mathematical, deductive reasoning to provide the formalism; but the interpretation as a statement about the unknown quantity p based on experience (the data) is usually called "inductive."

In fact, as a conceptual advance, this expression of knowledge and uncertainty via probability is highly nontrivial: despite quite a bit of earlier mathematical attention to games of chance, it was not until the late 1700s that there emerged any clear notion of inductive, or what we would now call *statistical* reasoning; it was not until the first half of the twentieth century that modern methods began to be developed systematically; and it was only in the second half of the twentieth century that their properties were fully understood. From a contemporary perspective the key point is that the confidence interval is achieved by uniting two distinct uses of probability. The first is descriptive (which philosophers sometimes call "phenomological"): saying P.S. will choose the non-burning house with probability p is analogous to saying the probability of rolling a 3 with an apparently fair die is 1/6. The second use of probability is often called "epistemic," and involves a statement of knowledge: saying we have 95% confidence that p is in the interval (.64,1.0) is analogous to someone saying they are 90% sure that the capital of Louisiana is Baton Rouge. The fundamental insight, gained gradually over many years, is that the descriptive probability in statistical models may used to produce epistemic statements for scientific inference. Technically, there are alternative frameworks for bringing phenomenological and epistemic probability together, the two principal ones being *Bayesian* and *frequentist*. We will discuss the distinction briefly in Chapter **??** and at somewhat greater length in Chapter **??**.

While we wish to stress the importance of statistical models in data analysis, we also want to issue several qualifications and caveats: first, the notion of "model" we intend here is quite general, the only restriction being that it must involve a probabilistic description of the data; second, modeling is done in conjunction with summaries and displays that do not introduce probability explicitly; third, it is very important to assess the fit of a model to a given set of data; and, finally, statistical models are mathematical abstractions, imposing structure on the data by introducing external assumptions. The

next three subsections explain these points further.

1.2.3 Statistical models may be either parametric or nonparametric

In emphasizing statistical models, our only restriction is that probability must be used to express the way regularity and variability in the data are to be understood. One very important distinction is that of *parametric* versus *nonparametric* models.

Example 1.1.1 (continued) In the SEF experiment introduced on page 4, the neural activity increased under the pattern condition, compared to that in the spatial condition, toward the end of the trial. However, there were a limited number of trials in the experiment, and perhaps this increase might possibly have been merely a coincidental fluctuation. The spike counts from 200 to 400 milliseconds post cue, across the 15 trials, gave firing rates of 48 spikes per second for the spatial condition versus 70 spikes per second for the pattern condition. As in the case of P.S. responding 14 out of 17, this looks like a substantial effect. But looking carefully, there is a fair bit of variability in spike counts across trials. It is therefore helpful to include uncertainty in the comparison. Using the standard parametric approach discussed in Chapter 3 we obtain 95% confidence intervals (40,56) spikes per second for the spatial condition versus (58,82) spikes per second for the pattern condition. Thus, based on this analysis, after taking account of the variability we continue to find a substantially elevated firing rate in the pattern condition.

The standard parametric approach, discussed in Chapter 7, assumes the data are normally distributed. An alternative, nonparametric method is presented in Chapter 9. It produces confidence intervals (42,56) spikes per second for the spatial condition versus (59,81) for the pattern condition. Here the parametric and nonparametric confidence intervals are slightly different but they lead to the same conclusions. □

The terminology comes from the representation of a probability distribution in terms of an unknown parameter. In Example 1.2.1 the parameter is

p. In Example 1.2.2 the parameter includes the pair (β_0, β_1), together with a noise variation parameter σ, explained in Chapter 13. In both of these cases knowing the values of the unknown parameters completely determines the probability distribution.

A related distinction arises in the context of y vs. x models of the type considered in Example 1.2.2. That example involved a linear relationship. As we note in Chapters **??** and **??**, the methods used to fit linear models can be generalized for nonlinear relationships. The methods in Chapter **??** are also called nonparametric because the fitted relationship is not required to follow a pre-specified form.

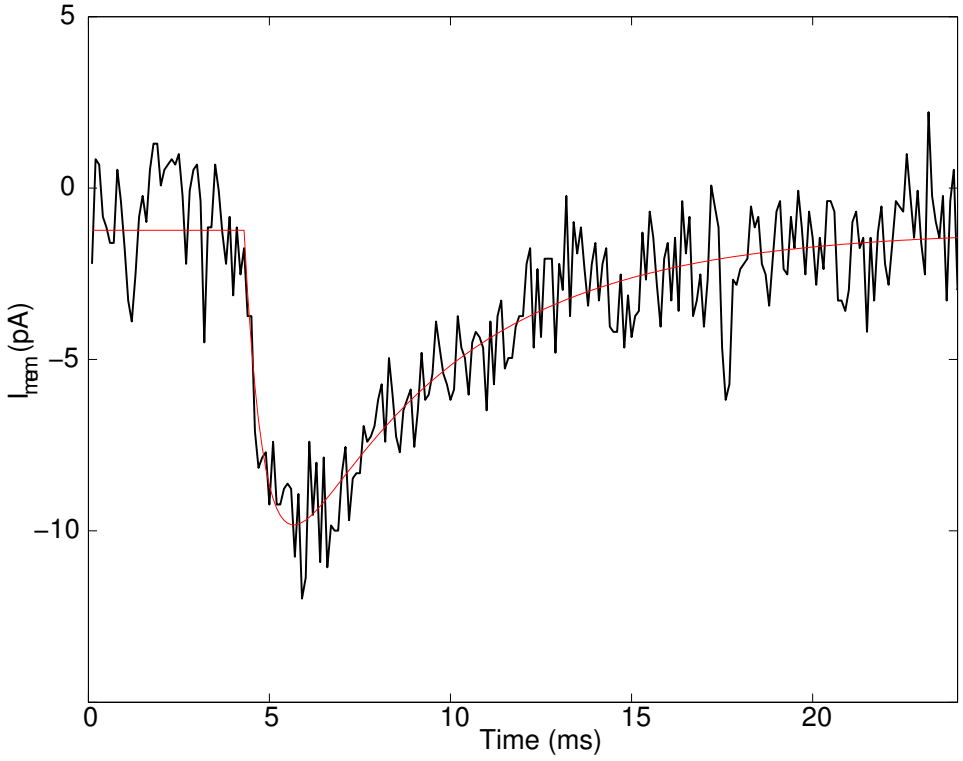


Figure 1.6: *Excitatory post-synaptic current. Current recorded from a rat hippocamal neuron, together with smoothed version (shown in red) obtained by fitting a suitable time-dependent function.*

Example 1.2.3 Excitatory Post-Synaptic Current As part of a study

on spike-timing-dependent plasticity (Dr. David Nauen, personal communication), rat hippocampal neurons were held in voltage clamp and post-synaptic currents were recorded following an action potential evoked in a presynaptic cell. Figure 1.6 displays a plot of membrane current as a function of time. One measurement of size of the current is the area under the curve, which represents the total charge transmitted. Other quantities of interest include the onset delay, the rate at which the curve "rises" (here, a negative rise) from onset to peak current, and the rate at which the curve decays from peak current back toward steady state. The current trace is clearly subject to measurement noise, which would contaminate the calculations. A standard way to reduce the noise is to fit a suitable time-dependent function of time. The fit is also shown in the figure. It may be used to produce values for the various constants needed in the analysis.

In this case, a function $y = f(t)$, with y being post-synaptic current and t being time, where

$$f(t) = A_1(1-\exp((t-t_0)/\tau_1))\,(A_2\exp((t-t_0)/\tau_2) - (1-A_2)\exp((t-t_0)/\tau_3))$$

was fitted, based on a suggestion by Nielsen *et al.* (2004) (Nielsen TA, DiGregorio DA, Silver RA (2004) Modulation of glutamate mobility reveals the mechanism underlying slow-rising AMPAR EPSCs and the diffusion coefficient in the synaptic cleft. Neuron 42: 757-771.) The fit is good, though it distorts slightly the current trace in the dip and at the end. The advantage of using this function is that its coefficients may be interpreted and compared across experimental conditions. □

The simple linear fit in Example 1.2.2 and the fit based on a somewhat complicated combination of exponential functions in Example 1.2.3 are both examples of parametric regression because both use specific functions. *Nonparametric regression* is used when the formula for the function is not needed. Nonparametric regression is a central topic of Chapter **??**. Here is an example.

Example 1.2.4 Electrooculogram smoothing for EEG artifact removal EEG recordings suffer from a variety of artifacts, one of which is their response to eye blinks. A good way to correct for eye-blink artifacts

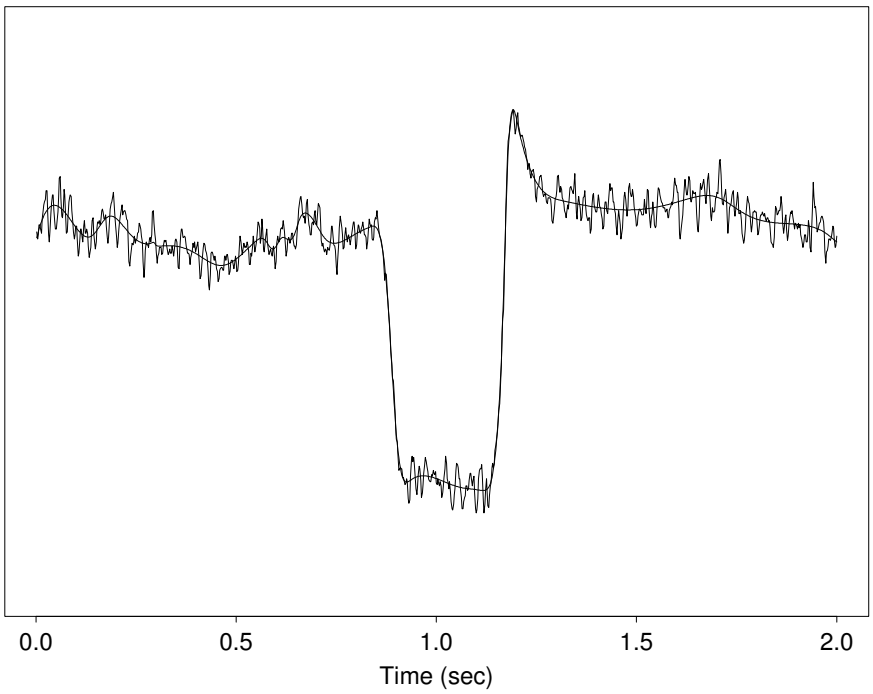


Figure 1.7: *Electrooculogram together with a smoothed, or "filtered" version that removes the noise. The method used for smoothing is an example of nonparametric regression.*

is to record potentials from additional leads in the vicinity of the eyes; such electrooculograms (EOGs) may be used to identify eye blinks, and remove their effects from the EEGs. Wallstrom *et al.* (2002, 2004) (Wallstrom, G.L., Kass, R.E., Miller, A., Cohn, J.F., and Fox, N.A. (2002) Correction of ocular artifacts in the EEG using Bayesian adaptive regression splines, in *Case Studies in Bayesian Statistics, Vol. VI*, edited by C. Gatsonis, A. Carriquiry, D. Higdon, R.E. Kass, D. Pauler, and I. Verdinelli. pp. 91–136, Springer-Verlag. Wallstrom, G.A., Kass, R.E., Miller, A., Cohn, J.F., and Fox, N.A. (2004) Automatic correction of ocular artifacts in the EEG: A comparison of regression-based and component-based methods, *Internat. J. of Psychophys.*, 53: 105–119.) investigated methods for removing ocular artifacts from EEGs using the EOG signals. In Chapter **??** it will become clear how to use a general smoothing method to remove high-frequency noise. This

does not require the use of a function having a specified form. Figure 1.7 displays an EOG recording together with a smoothed version of it, obtained using a nonparametric regression method known as BARS. □

1.2.4 Statistical model building is an iterative process that incorporates assessment of fit and is preceded by exploratory methods.

Another general point about the statistical paradigm is illustrated in Figure 1.8. This figure shows where the statistical work fits in. Real investigations are far less sequential than as depicted here, but it does provide a way of emphasizing two components of the process that go hand-in-hand with statistical modeling: exploratory analysis and assessment of fit. Exploratory analysis involves informal investigation of the data based on numerical or graphical summaries of the data, such as a histogram. Exploratory results, together with judgment based on experience, help guide construction of an initial probability model to represent variability in observed data. Every such model, and every statistical method, makes some assumptions, leading, as we have already seen, to a reduction of the data in terms of some small number of interpretable quantities. As shown in Figure 1.8, the data may be used, again, to check the probabilistic assumptions, and to consider ramifications of departures from them. Should serious departures from the assumptions be found, a new model may be formed. Thus, probability modeling and model assessment are iterative, and are followed by statistical inference. This process is imbedded into the production of scientific conclusions from experimental results (Box *et al.*, 1978).

1.2.5 All models are wrong, but some are useful.

The simple representation in Figure 1.8 is incomplete and may be somewhat misleading. Most importantly, while it is true that there are standard procedures for model assessment, some of which we will discuss in Chapter **??**, there is no uniformly-applicable rule for what constitutes a good fit. Statistical models, like scientific models, are abstractions and should not

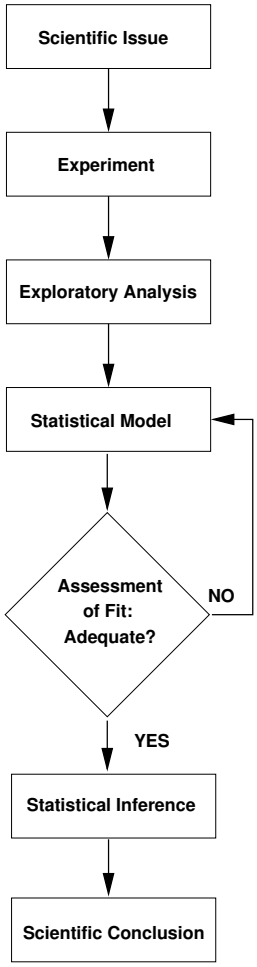


Figure 1.8: *Formal statistical inference within the process of drawing scientific conclusions. Statistical model building is a prerequisite to formal inference procedures. Model building is iterative in the sense that tentative models must be assessed and, if necessary, improved or abandoned. The figure is something of a caricature because the process is not as neat as depicted here. Furthermore, there are typically many aspects of the data, which bear on several different issues, leading to multiple inferences and conclusions, all of which are synthesized in a single scientific paper.*

be considered perfect representations of the data. As examples of scientific models in neuroscience we might pick, at one extreme, the Hodgkin-Huxley

model for action potential generation, and at the other extreme, being much more vague, the theory that vision is created via separate ventral and dorsal streams corresponding loosely to "what" and "where." Neither model is perfectly accurate—in fact, every scientific model fails[2] under certain conditions. Models are helpful because they capture important intuitions and can lead to specific predictions and inferences. The same is true of statistical models. On the other hand, statistical models are very often driven primarily by raw empiricism—they are produced to fit data and may have little or no other justification. Thus, experienced data analysts carry with them a strong sense of both the inaccuracies in statistical models and their lingering utility. This sentiment is captured well by the famous quote, "All models are wrong, but some are useful." (Box, G.E.P. (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics*, ed. by R.L. Launer and G.N. Wilkinson, NY: Academic Press.)

The schematic diagram in Figure 1.9 may help clarify the way we tend to think about statistical models. Pictured in the left column is the "real world" of data, i.e., the observables, obtained by recording in some form, often by measurement. In the right column is the "theoretical world" where both scientific and statistical models live. Scientific models help us organize facts with explanations. They can be high-level or detailed, but they should not, at least in principle, be confused with the observations themselves. The theoretical world seeks to make statements and predictions, often using a precise but abstract mathematical framework, which may be applied to things in the real world that may be observed. In a domain where theory works well, the theoretical world would be judged to be very close to the real world and, therefore, its predictions would be highly trustworthy. Statistical models are used to describe the imperfect predictability of phenomena, the regularity and variability of data, in terms of probability distributions. In Equation (1.3), for example, the regularity is represented by $f(x_i) = \beta_0 + \beta_1 x_i$ while the variability is represented by the noise random variable ϵ_i, as illustrated in Example 1.2.2. Such statistical models are abstract in the sense that the noise in the data is not actually generated by some random mechanism that follows a probability distribution; instead, probability is used to describe the

[2]For a discussion of some ways that great equations of physics remain fundamental while only approximating the real world, see Weinberg (2002). (Weinberg, S. (2002) How great equations survive, in Fermelo, G., ed., *It Must Be Beautiful: Great Equations of Modern Science*, Granta Press, pp. 253–257.)

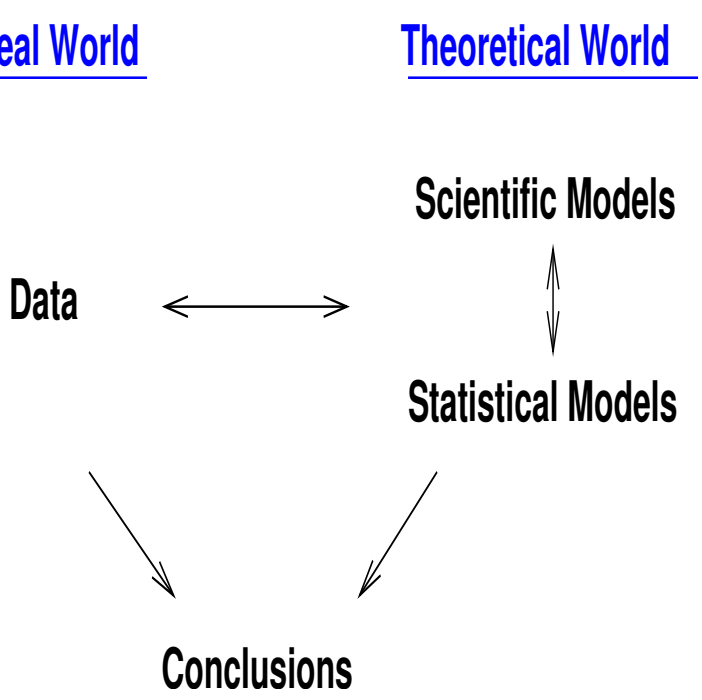


Figure 1.9: *The role of statistical models and methods in scientific inference. Statistical procedures are abstractly defined in terms of mathematics, but are used, in conjunction with scientific models and methods, to explain observable phenomena.*

variation that arises from both the natural process (the neural propagation of the action potential) and the measurement process. Because a random variable is thus a mathematical abstraction, we say that statistical models live in the theoretical world. But, like scientific models, when they are successful they do a very good job of describing the data. As illustrated in the diagram, the conclusions that are drawn, based on both the data and the scientific and statistical modeling, effectively straddle these worlds—specific predictions may speak concretely about future measurements, but often what is learned remains more general and may modify theories in important ways. In drawing conclusions from a scientific model one effectively assumes the phenomenon being described to be just like the model. Similarly, when applying statistical inferences to draw scientific conclusions one assumes that the phenomena

being described behave just like mathematical variables that have probability distributions posited by the statistical model.

In the case of Example 1.2.1, the statistical model contained the quantity p that could be *estimated* by a formal statistical procedure, and a confidence interval could be constructed. The confidence interval actually lives in the abstract mathematical world of statistical models and methods. Under our assumptions about the experiment we are able to conclude that the probability of P.S. choosing the non-burning house is clearly above .5. If these assumptions match reality reasonably closely, then the data do indeed provide a good indication that P.S. was not merely "guessing" and was, instead, exhibiting the phenomenon of blindsight.

A second aspect of the flow diagram in Figure 1.8 may be misleading. The diagram fails to highlight the way the judgment of adequate fit depends on context. When we say "All models are wrong, but some are useful," part of the point is that a model can be useful *for a specified inferential purpose*. Thus, in judging adequacy of a model, one must ask, "How might the reasonably likely departures from this model affect scientific conclusions?"

In Example 1.1.1, when we analyzed the SEF spike counts, we pointed out that the standard parametric approach to confidence intervals assumes the data follow a Normal distribution. An examination of the counts indicates this statistical model is probably somewhat inaccurate. However, as we point out in Chapter 7, the confidence intervals produced by the standard procedure are likely to provide reliable inferences when the departures from normality are not large and the sample sizes are not very small. For the SEF spike count data, the apparent departures from normality do not strongly affect the pertinent scientific conclusions. Indeed, the parametric and nonparametrical confidence intervals we obtained were not very different, which is an additional indication that the parametric model remains useful for this particular purpose.

1.2.6 Statistical theory is used to understand the behavior of statistical procedures under various probabilistic assumptions.

The second of the two major components of the statistical paradigm is that methods may be *analyzed* to determine how well they are likely to perform. As described especially in Chapter **??**, and in Chapter **??**, a series of general principles and criteria are widely used for this purpose. Statistical theory has been able to establish good performance of particular methods under certain probabilistic assumptions. We provide the necessary background for Chapter **??** in Chapters 3–6. When we wish to add arguments that are not essential to the flow of material we indent them, as follows.

> *Details:* We indent, like this, the paragraphs containing mathematical details we feel may be safely skipped. □

One easy and useful method of checking the effectiveness of a procedure, which is applicable in certain predictive settings, is *cross-validation*. The simplest form of cross-validation involves splitting the data set into two subsets, applying and refining a method using one of the subsets, and then judging its predictive performance (predicting the value of some response) on the second subset. Sometimes the second subset involves entirely new data. For example, in a behavioral study, a new set of subjects may be recruited and examined. Methods that perform well with this kind of cross-validation are often quite compelling. In addition to being intuitive, cross-validation has a theoretical justification discussed briefly in Chapter **??**. Data splitting is also sometimes advocated as a way to guard against certain kinds of misleading results from significance tests. We discuss this in Chapter **??**.

1.2.7 Measuring devices often pre-process the data.

Measurements of neural signals are often degraded by noise. A variety of techniques are used to reduce the noise and increase the relative strength of the signal, some of which will be discussed in Chapter 7. In many cases, methods such as these are applied by the measurement software to produce

the data the investigator will analyze. Functional MRI software, for example, collects data in terms of frequency and reconstructs a signal in time; on any given day each MEG sensor must be adjusted to ensure detection above background noise, which of course affects the response characteristics; an accurate characterization of background noise is essential for localization methods; and extracellular electrode signals are thresholded and filtered to isolate action potentials, which then must be "sorted" to identify those from particular neurons. In each of these cases the data that are to be analyzed are not in the rawest form possible. Such pre-processing is often extremely useful, but its effects are not necessarily benign. Inaccurate spike sorting, for example, is a notorious source of problems in some contexts. (See Bar-Gad I, Ritov Y, Vaadia E, Bergman H. (2001) *J Neurosci Methods.* 107:1–13, and Wood F, Black MJ, Vargas-Irwin C, Fellows M, Donoghue JP. (2004), *IEEE Trans Biomed Eng.* 51: 912–8.) The wise analyst will be aware of possible distortions that might arise before the data have been examined.

1.2.8 Data analytic techniques are rarely able to compensate for deficiencies in data collection.

A common misconception is that flaws in experimental design, or data collection, can be fixed by statistical methods after the fact. It is true that an alternative data analytic technique can help avoid some presumed difficulty an analyst may face in trying to apply a particular method—especially when associated with a particular piece of software. But when a measured variable does not properly capture the phenomenon it is supposed to be measuring, post hoc manipulation will be almost never be able to rectify the situation; in the rare cases that it can, much effort and very strong assumptions will typically be required. For example, we already mentioned that inaccurate spike sorting can create severe problems. When these problems arise, no post-hoc statistical manipulation is likely to fix them.

1.2.9 Simple methods are essential.

Another basic point concerning analytical methods is that simple, easily-understood data summaries, particularly visual summaries such as the PSTH, are essential components of analysis. These fit into the diagram of Figure 1.8 mainly under the heading of exploratory data analysis, though sometimes inferential analyses from simple models are also used in conjunction with those from much more elaborate models. When using a complicated procedure, it is important to understand the way results agree, or disagree, with those from simpler methods.

1.2.10 It is convenient to classify data into several broad types.

When spike train data, like those in Example 1.1.1, are summarized by spike counts occurring in particular time intervals, the values taken by the counts are necessarily non-negative integers. Because the integers are separated from each other, such data are called *discrete*. On the other hand, many recordings, such as MEG signals, or EEGs, can take on essentially all possible values within some range—subject only to the accuracy of the recording instrument. These data are called *continuous*. This is a very important distinction because specialized analytical methods have been developed to work with each kind of data.

Count data is an important subclass within the general category of discrete data. Within count data, a further special case occurs when the only possible counts are 0 or 1. These are *binary* data. The key characterization is that there are only two possible values; it is a matter of analytical convenience to consider the two values to be 0 or 1. As an example, the response of patient P.S. on each trial was binary. By taking the response "non-burning house" to be 1 and "burning house" to be zero, we are able to add up all the coded values (the 1s and 0s) to get the total number of times P.S. chose the non-burning house. This summation process is easy to deal with mathematically. A set of binary data would almost always be assumed to consist of 0s and 1s.

Two other kinds of data arising in neuroscience deserve special mention here. They are called *time series* and *point processes*. Both involve sequential observations made across time. Imaging signals are good examples of time series: at each of many successive points in time, a measurement is recorded. Spike trains are good examples of point processes: neuronal action potentials are recorded as sequences of event times. In each case, the crucial fact is that an observation at time t_1 is related to an observation made at time t_2 whenever t_1 and t_2 are close to each other. Because of this temporal relationship time series and point process data must be analyzed with specialized methods. Statistical methods for analyzing time series and point processes are discussed in Chapters **??** and **??**.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 2

Manipulating Data

Data analysis comprises two interrelated activities: manipulation and interpretation. Interpretation is based on the logic of statistical inference, which is discussed in Chapters 7–10. Manipulation includes the mechanics of statistical inference, that is, its formulas and computations. Another important kind of data manipulation has no explicit connection with inference, but is instead devoted to summarizing and visualizing data so that they may be more easily comprehended. We describe a few basic ideas below.

The term "data analysis" was coined by John Tukey (see Brillinger, 2002, Appendix D). (Brillinger, D.R. (2002) John W. Tukey: His life and professional contributions, *Annals of Statistics*, 30: 1535–1575.) Tukey emphasized the distinction between formal methods, based on the logic of statistical inference, and informal manipulations—which he called *exploratory*, having a role we indicated in Section 1.2.4. The informality of exploratory data analysis (EDA), however, should not be confused with algebraic simplicity. As we indicate in Section 2.1.2, the manipulations behind many EDA methods are quite complicated. Tukey's large and lingering influence came from demonstrating the power of mathematical, computational, and statistical insight in producing useful displays and summaries of data.

2.1 Describing Central Tendency and Variation

2.1.1 Alternative displays and summaries provide different views of the data.

Alternative displays and summaries may emphasize different aspects of the data. While certain data summaries may be well suited for particular purposes, there is never a uniquely "right" way to collapse the data. A multiplicity of possible data features is inherent to the data analytic process. Furthermore, the details of data summary can be important. A histogram displays the distribution of data values, but the way it does so depends on the way its bins are defined. This is illustrated in the next example.

Example 2.1.1 Saccadic reaction time in hemispatial neglect. Let us consider saccadic reaction times from a single patient in the study of hemispatial neglect by Behrmann *et al.* (2000). (Behrmann, M., Ghiselli-Crippa, T., Sweeney, J.A., DiMatteo, I., and Kass, R. (2000) Mechanisms underlying spatial representation revealed through studies of hemispatial neglect, *J. Cognitive Neurosci.*, 14: 272–290.) Each measured value is the time (in seconds) to complete an eye saccade. The data have been aggregated across several conditions for pedagogical purposes. There are 119 reaction times, which range from .072 to .988 seconds, or 72 to 988 milliseconds. The lower quartile (below which lie 25% of the data) is 140 milliseconds, the median (below which lie 50% of the data) is 188 milliseconds, and the upper quartile (below which lie 75% of the data) is 252 milliseconds. Thus, the fast reaction times (72 to 140 milliseconds) are bunched much closer to the middle reaction of 188 milliseconds than the slow reaction times (252 to 988 milliseconds). We refer to this feature of the distribution as *skewness* toward high values.

Four histograms of the data are shown in Figure 2.1. Although the same 119 values are used in each, the four histograms give somewhat different impressions of the data. In particular, the first histogram (top left) makes the distribution look *unimodal*, i.e., it looks like it has a single peak, while the second (top right) makes the distribution look *bimodal* (two peaks) or even

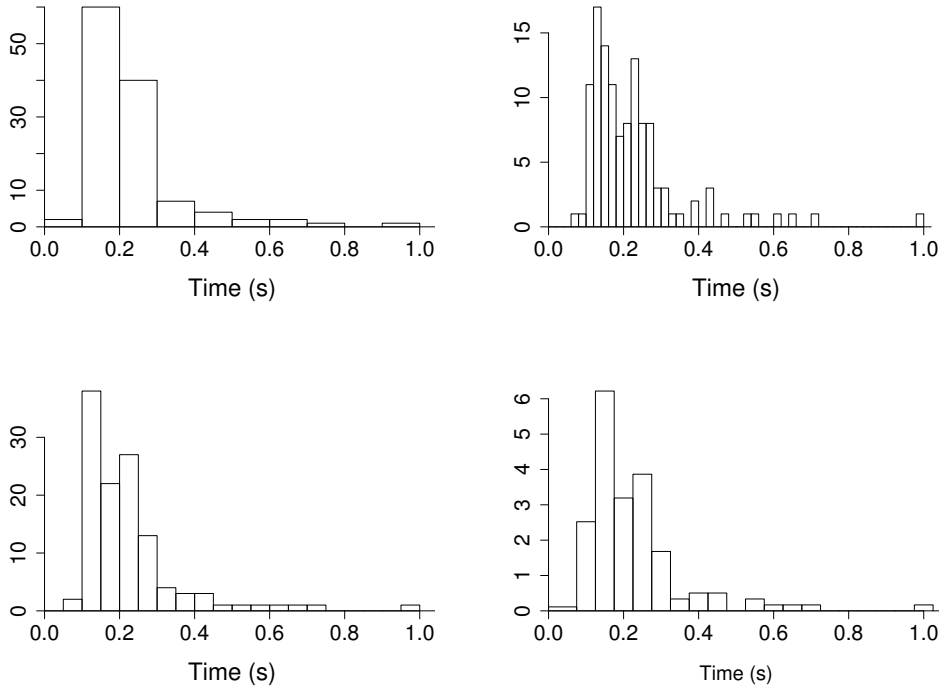


Figure 2.1: *Four histograms of saccadic reaction time data. The same data are used in each histogram. The appearance of the data distribution depends on details of histogram creation. The first three histograms have different bin sizes. The fourth histogram (bottom right) uses the same bin size as the third (bottom left) but shifts the bin locations slightly.*

multimodal (multiple peaks). However, all four give the clear impression of skewness toward high values. □

In discussing histograms it is important to distinguish this informal use of "distribution" from the mathematical use when we speak about a *probability distribution*. We will, beginning in Chapter 3, use probability distributions to describe data, but that should be recognized as a conceptual leap: data are observed, and part of the "real world" of Figure 1.9, while probability distributions are part of the "theoretical world." The word "distribution" is used in both contexts, and we typically hope that a particular probability

distribution will do a good job describing a data distribution. As an example, sometimes data distributions—as represented by histograms—are unimodal and more-or-less symmetrical about the median, i.e., the relative frequency of data higher than the median is about the same as that of corresponding data lower than the median by an equal amount. Symmetric and unimodal data distributions are easier to describe concisely with probability distributions and the *normal distribution*, discussed in Chapter 5, is unimodal and symmetrical (it is often called "the bell-shaped curve"). It is very rare to find a set of data that, on close inspection, may be described accurately by a normal distribution, but it is common to find unimodal and symmetric data distributions that are roughly normal-looking. A great deal of emphasis is placed on the normal distribution, in large part because of its appearance as a basic assumption of many formal statistical procedures and because such statistical procedures typically remain useful for modest departures from normality. When departures from normality become large, however, they can materially affect the behavior of the procedures. A standard practice, therefore, is to examine data via displays such as histograms, looking especially for substantial skewness.

The saccadic reaction time data are substantially skewed. One effect of this is that the mean (the arithmetic average) is substantially higher than the median: the mean reaction time is 226 milliseconds, while the median is 188 milliseconds. This is because the mean is affected much more strongly by values that are far away from the middle of the distribution. Data values that are very far from the middle of the distribution are called *outliers*, and the sensitivity of the mean to outliers is one reason it is often replaced by the median as a summary of central tendency. In addition to the mean and median, the *mode*, which is the value occuring most frequently, is sometimes mentioned in this context. However, the term "mode" is not used in a precise way very often when describing a bunch of numbers. The concept of a mode applies better to the theoretical setting of probability densities, where it is the value at which the density is maximized. For a bunch of numbers we typically speak, instead, informally and approximately, of "the mode" as being the rough location of the peak of the distribution.

Just as central tendency in data may be summarized by mean or median, variability may be summarized by more than one measure. We might ask, for example, how much the saccade times vary. For instance, if we were to

look at a control subject might we expect less variability? How should we quantify this?

The most widely used summary of variability is the *standard deviation*:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $x_1, x_2, \ldots, x_n$ are the observations and $\bar{x}$ is their mean. We may think of s as an "average deviation" of the values from their mean. For the saccadic reaction time data we find the standard deviation to be $s = .134$, or 134 milliseconds. The use of $n-1$ rather than n in the formula for s comes from certain theoretical arguments.

> *Details:* In Chapter 7 we will consider a random sample $X_1, X_2, \ldots, X_n$ and define the *sample variance* as
>
> $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$
>
> We will see that the theoretical average value of S^2, known as its *expectation*, is $E(S^2) = \sigma_X^2$. Because its expectation is equal to the quantity it is estimating, S^2 is called *unbiased*. If we instead used n in the denominator the expectation would be $(n-1)/n$ times σ_X^2, and for small n this can make S^2 slightly less accurate as an estimator of σ_X^2. Furthermore, in the related context of linear regression this kind of consideration becomes more consequential: if a response variable y is being related to regression variables $x_1, \ldots, x_{p-1}$ the denominator of the unbiased estimator of σ^2 becomes $n-p$.

An alternative to the standard deviation would be the mean absolute deviation $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$ but this turns out to be mathematically less convenient. In some contexts the median absolute deviation is used as this is not affected by outliers. If $\tilde{x} = median(x_1, x_2, \ldots, x_n)$ is the median of the n data values x_i then the median absolute deviation is $median(x_1 - \tilde{x}, x_2 - \tilde{x}, \ldots, x_n - \tilde{x})$. Sometimes the difference between the quartiles is used. This is called the *interquartile range*.

2.1.2 Exploratory methods can be sophisticated.

As we said in Section 1.2.4, exploratory data analysis (EDA) refers to the collection of methods that are relatively informal, based not on a cohesive logical framework built around statistical models but rather on tools that seem to illuminate interesting features of the data. The informal methods of EDA can be extremely useful. In this section we have mentioned a couple of very elementary descriptive methods, but in some cases informal techniques can draw on quite sophisticated ideas. The next example involves a method we will outline in Chapter **??**.

Example 2.1.2 EEG spectrogram under anesthesia When patients undergo general anesthesia for certain surgical procedures EEGs are collected to monitor brain activity. These recordings provide a comparison of various states of consciousness. A set of EEG traces for a patient during carotid endarterectomy surgery at the Massachusetts General Hospital is displayed in Figure 2.2. The figure shows EEGs and spectrograms during an initial awake phase, an anesthesia induction phase, the surgical phase, and the recovery phase. Spectrograms are made by taking the signal within successive time bins (here, 1 second bins) and using *Fourier analysis* to decompose the signal into trigonometric components that represent oscillations at varying frequencies. On the x-axis is time and on the y-axis is the frequency. The plotted spectrogram is the power (squared amplitude) of the signal at each frequency, for each time bin, indicated in the figure by three different colors representing low, medium, and high power. In Figure 2.2 the most easily visible oscillations are the alpha rhythm in the second half of the EEG trace in the awake phase (when the eyes are closed) and the delta rhythm during the surgical phase. Precise scientific statements often require statistical inferences (indications of uncertainy or tests of hypotheses), but spectrograms are very useful in displaying time-frequency information even without formal inferential assessments. □

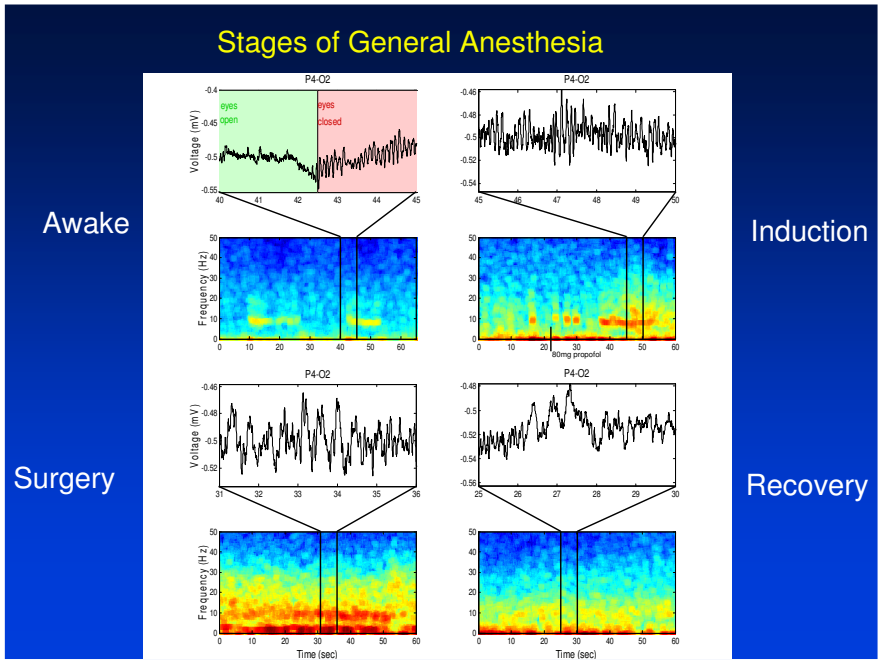


Figure 2.2: *EEG spectrograms for a subject in various stages of general anesthesia. In each of four stages an EEG voltage tracing is shown, and below it a spectrogram. The EEG tracings are for the P4 (right parietal) lead in an array of 16 leads (it is taken with O2 as reference lead). The spectrogram decomposes the voltage signal into frequency components across successive time bins. Red indicates high magnitudes, yellow medium magnitudes, and blue low magnitudes. Each displayed tracing corresponds to several successive time bins in the spectrogram, as indicated by the black lines. Two prominent features are the roughly 10 Hz "alpha rhythm" and the slower 1-4 Hz delta rhythm. Both sets of oscillations are visible in the EEG tracings, and their temporal presence or absence is indicated in the spectrogram. During the awake phase the alpha rhythm is absent when the eyes are open and present when the eyes are closed; the delta rhythm is also present, but only weakly. During surgery the delta rhythm is very strong, and the alpha rhythm is also stronger than in the awake phase.*

2.2 Data Transformations

2.2.1 Positive values are often transformed by logarithms.

Measurement scales arise from convenience, and need not be considered in any way absolute or immutable: changing the scale often produces a more elegant description. A canonical example involves the acidity of a dilute aqueous solution, which is determined by the concentration of hydrogen ions. The larger the concentration $[H^+]$ of hydrogen ions, the more acidity. Rather than using $[H^+]$ to measure acidity, we use its *log* which is known as pH. (Specifically, $pH = -\log_{10}([H^+])$, so that an increase in $[H^+]$ corresponds to a decrease in pH.) Because the defining property of the logarithm is

$$\log ab = \log a + \log b, \tag{2.1}$$

log transformations are used when multiplicative effects seem more natural than additive. In the case of pH, a solution having a hydrogen ion concentration of 10^{-5} moles per liter is 1 unit greater pH (less acidic) than a solution having a concentration of 10^{-4} moles per liter. Similarly, a solution having a hydrogen ion concentration of 10^{-9} moles per liter is 1 unit greater pH than a solution having a concentration of 10^{-8} moles per liter. In both cases, a 1 unit increase in pH corresponds to a 10-fold decrease in hydrogen ion concentration, regardless of the concentration we started with. In chemical calculations, the log concentration scale is simpler to work with than the concentration scale.

Many other familiar scales are logarithmic. One example is the use of decibels to measure the strength of an auditory signal.

Not only are log scales familiar and intuitive, data are often better behaved following a log transformation. In particular, it frequently happens that a batch of data look highly skewed in a given measurement scale, but are much closer to being symmetric in the log scale.

Example 2.1.1 (continued) Figure 2.3 displays the saccadic reaction time data in both the original scale and the log transformed scale. To transform the data to the log scale we have replaced $x =$ reaction time by $\log_{10}(x)$

for each of the 119 values. In the log scale the distribution is more symmetrical. In addition, the potential bimodality, or possibly even multimodality of the distribution is also evident in the log scale. The data shown here were aggregated by combining conditions in which the eyes began fixating centrally, to the right, or to the left, so it is not be surprising that the distribution appears non-unimodal. When the data are disaggregated into single conditions, in the log scale they do appear unimodal and roughly symmetrical. For this reason, Behrmann *et al.* chose to perform many of their analyses in the log scale. □

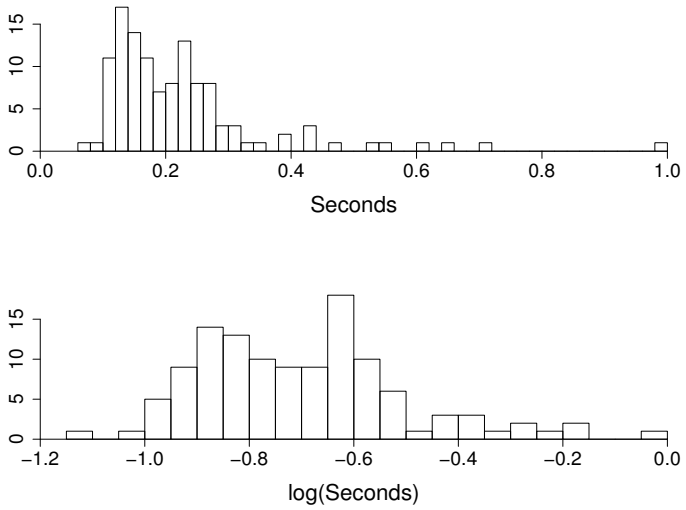


Figure 2.3: *Histograms of eye saccade data. Top display is for data in the original scale, bottom display is for the same data after being transformed by* $\log_{10}$. *The data are distributed more symmetrically in the log scale.*

Example 2.2.1 Log transform of the BOLD signal Lewis *et al.* (2005) have argued that for some purposes it may be advantageous to transform the BOLD signal in fMRI data by taking logarithms, at least in the case of high-field signals. (Lewis SM, Jerde TA, Tzagarakis C, Gourtzelidis P, Georgopoulos MA, Tsekos N, Amirikian B, Kim SG, Ugurbil K, Georgopoulos AP. (2005) Logarithmic transformation for high-field BOLD fMRI data. *Exp. Brain Res.*, 165:447-53.) Those authors examined the BOLD intensity for subjects during 4 Tesla imaging, with a simple visual stimulus. Figure 2.4

displays a histogram (with dots replacing bin heights) of the BOLD values collected from 19,000 voxels for each of 15 subjects and 15 images under their control condition, during which the subjects were fixating on a central spot on the screen they were watching. It is apparent that this distribution across voxels is quite skewed. The authors produced various plots aimed at suggesting the log transformation could be useful. □

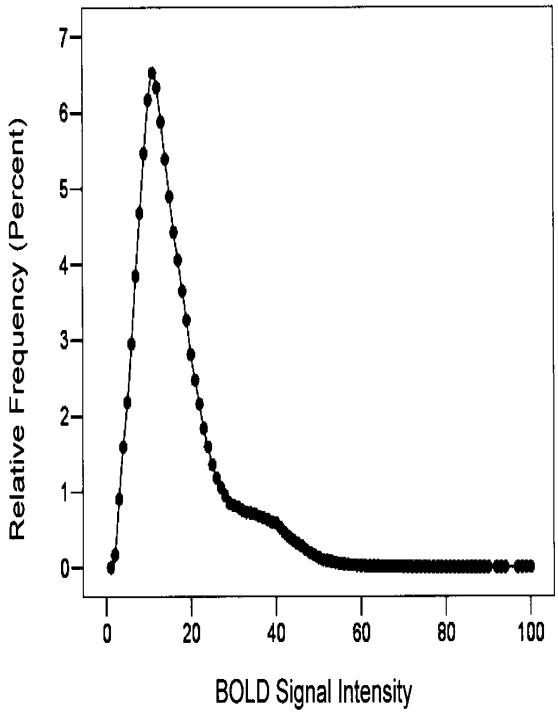


Figure 2.4: *High field of BOLD signal intensities. The frequencies are plotted as dots, rather than bin heights. The distribution across voxels is skewed toward high values. Modified from Lewis* et al. *(2005).*

The way we usually think of the log transformation is that it produces a more "natural" scale for measurements whenever they are necessarily positive and might reasonably be compared in proportional relationships. For instance, as we just described, various values of pH are quite naturally compared in proportional terms. Similarly, one might speak of eye saccades as

taking, say, 35% longer in hemispatial neglect patients than in control subjects. Growth phenomena, like lengths of some anatomical feature in some organism, are often considered in percentage terms. That is, in describing the variability in some set of organisms, we might find ourselves thinking in terms of proportional change from one to the next.

Transformations are important in data analysis. We can understand this a little more deeply by considering the way a bunch of numbers vary, and the resulting effects of taking logs. We have already mentioned that normal distributions for data are assumed by standard statistical procedures, that data distributions are rarely very close to normal, but that mild departures from normality are generally quite tolerable. Such mild departures are common: once we transform the data to a suitable scale, distributions are often unimodal and more-or-less symmetrical. Why? Presumably, this has to do with effects of the Central Limit Theorem. We will discuss this great theorem in Chapter 6. For now let us be content to state it this way: if we add up many small, independent effects their sum will be approximately normally distributed. The empirical observation of approximate normality may then be interpreted as follows: *if we choose the right scale*, the data values may be considered sums of many small, independent effects. With this in mind, let us return to the logarithmic relationship in Equation (2.1), and the role it may play when many small effects are combined to produce variability. The cases where the log transformation is valuable are those where it is natural to think in terms of proportionality. So suppose the reason that two measurements are different is that many small *proportional* effects, of somewhat different sizes in the two measurements, have been combined. For example, the length of a dendritic spine may depend on contributions to the cell membrane and its contents by vast numbers of lipid and protein molecules. If we break the growth process into many thousands of pieces, each might be considered a small effect, so that the net result is a composition of many, many small effects. When we see that one spine is longer than another, we might imagine that the many small effects in the longer spine tended to be *proportionally* larger than those in the shorter spine. Now consider two such small growth effects x_1 and x_2, occurring, respectively, in the shorter and longer dendrites. If we think of the variation as proportional, we may relate the values x_1 and x_2 by writing $x_2 = x_1(1 + \epsilon)$, where ϵ is a small number representing the proportional change (e.g., $\epsilon = .05$, or 5%) in going from x_1 to x_2. From Equation (2.1) together with a little calculus, for small ϵ we

have $\log(1+\epsilon) \approx \epsilon$ (see Appendix A). We then have

$$\begin{aligned} \log x_2 - \log x_1 &= \log(1+\epsilon) \\ &\approx \epsilon. \end{aligned}$$

In other words, when we add a small perturbation ϵ to $\log x_1$ we get $\log x_2$. Thus, if we wish to think of ϵ as a small random quantity that creates variability in the data multiplicatively, it does so additively on the log scale. When we consider a large number of such small effects acting proportionally, on the log scale the corresponding effects will be summed. Thus, following a log transformation, data that are approximately of this type will, according to the Central Limit Theorem, be described, approximately, by a normal distribution.

All of this is heuristic; there is no argument here that can be claimed formally correct—the Central Limit Theorem applies not to data but to mathematical quantities that live in the "theoretical world" described in Section 1.2.5. We are simply trying to provide a plausible explanation for the empirical fact that log-transformed growth measurements usually have fairly symmetrical distributions.

By the way, in transforming data it does not matter whether the "natural" log (the log-base-e, $\log_e(x)$) or the "common" log (the log-base-10, $\log_{10}(x)$) is used. The numbers will be multiplied by the same constant value (we have $\log_e(x) = \log_e(10)\log_{10}(x)$) for all of the measurements and therefore will not affect the appearance of the plots, nor the applicability of any subsequent statistical analysis.

Another motivation for logarithmic transformations is that they convert power laws, which are useful in describing many neuroscientific phenomena, to simpler linear forms: suppose

$$w = cv^b.$$

If we let $y = \log w$ and and $x = \log v$ then

$$y = a + bx,$$

where $a = \log c$. Such power laws may be summarized by saying that a proportional change in v produces a proportional change in w.

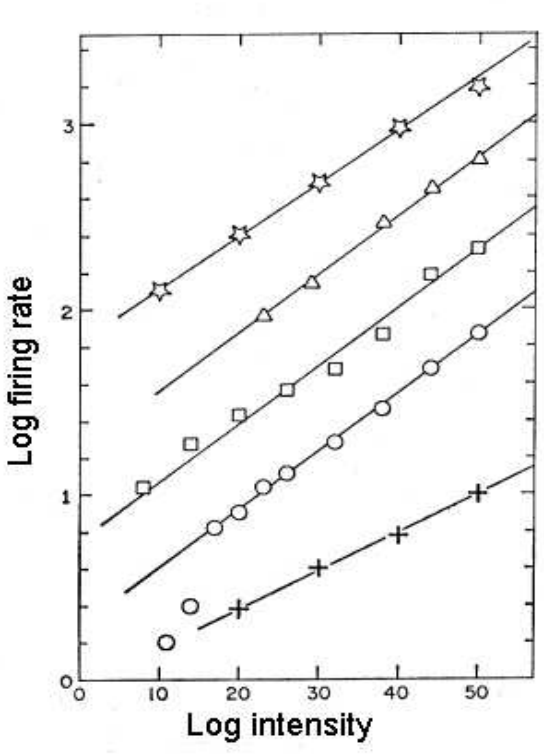


Figure 2.5: *Power function fits to firing-rate data, shown on log-log scale. On the y-axis are log firing rates, and on the x-axis is log intensity of light. The data are from three different sources, using three distinct methods of collection. Except for the deviation from the line at low intensities for the data set indicated by circles, the fits are quite good. Modified from Stevens (1970).*

Example 2.2.2 Stimulus-response power laws Power laws may be used to describe the way increases in stimulus intensity produce increased magnitudes of sensation (Stevens, 1961, *Science*, 133: 80–86) (where they replace the "Weber-Fechner" law $w = a + d \log v$), or increased neural firing rate (Stevens, 1970, *Science*, 170: 1043–1050). For example, Figure 2.5 displays five classic sets of data on neural responses from the eye of the horshoe crab *Limulus*. For each data set, the log of neural firing rate is plotted as a function of log of light intensity. In each case the function is approximately linear. In other words, in each case the relationship of firing rate to stimulus

intensity follows, approximately, a power law. □

Example 2.2.3 Power law for skill acquisition Power laws also arise in describing the effects of practice on recall or reaction time in memory and skill acquisition (Anderson, 1990, *Cognitive Psychology and its Implications*). An interesting set of data comes from Kolers (1976, Reading a year later, *J. Experimental Psychology: Human Learning and Memory*, 2: 554–565.) who investigated the learned skill of reading inverted text. As shown in Figure 2.6, he found two things. First, a decreasing power law describes the relationship of reading time to amount of practice. Second, when subjects were tested a year later, they had lost some of their ability to read the inverted text, and then regained it again according to a power law, though at a slower rate. The two relationships are shown in Figure 2.6 as a pair of lines with differing slopes and intercepts. (These studies are of great interest for education: they suggest that retained learning may be quantified by the decrease in training time required to achieve proficiency following re-training, compared to the original training time.) □

2.2.2 Non-logarithmic transformations are sometimes applied.

The log is by far the most common transformation, but there are others, too. The general method of transformations is to replace a measured variable like $x =$ reaction time with some $f(x)$ for every value of x. For example, reaction times and other time measurements are sometimes analyzed on the reciprocal scale $1/x$: the reciprocal transforms time to something proportional to speed (speed is distance/time). Square-root transformations are also sometimes used, especially for spike counts because the square-root can be a so-called *variance-stabilizing transformation*, as discussed in Chapter 9. Square-roots are also sometimes used for measurements of area and cube-root transformations are occasionally used for volumetric measurements. We may order these transformations by letting, for the moment, the symbol $<$ stand for "less strong than" and then writing them as follows:

$$x < x^{1/2} < x^{1/3} < \log(x) < 1/x.$$

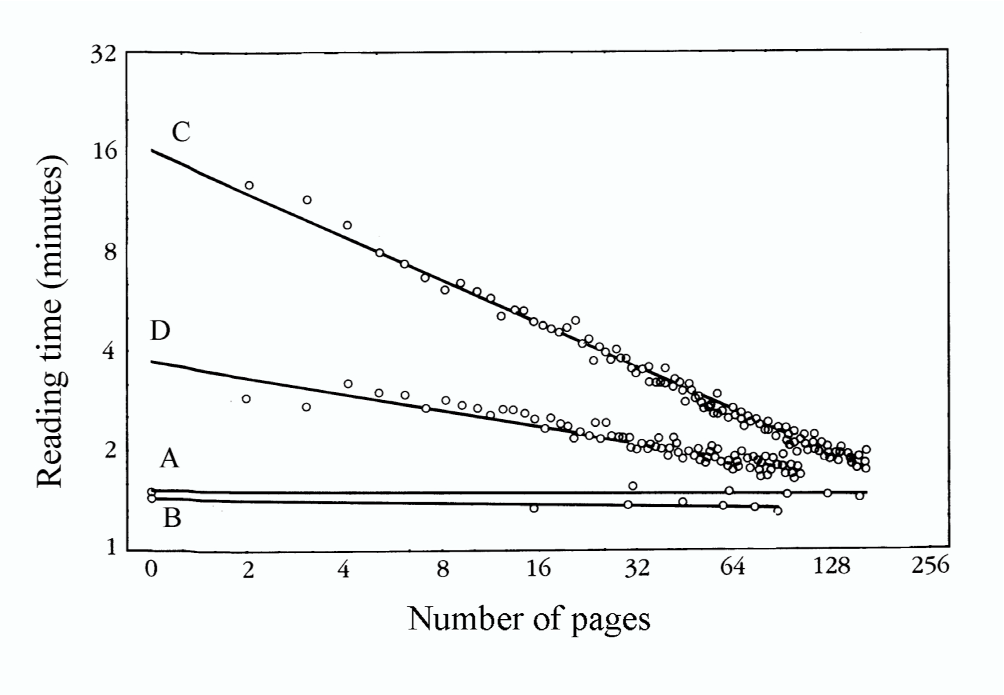


Figure 2.6: *Skill learning described by a power law, shown on a log-log scale. On the y-axis is the log (base 2) of the time taken to read a passage of inverted text (in minutes), and on the x-axis is log practice time (in pages). Four sets of data from multiple subjects are displayed. Data were obtained on two occasions, separated by a year, on both ordinary text and inverted text (creating a total of four conditions). Line A is fit to data based on ordinary text on the first occasion and line B is fit to data based on ordinary text on the second occasion. There is essentially no training effect. Lines C and D are the fits for inverted text. In both cases there is a clear power-law relationship, indicated by the good fit of the lines to the data. Substantively, after the delay by a year the subjects again improved with practice, but they had retained much of the skill of reading inverted text (line D is below line C) and needed only about 100 pages of training to reach the proficiency previously obtained after 200 pages. Modified from Kolers (1976).*

In each case we strengthen the transformation (make it pull in the right-hand tail more) as we decrease the power to which we raise x. Note that $1/x = x^{-1}$ and that, in this context, the log corresponds to using the power 0, so that increasing the strength of transformation corresponds to decreasing the exponent.

> *Details:* We may imbed the log in the power family of transformations by putting the power transformations in the normalized form
>
> $$f(x) = (x^\alpha - 1)/\alpha.$$
>
> By calculus (L'Hôpital's rule) it then follows that the log corresponds to $\alpha = 0$.

In general, both distributional symmetry and interpretability are important in determining a scale for analysis.

These "power transformations" are all monotonic. Occasionally, non-monotonic transformations are used, as in the analysis of EMG recordings.

Example 2.2.4 EMG in frog movement An Electromyogram (EMG) is a recording of the electrial impulses transmitted through a group of muscle fibers, recorded as electrical potentials. Because the instantaneous potential is generated from both agonist and antagonist muscle fibers, it is recorded as both positive and negative. This is shown in the top panel of Figure 2.7, which is a display of an EMG taken from a frog during a leg extension. Because the force generated by a muscle is only positive, the standard convention is to analyze the rectified signal, i.e., the absolute value of the potential. This is shown in the bottom panel of Figure 2.7. □

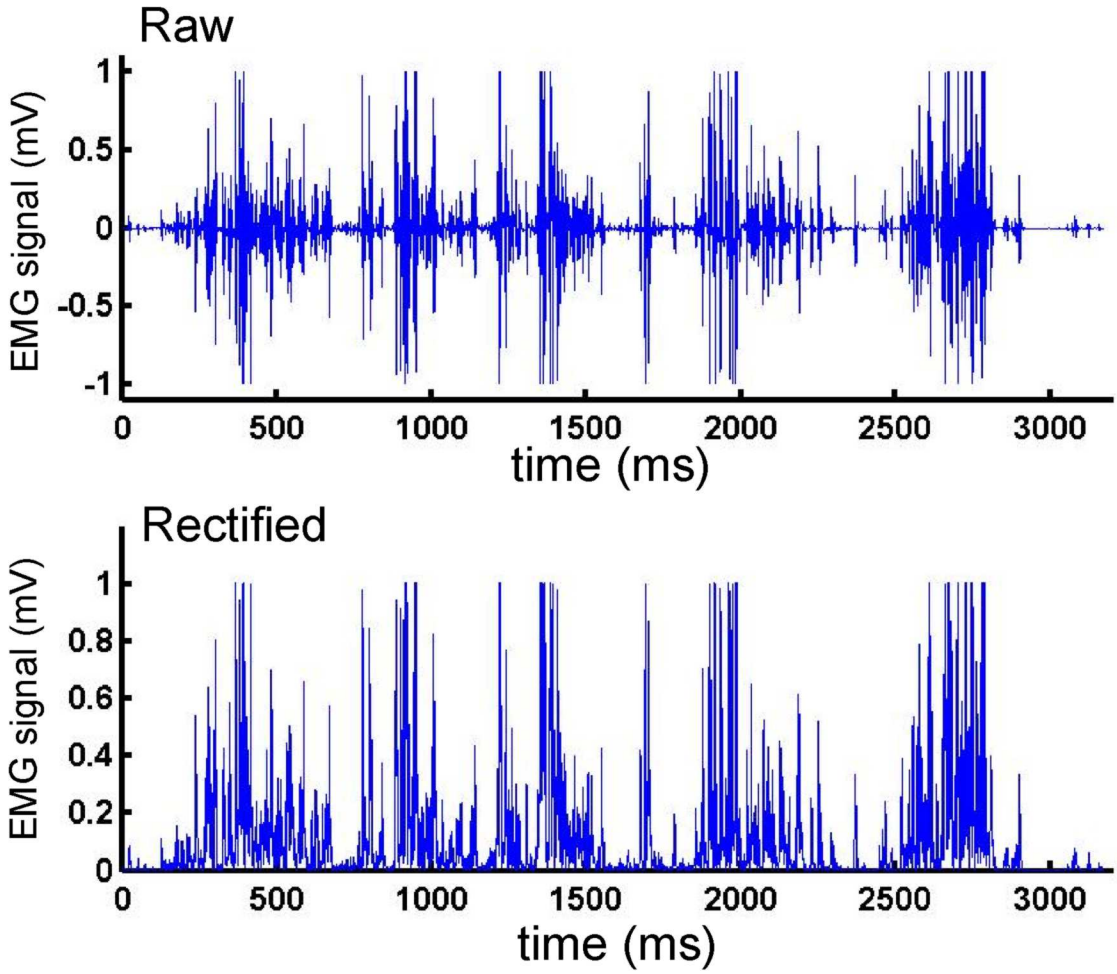


Figure 2.7: *EMG from the leg of a frog during a swimming motion. Top panel shows raw signal. Bottom panel shows the rectified signal.*

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 3

Probability and Random Variables

Probability is a rich and beautiful subject, a discipline unto itself. Its origins were concerned primarily with games of chance, and many lectures on elementary probability theory still contain references to dice, playing cards, and coin flips. These lottery-style scenarios remain useful because they are evocative and easy to understand. On the other hand, they give an extremely narrow and restrictive view of what probability is about: lotteries are based on elementary outcomes that are equally likely, but in many situations where quantification of uncertainty is helpful there is no compelling way to decompose outcomes into equally-likely components. In fact, the focus on equally-likely events is characteristic of pre-statistical thinking.[1] The great leap forward toward a more general notion of probability was slow, requiring over 200 years for full development.[2] This long, difficult transition involved

[1]See Stigler, S.M. (1986) *The History of Statistics: Measurement of Uncertainty before 1900*, Harvard.

[2]Its beginning point is usually traced to a text by Jacob Bernoulli, posthumously-published in the early 1713 (Bernoulli, J. (1713) *Ars Conjectandi* Basil: Thurnisiorum.), and its modern endpoint was reached in 1933, with the publication of a text by

a deep conceptual shift. In modern texts equally-likely outcomes are used to illustrate elementary ideas, but they are relegated to special cases. It is sometimes possible to compute the probability of an event by counting the outcomes within that event, and dividing by the total number of outcomes. For example, the probability of rolling an even number with a fair six-sided die, i.e., of rolling any of the three numbers 2, 4, or 6, out of the 6 possibilities, is $\frac{3}{6} = \frac{1}{2}$. In many situations, however, such reasoning is at best a loose analogy. To quantify uncertainty via statistical models a more general and abstract notion of probability must be introduced.

This chapter begins with the axioms and elementary laws of probability, and then discusses the way probability is used to describe variability. The key concept of *independence* is defined in Section 3.1.3. Quantities that are measured but uncertain are formalized in probability theory as *random variables*. More specifically, we set up a theoretical framework for understanding variation based on probability distributions of random variables, and the variation of random variables is supposed to be similar to real-world variation observed in data. Many families of probability distributions are used throughout the book. The most common ones are discussed in Chapter 5.

One quick note on terminology: the word *stochastic* connotes variation describable by probability. Within statistical theory it is often used in specialized contexts, but it is almost always simply a synonym for "probabilistic." We occasionally use this word ourselves.

3.1 The Calculus of Probability

3.1.1 Probabilities are defined on sets of uncertain events.

The calculus of probability is defined for *sets*, which in this context are called *events*. That is, we speak of "the probability of the event A" and we will write this as $P(A)$. Events are considered to be composed of *outcomes* from some

Kolmogorov (Kolmogorov, A.N. (1933) *Grundgebriffe der Wahrscheinlichkeithsrechnung,* Berlin: Springer-Verlag. English translation: 1950, *Foundations of the Theory of Probability*, New York: Chelsea.)

experiment or observational process. The collection of all possible outcomes (and, therefore, the union of all possible events) is called the *sample space* and will be denoted by $\mathcal{S}$. Because $\mathcal{S}$ is a set, we might also say that $\mathcal{S}$ is made up of elements (each of which is an outcome) and to indicate that ω is an element of $\mathcal{S}$ we would write $\omega \in \mathcal{S}$. Recall the definitions of *union* and *intersection*: for events A and B the union $A \cup B$ consists of all outcomes that are either in A or in B or in both A and B; the intersection $A \cap B$ consists of all outcomes that are in both A and B. The *complement* A^c of A consists of all outcomes that are *not* in A. We say two events are *mutually exclusive* or *disjoint* if they have empty intersection.

Example 3.1.1 Two neurons from primary visual cortex In an experiment on response properties of cells in primary visual cortex, Ryan Kelly and colleagues recorded approximately 100 neurons simultaneously from an anesthetized macaque monkey while the animal's visual system was stimulated by highly irregular random visual input. (Kelly *et al.*, 2007). (Kelly, R.C., Smith, M.A., Samonds, J.M., Kohn, A., Bonds, A.B., Movshon, J.A., and Lee, T.-S. (2007) Comparison of recordings from microelectrod arrays and single electrodes in the visual cortex, *J. Neurosci.*, 27: 261–264.) The stimulus they used is known as *white noise*, which will be defined in Chapter **??**. Kelly examined the response of two neurons during 100 milliseconds of the stimulus. Let A be the event that the first neuron fires at least once within the 100 millisecond time interval and B the event that the second neuron fires at least once during the same time interval. Here, $A \cup B$ is the event that at least one of the 2 neurons fires at least once, while $A \cap B$ is the event that both neurons fire at least once. Because it is possible that both neurons will fire during the time interval, the events A and B are *not* mutually exclusive. □

We now state the axioms of probability.

Axioms of probability:

1. For all events A, $P(A) \geq 0$.

2. $P(\mathcal{S}) = 1$.

3. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, then $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$.

A technical point is that in advanced texts, Axiom 3 would instead involve infinitely many events, and an infinite sum:

3′. If $A_1, A_2, \ldots,$ are mutually exclusive events (possibly infinitely many events), then $P(\cup_i A_i) = \sum_i P(A_i)$

where the notations mean that the union and summation extend across all events.

Regardless of whether one worries about the possibility of infinitely many events, it is easy to deduce from the axioms the elementary properties we need.

Theorem: Three Properties of Probability For any events A and B we have

(i) $P(A^c) = 1 - P(A)$, where A^c is the complement of A.

(ii) If A and B are mutually exclusive, $P(A \cap B) = 0$.

(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: To prove (i) we simply note that $\mathcal{S} = A \cup A^c$. From axiom (2) we then have $P(A \cup A^c) = 1$ and because A and A^c are mutually exclusive axiom (3) gives $P(A) + P(A^c) = 1$, which is the same as (i). It is similarly easy to prove (ii) and (iii). □

These facts are often illustrated by analyzing games of chance, which is the context in which many of the basic methods of probability were first worked out. For instance, in picking at random a playing card from a standard 52-card deck, we may compute the probability of drawing a spade or a face card, meaning either a spade that is not a face card, or a face card that is not a spade, or a face card that is also a spade. We take A to be the event that we draw a spade and B to be the event that we draw a face card. Then, because there are 3 face cards that are spades we have $P(A \cap B) = \frac{3}{52}$,

and, appying the last formula above, we get $P(A \cup B) = \frac{1}{4} + \frac{3}{13} - \frac{3}{52} = \frac{11}{26}$. This matches a simple enumeration argument: there are 13 spades and 9 non-spade face cards, for a total of 22 cards that are either a spade or a face card, i.e., $P(A \cup B) = \frac{22}{52} = \frac{11}{26}$. The main virtue of such formulas is that they also apply to contexts where probabilities are determined without reference to a decomposition into equally-likely sub-components.

Example 3.1.1 (Continued) From 1200 replications of the 100 millisecond stimulus Kelly calculated the probability that the first neuron would fire at least once was $P(A) = .13$ and the probability that the second neuron would fire at least once was $P(B) = .22$, while the probability that both would fire at least once was $P(A \cap B) = .063$. Applying the formula for the union (property (iii) above), the probability that at least one neuron will fire is $P(A \cup B) = .287$. □

3.1.2 The conditional probability $P(A|B)$ is the probability that A occurs given that B occurs.

We often have to compute probabilities under an assumption that some event has occurred. For instance, one may be interested in the probability that a neuron will fire in an interval of time $(t, t+\Delta t)$ given that it has already fired at a previous time t_0. If we let A be the event we are interested in and B the event that is assumed to have occurred, then we write[3] $P(A|B)$ for the *conditional probability of A given B*. From a Venn diagram (see Figure 3.1) it is easy to visualize the calculation required: we limit the universe to B and ask for the relative probability assigned to the part of A that is contained in B. Algebraicly, the formula is the following:

Definition: Conditional Probability Assume $P(B > 0)$. The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

[3]This notation is due to Jeffreys (1931); see his page 15. (Jeffreys, H. (1931) *Scientific Inference*, Cambridge)

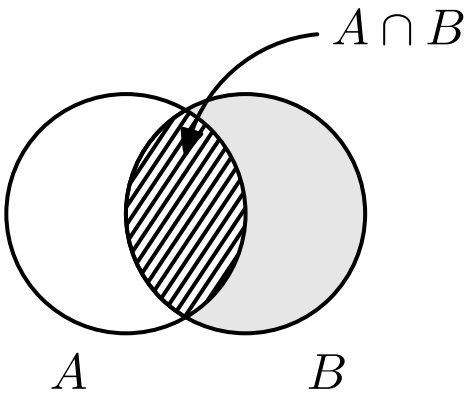


Figure 3.1: *Venn diagram showing the intersection of A and B. The events A and B are depicted as open and filled-in circles, respectively, while $A \cap B$, the portion of B that is also in A, is shown with diagonal lines. The conditional probability of A given B is the relative amount of probability assigned to A within the probability assigned to B, i.e., the probability assigned to the region having diagonal lines divided by the probability assigned to the whole of B.*

Again, using draws from a deck of cards, the probability of drawing a Jack given that we draw a face card is $P(A|B) = \frac{4/52}{12/52} = \frac{1}{3}$.

A rewriting of the definition of conditional probability is also sufficiently useful to have a name:

Multiplication rule If $P(B) > 0$ we have $P(A \cap B) = P(A|B) \cdot P(B)$.

Although conditional probability calculations are pretty straightforward, problems involving conditioning can be confusing. The trick to keeping things straight is to be clear about the event to be conditioned upon. Here is one standard example.

Illustration: The boy next door Suppose a family moves in next door to you and you know they have two children, but you do not know whether the children are boys or girls. Let us assume the probability that either particular child is a boy is $\frac{1}{2}$. We might label them Child 1 and Child 2 (e.g., Child 1 could be the older of the two). Thus, P(Child 1 is a boy) =

$P(\text{Child 2 is a boy}) = \frac{1}{2}$. Now suppose you find out that one of the children is a boy. What is the probability that the other child is also a boy?

It may seem that the answer is $\frac{1}{2}$ but, if we assume that "you find out one of the children is a boy" means *at least one of the children is a boy*, then the correct answer is $\frac{1}{3}$. Here is the argument. When you find out that one of the children is a boy you don't know whether Child 1 is a boy, nor whether Child 2 is a boy, but you do know that one of them is a boy—and possibly both are boys. This information amounts to telling you it is impossible that both are girls. Let A be the event that both children are boys and B the event that at least one child is a boy. We want $P(A|B)$. Note that there are four equally-likely possibilities:

$$\begin{aligned} & P(\text{Child 1 is a boy and Child 2 is a boy}) \\ = \; & P(\text{Child 1 is a boy and Child 2 is a girl}) \\ = \; & P(\text{Child 1 is a girl and Child 2 is a boy}) \\ = \; & P(\text{Child 1 is a girl and Child 2 is a girl}). \end{aligned}$$

Thus, we compute $P(A \cap B) = P(A) = \frac{1}{4}$ and $P(B) = \frac{3}{4}$. Plugging these numbers into the formula for conditional probability we get $P(A|B) = \frac{1}{3}$. □

3.1.3 Probabilities multiply when the associated events are independent.

Intuitively, two events are *independent* when the occurrence of one event does not change the probability of the other event. This intuition is captured by conditional probability: the events A and B are independent when knowing that B occurs does not affect the probability of A, i.e., $P(A|B) = P(A)$. This statement of independence is symmetrical: A and B are also independent if $P(B|A) = P(B)$. However, these statements are not usually taken as the definition of independence because they require the events to have nonzero probabilities (otherwise, conditional probability is not defined). Instead, the following is used as a definition.

Definition: Independence Two events A and B are independent if and only if $P(A \cap B) = P(A) \cdot P(B)$.

Note that from this definition, when A and B are independent and $P(B) > 0$ we have, as a consequence, $P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A)$.

Multiplication of probabilities should be very familiar. If a coin has probability .5 of coming up heads when flipped, then we usually say the probabiilty of getting two heads is $.25 = .5 \times .5$, because we usually assume that the two flips are independent.

Example 3.1.1 (Continued) For the probabilities $P(A)$, $P(B)$ given on page 55 we have $P(A)P(B) = .029$ while the probability of the intersection was reported to be $P(A \cap B) = .063$. The latter is more than double the product $P(A)P(B)$. We conclude that the two neurons are not independent. Their tendency to fire much more often together than they would if they were independent could be due to their being connected, to their having similar response properties, or to their both being driven by network fluctuations (see also Kelly *et al.*, 2010). (Kelly, R.C., Smith, M.A., Kass, R.E., T.S. Lee (2010) Local field potentials indicate network state and account for neuronal response variability, *J. Computational Neuroscience*, to appear.) □

This definition of independence extends immediately to more than two events. Independence is extremely useful because it greatly simplifies calculations (by avoiding specification of potentially complicated dependencies that would have to be represented by conditional probabilities) and, as a result, it is often assumed in statistical models and methods. On the other hand, as the simple example above illustrates, it is treacherous because the calculations can be way off if the assumption is wrong: in Example 3.1.1, the probability $P(A \cap B)$ based on independence would be off by a factor of more than 2. In many situations independence is the most consequential statistical assumption, and therefore must be considered carefully.

3.1.4 Bayes' Theorem for events gives the conditional probability $P(A|B)$ in terms of the conditional probability $P(B|A)$.

Bayes' Theorem is a very simple identity, which we derive easily below. Yet, it has profound consequences. We can state its purpose formally, without regard to its applications: Bayes' Theorem allows us to compute $P(A|B)$ from the reverse conditional probability $P(B|A)$, if we also know $P(A)$. As we will see below, and in Chapter **??**, there are more complicated versions of the theorem, and it is especially these that produce the wide range of applications. But the power of the result becomes apparent immediately when we take B to be some data and A to be a scientific hypothesis. In this case, we can use the statistical model $P(\text{data}|\text{hypothesis})$ to obtain the scientific inference $P(\text{hypothesis}|\text{data})$. In the words used in Chapter 1, Bayes' Theorem provides a vehicle for obtaining epistemic probabilities from descriptive probabilities. The inverting of conditional probability statements, together with the recognition that a different notion of probability was involved, led to the name "inverse probablity" during the early 1800s. This has been replaced by the name "Bayes" in the theorem, and the adjective "Bayesian" to describe many of its applications.[4] To derive the theorem we need a preliminary result which is also important.

Theorem: Law of Total Probability For events A and B we have

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Proof: We begin by decomposing B into two pieces: $B = (B \cap A) \cup (B \cap A^c)$. Because A and A^c are disjoint, $(B \cap A)$ and $(B \cap A^c)$ are disjoint. We then have $P(B) = P(B \cap A) + P(B \cap A^c)$. Applying the multiplication rule to $P(B \cap A)$ and $P(B \cap A^c)$ gives the result. □

Bayes' Theorem in the Simplest Case If $P(B) > 0$ then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \tag{3.1}$$

[4]For historical comments see Stigler (1986) and Fienberg (2006). (Fienberg, S.E. (2006) When did Bayesian inference become "Bayesian"? *Bayesian Analsysis*, 1:1–40.)

Proof: We begin with the definition of conditional probability and then use the multiplication rule in the numerator and the law of total probability in the denominator:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \end{aligned}$$

□

The "simplest case" modifier here refers to the statement of the theorem using only A and A^c as conditioning events. One interesting class of problems where this simple case is useful is in the interpretation of clinical diagnostic screening tests. These tests are used to indicate that a patient may have a particular disease, but they are not definitive. Bayes' Theorem serves as a quantitative reminder that when a disease is rare, screening tests are preliminary, and other information will be needed to provide a diagnosis. A famous example involves screening for prostate cancer based on the radioimmunoassay prostatic acid phosphatase (PSA). Even though the test is reasonably accurate, the disease remains sufficiently rare among young men that a random male who tests as positive will still have a low probability of actually having prostate cancer. An application of Bayes' Theorem (with A being the event that a randomly chosen man will have the disease and B the event that he tests positive) to data from Watson and Tang (1980), places the probability of disease given a positive test at about 1/125. The intuition comes from recognizing that, among men under age 65 in the United States, the disease has a prevalence of about 1/1500. Suppose we were to examine 1500 men, 1 of whom actually had the disease. If the test were 90% accurate, a 10% false positive rate would mean that about 150 men would test positively. In other words, about 1/150 of the positively tested men would actually have the disease. Bayes' Theorem refines this very crude calculation. Here is an example drawn from neurology.

Example 3.1.2 Diagnostic test for vascular dementia Vascular dementia (VD) is the second leading cause of dementia. It is important that it be distinguished from Alzheimer's disease because the prognosis and treatments are different. In order to study the effectiveness of clinical tests for

vascular dementia, Gold *et al.* (1997) examined 113 brains of dementia patients post mortem. (Gold G; Giannakopoulos P; Montes-Paixao Junior C; Herrmann FR; Mulligan R; Michel JP; Bouras C. (1997) Sensitivity and specificity of newly proposed clinical criteria for possible vascular dementia. *Neurology,* 49:690–4.) One of the clinical tests these authors considered was proposed by the National Institute for Neurological Disorders (NINDS, an institute of NIH). Gold *et al.* found that the proportion of patients with VD who were correctly identifed by the NINDS test was .58, while the proportion of patients who did not have VD who were correctly so identified by the NINDS tests was .80. These proportions are usually called the *sensitivity* and *specificity* of the test. Using these results, let us consider an elderly patient who is identified as having VD by the NINDS test, and compute the probability that this person will actually have the disease. Let A be the event that the person has the disease and B the event that the NINDS test is positive. We want $P(A|B)$, and we are given $P(B|A) = .58$ and $P(B^c|A^c) = .8$. To apply Bayes' Theorem we need $P(A)$. Let us take this probability to be $P(A) = .03$ (which seems a reasonable value based on Hebert and Brayne, 1995; (Hebert R; Brayne C, Epidemiology of vascular dementia, *Neuroepidemiology* 14:240–57.)). We then also have $P(A^c) = .97$ and, in addition, $P(B|A^c) = 1 - P(B^c|A^c) = .2$. Plugging these numbers into the formula gives us

$$P(A|B) = \frac{(.58)(.03)}{(.58)(.03) + (.2)(.97)} = .082$$

or, approximately, 1/12. Thus, based on the Gold *et al.* study, because VD is a relatively rare disease, without additional evidence, even when the NINDS test is positive it remains unlikely that the patient has VD. □

As in Example 3.1.2, this form of Bayes' Theorem requires probabilities $P(B|A)$, $P(B|A^c)$ and $P(A)$ which must come from some background information. All applications of Bayes' Theorem are analogous in needing background information as inputs in order to get the desired conditional probability as output.

To generalize Bayes' Theorem from the simplest case we need the law of total probability, which gives a formula for $P(B)$ in terms of a decomposition of $\mathcal{S}$: Given mutually exclusive events $A_1, A_2, \ldots, A_n$ that are exhaustive in

the sense that $\mathcal{S} = A_1 \cup A_2 \cup \cdots \cup A_n$, we have

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

with the sets $B \cap A_i$ being mutually exclusive. We then have

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i). \end{aligned}$$

From this we obtain a more general form of the theorem.

Bayes' Theorem Suppose $A_1, A_2, \ldots, A_n$ are mutually exclusive with $P(A_i) > 0$, for all i, and $A_1 \cup A_2 \cup \cdots \cup A_n = \mathcal{S}$. If $P(B) > 0$ then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)}.$$

Example 3.1.3 Decoding of saccade direction from SEF spike counts Bayes' Theorem is frequently used to study the ability of the relatively small networks of neurons to identify a stimulus or determine a behavior. As an example, Olson *et al.* (2000, *J. Neurophysiol.*) reported results from a study of supplementary eye field neurons during a delayed-saccade task. In this study, which we described previously in Chapter 1, there were four possible saccade directions: up, right, down, and left. For each direction, and for each neuron, spike counts in fixed pre-saccade time intervals were recorded across multiple trials. From a combination of data analysis, and assumptions, the probability distribution of various spike counts could be determined for each of the four directions. If we consider a single neuron, we may then let B be the event that a particular spike count occurs, and the events A_1, A_2, A_3, and A_4 be the saccade directions up, right, down, left. Assuming the four directions are equally likely, from the probabilities $P(B|A_k)$ together with Bayes' Theorem, we may determine from the spike count B the probability that the saccade will be in each of the four directions. In Bayesian decoding, the signals from many neurons are combined, and the direction A_k having the largest probability $P(B|A_k)$ is considered the "predicted" direction. In

unpublished work, Kass and Ventura found that from 55 neurons (many of which contributed relatively little information), Bayesian decoding was able to predict the correct direction more than 95% of the time. □

3.2 Random Variables

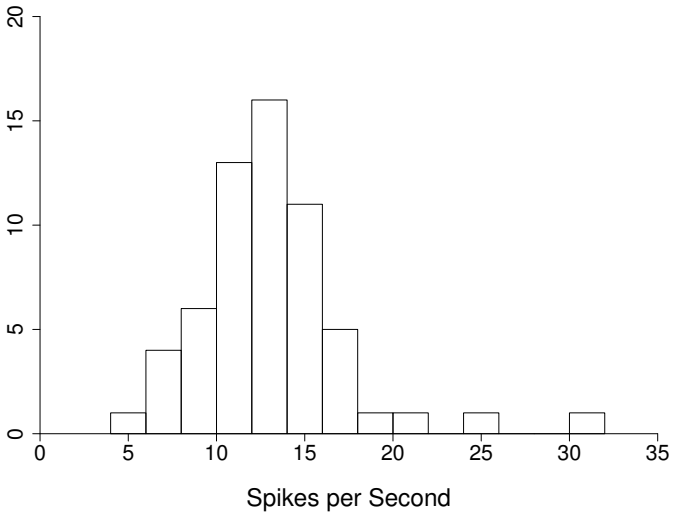


Figure 3.2: *Histogram of spike counts from a motor cortical neuron. The histogram displays 60 spike counts from a particular neuron recorded in primary motor cortex across 60 repititions of the practiced condition.*

So far we have discussed the basic rules of probability, which apply to sets representing uncertain events. A far more encompassing framework is obtained when we consider quantities measured from those events. For example, the number of times a neuron fires during a particular task may be observed, yielding a spike count. When the behavior is repeated across many trials, the spike counts will vary.

Example 3.2.1 Spike counts from a motor cortical neuron Matsuzaka *et al.* (2006) (Matsuzaka, Y., Picard, D., and Strick, P. (2006) Skill

representation in the primary motor cortex after long-term practice, *J. Neurophys.*, 97: 1819–1832.) studied cortical correlates of practicing a movement repeatedly by comparing the firing of neurons in primary motor cortex during two sequential button-pressing tasks: one in which the sequence was highly practiced, and the other in which the sequence was determined at random. Figure 3.2 displays spike counts from a single neuron across 60 repititions of the practiced condition. The histogram displays substantial variation among the counts. □

Probability may be used to describe variation among quantitative measurements, as seen in Figure 3.2. To extend the formalism we introduce mathematical objects called *random variables*, which assign to each outcome (e.g., neuronal spiking behavior on a particular trial) a number (the spike count). In this section we develop some of the basic attributes and properties of random variables, and we will use probability distributions to describe the way they vary.

At the outset it is important to emphasize the abstraction involved in using a random variable to describe observed data. Strictly speaking, random variables and their probability distributions live in the theoretical world of mathematics, while data live in the real world of observations. When we speak of the distribution of some data, as in the histogram in Figure 3.2, we are talking about observed variation. On the other hand, if we use a probability distribution (such as a normal distribution or a Poisson distribution, both discussed in Chapter 5), to describe some data, we are imposing a mathematical structure. To be useful, such a structure must capture dominant features that drive scientific inferences, and a fundamental part of data analytic expertise involves appreciation of the ways inaccuracies in probabilistic description may or may not lead to misleading inferences. We discuss assessments of probability distributions, and consequences of incorrect assumptions, throughout the book. In this chapter we concentrate on essential mathematical definitions and results.

3.2.1 Random variables take on values determined by events.

Let us start by considering the Hardy-Weinberg distribution, which is fundamental to population genetics because it describes the relative frequency of genotypes in equilibrium. In the simplest case, we assume each parent contributes one allele to an offspring. If we write this in the form

(allele from parent 1, allele from parent 2) $\rightarrow$ *offspring pair of alleles*

then the possibilities are denoted by

$$\begin{gathered}(A, A) \rightarrow AA \\ (A, a) \rightarrow Aa \\ (a, A) \rightarrow Aa \\ (a, a) \rightarrow aa.\end{gathered}$$

For instance, according to the classic Mendelian story, a garden pea might be wrinkled if aa but smooth otherwise. We assume that, in the population, the probability of allele A being inherited by the offspring is $P(A)$, and we will also write this number as p, so that $p = P(A)$. That is, if we were to select some offspring at random from the population, the probability that the offspring would have allele A is p. We also assume the inherited alleles are independent, meaning that the allele inherited from parent 1 does not affect the probability of inheriting allele A from parent 2. As we have already seen, the assumption of independence implies that the probabilities may be multiplied:

$$\begin{gathered}P(AA) = p^2, \\ P(Aa) = p(1-p) + (1-p)p = 2p(1-p), \\ P(aa) = (1-p)^2.\end{gathered}$$

Now take X to be the number of A alleles. Then

$$P(X = 2) = p^2,$$
$$P(X = 1) = 2p(1 - p),$$
$$P(X = 0) = (1 - p)^2.$$

In this situation X is a random variable and it has a *Binomial distribution.* More generally, given a sample space $\mathcal{S}$, a *random variable* is a mapping that assigns to every element of $\mathcal{S}$ a real number. That is, if $\omega \in \mathcal{S}$ (see page 53) then $X(\omega) = x$ is the value of the random variable X when ω occurs. In the simple genetics context above, $\mathcal{S} = \{AA, Aa, aa\}$ and $X(AA) = 2, X(Aa) = 1, X(aa) = 0$.

In Chapter 1 we discussed the distinction between continuous and discrete data. We may similarly distinguish continuous and discrete random variables: a random variable is continuous if it can take on all values in some interval (A, B), where it is possible that either $A = -\infty$ or $B = \infty$ (or both). The mathematical distinctions between discrete and continuous distributions are that (i) discrete distributions taken on only certain specific values (such as non-negative integers) that can be separated from each other and (ii) wherever summation signs appear for discrete distributions, integrals replace them for continuous distributions.

3.2.2 Distributions of random variables are defined using cumulative distribution functions and probability density functions, from which theoretical means and variances may be computed.

There are several definitions we need, which will apply to other probability distributions besides the Binomial. In the Hardy-Weinberg example, the values $P(X = 0)$, $P(X = 1)$, and $P(X = 2)$ form the *probability mass function.* For convenience, as indicated in Section 3.2.3, we generally instead call the probability mass function a *probability density function (pdf).* We would typically write $P(X = x)$, with x taking the values $0, 1, 2$. The shorthand $p(x) = P(X = x)$ is also often used. The function $F(x) = P(X \leq x)$ is called the *cumulative distribution function (cdf).* Thus, in the Hardy-Weinberg example we have $F(1) = P(X \leq 1) = P(X = 0) + P(X = 1)$. From the pdf

we can obtain the cdf, and vice-versa.

Illustration: Litter sizes of mice As a simple non-Binomial example, useful for pedagogical purposes, suppose that 50 female mice were maintained in a facility, that each gave birth to a litter, and that the litter sizes may be summarized in the following table:

size	3	4	5	6	7	8
count	3	7	12	14	10	4

Let us consider choosing a mouse at random from among the 50 that gave birth, and let X be the litter size for that mouse. By dividing each count in the table above by 50 we get the following table for the probability distribution of X:

x	3	4	5	6	7	8
$p(x)$	.06	.14	.24	.28	.20	.08

Thus, $p(3) = 3/50 = .06$ signifies a probability that a randomly drawn mouse that will have litter size 3. □

Notice that a plot of the counts (against x) would be a histogram of the 50 litter sizes. Aside from the divisor of 50 used in getting each probability from the corresponding count, a plot of $p(x)$ against x would look the same as the histogram of the counts; this would, instead, be a plot of the *relative frequencies*.[5] More generally, a plot of a probability distribution looks something like a histogram, except that the total amount of probability must equal 1.

One way to understand *any* specification of probabilities $p(x)$ is to consider them to represent relative frequencies among a population of individuals. However, in many cases the idea of a random drawing from a population is an abstraction, and may be rather unrealistic. This is actually an important philosophical point that has been argued about a great deal, but we will not go into it.

> In experimental settings, it is quite artificial to imagine that the repeated measurements (trials) of an experiment are being drawn

[5] In this context terminology is inconsistent: "frequency" can mean either "count" or "relative frequency."

at random from some population of such things. Similarly, when there is a single unique event, such as the outcome of a football game, or the flip of a fair coin, we can be comfortable speaking about the probability of the outcome without any need for a population. In the case of the coin, if we let $X = 1$ if it comes up heads and $X = 0$ if it comes up tails, we would take $p(1) = P(X = 1) = .5$ and $p(0) = P(X = 0) = .5$. We could, if we wished, imagine some very large population of fair coins, just like the one we are going to flip, among which, if flipped in just the same way, half would come up heads and half would come up tails. But we don't really need this imaginary device: thinking only about one single coin it remains easy enough to understand the idea that it is "fair" precisely when $p(1) = .5$ and $p(0) = .5$. That is, the notion that it is equally likely to be heads and tails does not require further elaboration. If we wished to have an operational meaning to "fair" we could take it to mean that we are willing to accept a fair bet, i.e., one in which we would win the same amount if heads as we would lose if tails.

For our purposes, what is important is that relative frequencies sometimes define probabilities, and more generally provide a useful analogy for thinking about probability.

Now, let us go on to the concepts of mean and variance. For the 50 litter sizes in the table on page 67 we would compute the mean as

$$\text{mean} = \frac{3(3) + 7(4) + 12(5) + 14(6) + 10(7) + 4(8)}{50} = 5.66.$$

Alternatively, we could write

$$\text{mean} = 3(\frac{3}{50}) + 4(\frac{7}{50}) + 5(\frac{12}{50}) + 6(\frac{14}{50}) + 7(\frac{10}{50}) + 8(\frac{4}{50}) = 5.66$$

which, from the table on page 67 is the same as

$$\text{mean} = 3 \cdot p(3) + 4 \cdot p(4) + 5 \cdot p(5) + 6 \cdot p(6) + 7 \cdot p(7) + 8 \cdot p(8) = 5.66.$$

This latter form may be interpreted as the litter size we would expect to see ("on average") for a randomly drawn mouse, and it is an instance of

the general expression for the *mean* or *expected value* or *expectation* of the random variable X:

$$\mu_X = E(X) = \sum_x x \cdot p(x). \tag{3.2}$$

Correspondingly, the *variance* of X is

$$\sigma_X^2 = V(X) = \sum_x (x - \mu_X)^2 \cdot p(x)$$

and the *standard deviation* is $\sigma_X = \sqrt{\sigma_X^2}$. The subscript X is often dropped, leaving simply μ and σ. The standard deviation summarizes the magnitude of the deviations from the mean; roughly speaking, it may be considered an average amount of deviation from the mean. It is thus a measure of the spread, or variability, of the distribution. There are alternative measures (such as $\sum_x |x - \mu| p(x)$), and these are used in special circumstances, but the standard deviation is the easiest to work with mathematically. It is, therefore, the most common measure of spread.

Note that μ_X and σ_X are theoretical quantities defined for *distributions*, and are analogous to the mean and standard deviation defined for data. In fact, if there are n values of x and we plug into (3.2) the special case $p(x) = \frac{1}{n}$ (which states that all n values of x are equally likely) we get back[6] $\mu_X = \bar{x}$. Because data are often called *samples*, the data-based mean and standard deviation are often called the *sample mean* and the *sample standard deviation* to differentiate them from μ_X and σ_X, which are often called the *population mean and standard deviation.* This terminolgy distinguishes samples from "populations," rather than distributions, with the word "sample" connoting a batch of observations randomly selected from some large population. Sometimes there is a measurement process that corresponds to such random selection. However, as we have already mentioned, probability is much more general than the population/sample terminology might lead one to expect; specifically, we do not need to have a well-defined population from which we are randomly sampling in order to speak of a probability distribution. So, at least in principle, we might rather avoid calling μ_X a population mean. On

[6]We also get $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \mu_X)^2}$ which, when we replace μ_X with $\bar{X}$, is not quite the same thing as the *sample standard deviation*; the latter requires a change from n to $n-1$ as the divisor for certain theoretical reasons, including that the sample variance then becomes an *unbiased* estimator of σ_X^2. See Chapter **??**.

the other hand, the "sample" terminology is useful for emphasizing that we are dealing with the observations, as opposed to the theoretical distribution, and it is deeply imbedded in statistical jargon. Similarly, the "population" identifier is frequently used rather than "theoretical." The crucial point is that one must be careful to distinguish between a theoretical distribution and the actual distribution of some sample of data. Many analyses assume that data follow some particular theoretical distribution, and in doing so *hope* that the match between theory and reality is pretty good. We will look at ways of assessing this match in Section 3.3.1.

The following properties are often useful.

Theorem For a discrete random variable X with mean μ_X and standard deviation σ_X we have

$$E(a \cdot X + b) = a \cdot \mu_X + b \tag{3.3}$$

and

$$\sigma_{aX+b} = |a| \cdot \sigma_X \tag{3.4}$$

Proof: These are easy to prove. For instance,

$$\begin{aligned} E(aX + b) &= \sum_x (ax + b)p(x) \\ &= a(\sum_x xp(x)) + b\sum_x p(x) \\ &= aE(X) + b. \end{aligned}$$

The proof of the variance formula is similar. □

3.2.3 Continuous random variables are similar to discrete random variables.

Suppose X is a continuous random variable on an interval (A, B), with $A = -\infty$ and $B = \infty$ both being possible. The *probability density function* (pdf) of X will be written as $f(x)$ where now

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

and

$$\int_A^B f(x)dx = 1.$$

Note that in this continuous case there is no distinction between $P(a \leq X)$ and $P(a < X)$ (we have $P(X = a) = 0$). We may think of $f(x)$ as the probability per unit of x; $f(x)dx$ is the probability that X will lie in an infinitesimal interval about x: $f(x)dx = P(x \leq X \leq x + dx)$. In some contexts there are various random variables being considered and we write the pdf of X as $f_X(x)$.

A technical point is that when either $A > -\infty$ or $B < \infty$, by convention, the pdf $f(x)$ is extended to $(-\infty, \infty)$ by setting $f(x) = 0$ outside (A, B). Typically, when we say that X is a continuous random variable on an interval (A, B) we will mean that $f(x) > 0$ on (A, B). We next give several examples of continuous distributions.

Illustration: Uniform distribution Perhaps the simplest example is the *uniform distribution*. For instance, if the time of day at which births occurred followed a uniform distribution, then the probability of a birth in any given 30 minute period would be the same as that for any other 30 minute period throughout the day. In this case the pdf $f(x)$ would be constant over the interval from 0 to 24 hours. Because it must integrate to 1, we must have $f(x) = 1/24$ and the probability of a birth in any given 30 minute interval starting at a hours is $\int_a^{a+.5} f(x)dx = 1/48$. When a random variable X has a uniform distribution on a finite interval (A, B) we write this as $X \sim U(A, B)$ and the pdf is $f(x) = \frac{1}{B-A}$. □

Figure 3.3 displays pdfs for four common distributions. For the two in the top panels, exponential and gamma distributions, X may take on all positive values, i.e., values in $(0, \infty)$. The lower left panel shows a beta distribution, which is confined to the interval (0,1). A normal distribution, which ranges over the whole real line, is shown in the bottom right panel. We discuss the exponential and normal distributions briefly below and return to them, and to the beta and gamma disributions in Chapter 5.

Illustration: Normal distribution The normal distribution (also called the Gaussian distribution) is the most important distribution in Statistics. The reason for this, however, has little to do with its ability to describe data:

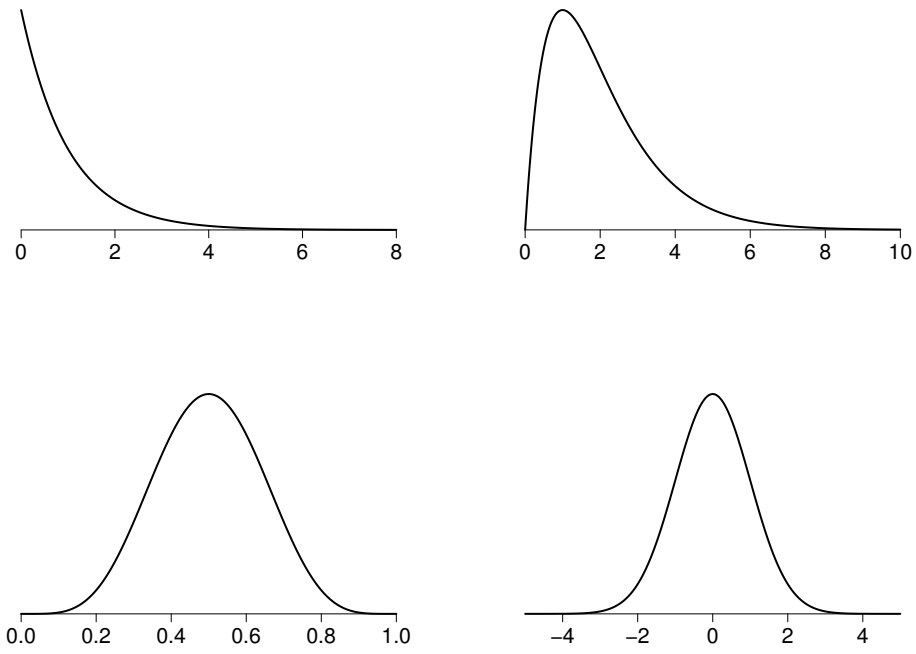


Figure 3.3: *Plots of pdfs for four continuous distributions. Top left: Exponential. Top right: Gamma, with shape parameter 2. Bottom left: Beta. Bottom right: Normal. See Chapter 5 for the explanation of the latter three distributions.*

Example **??** presents one of the few examples we know in which the data really appear normally distributed to a high degree of accuracy; it is rare for a batch of data *not* to be detectably non-normal. Instead, in statistical inference, the normal distribution is used to describe the variability in quantities *derived from* the data as functions of a sample mean. As we discuss in Chapter 6, according to the Central Limit Theorem, sample means are approximately normally distributed and in Chapter 9, we will also see that functions of a sample mean are approximately normally distributed.

The normal distribution is characterized by two parameters: the mean and the standard deviation (or, equivalently, its square, the variance). When a random variable X is normally distributed we write $X \sim N(\mu, \sigma^2)$. Both

in most software and in most applications, one speaks of the parameters μ and σ rather than μ and σ^2. The pdf for the normal distribution with mean μ and standard deviation σ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2).$$

This pdf can be hard to use for analytic calculations because its integral can not be obtained in explicit form. Thus, probabilities for normal distributions are almost always obtained numerically. Because of its shape the normal pdf is often called "the bell-shaped curve." We exemplify this in the next example. □

Example 1.1.2 (continued) We previously noted that the SQUID detectors in MEG are extremely sensitive, and that there is nontrivial background noise that is detected in the absence of any brain signal. Figure 3.4 shows a histogram of the signal at one detector during a short period with nothing in the machine. The noise histogram is very well approximated by a normal pdf. □

In fact, the general bell shape of the distribution is not unique to the normal distribution. On the other hand, the normal is very special among bell-shaped distributions. The most important aspect of its being very special is its role in the Central Limit Theorem, which we'll come back to in Chapter 6. We also describe additional important properties of normal distributions in Chapter 5.

The *cumulative distribution function (cdf)*, or simply *distribution function*, is written again as $F(x)$ and is defined as in the discrete case: $F(x) = P(X \leq x)$. If $A = -\infty$ and $B = \infty$ this becomes

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

If A is a number, i.e., $-\infty < A$, then $F(x) = 0$ when $x < A$ and

$$F(x) = \int_{A}^{x} f(t)dt,$$

while if B is a number ($B < \infty$) then $F(x) = 1$ when $x > B$.

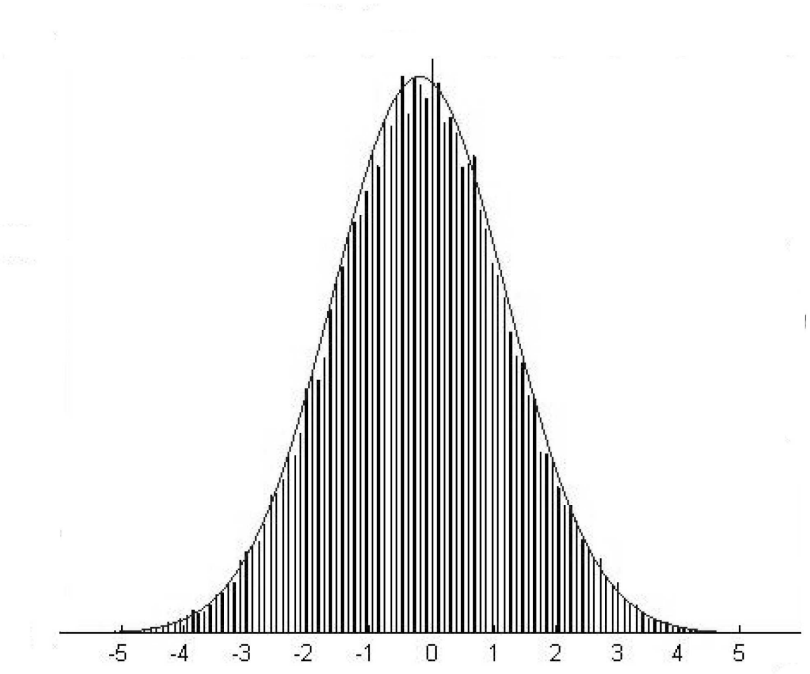


Figure 3.4: *A histogram of MEG noise at a SQUID sensor, overlaid with a normal density function (the "bell-shaped curve").*

Theorem Suppose $f(x)$ is a continuous pdf that is positive on (A, B). Then $F(x)$ is a non-decreasing function and it is strictly increasing ($F'(x) > 0$) on (A, B). In addition we have $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$.

Proof: By differentation (the Fundamental Theorem of Calculus) we have $F'(x) = f(x)$, which implies $F'(x) \geq 0$ and, by assumption, $F'(x) > 0$ on (A, B). Because $f(x)$ integrates to 1 on the interval (A, B), when $A = -\infty$ and $B = \infty$ we must have $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$. When A is a number, from the integral form of $F(x)$, $F(A) = 0$ and $F(x) \to 0$ as $x \to A$. Similarly, we get $F(B) = 1$ and then $F(x) \to 1$ as $x \to B$. □

In the continuous case, the *expected value* of X is

$$\mu_X = E(X) = \int_A^B x f(x) dx$$

and the *standard deviation* of X is $\sigma_X = \sqrt{V(X)}$ where

$$V(X) = \int_A^B (x - \mu_X)^2 f(x) dx$$

is the *variance* of X. Note that in each of these formulas we have simply replaced sums by integrals in the analagous definitions for discrete random variables. Note, too, that *pdf* and *cdf* values for certain continuous distributions may be computed with statistical software.[7] We again have

$$\mu_{a \cdot X + b} = a \cdot \mu_X + b \tag{3.5}$$

$$\sigma_{a \cdot X + b} = |a| \cdot \sigma_X. \tag{3.6}$$

These formulas are just as easy to prove as (3.3) and (3.4).

The *quantiles* or *percentiles* are often used in working with continuous distributions: for p a number between 0 and 1 (such as .25), the pth quantile or 100p-th percentile (e.g., the .25 quantile or the 25th percentile) of a distribution having cdf $F(x)$ is the value η such that $p = F(\eta)$. Thus, we may write $\eta = F^{-1}(p)$, where F^{-1} is the inverse cdf.

[7] The definitions of expectation and variance assume that the integrals are finite; there are, in fact, some important probability distributions that do not have expectations or variances because the integrals are infinite.

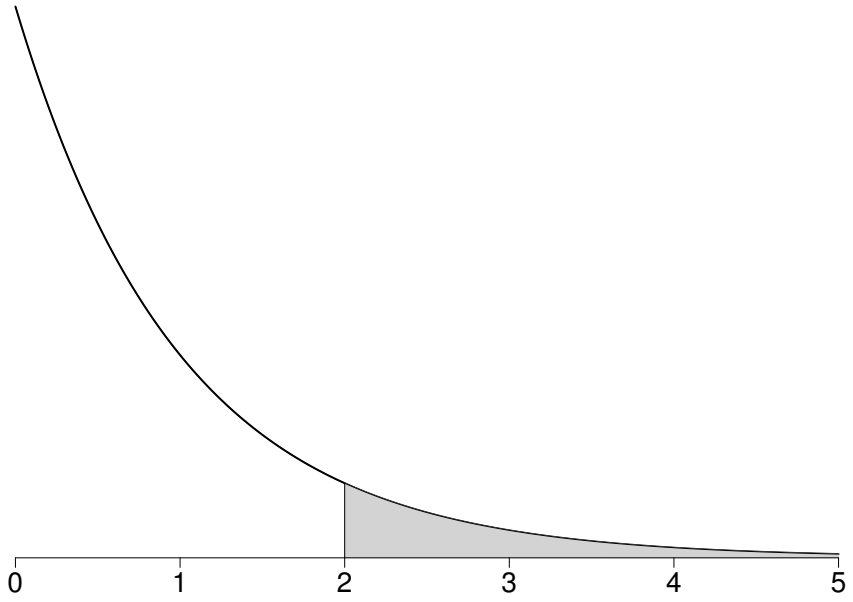


Figure 3.5: *The pdf of a random variable X having an exponential distribution with $\lambda = 1$. The shaded area under the pdf gives $P(X > 2)$.*

Illustration: Exponential distribution Let us illustrate these ideas in the case of the exponential distribution, which is special because it is easy to handle and also because of its importance in applications. We provide an interesting application in Example 3.2.2

A random variable X is said to have an exponential distribution with parameter λ when its pdf is

$$f(x) = \lambda e^{-\lambda x} \tag{3.7}$$

for $x > 0$, and is 0 for $x \leq 0$. We will then say that X has an $Exp(\lambda)$ distribution and we will write $X \sim Exp(\lambda)$ where the sign $\sim$ means "is distributed as". The pdf of X when $X \sim Exp(1)$ is shown in Figure 3.5. Also illustrated in that figure is computation of probabilities as areas under

the pdf. For example,

$$P(X > 2) = \int_2^\infty f(x)dx$$

which means we compute the area under the curve to the right of $x = 2$. For the exponential distribution this value is easy to compute using calculus. The cdf of an exponential distribution is

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t} dt \\ &= -e^{-\lambda t}\Big|_0^x \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Thus, when $X \sim Exp(\lambda)$ we also have

$$P(X > x) = e^{-\lambda x} \tag{3.8}$$

and if $\lambda = 1$

$$P(X > 2) = 1 - F(2) = e^{-2}.$$

The quantiles are also easily obtained. For example, if $X \sim Exp(\lambda)$ the .95 quantile of X is the value $\eta_{.95}$ such that $P(X \leq \eta_{.95}) = F(\eta_{.95}) = .95$. We have

$$.95 = F(\eta_{.95}) = 1 - e^{-\lambda \eta_{.95}}$$

and solving for $\eta_{.95}$ gives $\eta_{.95} = -\log_e(.05)/\lambda$.

If $X \sim Exp(\lambda)$ then, by similar calculations, we obtain

$$\begin{aligned} E(X) &= 1/\lambda \\ V(X) &= 1/\lambda^2 \\ \sigma_X &= 1/\lambda. \end{aligned}$$

We omit the details. □

If $X_1, X_2, \ldots, X_n$ are independently distributed as $Exp(\lambda)$ then their sum $Y = X_1 + X_2 + \cdots + X_n$ follows a *gamma* distribution with shape parameter n, written $Y \sim G(n, \lambda)$. The exponential is often used to describe event durations, and the gamma then becomes a sum of event durations, as illustrated in the next example.

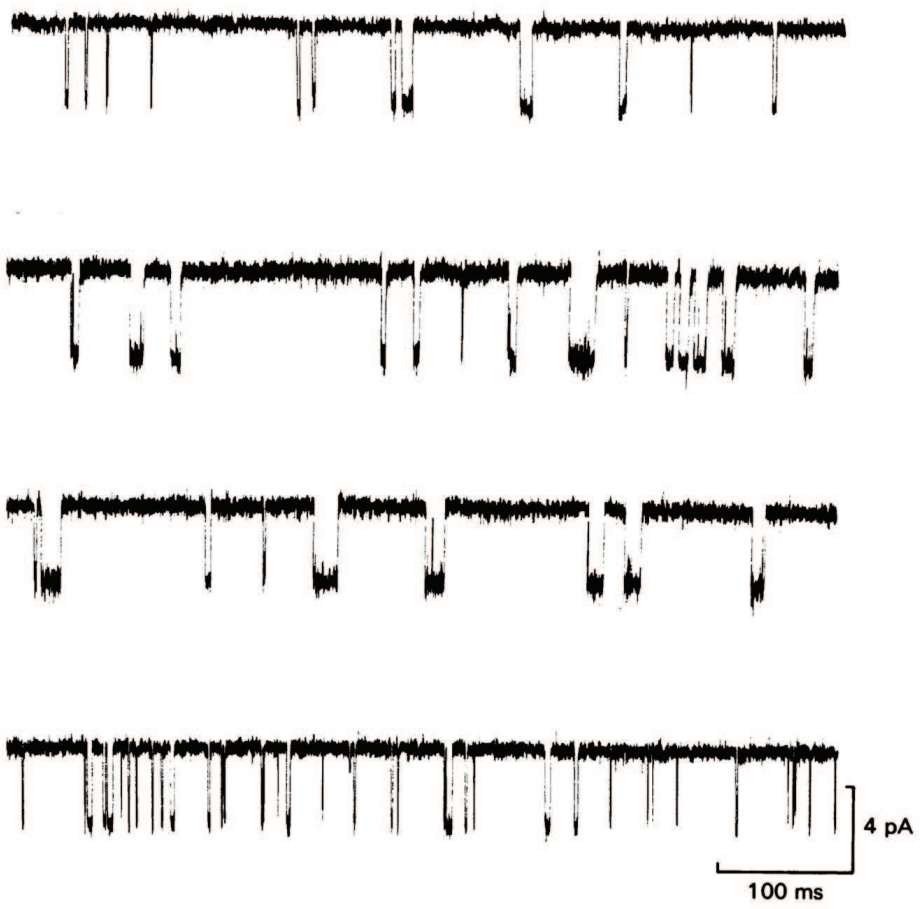


Figure 3.6: *Current recordings from individual ion channels in the presence of acetylcholine-type agonists. The records show the opening (higher current levels) and closing (lower current levels), with the timing of opening and closing being stochastic. Modified from Colquhoun and Sakmann (1985).*

Example 3.2.2 Duration of Ion Channel Activation To investiage the functioning of ion channels, Colquhoun and Sakmann (1985) used patch-clamp methods to record currents from individual ion channels in the presence of various acetylcholine-like agonists. (Colquhoun, D. and Sakmann, B. (1985), Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate, *J. Physiology*, 369: 501–557; see also Colquhoun, D. (2007) Classical Perspective: What have we learned from

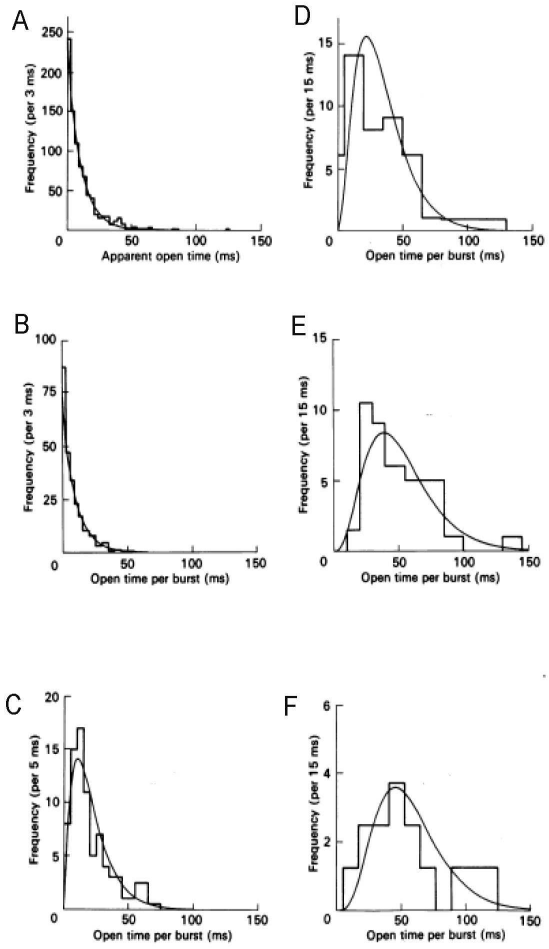


Figure 3.7: *Duration of channel openings. Panel A depicts the distribution of burst durations for a particular agonist. Panel B displays the distribution of bursts for which there was only 1 opening, with an exponential pdf overlaid. This illusrates the good fit of the exponential distribution to the durations of ion channel opening. Panels C displays the distributions of bursts for which there were 2 apparent openings, with a gamma pdf, with shape parameter 2, overlaid. Panel C again indicates good agreement. Panels D-F show similar results, for bursts with 3-5 openings. Modified from Colquhoun and Sakmann (1985).*

single ion channels? *J. Physiology*, 581: 425–427.) A set of their recordings is

shown in Figure 3.6. One of their main objectives was to describe the opening and closing of the channels in detail, and to infer mechanistic actions from the results. Colquhoun and Sakmann found that channels open in sets of activation "bursts" in which the channel may open, then and shut again and open again in rapid succession, and this may be repeated, with small gaps of elapsed time during which the ion channel is closed. A burst may thus have 1 or several openings. As displayed in Figure 3.7, Colquhoun and Sakmann examined separately the bursts having a single opening, then bursts with 2 openings, then bursts 3, 4, and 5 openings. Panel B of Figure 3.7 indicates that, for bursts with a single opening, the opening durations follow closely an exponential distribution. As discussed in Chapter 5, in the case of bursts with 2 openings, if each of the two opening durations were exponentially distributed, and the two were independent, then their sum—the total opening duration—would be gamma with shape parameter $\alpha = 2$. Panel C of Figure 3.7 indicates the good agreement of the gamma with the data. The remaining panels show similar results for the other cases. □

The formulas and concepts that apply to random variables are usually stated with the notation of integrals rather than sums. This is partly because it is cumbersome to repeat everything for both continuous and discrete random variables, when the results are in essence the same. In fact, there is an elegant theory of integration[8] that, among other things, treats continuous and discrete random variables together, with summations becoming special cases of integrals. Throughout our presentation we will, for the most part, discuss the continuous case with the understanding that the analogous results follow for discrete random variables. For example, we will freely use the terminology *pdf* for both continuous and discrete random variables, where for the latter it will refer to a probability mass function.

For many purposes we don't actually need formulas such as those derived for the exponential distribution. Most statistical software contains routines to generate random observations artificially[9] from standard distributions, such as those presented below, and the software will typically also provide

[8]Lebesgue integration is a standard topic in mathematical analysis; see for example, Billingsley, P. (1995) *Probability and Measure*, Third Edition.

[9]The numbers generated by the computer are really *pseudo*-random numbers because they are created by algorithms that are actually deterministic, so that in very long sequences they repeat and their non-random nature becomes apparent. However, good

pdf values, probabilities, and quantiles. Indeed, as we note below, random variables having essentially any continuous distribution may be generated on a computer from a program that generates $U(0,1)$ random variables. In showing this we will have to use the cdf, which is given next.

Illustration: Uniform distribution (continued) If a continous random variable X has cdf $F(x) = x$ on the interval $(0,1)$ we may differentiate to get the $U(0,1)$ pdf $f(x) = 1$. On the other hand, if $X \sim U(0,1)$ we integrate $f(x) = 1$ to get

$$F(x) = \int_0^x 1 \cdot dx = x.$$

In other words, X has a $U(0,1)$ distribution if and only if its cdf is $F(x) = x$ on the interval $(0,1)$. □

3.2.4 The hazard function of a random variable X at x is its conditional probability density, given that $X \geq x$.

Another useful characterization of a probability distribution arises in some specialized contexts, including the analysis of spike train data. The *hazard function* of a continuous random variable X is

$$\lambda(x) = \frac{f(x)}{1 - F(x)}.$$

Note that if $P(B) > 0$ then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Applying this to a continuous random variable X we have

$$P(X \in (x, x+h)|X > x) = \frac{F(x+h) - F(x)}{1 - F(x)}.$$

computer simulation programs use good random number generators, which take an extremely long time to repeat, so this is rarely a practical concern.

Therefore,

$$\lim_{h\to 0} \frac{P(X \in (x, x+h)|X > x)}{h} = \frac{f(x)}{1-F(x)} = \lambda(x)$$

which provides the fundamental interpretation of $\lambda(x)dx$, that it is the probability $X \in (x, x+dt)$ given $X > x$. For example, if X is the elapsed time that an ion channel is open then $\lambda(t)dt$ becomes the probability the ion channel will close in the interval $(t, t+dt)$, given that it has remained open up to time t. Similarly, if X is the elapsed time since a neuron last fired an action potential then $\lambda(t)dt$ becomes the probability the neuron will fire in the interval $(t, t+dt)$, given that it has not yet fired again before elapsed time t. As we will see in Chapter **??**, the hazard function is fundamental to spike train analysis.

The "hazard" terminology comes from lifetime analysis, where the random variable X is the lifetime of a unit (a lightbulb; a person) in units of time t and $\lambda(t)dt$ is the probability of failure (death) in the interval $(t, t+dt)$ given that failure has not yet occurred.

3.2.5 The distribution of a function of a random variable is found by the change of variables formula.

There are many situations in which we begin with a random variable X that has a particular distribution and we want, in addition, to obtain the distribution of another random variable $Y = g(X)$ for some function $g(x)$. This arises in the context of data transformations (discussed in Chapter 2) and it is also important in various theoretical derivations. In the simplest cases there is no need for any special formula.

Illustration: Hardy-Weinberg distribution Let us return to the genetics context where X is the number of A alleles and $P(X = 2) = p^2$, $P(X = 1) = p(1-p)$, $P(X = 0) = (1-p)^2$. Suppose $g(x) = 10^x$. Then we have $P(Y = 100) = p^2$, $P(Y = 10) = p(1-p)$, $P(Y = 1) = (1-p)^2$. It would be easy to calculate the mean and variance of Y from these probabilities. □

When X has a continuous distribution we may obtain the pdf $f_Y(y)$ of

$Y = g(X)$ using the change-of-variables formula from calculus—which follows from the chain rule.

Theorem: Pdf of a Function of a Random Variable Suppose X is a continuous random variable having pdf $f_X(x)$ for which $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise; suppose further that $g(x)$ is a differentiable function and $g'(x) \neq 0$ for $x \in (A, B)$. Then the random variable $Y = g(X)$ has pdf given by

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$$

wherever $y = g(x)$ for some x, and $f_Y(y) = 0$ elsewhere.

Proof: Let us consider $x \in (A, B)$. Because $g'(x) \neq 0$, $g'(x)$ is either always positive, in which case $g(x)$ is monotonically increasing, or always negative in which case $g(x)$ is monotonically decreasing. Let us assume $g'(x) > 0$. Because $g(x)$ is monotonically increasing we then have $x \leq c \iff g(x) \leq g(c)$. We will obtain the pdf $f_y(y)$ by differentiating the cdf $F_Y(y)$, using $f_y(y) = F'_Y(y)$. Suppose $y = g(x)$ for some x. Then

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

where the second equality used $x \leq c \iff g(x) \leq g(c)$. Now, by the chain rule, differentiation gives

$$f_y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

The case in which $g'(x) < 0$ requires a small modification of the argument above (which we leave to the attentive reader). □

Another result that will be used later in the book provides a way of reducing the distribution of X to a uniform distribution.

Theorem: The Probability Integral Transform, Part 1 Suppose X is a continuous random variable having pdf $f_X(x)$ for which $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise and let $Y = F_X(X)$. Then Y has a $U(0, 1)$ distribution.

Proof: First, let us note that $F_X(x)$ is strictly increasing on (A, B). It therefore has a well-defined, strictly increasing inverse function $F_X^{-1}(y)$ satisfying $F_X^{-1}(y) = x$ whenever $F_X(x) = y$. Furthermore, $x \leq c \Longleftrightarrow F_X^{-1}(x) \leq F_X^{-1}(c)$ and $F_X(F_X^{-1}(y)) = y$. We must show that $P(Y \leq y) = y$. We have

$$\begin{aligned} P(Y \leq y) = P(F_X(X) \leq y) &= P(X \leq F_X^{-1} y) \\ &= F_X(F_X^{-1}(y) \\ &= y. \end{aligned}$$

□

Theorem: The Probability Integral Transform, Part 2 Suppose X is a continuous random variable having pdf $f_X(x)$ for which $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise. If $U \sim U(0, 1)$ then $F_X^{-1}(U)$ has a distribution with cdf F_X.

Proof: The proof involves manipulations similar to those of part 1. □

This result gives a general method of generating a random variable that has a distribution with a given distribution function $F(x)$: we generate a $U(0, 1)$ random variables U and apply the transformation $F^{-1}(U)$.

3.3 The Empirical Cumulative Distribution Function

One way to check the accuracy with which a probability distribution fits the data is to overlay a pdf on a histogram, as in Figures 3.4 and 3.7. (In Chapter 7 we discuss how to choose the parameter values for the pdf, e.g., the λ in an exponential.) In this section we consider another pair of graphical

techniques, called Q-Q and P-P plots, which can be somewhat more sensitive than plotting the pdf.

The difficulty in examining the pdf is that its values cover a large range: it can be hard to judge deviations from a curving trend, especially when some of the values are close to zero. An alternative is to straighten things out so that a perfect fit is represented by a straight line. The two ways to accomplish this, employed with Q-Q and P-P plots, are based on the cdf. We begin by defining the data-based counterpart of the theoretical cdf.

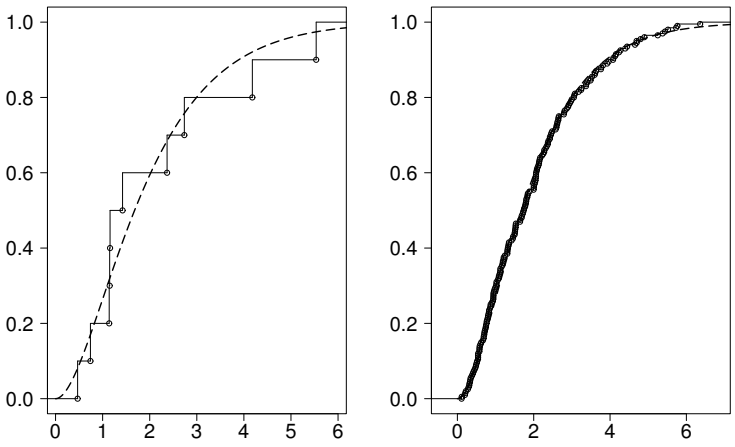


Figure 3.8: *Convergence of the empirical cdf to the theoretical cdf. The left panel shows the empirical cdf for a random sample of size 10 from a Gamma*$(2, 1)$ *distribution, together with the gamma cdf (dashed line). The right panel shows the same for a random sample of size 200. In the right panel the emprical cdf is quite close to the theoretical gamma cdf.*

Let $X_1, \ldots, X_n$ be independent random variables all having the same distribution function $F(x)$. The *empirical cumulative distribution function*, written $\hat{F}_n(x)$, is the cdf for the discrete probability distribution that puts mass $1/n$ on each value $X_1, \ldots, X_n$, i.e.,

$$\hat{F}_n(x) = \frac{\text{number of indices } i \text{ for which } X_i \leq x}{n}.$$

That is, $\hat{F}_n(x)$ provides the proportion of the random variables, out of n, that

are less than or equal to x. When n is large, we might expect this proportion to be close to the theoretical probability that each random variable is less than or equal to x, i.e., we might expect $\hat{F}_n(x)$ to be close to $F(x)$. We will see in Chapter 6 that this is necessarily so, for sufficiently large n. Figure 3.8 illustrates this in the case of a Gamma(2,1) distribution, for samples of size $n = 10$ and $n = 200$. The rough agreement of $\hat{F}_n(x)$ and $F(x)$ suggests ways to check whether a sample of data resembles a random sample drawn from a particular probability distribution.

3.3.1 Q-Q and P-P plots provide graphical checks for gross departures from a distributional form.

Suppose we wish to compare a cdf $\tilde{F}(x)$ with another, similar cdf $F(x)$. If $\tilde{F}(x) \approx F(x)$, we could simply plot $v = F(x)$ against $u = \tilde{F}(x)$ over the range of values of x, and judge the accuracy of the approximation by the deviation of this plot from the line $v = u$. In other words, we could plot probabilities against probabilities. This is the idea behind the P-P plot (P-P for Probability-Probability) except that in examining data it is performed with the empirical cdf $\hat{F}_n(x)$ replacing $F(x)$. Specifically, to examine the fit of $\tilde{F}(x)$ to some data, we pick suitable values of x and for these we plot $\tilde{F}(x)$ against $\hat{F}_n(x)$.

Which values of x should we pick? The usual set of x values chosen are the quantiles, or a slight variation on these. On page 75 we defined the quantiles of a probability distribution. The data quantiles (or observed quantiles, or sample quantiles) are analogous. To compute them from a sample of observations $x_1, x_2, \ldots, x_n$ we first put the data in ascending order according to the size of each observation: we write $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ is the smallest value, $x_{(2)}$ is the second-smallest, and $x_{(n)}$ is the largest. We then say that $x_{(r)}$ is the $\frac{r}{n+1}$ *quantile*, or the $100r/(n+1)$ *percentile*. The denominator is $n+1$ because the n data values correspond to n equally-spaced points in the interval (0,1). Thus, if we were to use the sample quantiles in the P-P plot, to check the fit of $\tilde{F}(x)$ to the data, we would plot $\tilde{F}(\frac{r}{n+1})$ against $\frac{r}{n+1}$. Instead, a slightly different version of the plot is commonly used: $\tilde{F}(\frac{r-.5}{n})$ is plotted against $\frac{r-.5}{n}$.

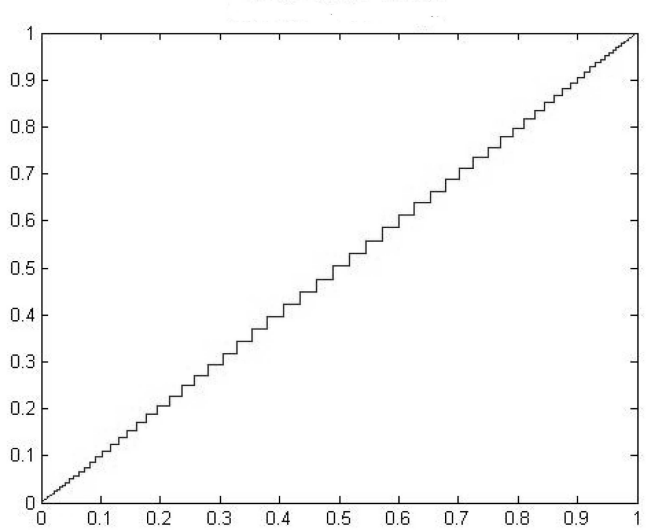


Figure 3.9: *A P-P plot of the MEG noise data from Figure 3.4. The straightness of the plot indicates excellent agreement with the Normal distribution.*

Example 1.1.2 (continued) A P-P plot of the data shown in Figure 3.4 is given in Figure 3.9, where we have used a normal distribution as our theoretical $\tilde{F}(x)$. The plot follows extremely closely the line $y = x$. □

Note that $F(x)$ and $\hat{F}_n(x)$ take values on (0,1). For the Q-Q plot (Q-Q for Quantile-Quantile) we consider examining $\tilde{F}^{-1}(y) \approx F^{-1}(y)$ by plotting $v = F^{-1}(y)$ against $u = \tilde{F}^{-1}(y)$, again seeing deviations between the two cdfs as deviations from the line $v = u$. Again, in order to examine the fit of $\tilde{F}(x)$ to data we replace $F(x)$ with the empirical cdf $\hat{F}_n(x)$. As values of x for the plot we again use a version of the quantiles: to check the fit of $\tilde{F}(x)$ to the data we compute

$$\eta_{(r)} = \tilde{F}^{-1}(\frac{r}{n+1})$$

for $r = 1, \ldots, n$ and plot the ordered data against these values. That is, we plot the points $(x_{(1)}, \eta_{(1)}), \ldots, (x_{(n)}, \eta_{(n)})$.

Figure 3.10 displays three Q-Q plots, where the theoretical quantiles are based on the normal distribution. Thus, we would make these plots in order to check whether the data could reasonably be described by a normal distri-

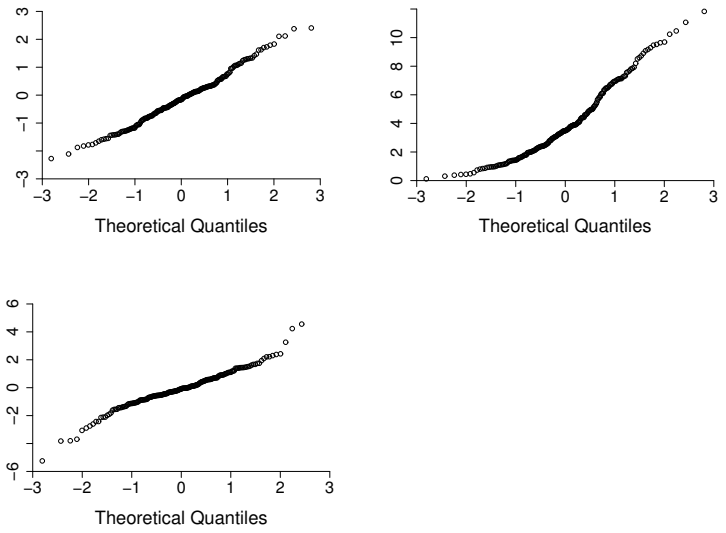


Figure 3.10: *Q-Q plots for 200 randomly-drawn observations from a* $N(0, 1)$ *(top left), a* χ^2_4 *(top right), and a* t_3 *(bottom left). In each case the theoretical quantiles come from a Normal distribution.*

bution. The three data sets were generated on the computer from three very different probability distributions. The first comes from a normal distribution, the second from a gamma distribution, which is skewed toward high values, and the third from a t-distribution, which has heavy tails in both directions. The first plot shows a adherence to a linear relationship between the observed and theoretical quantiles. The second, for skewed data, shows strong upward curvature: the points on the far right-hand side of the plot correspond to data values that are farther from the middle than would be expected if normal (the observed quantiles for those points are too large for the theoretical quantiles—the data should have been pulled in toward the middle—so the points appear too high) and those on the far left-hand side are too close to the middle (the observed quantiles are again too large—the data should now be pushed away from the middle—and the points are again too high). The third plot, for symmetrical but heavy-tailed data, has a striking S-shaped tendency (the observed quantiles are too large on the far right-hand side and too small on the left; on both extremes, to look more normal, the data should be pushed back toward the middle).

Although such plots are very useful for revealing serious departures from normality, small wiggles in these plots are very common even for computer-generated normal data. Thus, strong nonlinearities are what we look for, and even these are sometimes a bit subtle. Figure 3.11 shows Q-Q plots based on 30 randomly drawn observations from a $N(0, 1)$ distribution; the 8 plots show 8 replications of this random number generation and plotting. The departures from linearity indicate that randomly drawn observations fluctuate; they do not conform perfectly to what is theoretically "expected." Or, put differently, what we *should expect* is that small samples of truly normal data will be somewhat erratic and less regular than the theoretical curve based on infinitely much data. This basic lesson applies to all probability distributions, and applies to many situations other than examination of Q-Q plots. It is something we must keep in mind when using our personal perceptions to judge random quantities.[10]

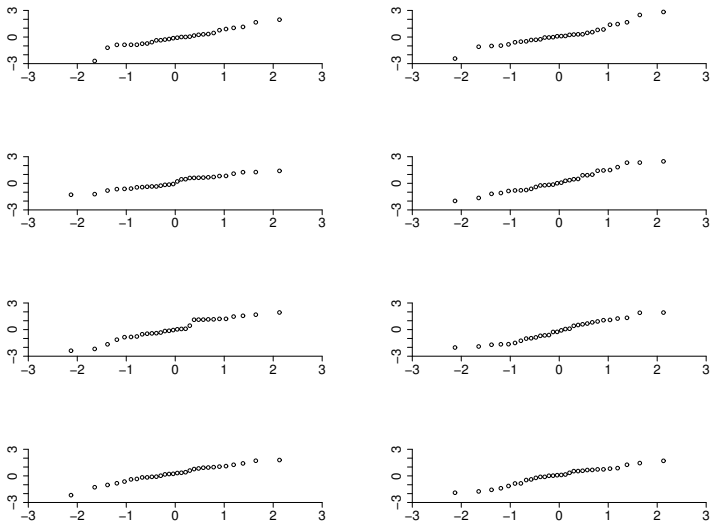


Figure 3.11: *Normal Q-Q plots for 30 randomly-drawn observations from a* $N(0, 1)$*, repeated 8 times. The plots are more or less linear, but display mild departures (wiggles, etc.) from linearity.*

[10]The cognitive psychology of perception of randomness has been studied quite extensively. See, for instance, Gilovich, T., Vallone, R., and Tversky, A. (1985) The hot hand in basketball: on the misperception of random sequences, *Cognitive Psychology, 17*, 295-314.

3.3.2 Q-Q and P-P plots may be used to judge the effectiveness of transformations

In Chapter 2 we discussed transformations of data, especially to improve symmetry. There we used histograms as displays. An alternative is to use Q-Q or P-P plots.

Example 2.1.1 (continued) Figure 3.12 provides Q-Q plots for the human eye saccade data shown in Chapter 2. The logarithm makes the distribution more symmetrical, and the reciprocal does an even better job. An outlier becomes apparent in the latter plot.

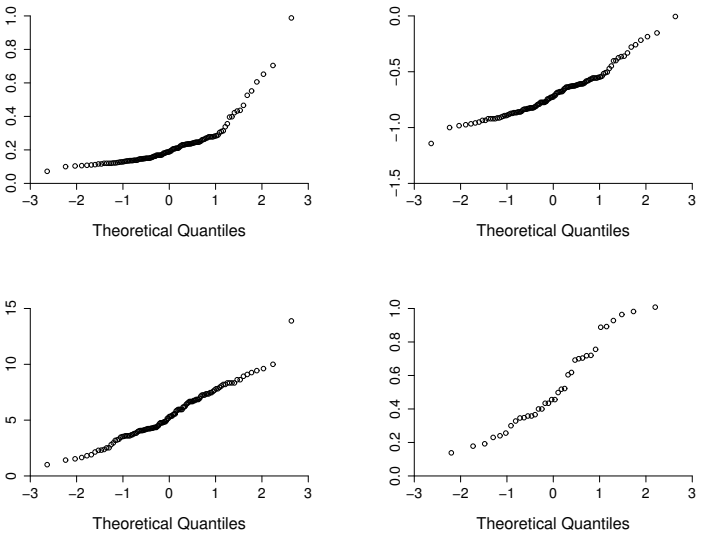


Figure 3.12: *Q-Q plots: The upper left is the Q-Q plot for the data from a particular patient, shown in Chapter 1, from the study by Behrmann* et al. *(2000). The upper right displays a Q-Q plot of the same data following a log transformation, and and the lower left a Q-Q plot following a reciprocal transformation. The plot for the log-transformed data is straighter than that for the raw data; the plot for the reciprocal-transformed data is straighter still. The lower right displays a Q-Q plot of data from a different patient, which exhibits an S shape.*

On the bottom right of Figure 3.12 is a Q-Q plot from a different patient,

for whom much of the data were unusable. We have included this because the plot has the classic S-shape, indicating a "heavy-tailed" distribution. Power transformations do not fix this problem. If one wishes to analyze data of this sort it is important to use a statistical procedure either specifically designed for such situations or having well-understood behavior in the presence of heavy-tailed distributions. We discuss nonparametric procedures in Chapters 9 and 11.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 4

Random Vectors

In most experimental settings data are collected simultaneously on many variables, and the statistical modeling problem is to describe their *joint* variation, meaning their tendency to vary together. The starting point involves m-dimensional *random vectors* (where m is some positive integer), which are the natural multivariate extension of random variables. The fundamental concepts of distribution, expectation, and variance discussed in Chapter 3 extend fairly easily to m dimensions. We review the essentials in this chapter.

4.1 Two or More Random Variables

Let us begin with a motivating example.

Example 4.1.1 Tetrode spike sorting One relatively reliable method of identifying extracellular action potentials *in vivo* is to use a "tetrode." As pictured in panel A of Figure 4.1, a tetrode is a set of four electrodes that

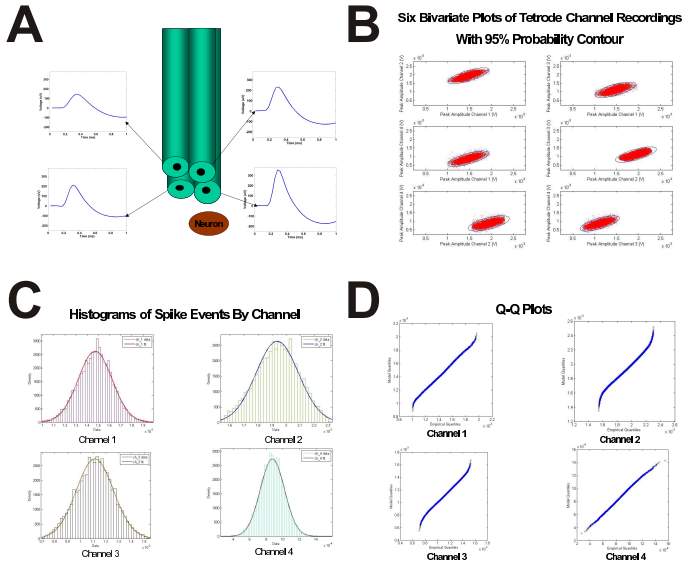


Figure 4.1: *Spike sorting from a tetrode recording. Panel A is a diagram of a tetrode recording device, which is a set of four electrodes; also shown there are signals being recorded from a particular neuron that is sitting near the tetrode. Panel B displays the six pairs of plots of event amplitudes. For instance, the top left plot in panel B shows the event amplitudes for channel 1 (x-axis) and channel 2 (y-axis). Also overlaid (in blue) on the data in panel B are 95% probability contours found from a suitable bivariate normal distribution. Panel C displays histograms for the event amplitudes on each channel, together with fitted normal pdfs, and panel D provides the corresponding normal Q-Q plots.*

sit near a neuron and record slightly different voltage readings in response to an action potential. The use of all four recordings allows more accurate discrimination of a particular neuronal signal from the many others that affect each of the lectrodes. Action potentials corresponding to a particular neuron are identified from a complex voltage recording by first "thresholding" the recording, i.e., identifying all events that have voltages above the threshold. Each thresholded event is a four-dimensional vector (x_1, x_2, x_3, x_4), with x_i being the voltage amplitude (in millivolts) recorded at the ith electrode or

"channel." Panels B-D display data from a rat hippocampal CA1 neuron. Because there are six pairs of the four tetrodes (channel 1 and channel 2, channel 1 and channel 3, etc.) six bivariate plots are shown in panel B. The univariate distributions are displayed in panel C and Q-Q plots are in panel D. We return to this figure in Chapter 5. □

Particularly for $m > 2$ it becomes hard to visualize multidimensional variation. Some form of one and two-dimensional visualization is usually the best we can do, as illustrated in Figure 4.1 of Example 4.1.1. As we contemplate theoretical representations, the possibilities for interactions among many variables quickly become quite complicated. Typically, simplifications are introduced and an important challenge is to assess the magnitude of any distortions they might entail. We content ourselves here with a discussion of multivariate means and variances, beginning with the bivariate case.

4.1.1 The variation of several random variables is described by their joint distribution.

If X and Y are random variables, their *joint distribution* may be found from their *joint pdf*, which we write as $f(x, y)$:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x, y) dx dy.$$

In the discrete case the integrals are replaced by sums. Each individual or *marginal* pdf is obtained from the joint pdf by integration (or, in the discrete case, summation): if $f_X(x)$ is the pdf of X then

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Illustration: Spike Count Pairs Suppose we observe spike counts for two neurons recorded simultaneously over an interval of 100 milliseconds. Let X and Y be the random variables representing the two spike counts. We may specify the joint distribution by writing down its pdf. Suppose it is

given by the following table:

$$\begin{array}{cc|ccc} & 2 & .03 & .07 & .10 \\ Y & 1 & .06 & .16 & .08 \\ & 0 & .30 & .15 & .05 \\ \hline & & 0 & 1 & 2 \\ & & & X & \end{array}$$

This means that the probability that the first neuron will spike once and the second neuron will spike twice, during the observation interval, is $P(X = 1, Y = 2) = .07$. We may compute from this table all of the marginal probabilities. For example, we have the following marginal probabilities: $P(X = 1) = .07 + .16 + .15 = .38$ and $P(Y = 2) = .03 + .07 + .10 = .2$. □

The example above explains some terminology. When we compute $P(Y = 2)$ we are finding a probability that would naturally be put in the *margin* of the table; thus, it is a marginal probability.

More generally, if $X_1, X_2, \ldots, X_n$ are continuous random variables their joint distribution may be found from their joint pdf $f(x_1, x_2, \ldots, x_n)$:

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_n \leq X_n \leq b_n) = \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

and the marginal pdf of the ith random variable X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

where all the variables other than x_i are integrated out. Once again, the formulas for discrete random variables are analagous.

Let us introduce a general notation. Sometimes we will write $X = (X_1, X_2, \ldots, X_n)$, so that X becomes a *random vector* with pdf (really, a joint pdf for its components) $f_X(x) = f_{(X_1, X_2, \ldots, X_n)}(x_1, x_2, \ldots, x_n)$. When we must distinguish row vectors from column vectors we will usually want X to be an $n \times 1$ column vector, so we would instead write $X = (X_1, X_2, \ldots, X_n)^T$, where the superscript T denotes the transpose of a matrix.

A very useful and important fact concerning two or more random variables is that their expectation is "linear" in the sense that the expectation of a linear combination of them is the corresponding linear combination of their expectations.

Theorem: Linearity of Expectation For random variables X_1 and X_2 we have

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2).$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ we have

$$E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i). \tag{4.1}$$

Proof: Consider the case of two random variables and assume X_1 and X_2 are continuous. Let $f_1(x_1), f_2(x_2)$, and $f_{12}(x_1, x_2)$ be the marginal and joint pdfs of X_1 and X_2, and assume these random variables take values in the respective intervals (A_1, B_1) and (A_2, B_2) (which could be infinite). We have

$$\begin{aligned}
E(aX_1 + bX_2) &= \int_{A_2}^{B_2} \int_{A_1}^{B_1} (ax_1 + bx_2) f_{12}(x_1, x_2) dx_1 dx_2 \\
&= a \int_{A_1}^{B_1} x_1 f_{12}(x_1, x_2) dx_1 dx_2 + b \int_{A_1}^{B_1} x_2 f_{12}(x_1, x_2) dx_1 dx_2 \\
&= a \int_{A_1}^{B_1} x_1 \int_{A_2}^{B_2} f_{12}(x_1, x_2) dx_2 dx_1 + b \int_{A_2}^{B_2} x_2 \int_{A_1}^{B_1} f_{12}(x_1, x_2) dx_1 dx_2 \\
&= a \int_{A_1}^{B_1} x_1 f_1(x_1) dx_1 + b \int_{A_2}^{B_2} x_2 f_2(x_2) dx_2 \\
&= aE(X_1) + bE(X_2).
\end{aligned}$$

The proof in the discrete case would replace the integrals by sums, and the proof in the general case of n variables follows the same steps. □

4.1.2 Random variables are independent when their joint pdf is the product of their marginal pdfs.

We previously said that two events A and B are independent if $P(A \cap B) = P(A)P(B)$, and we used this in the context of random variables that identify dichotomous events. There, for example, if the probability of a pea being wrinkled were p, and two peas with independently-sorting alleles were observed, the probability that both of the 2 peas would be wrinkled was p^2. Generally, we say that two random variables X_1 and X_2 are *independent* if

$$P(a \leq X_1 \leq b \text{ and } c \leq X_2 \leq d) = P(a \leq X_1 \leq b)P(c \leq X_2 \leq d) \qquad (4.2)$$

for all choices of a, b, c, d. It follows that when X and Y are independent we also have

$$f(x, y) = f_X(x)f_Y(y). \qquad (4.3)$$

Indeed, when X and Y are random variables with pdf $f(x, y)$, they are independent if and only if Equation (4.3) holds.

Illustration: Spike Count Pairs (continued) We return once again to the joint distribution of spike counts for two neurons, given by the table on page 96. Are X and Y independent?

The marginal pdf for X is $f_X(0) = .39$, $f_X(1) = .38$, $f_X(2) = .23$ and the marginal pdf for Y is $f_Y(0) = .50$, $f_Y(1) = .30$, $f_Y(2) = .20$. We thus obtain $f_X(0)f_Y(0) = .195 \neq .30 = f(0, 0)$, which immediately shows that X and Y are not independent. □

We may generalize the definition of independence to multiple random variables: we say that $X_1, X_2, \ldots, X_n$ are independent random variables if their joint pdf $f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n)$ is equal to the product of their marginal pdfs,

$$f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

In the previous subsection we showed that the expectation of a sum is always the sum of the expectations. In general, it is not true that the variance

of a sum of random variables is the sum of their variances, but this *is* true under independence.

Theorem: Variance of a Sum of Independent Random Variables For independent random variables X_1 and X_2 we have

$$V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2).$$

More generally, for independent random variables $X_1, X_2, \ldots, X_n$ we have

$$V(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} a_i^2V(X_i). \tag{4.4}$$

Proof: The proof is similar to that of the theorem on linearity of expectations, except that the factorization of the joint pdf, due to independence, must be used. □

As a simple example of failure of (4.4) when X_1 and X_2 are not independent, let $X_2 = -X_1$. Then $X_1 + X_2 = 0$ and $V(X_1 + X_2) = 0$.

4.2 Bivariate Dependence

In Section 4.1.2 we said that random variables $X_1, X_2, \ldots, X_n$ are independent if their joint pdf $f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n)$ is equal to the product of their marginal pdfs. We now consider the possibility that $X_1, X_2, \ldots, X_n$ are *not* independent and develop some simple ways to quantify their dependence. In the case of two random variables the most common way to measure dependence is through their correlation, which is discussed in Section 4.2.1. We provide some additional interpretation of the correlation, first in the context of the bivariate normal distribution in Section **??**, and then also using conditional expectation in Section 4.2.4 after we discuss conditional densities in Section 4.2.3. We then turn to the case of arbitrarily many random variables ($X_1, \ldots, X_n$ with $n \geq 2$), providing results in Section 4.3 that will be useful later on, and we discuss general multivariate normal distributions in Chapter 5.

4.2.1 The linear dependence of two random variables may be quantified by their correlation.

When we consider X and Y simultaneously, we may characterize numerically their joint variation, meaning their tendency to be large or small together, This is most commonly done via the *covariance* of X and Y which, for continuous random variables, is

$$\begin{aligned} Cov(X,Y) &= E(X-\mu_X)(Y-\mu_Y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy \end{aligned}$$

and for discrete random variables the integrals are replaced by sums. The covariance is analagous to the variance of a single random variable. We now generalize Equation (4.4) to the case in which the random variables may not be independent.

Theorem: Variance of a Sum of Random Variables For random variables X_1 and X_2 we have

$$V(aX_1+bX_2) = a^2V(X_1)+b^2V(X_2)+2abCov(X_1,X_2).$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ we have

$$V(\sum_{i=1}^{n} a_iX_i) = \left(\sum_{i=1}^{n} a_i^2V(X_i)\right) + 2\sum_{i<j} a_ia_jCov(X_i,X_j). \qquad (4.5)$$

Proof: The proof follows from the definition by straightforward algebraic manipulations and is omitted. □

The covariance depends on the variability of X and Y individually, as well as their joint variation, and therefore depends on scaling. For instance, as is immediately verified from the definition, $Cov(3X,Y) = 3Cov(X,Y)$. To obtain a measure of joint variation that does not depend on the variance of X and Y, we standardize. The *correlation* of X and Y is

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}.$$

It is easy to check that $Cor(X, Y)$ is invariant to linear rescaling of X and Y and it may be shown that $-1 \leq Cor(X, Y) \leq 1$. The latter is an instance of what is known in mathematical analysis as the Cauchy-Schwartz inequality. When X and Y are independent their covariance, and therefore also their correlation, is zero.

Details: This last fact follows from the definition of covariance: if X and Y are independent we have $f(x, y) = f_X(x)f_Y(y)$ and then

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)f_X(x)(y-\mu_Y)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy\end{aligned}$$

but from the definition of μ_Y

$$\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy = 0$$

(and similarly the integral over x is zero). □

We now illustrate the calculation of correlation in a simple example, introduced earlier.

Illustration: Spike count pairs (continued) We return to the joint distribution of spike counts for two neurons, discussed on page 95, given by the following table:

	2|	.03	.07	.10
Y	1|	.06	.16	.08
	0|	.30	.15	.05
		0	1	2
			X	

We may compute the covariance and correlation of X and Y as follows:

$$\begin{aligned}\mu_X &= 0 + 1\cdot(.38) + 2\cdot(.23) \\ \mu_Y &= 0 + 1\cdot(.30) + 2\cdot(.2) \\ \sigma_X &= \sqrt{.39\cdot(0-\mu_X)^2 + .38\cdot(1-\mu_X)^2 + .23\cdot(2-\mu_X)^2} \\ \sigma_Y &= \sqrt{.5\cdot(0-\mu_Y)^2 + .3\cdot(1-\mu_Y)^2 + .2\cdot(2-\mu_Y)^2}\end{aligned}$$

which gives

$$\begin{aligned} \mu_X &= .84 \\ \mu_Y &= .7 \\ \sigma_X &= .771 \\ \sigma_Y &= .781. \end{aligned}$$

We then get

$$\sum f(x,y)(x-\mu_X)(y-\mu_Y) = .272$$

and

$$\frac{\sum f(x,y)(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} = .452.$$

Thus, the correlation is $\rho \approx .45$. □

The correlation is undoubtedly the most commonly used measure of association between two random variables, but it is rather special. For one thing, $Cor(X,Y) = 0$ does *not* imply that X and Y are independent. Here is a counterexample.

Illustration: Dependent variables with zero correlation. Suppose X is a continuous random variable that has a distribution that is symmetric about 0, meaning that for all x we have $f_X(-x) = f_X(x)$, and let us assume that $E(X^4)$ is a number (i.e., $E(X^4) < \infty$). From symmetry we have

$$\int_{-\infty}^{0} x f_X(x) dx = -\int_{0}^{\infty} x f_X(x) dx$$

so that

$$\begin{aligned} E(X) &= 0 \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{0} x f_X(x) dx + \int_{0}^{\infty} x f_X(x) dx = 0 \end{aligned}$$

and, similarly, $E(X^3) = 0$. Now let $Y = X^2$. Clearly X and Y are not independent: given $X = x$ we know that $Y = x^2$. On the other hand,

$$Cov(X,Y) = E(X(Y-\mu_Y)) = E(X^3) - E(X)\mu_Y = 0.$$

Therefore, $Cor(X, Y) = 0$. □

A more complete intuition about correlation may be found from the next result. Suppose we wish to predict a random variable Y based on another random variable X. That is, suppose we take a function $f(x)$ and apply it to X to get $f(X)$ as our prediction of Y. To evaluate how well $f(X)$ predicts Y we can examine the average size of the error, letting under-prediction ($f(x) < y$) be valued the same as over-prediction ($f(x) > y$). A mathematically simple criterion that accomplishes this is expected squared error $E((Y - f(X))^2)$. Thus, we want to find the form of $f(x)$ that minimizes expected squared error. There is a general solution to this problem, which we give in Section 4.2.4. For now we consider the special case in which $f(x)$ is linear, and find the best linear predictor in the sense of minimizing expected squared error.

Theorem: Linear prediction Suppose X and Y are random variables having variances σ_X^2 and σ_Y^2 (with $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$). In terms of expected squared error, the best linear predictor of Y based on X is $\alpha + \beta X$ where

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \tag{4.6}$$

$$\alpha = \mu_Y - \beta \mu_X \tag{4.7}$$

where $\rho = Cor(X, Y)$. In other words, the values of α and β given by (4.7) and (4.6) minimize $E((Y - \alpha - \beta X)^2)$. With α and β given by (4.7) and (4.6) we also obtain

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2(1 - \rho^2). \tag{4.8}$$

Proof Details: Write

$$Y - \alpha - \beta X = (Y - \mu_Y) - (\alpha - \beta(X - \mu_X)) + \mu_Y - \beta \mu_X$$

then square both sides, take the expected value, and use the fact that for any constants c and d, $E(c(X - \mu_X)) = 0 = E(d(Y - \mu_Y))$. This leaves

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2 + \beta^2 \sigma_X^2 - 2\beta\rho\sigma_X\sigma_Y + (\mu_Y - \alpha - \beta\mu_X)^2. \tag{4.9}$$

Minimizing this quantity by setting

$$0 = \frac{\partial}{\partial \alpha} E\left((Y - \alpha - \beta X)^2\right)$$

and

$$0 = \frac{\partial}{\partial \beta} E\left((Y - \alpha - \beta X)^2\right)$$

and then solving for α and β gives (4.7) and (4.6). Inserting these into (4.9) gives (4.8). □

Let us now interpret these results by considering how well $\alpha + \beta X$ can predict Y. From (4.8) we can make the prediction error (expected squared error) smaller simply by decreasing σ_Y. In order to standardize we may instead consider the ratio $E((Y - \alpha - \beta X)^2)/\sigma_Y^2$. Solving (4.8) for ρ^2 we get

$$\rho^2 = 1 - \frac{E\left((Y - \alpha - \beta X)^2\right)}{\sigma_Y^2}.$$

We now see that the better the linear prediction is, the closer to 1 will ρ^2 be; and, conversely, the prediction error is maximized when $\rho = 0$. Furthermore, we have $\rho > 0$ for positive association, i.e, $\beta > 0$, and $\rho < 0$ for negative association, i.e, $\beta < 0$. In this sense we may say that correlation is a measure of linear association between X and Y. Note that the counterexample on page 102, in which X and Y were perfectly dependent yet had zero correlation, is a case of nonlinear dependence.

4.2.2 A bivariate normal distribution is determined by a pair of means, a pair of standard deviations, and a correlation coefficient.

As you might imagine, to say that two random variables X and Y have a bivariate normal distribution is to imply that each of them has a (univariate) normal distribution and, in addition, they have some covariance. Actually, there is a mathematical subtlety here: the requirement of bivariate normality is much more than that each has a univariate normal distribution. We return

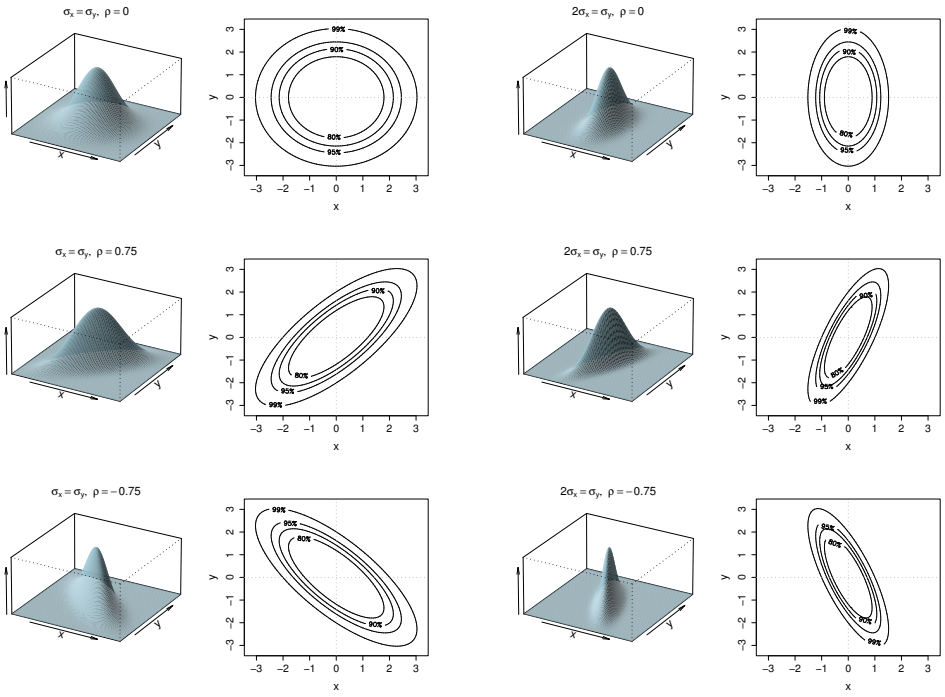


Figure 4.2: *The bivariate normal pdf. Perspective plots and contour plots are shown for various values of* σ_X, σ_Y *and* ρ, *with* $(\mu_X, \mu_Y) = (0, 0)$. *Left column has* $\sigma_X = \sigma_Y$ *and right column has* $2\sigma_X = \sigma_Y$. *First, second, and third rows correspond to* $\rho = 0$, $\rho = .75$, $\rho = -.75$. *Contours enclose probability equal to .8, .9, .95, and .99.*

to this technical point later in this section. For now, we will say that X and Y have a bivariate normal distribution when they have a joint pdf

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2\sqrt{1-\rho^2}}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}$$

where $\rho = Cor(X, Y)$ and we assume that $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$. We may also write this pdf in the form

$$f(x,y) \propto e^{-Q(x,y)}$$

where

$$Q(x,y) = \frac{1}{2\sqrt{1-\rho^2}}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)$$

the proportionality indicating that a constant has been omitted that does not depend on either x or y. Here, $Q(x,y)$ is a quadratic centered at the mean vector and we have inserted the minus sign as a reminder that the density has a maximum rather than a minimum. An implication involves the contours of the pdf. In general, a *contour* of a function $f(x,y)$ is the set of (x,y) points such that $f(x,y) = c$ for some particular number c. When the graph $z = f(x,y)$ is considered, a particular contour represents a set of points for which the height of $f(x,y)$ is the same. The various contours of $f(x,y)$ are found by varying c. The contours of a bivariate normal pdf satisfy $Q(x,y) = c^*$, for some number c^*, and it may be shown that the set of points (x,y) satisfying such a quadratic equation form an ellipse. Therefore, the bivariate normal distribution has elliptical contours. See Figure 4.2. The orientation and narrowness of these elliptical contours are governed by σ_X, σ_Y, and ρ. When $\sigma_X = \sigma_Y$ the axes of the ellipse are on the lines $y = x$ and $y = -x$. As ρ increases toward 1 (or decreases toward -1) the ellipse becomes more tightly concentrated around $y = x$ (or $y = -x$). When $\rho = 0$ the contours become circles. When $\sigma_X \neq \sigma_Y$ the axes rotate to $y = \frac{\sigma_Y}{\sigma_X}x$ and $y = -\frac{\sigma_X}{\sigma_Y}x$.

We have assumed here that $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$, which corresponds to "positive definiteness" of the quadratic, a point we return to in Section 4.3. Sometimes a more general definition of bivariate normality is needed: we say that (X,Y) is bivariate normal if every nonzero linear combination of X and Y has a normal distribution, i.e., for all numbers a and b that are not both zero, $aX + bY$ is normally distributed. This covers additional cases, such as when $\rho = 1$, and we mention it again in Chapter 5 when we discuss the general multivariate normal distribution. An important point is that joint normality is a stronger requirement than normality of the individual components. It is not hard to construct a counterexample in which X and Y are both normally distributed but their joint distribution is not bivariate normal.

Illustration: Marginal normality without joint normality. Here is an example in which X and Y are each normally distributed, but they

do not have a bivariate normal distribution. Let U and V be independent $N(0,1)$ random variables. Let $Y = V$ and for $U < 0, V > 0$ or $U > 0, V < 0$ take $X = -U$. This amounts to taking the probability assigned to (U, V) in the 2nd and 4th quadrants and moving it, respectively, to the 1st and 3rd quadrants. The distribution of (X, Y) is then concentrated in the 1st and 3rd quadrants ((X, Y) has zero probability of being in the 2nd or 4th quadrants), yet X and Y remain distributed as $N(0,1)$. □

While this counterexample is admittedly somewhat contrived, the logical inability to infer joint normality from marginal normality should be kept in mind. In practice, when we examine data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ to see whether their variation appears roughly to follow a bivariate normal distribution, the general result suggests one should plot them together as scatterplot pairs $(x_1, y_1), \ldots, (x_n, y_n)$, rather than simply examining $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ separately. In the multivariate case, however, one must rely on 1-dimensional and 2-dimensional visual representations of data, as in Figure 4.1.

4.2.3 Conditional probabilities involving random variables are obtained from conditional densities.

We previously defined the probability of one event conditionally on another, which we wrote $P(A|B)$, as the ratio $P(A \cap B)/P(B)$, assuming $P(B) > 0$. When we have a pair of random variables X and Y with $f(y) > 0$, the conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \tag{4.10}$$

with $f_{X|Y}(x|y)dx$ being interpreted as the probability $P(x \leq X \leq x+dx | y \leq Y \leq y+dy)$. When X and Y are independent we have

$$f_{X|Y}(x|y) = f_X(x).$$

Illustration: Spike count pairs (continued) We return to the joint distribution of spike counts for two neurons (see page 101). We may calculate the conditional distribution of X given $Y = 0$. We have $f_{X|Y}(0|0) =$

.30/.50 = .60, $f_{X|Y}(1|0) = .15/.50 = .30$, $f_{X|Y}(2|0) = .05/.50 = .10$. Note that these probabilities are different than the marginal probabilities .39, .38, .23. In fact, if $Y = 0$ it becomes more likely that X will also be 0, and less likely that X will be 1 or 2. □

4.2.4 The conditional expectation $E(Y|X = x)$ is called the regression of Y on X.

The conditional expectation of $Y|X$ is

$$E(Y|X = x) = \int y f_{Y|X}(y|x)dy$$

where the integral is taken over the range of y.

Illustration: Spike count pairs (continued) For the joint distribution of spike counts let us compute $E(X|Y = 0)$. We previously found $f_{X|Y}(0|0) = .60$, $f_{X|Y}(1|0) = .30$, $f_{X|Y}(2|0) = .10$. Then

$$E(X|Y = 0) = 0(.6) + 1(.3) + 2(.1) = .5.$$

□

Note that $E(Y|X = x)$ is a function of x, so we might write $M(x) = E(Y|X = x)$ and thus $M(X) = E(Y|X)$ is a random variable. An important set of result concerning $M(X)$ is often called the law of total expectation.

Theorem: Law of total expectation. Suppose X and Y are random variables and Y has finite expectation. Then we have

$$E(E(Y|X)) = E(Y).$$

Proof: From the definition we compute

$$\begin{aligned} E(E(Y|X = x)) &= \int \left(\int y f_{Y|X}(y|x)dy \right) f_X(x)dx \\ &= \int \int y f_{Y|X}(y|x) f_X(x)dxdy \end{aligned}$$

$$\begin{aligned} &= \int\int y f_{(X,Y)}(x,y)dxdy \\ &= \int y f_Y(y)dy = E(Y).\square \end{aligned}$$

There are also the closely-related law of total probability.

Theorem: Law of total probability. Suppose X and Y are random variables. Then we have

$$E(P(Y \leq y|X)) = F_Y(y).$$

Proof: The proof follows a series of steps similar to those in the proof of the law of total expectation. $\square$

We may also define the conditional variance of $Y|X$

$$V(Y|X=x) = \int (y - E(Y|X=x))^2 f_{Y|X}(y|x)dy$$

and then get the following, which has important applications.

Theorem: Law of total variance. Suppose X and Y are random variables and Y has finite variance. Then we have

$$V(Y) = V(E(Y|X)) + E(V(Y|X)).$$

Proof: The proof is similar to that of the law of total expectation. $\square$

In the spike count pairs illustration, we computed the conditional expectation $E(X|Y=y)$ for a single value of y. We could evaluate it for each possible value of y. When we consider $E(X|Y=y)$ as a function of y, this function is called the *regression* of X on Y. Similarly, the function $E(Y|X=x)$ is called the regression of Y on X. To understand this terminology, and the interpretation of the conditional expectation, consider the case in which (X,Y) is bivariate normal.

Example: Regression of son's height on father's height A famous data set, from Pearson and Lee (1903) (Pearson, K. and Lee, A. (1903) On

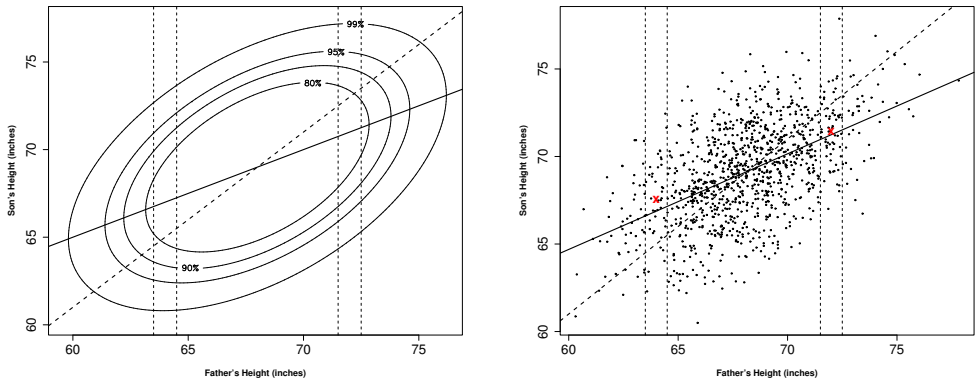


Figure 4.3: *Conditional expectation for bivariate normal data mimicking Pearson and Lee's data on heights of fathers and sons. Left panel shows contours of the bivariate normal distribution based on the means, standard deviations, and correlation in Pearson and Lee's data. The dashed vertical lines indicate the averaging process used in computing the conditional expectation when $X = 64$ or $X = 72$ inches: we average y using the probability $f_Y|X(y|x)$, which is the probability, roughly, in between the dashed vertical lines, integrating across y. In the right panel we generated a sample of 1,078 points (the sample size in Pearson and Lee's data set) from the bivariate normal distribution pictured in the left panel. We then, again, illustrate the averaging process: when we average the values of y within the dashed vertical lines we obtain the two values indicated by the red x. These fall very close to the least-squares regression line (the solid line).*

the laws of inheritance in man, *Biometrika*, 2: 357–462.), has been used frequently as an example of regression. (See Freedman, Pisani, and Purves (2007).) (Freedman, D., Pisani, R., and Purves, R. (2007) *Statistics*, Fourth Edition, W.W. Norton.) Figure 4.3 displays both a bivariate normal pdf and a set of data generated from the bivariate normal pdf—the latter are similar to the data obtained by Pearson and Lee (who did not report the data, but only summaries of them). The left panel of Figure 4.3 shows the theoretical regression line. The right panel shows the regression based on the data, fitted by the method of least-squares, was discussed briefly in Chapter 1 and will be discussed more extensively in Chapter 13. In a large sample like this one, the least-squares regression line (right panel) is close to the theoretical regression

line (left panel). The purpose of showing both is to help clarify the averaging process represented by the conditional expectation $E(Y|X = x)$.

The terminology "regression" is illustrated in Figure 4.3 by the slope of the regression line being less than that of the dashed line. Here, $\sigma_Y = \sigma_X$, because the variation in sons' heights and fathers' heights was about the same, while $(\mu_X, \mu_Y) = (68, 69)$, so that the average height of the sons was about an inch more than the average height among their fathers. The dashed line has slope $\sigma_Y/\sigma_X = 1$ and it goes through the point (μ_X, μ_Y). Thus, the points falling on the dashed line in the left panel, for example, would be those for which a theoretical son's height was exactly 1 inch more than his theoretical father. Similarly, in the plot on the left, any data points falling on the dashed line would correspond to a real son-father pair for which the son was an inch taller than the father. However, if we look at $E(Y|X = 72)$ we see that among these taller fathers, their son's height tends, on average, to be less than the 1 inch more than the father's predicted by the dashed line. In other words, if a father is 3 inches taller than average, his son will likely be *less than* 3 inches taller than average. This is the tendency for the son's height to "regress toward the mean." We understand this as follows: the father is tall partly for genetic reasons and partly by "lucky" environmental factors which pushed him to be taller. The son will inherit the genetic component, but will tend (on average) not to get as big a positive push toward being tall from the environmental factors. The same tendency, now in the reverse direction, is apparent when the father's height is $X = 64$. Regression to the mean is a ubiquitous phenomenon found whenever two variables vary together. □

In general, the regression $E(Y|X = x)$ could be a nonlinear function of x but in Figure 4.3 it is a straight line. This is not an accident: if (X, Y) is bivariate normal, the regression of Y on X is linear with slope $\rho \cdot \sigma_Y/\sigma_X$. Specifically,

$$E(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X). \tag{4.11}$$

We say that Y has a regression on X with regression coefficient $\beta = \rho\frac{\sigma_Y}{\sigma_X}$. This means that when $X = x$, the *average* value of Y is given by (4.11). We should emphasize, again, that we are talking about random variables, which are *theoretical* quantities, as opposed to observed data. In data-analytic contexts the word "regression" almost always refers to least-squares regression, illustrated in the right panel of Figure 4.3.

Compare Equation (4.11) to Equations (4.7) and (4.7). From the latter we have that the best linear predictor of Y based on X is $f(X)$ where

$$f(x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X). \tag{4.12}$$

In general, we may call this the *linear regression* of Y on X. In the case of bivariate normality, the regression of Y on X is equal to the *linear* regression of Y on X, i.e., the regression is linear. We derived (4.12) by minimizing expected squared error, restricting $f(x)$ to be linear. More generally, let us write the regression function as $M(x) = E(Y|X = x)$. It is easy to show that $M(X)$ is the best predictor of Y in the sense of minimizing expected squared error.

Theorem: Prediction The function $f(x)$ that minimizes $E((Y-f(X))^2)$ is the conditional expectation $f(x) = M(x) = E(Y|X = x)$.

> *Proof Details:* Note that $E(Y - M(X)) = E(Y) - E(E(Y|X))$ and by the law of total expectation (above) this is zero. Now write $Y - f(X) = (Y - M(X)) + (M(X) - f(X))$ and expand $E((Y - f(X))^2)$ to get
>
> $$E((Y - f(X))^2) = E((Y - M(X))^2) + 0 + E((M(X) - f(X))^2).$$
>
> The right-hand term $E((M(X) - f(X))^2)$ is always non-negative and it is zero when $f(x)$ is chosen to equal $M(x)$. Therefore the whole expression is minimized when $f(x) = M(x)$. □

4.3 Multivariate Dependence

4.4 The mean of a random vector is a vector and its variance is a matrix.

Now suppose we wish to consider the way m random variables $X_1, \ldots, X_m$ vary together. If we have $\mu_i = E(X_i)$, $\sigma_i^2 = V(X_i)$, and $\rho_{ij} = Cor(X_i, X_j)$,

for $i = 1, \ldots, m$ and $j = 1, \ldots, m$, we may collect the variables in an m-dimensional *random vector* $X = (X_1, \ldots, X_m)^T$, and can likewise collect the means in a vector

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}.$$

Similarly, we can collect the variances and covariances in a matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1m}\sigma_1\sigma_m \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2m}\sigma_1\sigma_m \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{m1}\sigma_1\sigma_m & \rho_{m2}\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{pmatrix}.$$

Note that $\rho_{ij} = \rho_{ji}$ so that Σ is a symmetric matrix (the element in its ith row and jth column is equal to the element in its jth row and ith column, for every i and j). We write the *mean vector* $E(X) = \mu$ and the *variance matrix* $V(Y) = \Sigma$. The latter is sometimes called the *covariance matrix.*

Let w be an m-dimensional vector. By straightforward matrix manipulations we obtain the mean and variance of $w^T X$ as

$$E(w^T X) = w^T \mu \tag{4.13}$$
$$V(w^T X) = w^T \Sigma w. \tag{4.14}$$

Equations (4.13) and (4.14) generalize (4.1) and (4.5).

Let us now recall that a symmetric $m \times m$ matrix A is positive semi-definite if for every m-dimensional vector v we have $v^T A v \geq 0$ and it is positive definite if for every nonzero m-dimensional vector v we have $v^T A v > 0$. From the definition of variance (involving the integral of a non-negative function), every variance is non-negative. Therefore, $V(w^T X) \geq 0$ so that the variance matrix Σ is necessarily positive semi-definite. However, a variance matrix may or may not be positive definite. The non-positive-definite case is the generalization of $\sigma_X = 0$ for a random variable X: in the non-positive-definite case the distribution of the random vector X "lives" on a subspace that has dimensionality less than m. For example, if X and Y are both normally distributed but $Y = X$ then their joint distribution "lives" on a 1-dimensional subspace $y = x$ of the 2-dimensional plane.

An important tool in analyzing a variance matrix is the spectral decomposition. The spectral decomposition of a positive semi-definite matrix A is $A = PDP^T$ where D is a diagonal matrix with diagonal elements $\lambda_i = D_{ii}$ for $i = 1, \ldots, m$, and P is an orthogonal matrix, i.e., $P^T P = I$, where I is the m-dimensional identity matrix. Here, $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A and the columns of P are the corresponding eigenvectors. See Appendix A.

Lemma If Σ is a symmetric positive definite matrix then there is a symmetric positive definite matrix $\Sigma^{\frac{1}{2}}$ such that

$$\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$$

and, furthermore, writing its inverse matrix as $\Sigma^{-\frac{1}{2}} = (\Sigma^{\frac{1}{2}})^{-1}$ we have

$$\Sigma^{-1} = \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}.$$

Proof: This follows from the spectral decomposition. See Appendix A. Writing $\Sigma = PDP^T$, with D being diagonal we simply define $D^{\frac{1}{2}}$ to be the diagonal matrix having elements $(\sqrt{D_{11}}, \ldots, \sqrt{D_{mm}})$ and take $\Sigma^{\frac{1}{2}} = PD^{\frac{1}{2}}P^T$. The stated results are easily checked. □

Theorem Suppose X is a random vector with mean μ and covariance matrix Σ. Define the random vector $Y = \Sigma^{-1/2}(X - \mu)$. Then $E(Y)$ is the zero vector and $V(Y)$ is the m-dimensional identity matrix.

Proof: This follows from the lemma. □

We will use this kind of standardization of a random vector in Chapter 6.

4.4.1 The dependence of two random vectors may be quantified by mutual information.

It often happens that the deviation of one distribution from another must be evaluated. Consider two continuous pdfs $f(x)$ and $g(x)$, both being positive on (A, B). The *Kullback-Leibler (KL) discrepancy* is the quantity

$$D_{KL}(f, g) = E_f\left(\log \frac{f(X)}{g(X)}\right)$$

where E_f refers to the expectation with respect to the pdf $f(x)$. In other words, we have

$$D_{KL}(f,g) = \int_A^B (\log f(x) - \log g(x)) f(x) dx.$$

The KL discrepancy may also be defined, analogously, for discrete distributions. There are other discrepancy functions, but for many purposes the KL discrepancy is quite useful. Part of its motivation comes from its application to normal distributions.

Illustration: Two normal distributions Suppose $f(x)$ and $g(x)$ are the $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ pdfs. Then, from the formula for the normal pdf we have

$$\log \frac{f(x)}{g(x)} = -\frac{(x-\mu_1)^2 - (x-\mu_2)^2}{2\sigma^2} = \frac{2x(\mu_1 - \mu_2) - (\mu_1^2 - \mu_2^2)}{2\sigma^2}$$

and substituting X for x and taking the expectation (using $E_X(X) = \mu_1$), we get

$$D_{KL}(f,g) = \frac{2\mu_1^2 - 2\mu_1\mu_2 - \mu_1^2 + \mu_2^2}{2\sigma^2} = \left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2.$$

That is, $D_{KL}(f,g)$ is simply the squared standardized difference between the means. This is a highly intuitive notion of how far apart these two normal distributions are. □

Example 4.4.1 Auditory-dependent vocal recovery in zebra finches Song learning among zebra finches has been heavily studied. When microlesions are made in the HVC region of an adult finch brain, songs become destabilized but the bird will recover its song within about 1 week. Thompson *et al.* (2007) ablated the output nucleus (LMAN) of the anterior forebrain pathway of zebra finches in order to investigate its role in song recovery. (Thompson, J.A., Wu, W., Bertram, R., and Johnson, F. (2007) Auditory-dependent vocal recovery in adult male zebra finches is facilitated by lesion of a forebrain pathway that includes basal ganglia, *J. Neurosci.*, 27: 12308–12320.) They recorded songs before and after the surgery. The multiple bouts of songs, across 24 hours, were represented as individual notes having a particular spectral composition and duration. The distribution of

these notes post-surgery was then compared to the distribution pre-surgery. In one of their analyses, for instance, the authors compared the distribution of pitch and duration before and after surgery. Their method of comparison was to compute the KL discrepancy. Thompson *et al.* found that deafening following song disruption produced a large KL discrepancy whereas LMAN ablation did not. This indicated that the anterior forebrain pathway is not the neural locus of the learning mechanism that uses auditory feedback to guide song recovery. □

The Kullback-Leibler discrepancy may be used to evaluate the association of two random vectors X and Y. We define the *mutual information* of X and Y as

$$I(X,Y) = D_{KL}(f_{(X,Y)}, f_X f_Y) = E_{(X,Y)} \log \frac{f_{(X,Y)}(X,Y)}{f_X(X) f_Y(Y)}.$$

In other words, the mutual information between X and Y is the Kullback-Leibler discrepancy between their joint distribution and the distribution they would have if they were independent. In this sense, the mutual information measures how far a joint distribution is from independence.

Illustration: Bivariate normal If X and Y are bivariate normal with correlation ρ some calculation following application of the definition of mutual information gives

$$I(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

Thus, when X and Y are independent, $I(X,Y) = 0$ and as they become highly correlated (or negatively correlated) $I(X,Y)$ increases indefinitely. □

Theorem For random variables X and Y that are either discrete or jointly continuous having a positive joint pdf, mutual information satisfies (i) $I(X,Y) = I(Y,X)$, (ii) $I(X,Y) \geq 0$, (iii) $I(X,Y) = 0$ if and only if X and Y are independent, and (iv) for any one-to-one continuous transformations $f(x)$ and $g(y)$, $I(X,Y) = I(f(X), g(Y))$.

Proof: Omitted. See, e.g, Cover and Thomas (1991). (Cover, T.M. and Thomas, J.Y. (1991) *Elements of Information Theory*, New York: Wiley.) □

Property (iv) makes mutual information quite different from correlation. We noted that correlation is a measure of *linear* association and, as we saw in the illustration on page 102, it is possible to have $Cor(X, X^2) = 0$. In contrast, by property (iv), we may consider mutual information to be a measure of more general forms of association, and we always have $I(X, X^2) = \infty$.

A popular interpretation of mutual information involves the notion of *entropy*.

Details: We say that the *entropy* of a continuous random variable X is

$$H(X) = -\int f_X(x) \log f_X(x) dx \tag{4.15}$$

We may also call this the entropy of the distribution of X. The entropy of a distribution was formalized analogously to Gibbs entropy in statistical mechanics by Claude Shannon in his development of communication theory. As in statistical mechanics, the entropy may be considered a measure of disorder in a distribution. For example, the distribution over the interval $(0, 1)$ having maximal entropy is the uniform distribution and, roughly speaking, as a distribution becomes concentrated near a point its entropy decreases. In the discrete case the integral becomes a sum. Suppose there are finitely many possible values of X, say $x_1, \ldots, x_m$, and someone picks one of these values with probabilities given by $f(x_i)$, then we try to guess which value has been picked by asking "yes" or "no" questions (e.g., "Is it greater than x_3?"). In this case the entropy may be interpreted as the minimum average number of questions that must be asked in order to determine the number, the average being taken over replications of the game.

Entropy may be used to characterize many important probability distributions in addition to the uniform. The distribution on the set of integers $0, 1, 2, \ldots, n$ that maximizes entropy subject to having mean μ is the binomial. The distribution on the set of all non-negative integers that maximizes entropy subject to having mean μ is the Poisson. The distribution on the positive real line that maximizes entropy subject to having mean μ is the exponential. The distribution on the positive real line that maximizes

entropy subject to having mean μ and variance σ^2 is the gamma. The distribution on the whole real line that maximizes entropy subject to having mean μ and variance σ^2 is the normal.

Now, if Y is another continuous random variable then the entropy in the conditional distribution of $X|Y = y$ may be written

$$H(X|Y = y) = -\int f_{X|Y}(x|y) \log f_{X|Y}(x|y) dx$$

and if we average this quantity over Y, by taking its expectation with respect to $f_Y(y)$, we get what is called the *conditional entropy* of X given Y:

$$H(X|Y) = \int \left(-\int f_{X|Y}(x|y) \log f_{X|Y}(x|y) dx \right) f_Y(y) dy.$$

Algebraic manipulation then shows that the mutual information may be written

$$I(X, Y) = H(X) - H(X|Y).$$

This says that the mutual information is the average amount (over Y) by which the entropy of X decreases given the additional information $Y = y$. For an extensive discussion of entropy, mutual information, and communication theory see Cover and Thomas (1991) or MacKay (2003). (Mackay, D. (2003) *Information Theory, Inference, and Learning Algorithms,* Cambridge.)

Mutual information has been used extensively to quantify the dependence of a neural response (Y) on a stochastic stimulus (X).

Example 4.4.2 Temporal coding in inferotemporal cortex In an influential paper, Optican and Richmond (1987) reported responses of single neurons in inferotemporal (IT) cortex of monkeys while the subjects were shown various checkerboard-style grating patterns as visual stimuli. (Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J. Neurophysiol.*, 57: 162–178.) Optican and Richmond

computed the mutual information between the 64 randomly-chosen stimuli (the random variable X here taking 64 equally-likely values) and the neural response (Y), represented by firing rates across multiple time bins. They compared this with the mutual information between the stimuli and a single firing rate across a large time interval and concluded that there was considerably more mutual information in the time-varying signal. Put differently, more information about the stimulus was carried by the time-varying signal than by the overall spike count. □

In both Example 4.4.1 and Example 4.4.2 the calculations were based on pdfs that were *estimated* from the data. We discuss probability *density estimation* in Chapter **??**.

4.4.2 Bayes' Theorem for random vectors is analogous to Bayes' Theorem for events.

Now suppose X and Y are random vectors with a joint density $f(x, y)$. Substituting $f(x, y) = f_{Y|X}(y|x)f(x)$ into (4.10), we have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{Y|X}(x, y)}{f_Y(y)} \\ &= \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \end{aligned}$$

If we then substitute $f_Y(y) = \int f(x, y)dx$ in the denominator and rewrite the joint density in the integrand using $f(x, y) = f_{Y|X}(y|x)f_X(x)$ we obtain Bayes' Theorem for random variables.

Theorem For continuous random vectors X and Y we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|x)f_X(x)dx}.$$

The resemblance of this result to Bayes' Theorem for events may be seen by comparing the formula (3.1), identifying X with A and Y with B.

The theorem also holds, as a special case, if X and Y are random variables.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 5

Important Probability Distributions

In Chapter 1 we said that a measurement is determined in part by a "signal" of interest, and in part by unknown factors we may call "noise." Statistical models introduce probability distributions to describe the variation due to noise, and thereby achieve quantitative expressions of knowledge about the signal—a process we will describe more fully in Chapters 7 and **??**. The essential ideas in statistical modeling are simple and very general, allowing modern methods to make flexible—and thus reasonably realistic—assumptions. Despite this wide-ranging generality, the models found in elementary statistics rely heavily on a small handful of probability distributions. For this reason alone, a beginning student must learn about the binomial model for binary observations, the Poisson model for counts, and the normal model for continuous measurements. But there are additional motivations for studying these and several other probability distributions. While it may be tempting to dismiss the ubiquity of these distributions as a historical quirk, a throwback to a pre-computer age in which simplicity was essential, a small number of distributions remain especially important in contemporary

practice. This is partly because many methods of statistical inference, when applied carefully, are remarkably robust in the face of modest deviations from theoretical assumptions. In addition, the simplest distributions often serve as a starting point when building more general and elaborate models. Furthermore, these distributions continue to be important because they arise in theoretical calculations. In this chapter we discuss at greater length some of the probability distributions we mentioned in Chapters 3 and 4. We also introduce several others.

5.1 Bernoulli Random Variables and the Binomial Distribution

5.1.1 Bernoulli random variables take values 0 or 1.

A random variable X that takes the value 1 with probability p and 0 with probability $1-p$ is called a *Bernoulli random variable*. We have already considered, as an example, the inheritance of allele A from one of an offspring's parents. Any dichotomous pair of events whose outcome is uncertain may considered a Bernoulli random variable. For example, patient P.S. in Example 1.2.1 made repeated choices of the "burning" or "non-burning" house. Each such choice could be considered a Bernoulli random variable by coding "burning" as 0 and "non-burning" as 1 (or vice-versa).

5.1.2 The binomial distribution results from a sum of independent and homogeneous Bernoulli random variables.

In the case of the binomial distribution arising from Hardy-Weinberg equilibrium, the two probabilistic assumptions were (i) *independence*, the pairs of alleles sorted independently, and (ii) *homogeneity*, the allele probabilities were the same across individuals. With X being the number of A alleles, these assumptions lead to X having a binomial distribution over the possible

values $0, 1, 2$, with $p = P(A)$. We would write this by saying the distribution of X is $B(2, p)$, or $X \sim B(2, p)$.

The binomial distribution is easy to generalize: instead of counting the number of outcomes of a certain type out of a maximal possible value of 2, we allow the maximal value to be any positive integer n; under assumptions of independence and homogeneity we then would say X has distribution $B(n, p)$, or simply $X \sim B(n, p)$. For example, if we were counting the number of wrinkled peas in a pod of only 3, with each pea having probability p of being wrinkled, then we would let X be the number of wrinkled peas and would say that X has a binomial distribution with $n = 3$ or $X \sim B(3, p)$. By a similar argument to that made in the Hardy-Weinberg example, we obtain $P(X = 3) = p^3$, $P(X = 2) = 3p^2(1 - p)$, $P(X = 1) = 3p(1 - p)^2$, $P(X = 0) = (1 - p)^3$.

Again, similarly, if we count the number of wrinkled peas out of a total of 4, then $X \sim B(4, p)$ and $P(X = 4) = p^4$, $P(X = 3) = 4p^3(1 - p)$, $P(X = 2) = 6p^2(1 - p)^2$, $P(X = 1) = 4p(1 - p)^3$, $P(X = 0) = (1 - p)^4$.

The general formula for n peas, with $X \sim B(n, p)$, is

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \tag{5.1}$$

for $x = 0, 1, 2, \ldots, n$, where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways of choosing x objects from n without regard to ordering. Equation (5.1) is the binomial probability mass function (or pdf). If $X \sim B(n, p)$ then

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1 - p) \\ \sigma_X &= \sqrt{np(1 - p)}. \end{aligned}$$

The individual binary observations, such as the outcomes for the individual peas, are independent Bernoulli random variables all having the same probability of taking the value 1, i.e., the Bernoulli random variables are both independent and *homogeneous*. Such random variables are often called *Bernoulli trials*. The sum of n Bernoulli trials has a $B(n, p)$ distribution. That is, in general, if $Y_1, Y_2, \ldots, Y_n$ are independent Bernoulli random variables and $P(Y_i = 1) = p$ for all i, and we define $X = \sum_{i=1}^{n} Y_i$, then $X \sim B(n, p)$.

Binomial distributions usually arise as the sum of Bernoulli trials. Thus, the binomial distribution is reasonable to assume if the Bernoulli random variables appear to be independent and homogeneous. It is important to consider both assumptions carefully. In a question-and-answer survey setting in which each Bernoulli random variable represents the "yes" or "no" response for an individual subject, if the subjects communicated with each other before answering then independence would be violated. The assumptions of independence and homogeneity are frequently violated when the Bernoulli random variables are observed across time.

Example 1.2.1 (continued) In judging the 14 out of 17 occasions on which P.S. chose the non-burning house by statistical methods we would assume that the set of 17 forced choices were Bernoulli trials. The independence assumption would be violated if P.S. had a tendency, say, to repeat the same response she had just given regardless of her actual perception. The homogeneity assumption would be violated if there were a drift in her response probabilities (e.g., due to fatigue) over the time during which the experiment was carried out. □

In the case of alleles contributed by parents in producing offspring the assumption of independence would be violated if somehow the two parents were coupled at the molecular level, so that the processes of separating the alleles in the two parents were connected; in most studies this seems very unlikely and thus the first assumption is quite reasonable. The second assumption is that there is a single, stable value for the probability of the allele A. This clearly could be violated: for instance, the population might actually be a mixture of two or more types of individuals, each type having a different value of $P(A)$; or, when the population is not in equilibrium due to such things as non-random mating, or genetic drift, we would expect deviations from the binomial prediction of the Hardy-Weingberg model. Indeed, in population genetics, a check on the fit of the Hardy-Weinberg model to a set of data is used as a prelimary test before further analyses are carried out.

Example 5.1.1 Nicotinic acetylcholine receptor and ADHD Attention deficit hyperactivity disorder (ADHD), a major psychiatric disorder among children, and has been the focus of much recent research. There is evidence of heritability of ADHD, and effective medications (such as Ritalin) involve inhibition of dopamine transport. There is also evidence of

involvement of the nicotine system, possibly due to its effects on dopamine receptors. Kent *et al.* (2001, *Psychiatric Gentics*, 11: 37–40) examined genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. At issue was the frequency of an $T \rightarrow C$ exchange in one base in the gene sequence. In order to carry out the standard analysis the authors first examined whether the population appeared to be in equilibrium. If so, the probabilities of the allele combinations TT, CT, CC would be given by Hardy-Weinberg model (which was discussed briefly in Chapter 3). The frequencies for the 136 parents in their study were as follows:

	TT	CT	CC
Number	48	71	17
Frequency	.35	.52	.13
Hardy-Weinberg Probability	.38	.47	.15

In this case, the probabilities determined from the Hardy-Weinberg model (how we obtain these will be discussed in Chapter 7) are close to the observed allele frequencies, and there is no evidence of disequilibrium in the population (we also discuss these details later). Kent *et al.* went on to find no evidence of an association between this genetic polymorphism and the diagnosis of ADHD. □

In some cases the probability p is not stable across repititions. Indeed, sometimes the change in probability is the focus of the experiment, as when learning is being studied.

Example 5.1.2 Learning impairment following NMDA antagonist injection Experiments on learning often record responses of subjects as either correct or incorrect on a sequence of trials during which the subject is given feedback as to whether their response was correct or not. The subjects typically begin with probability of being correct that is much less than 1, perhaps near the guessing value of .5, but after some number of trials they get good at responding and have a high probability of being correct, i.e., near 1. An illustration of this paradigm comes from Smith *et al.* (2005),

who examined data from an experiment in rats by Stefani *et al.* (2003) demonstrating that learning is impaired following an injection of an NMDA antagonist into the frontal lobe. In a first set of trials, the rats learned to discriminate light from dark targets, then, in a second set of trials, which were the trials of interest, they needed to discrimate smooth versus rough textures of targets. In two groups of rats a buffered salt solution with the NMDA antagonist was injected prior to the second set of trials, and in two other groups of rats the buffered salt solution without the antagonist was injected. Figure 5.1 displays the responses across 80 learning trials for set 2. It appears from the plot of the data that the groups of rats without the NMDA antagonist did learn the second task more quickly than the second group of rats, as expected.

The Smith *et al.* analysis was based on the method of maximum likelihood, which we will discuss subsequently. For now, however, we may use the example to consider the possibility of aggregating the responses within groups of rats. Two possible ways to aggregate would be either across rats or across trials, the latter producing blocks of trials (e.g., 10 blocks of 8 trials). In each case, aggregation would produce a number X of correct responses out of a possible number n. We would then be able to plot the value of X across time in order to help examine the differences among the groups. If we were to assume $X \sim B(n, p)$, in each case, what would we be assuming about the trials themselves? If we were to aggregate across rats we would be assuming that the rats' responses were independent, which is reasonable, and that the rats all had the same probability of responding correctly, which is dubious. Making this kind of dubious assumption is often a useful first step, and in fact can be innocuous for certain analyses, but it must be considered critically. After aggregating trials into blocks, the binomial assumption would be valid if the trials were independent and had the same probability of correct response, both of which would be dubious—though again potentially useful if its effects were examined carefully. In situations such as these it would be incumbant upon the investigator to show that aggregation would be unlikely to produce incorrect analytical results. □

Before leaving the binomial distribution, let us briefly examine one further application.

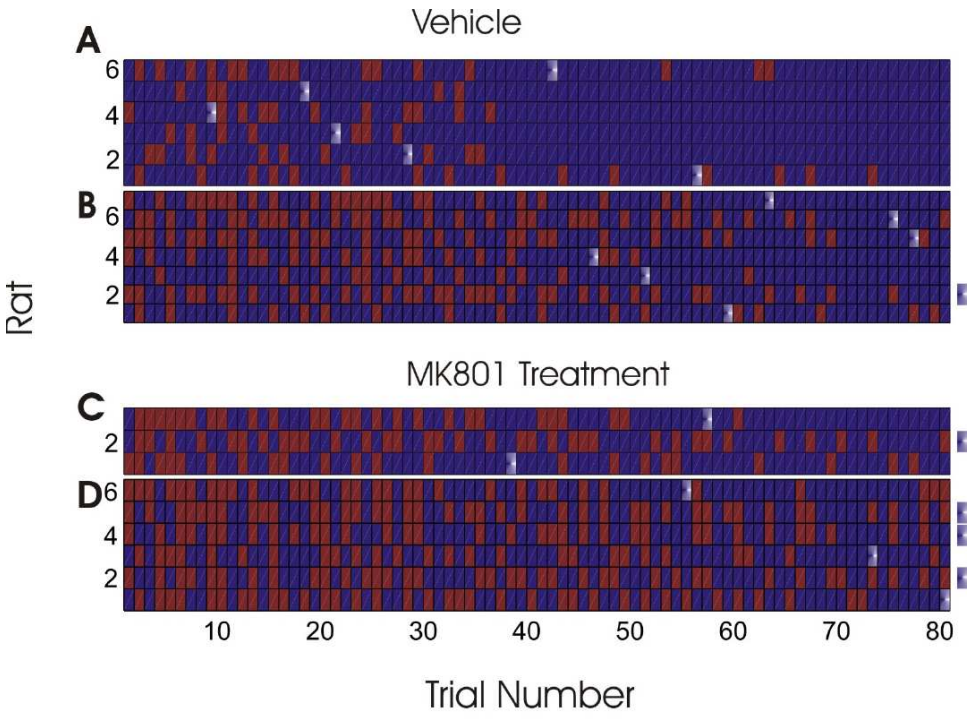


Figure 5.1: *Responses for 13 rats in the placebo group (labeled "Vehicle," in reference to the buffered solution vehicle) and 9 rats in the treatment group ("MK801 Treatment") for set 2. Blue and red indicate correct and incorrect responses, respectively. Each row displays responses for a particular rat across 80 trials. Light blue squares indicate that the rat had 8 correct trials in a row. A bar appearing after the end of the trials, to the right, indicates that the rat did not achieve 8 correct trials in a row by the end of the 80 trials. Groups A and C were rewarded for dark arm on set 1 while groups B and D were rewarded for light arm on set 1. The rats in group A clearly learned the discrimination task relatively quickly.*

Example 5.1.3 Membrane conductance Anderson and Stevens (1983, *J. Physiology*, 235: 655–691) were able to estimate single-channel membrane conductance by measuring total conductance at a frog neuromuscular junction. Their method relied on properties of the binomial distribution. Suppose that there are n channels, each either open or closed, all acting independently, and all having probability p of being open. Let X be the number of channels that are open, and γ be the single-channel conductance. Then the measured membrane conductance G satisfies $G = \gamma X$ where $X \sim B(n, p)$. From formulas (3.3) and (3.4) it follows that the mean and variance of G are given by

$$E(G) = \gamma np$$

and

$$V(G) = \gamma^2 np(1-p).$$

Now, assuming that p is small, we have $1 - p \approx 1$ so that γ satisfies

$$\gamma = \frac{V(G)}{E(G)}.$$

Anderson and Stevens made multiple measurements of the membrane conductance at many different voltages, obtaining many estimates of $V(G)$ and $E(G)$. The slope of the line through the origin fitted to a plot of $V(G)$ against $E(G)$ thereby furnished an estimate of the single-channel conductance.[1] □

The Anderson and Stevens estimate of single-channel conductance is based on the approximate proportionality of the variance and mean across voltages. In the derivation above this was justified from the binomial, for small p. The small-p case of the binomial is very important and, in general, when p is small while n is large, the binomial distribution may be approximated by the Poisson distribution.

[1] Additional comments on this method, and its use in analysis of synaptic plasticity, may be found in Faber and Korn (1991). Faber, D.S. and Korn, H. (1991) Applicability of the coefficient of variation method for analyzing synaptic plasticity, *Biophysical J.*, 60: 1288–1294.

5.2 The Poisson Distribution

5.2.1 The Poisson distribution is often used to describe counts of binary events.

The Poisson distribution is the most widely-used distribution for *counts*. Strictly speaking, the Poisson distribution assigns a positive probability to every nonnegative integer $0, 1, 2, \ldots$, so that every nonnegative integer becomes a mathematical possibility. This may be contrasted with the binomial, which takes on numbers only up to some n, and leads to a proportion (out of n). The defining feature of the Poisson distribution, however, is that it arises as a small-p and large-n approximation to the binomial, which we discuss in Section 5.2.2. That mathematical characterization portrays the count, approximately, as a sum of many binary variables, each indicating whether an event occurs (perhaps across time or across space), with each event occurrence having a small probability p. For example, neural spike counts are sometimes modeled as Poisson random variables. This results from a characterization of the spike train as a sequence of discrete event times, and if we decompose time into small bins (e.g., having 1 millisecond width) we may consider each time bin to define a binary variable that indicates whether a spike occurs within that bin, as depicted in Figure 5.2. When we consider discrete events across time there is necessarily some time scale (corresponding to a small bin width) on which the events become rare, so that the probability p that any binary variable will take the value 1 becomes small. For a spiking neuron with a low or moderate firing rate (say 10 spikes per second or less), for example, a scale in milliseconds leaves large gaps (many milliseconds) between each spike and makes the probability of a spike within any 1 millisecond bin quite small (e.g., less than 10/1000=.01). For this reason the Poisson is often said to be a model for the variation in the number of occurrences of rare events.[2]

Counts of such "rare" events are common in neuronal data analysis, but it

[2]The derivation of the Poisson distribution as an approximation to the binomial is credited to Siméon D. Poisson, having appeared in his book, published in 1837. Bortkiewicz (1898, *The Law of Small Numbers*) emphasized the importance of the Poisson distribution as a model of rare events.

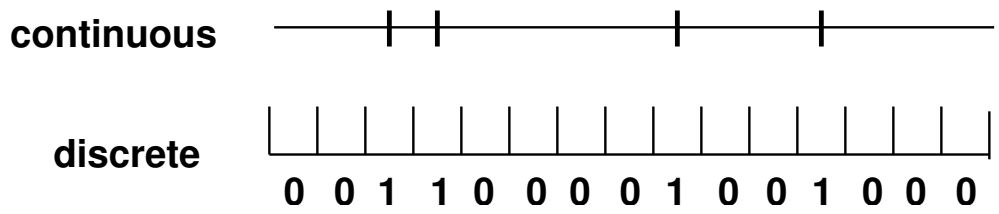


Figure 5.2: *Several event times displayed both in continuous time and in discrete time. In the discrete case time has been decomposed into bins and for each bin the presence or absence of an event is indicated by a 1 or 0.*

is important to recognize that many count distributions are discernibly *non-Poisson*. We begin our discussion with a classic data set from a situation where there are good reasons to think the Poisson distribution ought to provide an excellent description of the variation among counts. Although drawn from physics, this example helps to fix ideas about assumptions that generate Poisson variability. We then mention some situations in neural data analysis where Poisson distributions have been assumed. After that, we will elaborate on the motivation for the Poisson and then we will conclude with some discussion of frequently-occurring departures from Poisson variation among counts.

Example 5.2.1 Emission of α particles Rutherford and colleagues (1920) counted the number of α-particles emitted from a radioactive specimen during 2608 7.5 second intervals.[3] The data are summarized in the table below. The first column gives the counts $0, 1, 2, \ldots, 9, \geq 10$, and the second column gives number of times the corresponding count occurred. For example, in 383 of the 2608 intervals there were 2 particles emitted. The third colum provides the "expected" frequencies based on the Poisson distribution (obtained by a method to be explained later).

[3]Rutherford, Chadwick, and Ellis (1920) *Radiations from Radioactive Substances*, Cambridge, p. 172; cited in Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, Third Ed., New York: Wiley.

x	observed	expected
0	57	54.40
1	203	210.52
2	383	407.36
3	525	525.50
4	532	508.42
5	408	393.52
6	273	253.82
7	139	140.33
8	45	67.88
9	27	29.19
≥ 10	16	17.08

Here, the emission of any one particle is (on an atomic time scale) a "rare event" so that the number emitted during 7.5 seconds may be considered the number of rare events that occurred. □.

The Poisson pdf is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{5.2}$$

and we write $X \sim P(\lambda)$. The mean, variance, and standard deviation of X are given by

$$\begin{aligned} E(X) &= \lambda \\ V(X) &= \lambda \\ \sigma_X &= \sqrt{\lambda} \end{aligned}$$

The equality of variance and mean is highly restrictive and is often used to examine whether repeated series of observations depart from Poisson variation: a plot of variance vs. mean should fall approximately on the line $y = x$.

Here is a physiological setting involving particle emissions where the Poisson distribution was used much as in Example 5.2.1.

Example 5.2.2 Human detection of light Hecht, Schlaer, and Pirenne (1942, *J. Gen. Physiol.*, 25: 819–40) investigated the sensitivity of the human visual system to very dim light, and calculated the number of light

quanta required to drive perception. To do this, Hecht *et al.* constructed an apparatus that would emit very dim flashes of light, of 1 ms duration, in a darkened room; they presented these to several subjects and determined the proportion of times each subject would respond that he or she had seen a flash of light. In one part of their analysis, they assumed that the number of light quanta penetrating the retina would follow a Poisson distribution. If X is the number of quanta emitted, and if c is the number required for perception of the flash, then the probability of perception of flash is

$$P(X \geq c) = 1 - F(c-1) \tag{5.3}$$

where $F(x)$ is the Poisson cumulative distribution function. (Note that the argument $c-1$ appears because $P(X \geq c) = P(X > c-1) = 1 - F(c-1)$.) Using the formula for the Poisson cdf (i.e., the summed pdf), Hecht *et al.* fit this to observed data and found that, roughly, a minimum of 6 quanta must be absorbed by the retina in order for a human to detect light. □.

Not all applications of the Poisson distribution involve events across time. In the next example the events are distributed across space—on neural synaptic boutons.

Example 5.2.3 Quantal response in synaptic transmission The quantal response hypothesis is that neurotransmitter is released from a large number of presynaptic vesicles in packets, or "quanta," each of which has a small probability of being released. To test this, del Castillo and Katz (1954, *J. Physiol.*, 124: 560–573) recorded postsynaptic potentials, or end-plate potentials (EPPs), at a frog neuromuscular junction. By assuming a Poisson distribution for the number of quanta released following an action potential, the authors obtained good experimental support for the quantal hypothesis. □.

5.2.2 For large n and small p the binomial distribution is approximately the same as Poisson.

Example 5.2.3 (continued) Let us go a step further in examining the argument of del Castillo and Katz. Under behavioral conditions the EPP

would typically involve hundreds of quanta, but del Castillo and Katz used a magnesium bath to greatly decrease this number. In addition, they recorded spontaneous ("miniature") EPPs, which, according to the quantal hypothesis, should involve single quanta. They observed that this gave them two different ways to estimate the mean number of quanta released. The first method is to estimate the mean in terms of $P(X = 0)$ using the Poisson pdf formula $P(X = 0) = e^{-\lambda}$ or

$$\lambda = -\log P(X = 0). \tag{5.4}$$

To estimate $P(X = 0)$ they used the ratio D/C, where C was the total number of presynaptic action potentials and D was the number of times that the postsynaptic voltage failed to increase. Their second method used the ratio A/B, where A was the mean EPP voltage response following actional potentials and B was the mean spontaneous EPP voltage response. When the data from 10 experiments were plotted, the ten (x, y) pairs with $y = -\log D/C$ and $x = A/B$ were very close to the line $y = x$. □.

A major motivation for the Poisson distribution is that it approximates the binomial distribution as p gets small and n gets large (with $\lambda = np$). One way to express this is given by the theorem below, but the argument used by del Castillo and Katz, described above, highlights both the key assumptions and the key mathematical result. Under the quantal hypothesis that vesicle release is binary *together with* the Bernoulli assumptions of independence and homogeneity, we have

$$P(X = 0) = (1 - p)^n$$

where p is the probability that any given vesicle will release and n is the number of vesicles. We define $\lambda = np$ and make the substitution $p = \lambda/n$, then take logs of both sides to get

$$\log P(X = 0) = n \log(1 - \frac{\lambda}{n}).$$

Now, for large n, an elementary application of calculus (a first-order Taylor series approximation) gives $n \log(1 - \lambda/n) \approx -\lambda$. This says that Equation (5.4) becomes a good approximation for small p and large n. The rest of the argument is given below.

Theorem: Poisson pdf approximation to binomial pdf As

$n \to \infty$ we have

$$\binom{n}{k} p^k (1-p)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}. \tag{5.5}$$

Proof: To derive Equation (5.5), the key manipulation is an expansion of the log function. First, for t near 0 we may use a first-order Taylor series to expand the function $\log(1+t)$ as $\log(1+t) \approx t$. We may use this with $t = -\lambda/n$ for any fixed number λ when n is large (so that t is near zero) and get

$$\log(1 - \lambda/n) \approx -\lambda/n.$$

Multiplying by n we have

$$n \log(1 - \lambda/n) \approx -\lambda$$

and taking the exponential of both sides we get the equivalent approximation

$$(1 - \lambda/n)^n \approx e^{-\lambda}$$

which is formalized by saying that as $n \to \infty$,

$$(1 - \frac{\lambda}{n})^n \to e^{-\lambda}. \tag{5.6}$$

Now let $\lambda = pn$, substitute $p = \lambda/n$ into the binomial pdf,

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

and rearrange the terms to get

$$f(k) = \underbrace{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)}\ \underbrace{\left(\frac{\lambda^k}{k!}\right)}\ \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}\ \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}.$$

As $n \to \infty$, the expression over the first underbrace converges to 1; the second remains constant (n does not appear there); by (5.6) the third converges $e^{-\lambda}$; and the fourth expression converges to 1. This gives (5.5). □

5.2.3 The Poisson distribution results when the binary events are independent.

In thinking about binomial assumption for a random variable X one generally ponders whether it is reasonable to think of X as a sum of Bernoulli trials with the independence and homogeity assumptions. Similarly, in the Poisson case, one typically asks whether the count variable X could be considered a sum of Bernoulli trials for small p (and large n). The first requirement is that the counts really are sums of binary events. This means that a depiction as in Figure **??** is possible. In Example 5.2.1, each emission event corresonds to a state transition in the nucleus of a particular atom. It is reasonable to assume that it is impossible for two nuclei to emit particles at precisely the same time and, furthermore, that each Geiger-counter "click" corresponds to exactly one particle emission. Independence, usually the crucial assumption, here refers to the independence of the many billions of neuclei residing within the specimen. This is an assumption, apparently well justified, within the quantum-mechanical conception of radioactive decay. It implies, for example, that any tendency for two particles to be emitted at nearly the same time would be due to chance alone: because there is no interaction among the nuclei, there is no physical "bursting" of multiple particles. Furthermore, the probability of an emission would be unlikely to change over the course of the experiment unless the speciman were so tiny that its mass changed appreciably. To summarize, the Poisson distribution for counts of events across time results when we can form HERE (i) the events are discrete (no two events occur at the same time), (ii) the probability

Similar things could be said about quanta of light in the Hecht *et al.* experiment. What about the vesicles at the neuromuscular junction? Here, the quantal hypothesis is what generates the sequence of dichotomous events (release vs. no release). Is release at one vesicle independent of release at another vesicle? If neighboring vesicles tend to release in small clumps, then we would expect to see more variability in the counts than that predicted by the Poisson, while if release from one vesicle tended to inhibit release of neighbors we would expect to see more regularity, and less variability in the counts. It is reasonable to begin by assuming independence, but ultimately it is an empirical question whether this is justified. Homogeneity is suspect: the release probability at one vesicle may differ substantially from that at another

vesicle. However, as del Castillo and Katz realized, this is not an essential assumption. We elaborate when we return to the Poisson distribution, and its relationship to the *Poisson process*, below.

Neuronal spike counts are sometimes assumed to be Poisson-distributed. Let us consider the underlying assumptions in this case. First, if measurements are made on a single neuron to a resolution of 1 millisecond or less, it is the case that a sequence of dichotomous firing events will be observed: in any given time bin (e.g., any given millisecond) the neuron either will or will not have an action potential, and it can not have two. But are these events independent? The membrane of a neuron undergoes changes that alter its propensity to fire immediately after it has fired. In particular, there is a refractory period during which it can not fire again. This clearly violates the assumption of independence. In addition, there may be a build-up of ions, or activity in the local neural network, that makes a neuron more likely to fire if it has fired recently in the past (it may be "bursting"). This again would be a violation of independence. In many experiments such violations of independence produce markedly non-Poisson count distributions and turn out to have a substantial effect, but in others the effects are relatively minor and may be ignored. We indicated that, in the case of vesicle release of neurotransmitters, the homogeneity assumption is not needed in order to apply the Poisson approximation. The same is true for neuronal spike counts: the spike probabilities can vary across time and still lead to Poisson-distributed counts. The key assumption, requiring thought, is independence. On the other hand, the question of whether it is safe to assume Poisson variation remains an empirical matter, subject to statistical examination. As in nearly all statistical situations, judgment of the accuracy of the modeling assumptions—here, the accuracy of the Poisson distribution in describing spike count variation—will depend on the analysis to be performed.

5.3 The Normal Distribution

As we said in Chapter 3, the normal distribution plays a central role in statistical theory because of the Central Limit Theorem, which we state in Chapter 6. In Section 5.3.1 we review a property of the normal distribution that leads to interpretation of standard errors and confidence intervals,

and in Section 5.3.2 we note its relationship to the binomial and Poisson distributions.

5.3.1 Normal random variables are within 1 standard deviation of their mean with probability $2/3$; they are within 2 standard deviations of their mean with probability .95.

Let us be explicit about the two facts we'll call the "$\frac{2}{3}$ and 95% rule."

For a normal random variable X with mean μ and standard deviation σ,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \tfrac{2}{3}$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx .95$$

We also have $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx .997$, but this is less important.

Example: Ebbinghaus on human memory A very early experiment on human memory was reported by Ebbinghaus (1885). Ebbinghaus used himself as the sole subject of his investigation, and he taught himself to learn lists of nonsense syllables made up of consonant-vowel-consonant trigrams such as DAX. Ebbinghaus memorized relatively long lists (e.g., 16 trigrams) to the point of being able to recite them without error, twice consecutively, and kept track of the time it took for him to achieve this success. He then repeated the task using the same lists after a delay period, that is, he re-learned the lists, and he examined the way his re-learning time increased with the length of the delay period. This was a way to quantify his rate of forgetting. (Compare the experiment of Kolers at the end of Chapter 1.) The method Ebbinghaus used relied on the normal distribution. In one of his tabulations, he examined 84 memorization times, each obtained by averaging sets of 6 lists. He found the distribution of these 84 data values to be well approximated by the normal distribution, with mean 1,261 seconds and standard deviation 72 seconds.[4] This would mean that for about 2/3

[4]He actually found the "probable error," which is $.6745\sigma$ to be 48.4 seconds. See Stigler (1986) for a discussion of these data.

of the sets of lists his learning time was between 1,189 seconds and 1,333 seconds. It also would mean that a set-averaged learning time less than 1,117 seconds or greater than 1,405 seconds would be rare: each of these would occur for only about 2.5% of the sets of lists. □

It may seem odd that in examining the suitability of the normal distribution Ebbinghaus did not look at the distribution of learning times for lists, but rather chose to work with the distribution of *average* learning times across sets of 6 lists. This distribution of learning times was skewed. Only after averaging across several learning times did the distribution become approximately normal. This effect is due to the Central Limit Theorem, discussed in Section 6.3.1.

Normal distributions are often *standardized* so that $\mu = 0$ and $\sigma = 1$. In general, using Equation (3.5) if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$. As a special case, if $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$ then $Z \sim N(0, 1)$. This is often used for calculation: if we know probabilities for the $N(0, 1)$ distribution then we can easily obtain them for any other normal distribution. For example, we also have

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}).$$

Thus, the right-hand side may be found using a table in order to obtain the answer for the left-hand side. Many books do lots of calculations this way, but it's just as easy (or easier) to use statistical software.

5.3.2 Binomial and Poisson distributions are approximately normal, for large n or large λ.

The normal distribution may be used to approximate a large variety of distributions for certain values of parameters. In the case of the binomial with parameters n and p, we take the normal mean and standard deviation to be $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. An illustration is given in Figure 5.3. The approximation is generally considered to be quite accurate for most calculations when n is large and p is not close to its boundary values of 0 and 1; a commonly-used rule of thumb (which is somewhat conservative, at least for $.2 < p < .8$) is that it will work well when $np \geq 5$ and $n(1-p) \geq 5$.

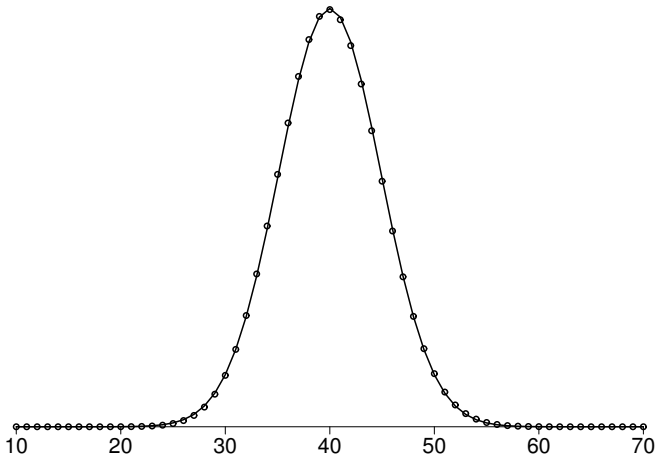


Figure 5.3: *The normal approximation to the binomial. Black circles are pdf values for a* $B(100, .4)$ *distribution; Curve is pdf of a normal having the same mean and variance.*

In the case of the Poisson with parameter λ we take the normal mean and standard deviation to be $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$; the approximation is generally considered to be acceptably accurate for many calculations[5] when $\lambda \geq 15$.

These approximations are a great convenience, especially in conjunction with the "$\frac{2}{3} - 95\%$ rule."

[5]Actually, different authors give somewhat different advice. The acceptability of this or any other approximation must depend on the particular use to which it will be put. For computing the probability that a Poisson random variable will fall within 1 standard deviation of its mean, the normal approximation has an error of less than 10% when $\lambda = 15$. However, it will not be suitable for calculations that go far out into the tails, or that require several digits of accuracy. In addition, a computational fine point is mentioned in many books. Suppose we wish to approximate a discrete cdf $F(x)$ by a normal, say $\tilde{F}(x)$. The the value $\tilde{F}(x + .5)$ is generally closer to $F(x)$ than is $\tilde{F}(x)$. This is sometimes called a continuity correction.

5.4 Some Other Common Distributions

5.4.1 The multinomial distribution extends the binomial to multiple categories.

In Section 5.1.2 we cited an application of the Hardy-Weinberg model in a study of genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. The three genotypes were labeled TT, CT, CC. This constitutes three distinct categories. For the ith individual in the study, let $Y_i = (1, 0, 0)$ if that individual has genotype TT, $Y_i = (0, 1, 0)$ if that individual has genotype CT, and $Y_i = (0, 0, 1)$ if that individual has genotype CC. The variable Y_i thus indicates the genotype of the ith individual, for $i = 1, 2, \ldots, n$. Let $p_1 = P(Y_i = (1, 0, 0))$, $p_2 = P(Y_i = (0, 1, 0))$, and $p_3 = P(Y_i = (0, 0, 1))$, where $p_1 + p_n + p_3 = 1$ and define $X = \sum_{i=1}^{n} Y_i$. Note that X gives the number of individuals, among a total of n, that have each of the three genotypes. In the Kent *et al.* data listed earlier, for instance, there were 136 individuals: 48 of genotype TT, 71 of genotype CT, and 17 of genotype CC, and we could write $X = (48, 71, 17)$. If we assume the $Y_1, Y_2, \ldots, Y_n$ are independent then X follows a *multinomial* distribution, written $X \sim M(n; p_1, p_2, p_3)$ with pdf

$$P(X = (x_1, x_2, x_3)) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}. \tag{5.7}$$

According to the Hardy-Weinberg model the probabilities (p_1, p_2, p_3) would be restricted to satisfy the binomial pdf $p_1 = p^2$, $p_2 = 2p(1-p)$ and $p_3 = (1-p)^2$. However, (5.7) holds regardless of the validity of the Hardy-Weinberg model, as long as the genotypes are independent and homogeneous across individuals.

More generally, a random variable is distributed as $X \sim M(n; p_1, p_2, \ldots, p_m)$ if its pdf is given by

$$P(X = (x_1, x_2, \ldots, x_m)) = \frac{n!}{x_1! x_2! \cdots x_m!} \prod_{j=1}^{m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}.$$

5.4.2 The exponential distribution is used to describe waiting times without memory.

We defined the exponential distribution in Equation (3.7), using it to illustrate calculations based on the pdf, and we showed how it may be applied to ion channel channel activation durations in Example 3.2.2. The exponential distribution is very special[6] because of its "memoryless" property. To understand this, let X be the length of time an ion channel is open, and let us consider the probability that the channel will remain open for the next time interval of length h. For example, h might be 5 milliseconds. How do we write this? If we begin the moment the channel opens, i.e., at $x = 0$, the next interval of length h is $(0, h)$ and we want $P(X > h)$. On the other hand, if we begin at time $x = t$, for some positive t, such as 25 milliseconds, the interval in question is $(t, t + h)$ and we are asking for a *conditional* probability: if the channel is open at time t we must have $X > t$, so we are asking for $P(X > t + h|X > t)$. We say that the channel opening duration is memoryless if

$$P(X > t + h|X > t) = P(X > h) \tag{5.8}$$

for all $t > 0$ and $h > 0$. That is, if $t = 25$ milliseconds, the channel does not "remember" that it has been open for 25 milliseconds already; it still has the same probability of remaining open for the next 5 milliseconds that it had when it first opened. And this is true regardless of the time t we pick. The exponential distributions are the *only* distributions that satisfy Equation (5.8).

Contrast this memorylessness with, say, a uniform distribution on the interval $[0, 10]$, measured in milliseconds. According to this uniform distribution, the event (e.g., the closing of the channel) must occur within 10 milliseconds and initially every 5 millisecond interval has the same probability. In particular, the probability the event will occur in the first 5 milliseconds, i.e., in the interval $[0, 5]$, is the same as the probability it will occur in the last 5 milliseconds, in $[5, 10]$. Both probabilities are equal to $\frac{1}{2}$. However, if at time $t = 5$ milliseconds the event has not yet occurred then we are *certain* it will occur in the next half second $[5, 10]$, i.e., this probability is 1, which is

[6] Another reason the exponential distribution is special is that among all distributions on $(0, \infty)$ with mean $\mu = 1/\lambda$, the $Exp(\lambda)$ distribution has the maximum entropy. See Equation (4.15).

quite different than $\frac{1}{2}$. In anthropomorphic language we might say the random variable "remembers" that no event has yet occurred, so its conditional probability is adjusted. For the exponential distribution, the probability the event will occur in the next 5 milliseconds, given that it has not already occurred, stays the same as time progresses.

Theorem A random variable X satisfies $X \sim Exp(\lambda)$ if and only if (5.8) is satisfied for all positive t and h.

Proof: Using Equation (3.8) we have

$$\begin{aligned} P(X > t+h | X > t) &= \frac{P(X > t+h, X > t)}{P(X > t)} \\ &= \frac{P(X > t+h)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} \\ &= e^{-\lambda h} \\ &= P(X > h). \end{aligned}$$

Thus, every exponential distribution is memoryless. On the other hand, let $G(x) = 1 - F(x)$ where $F(x)$ is the distribution function of X. Memorylessness implies

$$P(X > t+h) = P(X > t)P(X > h)$$

i.e.,

$$G(t+h) = G(t)G(h)$$

for all positive t and h. But (as mentioned in Appendix A), $G(x)$ can satisfy this equation for all positive t and h only if it has an exponential form $G(x) = ae^{bx}$. Because $F(x) = 1 - G(x)$ is a distribution function, it is constrained to asymptote to 1 as $x \to \infty$, which implies $b > 0$, and it must vanish at $x = 0$, which implies $a = 1$. Thus $F(x) = 1 - e^{-\lambda x}$ for some λ, i.e., $X \sim Exp(\lambda)$. □

An additional characterization of the exponential distribution is that it has a constant hazard function.

Theorem: A continuous random variable X satistfies $X \sim Exp(\lambda_0)$ if and only if its hazard function is $\lambda(x) = \lambda_0$.

Proof: First suppose $X \sim Exp(\lambda_0)$. The hazard function is easy to compute from the definition

$$\lambda(x) = \frac{f(x)}{1 - F(x)}.$$

Substituting $f(x) = \lambda_0 e^{-\lambda_0 x}$ and $F(x) = 1 - e^{-\lambda_0 x}$ we have

$$\begin{aligned} \lambda(x) &= \frac{\lambda_0 e^{-\lambda_0 x}}{e^{-\lambda_0 x}} \\ &= \lambda_0. \end{aligned}$$

On the other hand, if the hazard function is $\lambda(x) = \lambda_0$ we may rewrite the definition of $\lambda(x)$ and solve for $F(x)$,

$$F(x) = 1 - \lambda_0 f(x)$$

and then differentiate to get

$$f(x) = -\lambda_0 f'(x)$$

which implies $f(x) \propto e^{-\lambda_0 x}$ and because $f(x)$ must integrate to 1 we get $f(x) = \lambda_0 e^{-\lambda_0 x}$. □

The constant hazard of the exponential may be considered another way to view memorylessness: with constant hazard, given that the event has not already occurred at time t the probability that the event occurs in the next infinitesimal interval $(t, t + dt)$ is the same as it would be for any other infinitesimal interval $(t', t' + dt)$.

In Chapter **??** we will discuss the role played by the exponential distribution in *Poisson processes*, which are sometimes used to model spike trains.

5.4.3 Gamma distributions are sums of exponentials.

In Example 3.2.2 we illustrated a basic property of a gamma distribution: if $X_1, X_2, \ldots, X_n$ are distributed as $Exp(\lambda)$, independently, and $Y = X_1 +$

$\cdots + X_n$, then $Y \sim Gamma(n, \lambda)$. Note that a $Gamma(1, \lambda)$ distribution is the same as an $Exp(\lambda)$ distribution. More generally, a random variable X is said to have a $Gamma(\alpha, \beta)$ distribution when its pdf is

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

for $x > 0$ and is 0 when $x \leq 0$. Here, the function $\Gamma(a)$ is the gamma function:

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The gamma function is a variant of the factorial function; we have $\Gamma(n) = (n-1)!$ for any positive integer n. If $X \sim Gamma(\alpha, \beta)$ then

$$\begin{aligned} E(X) &= \frac{\alpha}{\beta} \\ V(X) &= \frac{\alpha}{\beta^2} \\ \sigma_X &= \frac{\sqrt{\alpha}}{\beta}. \end{aligned}$$

Plots of the gamma will be displayed for the special case of the chi-squared distribution, in the next subsection.

5.4.4 Chi-squared distributions are special cases of Gammas.

If $W \sim N(0, 1)$ then $X = W^2$ is said to have a *chi-squared distribution* on 1 degree of freedom, which is written $X \sim \chi^2_1$. If $W_i \sim \chi^2_1$ for all $i = 1, \ldots, n$, independently, and if $X = W_1 + W_2 + \cdots + W_n$, then X is said to have a chi-squared distribution on n degrees of freedom, written $X \sim \chi^2_n$. The most important way chi-squared distributions arise is as sums of squares of independent normal distributions. In general, a random variable X is said to have a chi-squared distribution with degrees of freedom ν, written χ^2_ν if it has a $Gamma(\alpha, \beta)$ distribution with $\alpha = \frac{\nu}{2}$ and $\beta = \frac{1}{2}$.

If $X \sim \chi^2_\nu$ then

$$E(X) = \nu$$

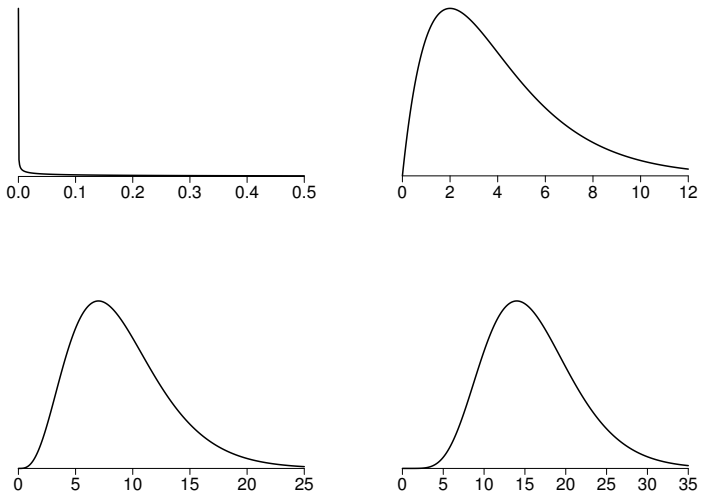


Figure 5.4: *Chi-squared pdfs for four values of the degrees of freedom: $\nu = 1$ (top left), 4 (top right), 9 (bottom left), and 16 (bottom right).*

$$\begin{aligned} V(X) &= 2\nu \\ \sigma_X &= \sqrt{2\nu}. \end{aligned}$$

Figure 5.4 shows several chi-squared density functions. Note that, for small degrees of freedom, the distribution is *skewed toward high values* (or *skewed to the right*). That is, it is not symmetrical, but rather large values distant from the middle (to the right) are more likely than small values distant from the middle (to the left). For the χ^2_4, the middle of the distribution is roughly between 1 and 6 but values less than 0 are impossible while values much greater than 7 have substantial probability. For large degrees of freedom ν the χ^2_ν becomes approximately normal. For $\nu = 16$ in Figure 5.4 there remains some slight skewness, but the distribution is already pretty close to normal over the plotted range.

5.4.5 The Beta distribution may be used to describe variation on a finite interval.

A random variable X is said to have a Beta distribution with parameters α and β if its pdf is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$ and is 0 otherwise. We then write $X \sim Beta(\alpha, \beta)$. Suppose $W_1 \sim Gamma(\alpha_1, \beta)$ and $W_2 \sim Gamma(\alpha_2, \beta)$, independently, and let $X = W_1/(W_1 + W_2)$. Then we have $X \sim Beta(\alpha, \beta)$.

If $X \sim Beta(\alpha, \beta)$ then $E(X) = \alpha/(\alpha+\beta)$ and $V(X) = \alpha + \beta + 1$. The beta distribution is sometimes written instead in terms of the parameters $\mu = E(X)$ and $\nu = V(X) - 1$, so that $\alpha = \mu\nu$ and $\beta = (1-\mu)\nu$. The beta distribution is commonly used to describe continuous variation that is confined to (0,1). By rescaling it is easy to obtain a distribution confined to any finite interval (a, b).

5.4.6 The Inverse Gaussian distribution describes the waiting time for a threshold crossing by Brownian motion.

A random variable X is said to have an inverse Gaussian distribution if its pdf is

$$f(x) = \sqrt{\lambda/(2\pi x^3)}\exp(-\lambda(x-\mu)^2/(2\mu^2 x))$$

for $x > 0$. Here, $E(X) = \mu$ and $V(X) = \mu^3/\lambda$.

The inverse Gaussian arises in conjunction with Brownian motion, where it is the distribution of "first passage time," meaning the time it takes for the Brownian motion (with drift) to cross a boundary. (See Whitmore, G.A. and Seshadri, V. (1987) A Heuristic Derivation of the Inverse Gaussian Distribution *The American Statistician*, 41: 280-281. Also, Mudholkar, G.S. and Tian, L. (2002) An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test, *J. Statist. Planning and Inference*, 102:

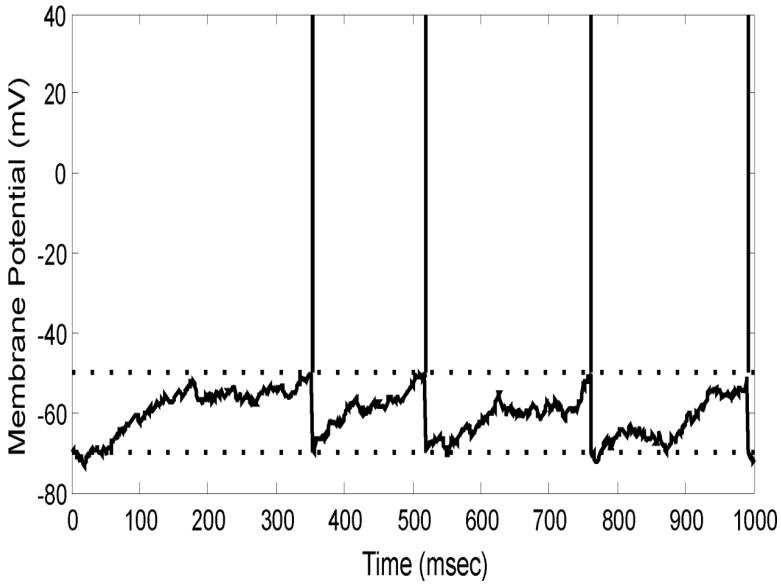


Figure 5.5: *Example of an integrate-and-fire neuron. At each time step there is either an EPSP or an IPSP, with probabilities p and $1-p$. For $p > 1-p$ this creates a stochastic upward "drift" of the voltage (as the inputs are summed or "integrated") until it crosses the threshold and the neuron fires. The neuron then resets to its baseline voltage. The resulting ISI distribution is approximately inverse Gaussian.*

211–221.) In theoretical neurobiology the interspike interval distribution for an integrate-and-fire neuron is inverse Gaussian when the subthreshold neuronal voltage is modeled as Brownian motion, with drift, and the "boundary" is the voltage threshold for action potential generation. The essential idea here is that excitatory and inhibitory post-synaptic potentials, EPSPs and IPSPs, are considered to arrive in a sequence of time steps of length δ, with each EPSP and IPSP contributing normalized voltages of +1 and -1, respectively, and with the probability of EPSP and IPSP being p and $1-p$, where $p > 1-p$ creates the upward "drift" toward positive voltages. Let $X_t, \ldots$ be the post-synaptic potential at time t with $t = 1, 2, \ldots$ and let $S_n = X_1 + X_2 + \cdots + X_n$. The variable S_n is said to follow a *random walk* and an action potential occurs when S_n exceeds a particular threshold value a. The behavior of an integrate-and-fire neuron based on such a random walk process is illustrated in Figure 5.5. The continuous-time stochastic pro-

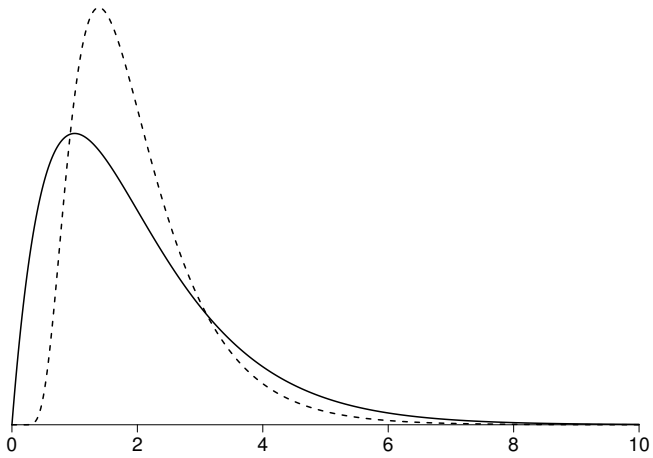


Figure 5.6: *Inverse Gaussian pdf plotted together with a Gamma*$(2, 1)$ *pdf. The Inverse Gaussian (blue) has the same mean and variance as the Gamma. Note its convexity near 0.*

cess known as Brownian motion with drift (and thus the inverse Gaussian distribution of the ISIs) results from taking $\delta \to 0$ and $n \to \infty$, while also constraining the mean and variance in the form $E(S_n) \to m$ and $V(S_n) \to v$, for some m and v.

Figure 5.6 gives an example of an inverse Gaussian pdf, with a Gamma pdf for comparison. Note in particular that when x is near 0 the inverse Gaussian pdf is very small. This gives it the ability to model, approximately, neuronal interspike intervals in the presence of a refractory period, i.e., a period at the beginning of the interspike interval (immediately following the previous spike) during which the neuron doesn't fire, or has a very small probability of firing.

5.4.7 The t and F distributions are defined from normal and chi-squared distributions.

Two distributions are used very frequently in statistical hypothesis testing. The first is the t distribution.

If $X \sim N(0,1)$ and $Y \sim \chi^2_\nu$, independently, then

$$T = \frac{X}{\sqrt{\frac{Y}{\nu}}}$$

is said to have a t distribution on ν degrees of freedom, which we write as $T \sim t_\nu$. This form of the T ratio arises in "t tests" and related procedures.

Note that T would be $N(0,1)$ if the denominator were equal to 1. The denominator is actually very close to one when ν is large: if $Y \sim \chi^2_\nu$ we have $E(Y/\nu) = 1$ while $V(Y/\nu) = 2\nu/\nu^2$ which becomes very close to zero for large ν. That is, the random variable Y/ν has a very small standard deviation and thus takes values mostly very close to its expectation of 1. Therefore, for large ν, the t_ν distribution is very close to a $N(0,1)$ distribution. One rule of thumb is that for $\nu > 12$, when computing probabilities in the middle of the distribution, the t_ν is may be considered essentially the same as $N(0,1)$. For small ν, however, the probability of large positive and negative values becomes much greater than that for the normal. For example, if $X \sim N(0,1)$ then $P(X > 3) = .0014$ whereas if $T \sim t_3$ then $P(T > 3) = .029$, about 20 times the magnitude. To describe this phenomenon we say that the t_3 distribution has much *heavier tails* (or *thicker tails*) than the normal.

If $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$, independently, then

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

is said to have an F distribution on ν_1 and ν_2 degrees of freedom, which are usually referred to as the numerator and denominator degrees of freedom. We may write this as $F \sim F_{\nu_1,\nu_2}$. This distribution arises in regression and analysis of variance, where ratios of sums of squares are computed and each sum of squares has (under suitable assumptions) a chi-squared distribution.

When $\nu_1 = 1$ the numerator is the square of a normal and $F = T^2$, where T is the ratio of a $N(0,1)$ and the square-root of a $\chi^2_{\nu_2}$. That is, the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution. Also, analagously to the situation with the t_ν distribution, when ν_2 gets large the denominator Y/ν_2 is a random variable that takes values mostly very close to 1 and F_{ν_1,ν_2} becomes close to a $\chi^2_{\nu_1}$.

5.5 Multivariate Normal Distributions

5.5.1 A random vector is multivariate normal if linear combinations of its components are univariate normal.

We described the bivariate normal distribution in Section 4.2.2. We say that an m-dimensional random vector X has an *m-dimensional multivariate normal distribution* if for every nonzero linear combination of its components is normally distributed. If μ and Σ are the mean vector and variance matrix of X we write this as $X \sim N_m(\mu, \Sigma)$. Using (4.13) and (4.14) we thus characterize $X \sim N_m(\mu, \Sigma)$ by saying that for every nonzero m-dimensional vector w we have $w^T X \sim N(w^T\mu, w^T\Sigma w)$. This definition does not require Σ to be positive definite. In discussing the bivariate normal pdf for (X, Y) we had to assume $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$. This is equivalent to saying that the variance matrix of the (X, Y) vector is positive definite. When we work with the multivariate normal distribution we usually assume the variance matrix is positive definite. Occasionally, however, one must deal with the non-positive definite case. This arises, for example, when ones wants to model the joint variation of m variables by assuming it is concentrated in fewer than m dimensions (analogously to the bivariate case with $\rho = 1$).

Notice that, just as the univariate normal distribution is completely characterized by its mean and variance, and the bivariate normal distribution is characterized by means, variances, and a correlation, the multivariate normal distribution is completely characterized by its mean vector and covariance matrix. In many cases, the components of a multivariate normal random

vector are treated separately, with each diagonal element of the covariance matrix furnishing a variance, and the off-diagonal elements being ignored. In some situations, however, the joint distribution, and thus all the elements of the covariance matrix, are important.

The multivariate normal distribution is even more prominent in multivariate data analysis than the normal distribution is for univariate data analysis. The main reason is that specifying only the first two moments, mean vector and variance matrix, is a huge simplification. In addition, there is a generalization of the Central Limit Theorem, which we give in Section 6.3.2.

5.5.2 The multivariate normal pdf has elliptical contours, with probability density declining according to a χ^2 pdf.

If X is m-dimensional multivariate normal, having mean vector μ positive definite covariance matrix Σ, then its pdf is given by

$$f(x) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}Q(x)}$$

where

$$Q(x) = (x - \mu_X)^T \Sigma^{-1} (x - \mu_X)$$

with $|\Sigma|$ being the determinant of Σ. We have labeled the exponent by $Q(x)$ to emphasize that it gives a quadratic in the components of x. The positive definiteness of this quadratic implies that the contours of $f(x)$ are multidimensional ellipses. (See Appendix A.) Importantly, every pair of components X_i and X_j has a bivariate normal distribution (this follows immediately from the definition of multivariate normality), which has elliptical contours for its joint distribution.

Using simple matrix multiplication arguments, it is not hard to show that if $X \sim N_n(\mu, \Sigma)$ then $Q(X)$ has a chi-squared distribution with m degrees of freedom.

> *Details:* Let Y be m-dimensional multivariate normal with the zero vector as its mean vector and the m-dimensional identity

matrix as its variance matrix. The components of Z follow $Z \sim N(0,1)$, independently. Thus, from the definition of the chi-squared distribution in Section 5.4.4, $Z^T Z \sim \chi^2_m$. Now, if $X \sim N_n(\mu, \Sigma)$ then, by the theorem on page 114, $Y = \Sigma^{-1/2}(X - \mu)$ satisfies $Y^T Y \sim \chi^2_m$. But $Y^T Y = Q(X)$. □

Taken together these results imply that each contour $\{x : Q(x) = c\}$ of the multivariate normal pdf is elliptical and encloses a region $\{x : Q(x) > c\}$ that has probability given by the χ^2_m distribution function.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 6

Sequences of Random Variables

One of the great ideas in data analysis is to base probability statements on large-sample approximations, which are often easy to obtain either analytically or numerically. This short chapter contains the two fundamental results that produce most of the methodology, the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). Both concern the behavior of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i$. These theorems form a foundation for much data analytic theory because many statistical functions may be either rewritten or approximated in terms of sample means.

6.1 Random Sequences and the Sample Mean

We need a crucial piece of preliminary terminology: if X_1, X_2, ..., X_n are drawn independently from the same distribution, then X_1, X_2, ..., X_n is said to form a *random sample* from that distribution, and the random variables X_i are said to be *independent and identically distributed (i.i.d.)*. This section is about means computed from random samples (sets of i.i.d. random

variables). Let $\mu = E(X_i)$. The LLN says that $\bar{X}$ gets arbitrarily close to μ as n increases indefinitely. The CLT says that the distribution of $\bar{X}$ becomes arbitrarily close to a normal distribution as n increases indefinitely. Similar results hold for many other data summaries, as well (because they may be written in terms of sample means). They are extremely important because they allow the properties in Section 5.3.1 to be applied, producing simple and useful probability statements.

In analyzing the behavior of the sample mean, the first point to recognize is that drawing a new sample would produce a new value of the sample mean, so that if we were to repeat the process of drawing a new sample many times, we would observe variability in the sample mean. On page137, for example, we described some data on re-learning time from Ebbinghaus (1885), and noted that he examined 84 means, each of which was obtained by averaging the re-learning time across 6 lists of trigrams. Each mean was slightly different: they exhibited variation. The second point is that there is, typically[1], the variation in the sample mean is smaller than that in the original data, and it decreases with increasing sample size.

Example 3.2.1 (continued) Figure 3.2 displays a histogram of 60 spike counts from a motor cortical neuron during a reaching task. The mean among these 60 counts is 13.6 spikes. (The time interval was 600 milliseconds, so this neuron's mean firing rate was 22 spikes per second.) Imagine drawing one spike count at random from among the 60, and doing this repeatedly. The histogram gives a sense of the variability we would see in these repeated random draws. Now suppose instead we were to draw 4 spike counts at random, and compute its mean, and then repeat this process many times. Because it would be likely that some of the 4 values would be bigger than 13.6, and some would be less, a mean of these 4 values would tend to be closer to 13.6 than any a single random value would be—in other words, the mean of 4 observations would tend to exhibit less variability than did the original observations themselves. We can see this by considering the first 12 of the spike counts:

16 12 14 9 9 4 12 14 13 13 17 16

[1]There are exceptions to this rule if the expectation does not exist, which can occur when the tails of the pdf fall to zero very slowly. An example is the Cauchy distribution, which is the t distribution on 1 degree of freedom.

The mean count among these 12 is 12.4 spikes and the standard deviation is 3.7 spikes. Now consider the remaining 48 spike counts:

21	16	16	10	12	15	11	11	8	26	12	12
18	13	13	12	8	16	14	12	7	13	12	14
14	16	10	11	7	17	15	14	16	10	13	13
14	10	14	15	16	17	12	18	32	11	19	13

The data have been arranged in 12 columns of length 4 in order to consider the column means. In this case, the mean of these 12 means is 13.9 spikes and the standard deviation is 1.8 spikes: we find that the variation among the 12 means is smaller than the variation among the 12 raw counts. □

The points illustrated by these motor cortical spike counts are (i) if we calculate the mean of a set of observations (a set of 4 trials) repeatedly for new data (12 repetitions of the sets of 4) we observe variation among the means, and (ii) the variation among the means (the standard deviation of 1.8) is smaller than the variation we would typically see among the raw spike counts (the standard deviation of 3.7). However, this illustration was intended only to set the stage for an entirely theoretical discussion. In this section we consider the *random variable* $\bar{X}$. Its variation may be quantified by its standard deviation $\sigma_{\bar{X}}$. Notice that this is not the same thing as the standard deviation σ_X of the original data. In fact, $\sigma_{\bar{X}}$ decreases as the sample size increases; qualitatively, the larger the sample size, the less variation in the sample mean. Specifically, we have $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$. After giving this result in Section 6.1.1 we present the law of large numbers in Section 6.2.1 and the Central Limit Theorem in Section 6.3.1. These theorems require the use of some mathematics for dealing with sequences of random variables, which is the topic of Section 6.1.2.

6.1.1 The standard deviation of the sample mean decreases as $1/\sqrt{n}$.

If we repeatedly draw random samples $X_1, \ldots, X_n$, and from them repeatedly compute $\bar{X}$, the value of $\bar{X}$ will fluctuate: it will be a random variable. The dominant features of the distribution of $\bar{X}$ are captured by its mean and variance, which may be computed easily from the formulas (**??**) and (**??**).

Theorem If $X_1, X_2, \ldots, X_n$ form a random sample from a distribution having mean μ_X and standard deviation σ_X then

(i) $E(\bar{X}) = \mu_X$, and

(ii) $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$.

Proof: The expectation $E(\bar{X})$ is immediate from (**??**). For the variance, in formula (**??**), plug in $V(X_i) = \sigma_X^2$ to get $V(X_1 + X_2 + \cdots + X_n) = n\sigma_X^2$. Then take square-roots and, remembering that $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, apply (3.4). □

The statement that $E(\bar{X}) = \mu_X$ says that the average amount by which $\bar{X}$ exceeds μ is equal to the average amount by which μ exceeds $\bar{X}$. The statement $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$ quantifies how rapidly the fluctuations in $\bar{X}$ diminish as a function of sample size. It is sometimes called "the square-root of n law." A consequence of diminishing fluctuations is that $\bar{X}$ must tend to get closer and closer to μ_X. This is the LLN, given in Section 6.2.1.

These results may be illustrated in the case of Bernoulli trials, where X_i is either 0 or 1. For the case $p = P(X_i = 1) = .4$ and $n = 4$, the sum $\sum_{i=1}^n X_i$ takes possible values of 0,1,2,3,4, with binomial probabilities .0625,.25,.375,.25,.0625. Thus, the mean $\bar{X}$ takes possible values of 0,.25,.5,.75,1, also with probabilities .0625,.25,.375,.25,.0625. The pdf is plotted in Figure 6.1. The pdfs when $n = 10, 25$ and 100 are also shown there. For $n = 4$ the distribution is relatively wide, but as n increases it gets more concentrated. Note that in the case of the binomial we may write $Y = \sum_{i=1}^n X_i$, so that $Y \sim B(n, p)$ and then $\bar{X} = Y/n$. Using the binomial formula $V(Y) = np(1-p)$ (see page **??**) together with the general formula $V(aY) = a^2V(Y)$ (see Equation (3.6)) we get $\sigma_{\bar{X}} = \sqrt{p(1-p)/n}$.

6.1.2 Random sequences may converge according to several distinct criteria.

In discussing the large-n behavior of a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ we need a formalism for two kinds of statements. First, we want to be able

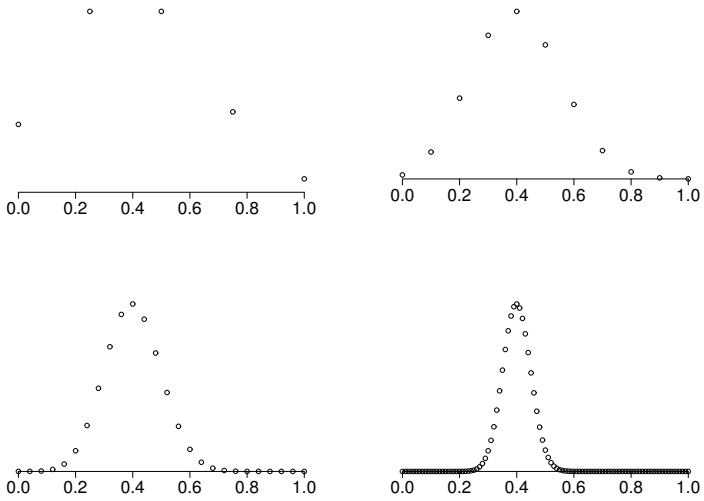


Figure 6.1: *The pdf of the binomial mean $\bar{X}$ when $p = .4$ for four different values of n. As n increases the distribution becomes concentrated ($\sigma_{\bar{X}}$ becomes small), with the center of the distribution getting close to $\mu_X = .4$ (the LLN). In addition, the distribution becomes approximately normal (the CLT).*

to say that the distribution of X_n is approximately of a particular form. We do this by examining the cdfs. Suppose that the variables $X_1, X_2, \ldots, X_n, \ldots$ have corresponding cdfs $F_1(x), F_2(x), \ldots, F_n(x), \ldots$ and suppose further that the particular distribution that we want to consider an approximating distribution has cdf $F(x)$. We may then formalize the approximation by giving a precise meaning to the expression $F_n(x) \approx F(x)$ for n large, meaning that $F_n(x)$ is approximately equal to $F(x)$ for n large. We do this using limits. Recall that a sequence of numbers x_n, for $n = 1, 2, \ldots$, converges to x if for every $\epsilon > 0$ we have $|x_n - x| < \epsilon$ for all sufficiently large n. This is written $\lim_{n\to\infty} x_n = x$.

Definition Suppose $X_1, X_2, \ldots,$ is a sequence of random variables and F_n is the cdf of X_n. We say that X_n *converges in distribution* to a continuous

random variable X with cdf F if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x. More generally, X_n converges in distribution to a random variable X with cdf F (which may or may not be continuous) if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x at which F is continuous.

The second kind of statement we want to make has to do with the case in which the sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ gets progressively closer to a number, i.e., a fixed constant c rather than having some probability distribution. This is needed for the LLN. We may think of the constant as a probability distribution that has collapsed down to a point: we say that a random variable X is *degenerate*, meaning that it is identically equal to a constant c, when $P(Y = c) = 1$. In this situation the cdf of X is $F(x) = 0$ for $x < c$ and $F(x) = 1$ for $x \geq c$.

Definition Suppose $X_1, X_2, \ldots,$ is a sequence of random variables and F_n is the cdf of X_n. We say that X_n *converges in probability* to c if X_n converges in distribution to the degenerate random variable X for which $P(X = c) = 1$.

The notion of convergence in probability is more general than the defintion above indicates, but we do not need the general definition. There are also two stronger notions of convergence, convergence in quadratic mean and convergence with probability one—but again we do not need these here.

Details: In applying convergence in probability, the criterion that is used is the following.

Theorem A sequence $X_1, X_2, \ldots$ converges in probability to c if and only if for every $\epsilon > 0$, $P(|X_n - c| > \epsilon) \to 0$ as $n \to \infty$.

Proof: This involves straightforward manipulations using the definition. The details are omitted. □

6.2 The Law of Large Numbers

6.2.1 As the sample size n increases, the sample mean converges to the theoretical mean.

The LLN is an accessible result, in the sense that its statement may be understood without advanced mathematics. The proof is not especially difficult, and we include it here, but we will regard it as an inessential detail.

Theorem: The Law of Large Numbers If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables having a distribution with mean μ_X and standard deviation σ_X, then $\bar{X}$ converges in probability to μ_X.

The form of the LLN given here is sometimes called the "weak" law of large numbers. The strong law instead says that convergence occurs with probability 1. However, considerably more machinery is needed in order to say this in precise mathematical terms. Intuitively, "with probability 1" means that the convergence is certain to occur.

The proof will require the following lemma.

Lemma (Markov's Inequality) Let Y be a positive random variable on (A, B) with $\mu_Y = E(x) < \infty$. Then for any positive α,

$$P(Y > \alpha) < \frac{\mu_Y}{\alpha}.$$

Proof of Lemma: Let us assume that Y is continuous. We have

$$P(Y > \alpha) = \int_{\alpha}^{B} f_Y(x)dy$$

and

$$\alpha \int_{\alpha}^{B} f_Y(x)dy \leq \int_{\alpha}^{B} y f_Y(x)dy.$$

Combining these, and continuing, we then have

$$\begin{aligned} \alpha P(Y > \alpha) &\leq \int_{\alpha}^{B} y f_Y(x) dy \\ &\leq \int_{A}^{\alpha} y f_Y(x) dy + \int_{\alpha}^{B} y f_Y(x) dy \\ &= \int_{A}^{B} y f_Y(x) dy = E(x). \end{aligned}$$

The case in which Y is not continuous may be handled by an analogous argument. □

Proof of Theorem: We need to show that for any positive ϵ we may find n sufficiently large that $P(|\bar{X} - \mu_X| > \epsilon)$ becomes arbitrarily close to 0. We have $P(|\bar{X} - \mu_X| > \epsilon) = P((\bar{X} - \mu_X)^2 > \epsilon^2)$. Let $Y = (\bar{X} - \mu_X)^2$, note that $E(Y) = \sigma_X^2/n$, and apply the Lemma to get

$$P(|\bar{X} - \mu_X| > \epsilon) < \frac{\sigma^2}{\epsilon^2 n}.$$

This shows that for sufficiently large n, $P(|\bar{X} - \mu_X| > \epsilon)$ becomes arbitrarily close to 0. □

6.2.2 The empirical cdf converges to the theoretical cdf.

We introduced the empirical cdf $\hat{F}_n(x)$ in Section **??** and noted there that, for large n, it approximates the cdf $F_X(x)$ and illustrated the phenomenon in Figure 3.8. We now relate this behavior to the LLN.

In the proof we need the following definition: for a random variable X, we let the *indicator variable* $I_{\{X \leq x\}}$ be 1 if $X \leq x$ and 0 otherwise.

Theorem If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables then, for every x, $\hat{F}_n(x)$ converges in probability to $F(x)$.

Proof: Another way to think about $\hat{F}_n(x)$ is that it counts the number of random variables X_i in the random sample $X_1, \ldots, X_n$ for which $X_i \leq x$,

and then divides by n. This is the same thing as adding $1/n$ for each of the X_i variables that are less than x. Mathematically, we express this counting operation using indicator variables. Consider a sequence $X_1, X_2, \ldots$ of i.i.d. random variables with cdf $F(x)$. We may write the empirical cdf in the form

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}.$$

We now use

$$\begin{aligned} E(I_{\{X_i \leq x\}}) &= 1 \cdot P(X_i \leq x) + 0 \cdot P(X_i > x) \\ &= P(X_i \leq x) = F(x) \end{aligned}$$

and apply the LLN. □

In addition to supplying the theoretical foundation for P-P and Q-Q plots, as discussed in Chapter 3, this result is also the starting point[2] for the *bootstrap* method of statistical inference, which we cover in Chapter 9.

6.3 The Central Limit Theorem

6.3.1 For large n, the sample mean is approximately normally distributed.

The LLN concerns only the large-sample tendency of $\bar{X}$ to get arbitrarily close to μ_X. The CLT describes the large-sample probability distribution of $\bar{X}$. Actually, we are speaking a bit loosely here: the LLN says that the distribution of $\bar{X}$ becomes degenerate at μ_X; to get fluctuations that are described, approximately, by a normal distribution we have introduce rescaling. Instead of $\bar{X}$, the CLT describes the behavior of the random sequence of variables Z_n, in which $\bar{X}$ is standardized by subtracting its mean and standard deviation (the standard deviation of $\bar{X}$ being $\sigma_X/\sqrt{n}$).

[2] Actually, a stronger result is needed: the convergence is uniform in the sense that $\sup_x |\hat{F}_n(x) - F(x)| \to 0$.

The Central Limit Theorem: Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables having mean μ_X and standard deviation σ_X, and let $Z_n = \sqrt{n}(\bar{X} - \mu_X)/\sigma_X$. Then Z_n converges in distribution to a normal random variable having mean 0 and variance 1.

The proof of this result is given in Appendix B. The effects of the LLN and CLT are illustrated in Figure 6.1. For $n = 4$ the distribution of $\bar{X}$ does not look very close to normal. However, as n increases the distribution of $\bar{X}$ gets more tightly concentrated near the mean $\mu_X = .4$ (a consequence of the LLN) and it looks more and more normal (the CLT).

What we've just done is looked at the distribution of $\bar{X}$ for Bernoulli trials for several values of n with $p = .4$. The distribution of $n\bar{X}$ is binomial and the picture of its distribution would look just like the pictures we had for the distribution of $\bar{X}$ except that the x-axis would be multiplied by n. In particular, as n gets large we see that the distribution looks normal. This effect of the CLT may be considered an explanation for the normal approximation to the Binomial.

In fact, there are much more general versions of the CLT. It is worth stating one of these in a somewhat vague form.

Roughly speaking, if $X_1, X_2, \ldots, X_n$ are independent random variables, possibly having different distributions but with no individual X_i making a dominant contribution to the mean $\bar{X}$, then for n sufficiently large, the distribution of $\bar{X}$ is approximately normal with mean $E(\bar{X})$ and standard deviation $\sqrt{V(\bar{X})}$.

The "no dominant contribution" phrase may be made precise as the *Lindeberg condition*, and the CLT then follows.[3] This version of the CLT helps to explain why the normal distribution arises so often in statistical theory, and also why it seems to fit, at least crudely, so many observed phenomena. It says that whenever we average a large number of small independent[4] effects, the result will be approximately normally distributed.

[3]See Section 27 of Billingsley, P. (1995) *Probability and Measure*, Third Edition, Wiley.

[4]There are also versions of the CLT for non-independent variables, though they are considerably more complicated. See Francq, C. and Zakoiän, J.-M. (2005) A central limit

A detail: Among all distributions having mean μ and standard deviation σ, the $N(\mu, \sigma^2)$ distribution is the most disorderly possible, in the sense of having maximal entropy, defined in Equation (4.15). The CLT says that as the sample size gets very large the distribution of the sample mean becomes as disorderly as possible. This characterization offers another way to state and prove the CLT. See Madiman, M. and Barron, A.R. (2007) Generalized entropy power inequalities and monotonicity properties of information, *IEEE Trans. Information Theory* 53: 2317–2329.

While the Central Limit Theorem involves the sample mean, it drives the large-sample behavior of most statistics: a statistic derived from a sample may usually be written, approximately, as some function (possibly a complicated function) of a sample mean, and that usually produces approximate normality of the statistic itself. An important example is the Maximum Likelihood Estimator, discussed in Chapters 7 and **??**.

6.3.2 For large n, the multivariate sample mean is approximately multivariate normal.

The multivariate version of the CLT is analogous to the univariate CLT. We begin with a set of multidimensional samples of size n: on the first variable we have a sample $X_{11}, X_{12}, \ldots, X_{1n}$, on the second, $X_{21}, X_{22}, \ldots, X_{2n}$, and so on. In this notation, X_{ij} is the jth observation on the ith variable. Suppose there are m variables in all, and suppose further that $E(X_{ij}) = \mu_i$, $V(X_{ij} = \sigma_i^2$, and $Cor(X_{ij}, X_{ik}) = \rho_{jk})$ for all $i = 1, \ldots, m$, $j = 1, \ldots, m$, and $k = 1, \ldots, m$. As before, let us collect the means into a vector μ and the variances and covariances into a matrix Σ. We assume, as usual, that the variables across different samples are independent. Here this means X_{ij} and X_{hk} are independent whenever $i \neq h$. The sample means

$$\bar{X}_1 \;=\; \frac{1}{n}\sum_{j=1}^{n} X_{1j}$$

theorem for mixing triangular arrays of variables whose dependence is allowed to grow with the sample size. *Econometric Theory*, 21: 1165–1171.

$$\begin{aligned} \bar{X}_2 &= \frac{1}{n}\sum_{j=1}^{n} X_{2j} \\ &\vdots \\ \bar{X}_m &= \frac{1}{n}\sum_{j=1}^{n} X_{mj} \end{aligned}$$

may be collected in a vector

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_m \end{pmatrix}.$$

Multivariate Central Limit Theorem: Suppose $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ are means from a set of m random samples of size n, as defined above, with the covariance matrix Σ being positive definite. For any m-dimensional vector w define

$$Z_n(w) = \sqrt{n} w^T \Sigma^{-\frac{1}{2}} (\bar{X} - \mu).$$

Then for every nonzero m-dimensional vector w, $Z_n(w)$ converges in distribution to a normal random variable having mean 0 and variance 1.

More loosely, the multivariate CLT says that $\bar{X}$ is approximately multivariate normal with mean μ and variance matrix $\frac{1}{n}\Sigma$. As in the univariate case, there are much more general versions of the multivariate CLT.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 7

Estimation and Uncertainty

7.1 Fitting Statistical Models

The examples in previous chapters, involving experimental settings ranging from human and animal behavior, to neuroimaging, EEG and EMG, neural spike trains, and *in vitro* recording, have all illustrated the way statistical models describe regularity and variability of neural data. All of these models involve free parameters. In Example 1.2.2 we reviewed the use of least squares in demonstrating an approximately linear relationship between conduction velocity and nerve diameter. Least squares is easy to understand and often works well for models of the form

$$Y_i = f(x_i) + \epsilon_i.$$

But what about other situations? In Example 3.2.2, for instance, we displayed fits of $Gamma(\alpha, \beta)$ distributions to histograms of ion-channel opening durations, but we did not say how the parameters α and β were chosen. A näive approach to the problem of using the data to determine suitable

values of parameters might propose a particular method and argue for it on intuitive grounds. According to the doctrine of statistics, however, principles may be introduced and used in analyzing the performance of alternative methods. By demonstrating the properties of solutions under general conditions, statistical theory brings coherence to an otherwise bewildering array of disparate problems. In this chapter, together with Chapters **??** and 9, we present the key ideas.

We start with a traditional, though somewhat artificial, separation of two aspects of the fitting problem that are intimately connected in practice: estimation of parameters and assessment of uncertainty. In Section 7.2 we formalize the process of estimation and then give two alternative methods, the *method of moments* and *maximum likelihood (ML)*. In the 1920s R.A. Fisher proposed maximum likelihood and demonstrated that it is optimal quite generally for large sample sizes. Fisher also showed how uncertainty about the answer can be assessed, and an alternative perspective was provided at about the same time by Harold Jeffreys using Bayes' Theorem. It took roughly 50 more years to refine the early concepts to its full-fledged modern incarnation and, in fact, new variants of algorithms continue to be developed so that it may be applied to ever more complicated situations. ML estimation is the dominant approach to parameter estimation in contexts where finitely-many parameter values[1] completely specify the statistical model. We will review the basic properties and uses of ML estimation in Chapter **??**. We discuss alternative, non-parametric methods of estimation in Chapter **??**.

In Section **??** we discuss confidence intervals. In Chapter 1, page 17, we described the use of a confidence interval to assess the uncertainty associated with responses of patient P.S. when forced repeatedly to choose between pictures of burning and non-burning houses: we noted that an approximate 95% confidence interval for her propensity to choose the non-burning house was (.64,1.0) and we concluded it was not very likely that she was choosing them with equal probabilities (a propensity of .5); instead, she apparently saw the two complete pictures without conscious awareness of processing their left ends, which is where the fire appeared. As a data-analytic tool, confidence intervals have become straightfoward to use in many, varied situations. We treat several simple yet important problems in Section 7.3 and

[1]From the point of view of the mathematical theory, a nonparametric method does not eliminate the parameters but rather makes them infinite dimensional.

supplement with more general methods in Chapters **??** and 9. As one thinks harder about interpretation, the subject gets murkier. We review the issues in Sections 7.3.8 and 7.3.9. On the other hand, confidence intervals are fundamental to statistical practice and, from a contemporary standpoint, they seem very natural. Seen in historical context, the introduction of confidence intervals by Jerzy Neyman in the 1930s was quite ingenious, and a giant leap forward.

One of the ways confidence intervals are found in conjunction with maximum likelihood is to apply the *bootstrap*, which is discussed in Chapter 9. As additional motivation for the discussion in this and subsequent chapters, here is a concrete example where these methods have been used in fitting a statistical model of mental processes.

Example 7.1.1 Visual Attention Experiments on visual attention often study the ability of subjects to see and remember multiple objects that are exposed to them for a very short time. Following Sperling (1967), Bundesen and colleagues developed a quantitative theory of visual attention (Bundesen, 1998) (Bundesen, C. (1998) A computational theory of visual attention, *Phil. Trans. Royal Soc. London, B*, 353: 1271–1281.) according to which, objects in the visual field are compared with representations in visual memory, and if the comparison is completed prior to the end of visual exposure, the object is recognized. In this theory the time taken to process and store an object identity is a random variable. For object i call this random variable X_i. The processing is considered to begin after a latency of length t_0, so that if t is the total time an object is displayed then the ith object is recognized if $X_i \leq t - t_0$. Bundesen assumed $X_i \sim Exp(\lambda_i)$. Letting $f_i(x)$ and $F_i(x)$ be the $Exp(\lambda_i)$ pdf and cdf, for exposure of length $x = t - t_0$, $F_i(t - t_0)$ is the probability of object recognition success and $1 - F_i(t - t_0)$ is the probability of object recognition failure. Suppose S is the stimulus set and let R denote some particular subset of objects that are recognized. If the subject's memory capacity is not exceeded, and if recognition of object i is independent of recognition of all other objects (and this is true for every i), then the probability that the subject will recognize all objects in R, and fail to recognize all objects not in R (i.e., fail to recognize those in the complement,

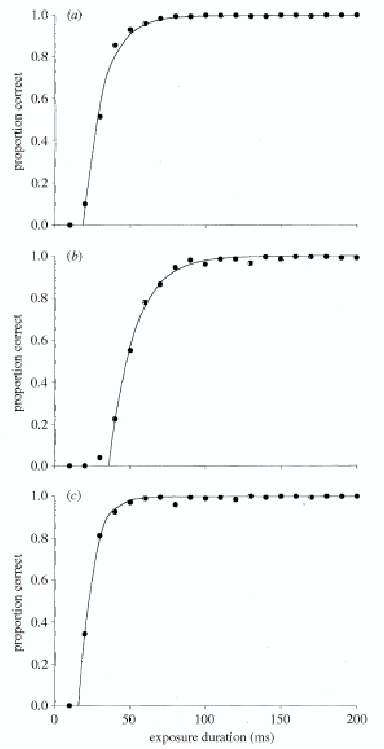


Figure 7.1: *Data from three subjects, together with fits of a model for probability of letter identification as a function of exposure duration. From Bundesen (1998).*

which may be written $S - R$), is given by

$$P_S(R) = \prod_{i \in R} F_i(t - t_0) \prod_{j \in S-R} (1 - F_j(t - t_0)).$$

This model has several unknown parameters (the encoding rates λ_i, the latency t_0, and the memory capacity) which must be determined in order to compute the probabilities and compare them to data. Figure 7.1 displays fits of the model to data from three subjects. The model fitting was performed by the method of maximum likelihood, and uncertainties associated with each of the parameters of interest may be obtained by bootstrap methods. (See Kullingsbaek, S. (2006) Modeling visual attention. *Behavioral Research Methods*, 38: 123–133.) □

7.2 The Problem of Estimation

In order to fit a model to data, a parameter or set of parameters needs to be determined. Following a convention in the statistical literature, we use θ to denote a generic parameter. In much of our initial discussion we will focus on the case of a single, scalar parameter, but in most real-world problems θ becomes a vector. For example, in fitting a $Gamma(\alpha, \beta)$ model we would be taking $\theta = (\alpha, \beta)$ and we would speak of "the parameter" θ in place of "the parameters" α and β. The problem of estimation is to determine a method of estimating θ from the data. To constitute a well-defined method we must have an explicit procedure (a formula or a rule) by which a set of data values $x_1, x_2, \ldots, x_n$ produces an estimate. We consider an *estimator* to have the form $T = T(X_1, X_2, \ldots, X_n)$, i.e., the estimator is a random variable derived from the random sample. The properties of an estimator may be described in terms of its probabilistic behavior.

Before presenting the method of moments and maximum likelihood, we need to make two comments on notation. First, when we write $T = T(X_1, \ldots, X_n)$ we are using capital letters to indicate clearly that we are considering the estimator to be a random variable, and the terminology distinguishes the random "estimator" from an "estimate," the latter being a value the estimator takes. Nonetheless, neither we nor others in the literature are systematically careful in making this distinction; it is important conceptually, but some sloppiness is tolerable. Second, we often write θ^* or $\hat{\theta}$ the value of an estimator, so we would have, say, $T = \hat{\theta}$. The latter notation, using $\hat{\theta}$ to denote an estimate, or an estimator, is very common in the statistical literature. Sometimes, however, $\hat{\theta}$ refers specifically to the maximum likelihood estimator (MLE). This is another potential source of confusion, which the context should clarify.

7.2.1 The method of moments uses the sample mean and variance to estimate the theoretical mean and variance.

In order to fit a probability distribution to a set of data, an especially simple idea is to equate the theoretical mean and variance to the sample mean and variance. The result is a *method of moments* estimator.

Illustration: Fitting a gamma distribution We noted above that the mean and variance of a $Gamma(\alpha, \beta)$ random variable are

$$\begin{aligned} \mu &= \frac{\alpha}{\beta} \\ \sigma^2 &= \frac{\alpha}{\beta^2}. \end{aligned}$$

We may solve these for β and α, dividing the first equation by the second, to get

$$\begin{aligned} \beta &= \frac{\mu}{\sigma^2} \\ \alpha &= \frac{\mu^2}{\sigma^2}. \end{aligned}$$

We then substitute $\bar{x}$ and s^2 for μ and σ^2 to get the method of moments estimator:

$$\begin{aligned} \beta^* &= \frac{\bar{x}}{s^2} \\ \alpha^* &= \frac{\bar{x}^2}{s^2}. \end{aligned}$$

□

The method of moments is, in some cases, like the gamma, quite easy to apply. In principle, higher-order moments could be used (e.g., $E(\sum(X_i - \mu)^3)$ could be equated to the sample analogue), though this is rare in practice.

7.2.2 The method of maximum likelihood maximizes the likelihood function, which is defined up to a multiplicative constant.

To introduce maximum likelihood estimation, let us begin by framing the estimation problem concretely, using the binomial, and let us write the binomial pdf in the form

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

which was previously denoted by $f(x) = P(X = x)$, with p replacing θ. Here the notation $f(x|\theta)$ is used to imply that we are examining the pdf of X given the value of θ. The binomial pdf describes the probabilities to be attached to varying possible values $X = x$ for a given fixed value of θ. That is, once we plug in a value of θ we have completely determined the pdf for all values of x. The problem of estimation, however, attempts to find a sensible guess at θ given that $X = x$ has been observed. It thus reverses the situation: instead of assuming a value for θ and finding values of x, we must assume a value of $X = x$ and come up with a value of θ. In this sense, it is involves an *inverse* or *inductive* form of reasoning. The method of maximum likelihood chooses the value $\hat{\theta}$ of θ that assigns to the observed data x the highest possible probability:

$$f(x|\hat{\theta}) = \max_{\theta} f(x|\theta).$$

In the binomial problem we will, below, show that $\hat{\theta} = x/n$. In other words, maximum likelihood estimates the theoretical proportion (or propensity) θ by the observed proportion x/n.

> *A detail:* Why do we call θ a theoretical proportion? We have that X/n is the mean of n Bernoulli trials, each having probability θ of being 1. By the law of large numbers
>
> $$\frac{X}{n} \xrightarrow{P} \theta$$
>
> so that θ is, roughly speaking, the proportion of 1s observed in infinitely many trials. In this sense we can say that θ is a theoretical proportion.

To understand the maximum likelihood idea better we consider what the pdf $f(x|\theta)$ tells us about the various possible values of θ. To do this we *invert* its functionality by thinking of $f(x|\theta)$ as a function of θ rather than of x. That is, having observed $X = x$, we fix x in the pdf $f(x|\theta)$ and then consider how each different choice of θ produces a different probability $f(x|\theta)$. We do not regard this, at all, as an intuitively obvious thing to do. It becomes much more intuitive from a Bayesian point of view, as we mention in Section 7.3.8. For now we ask the reader to bear with us and make sure to understand what we mean.

The distinction we are trying to draw here, between $f(x|\theta)$ as a function of x and $f(x|\theta)$ as a function of θ is illustrated in Figure 7.2, which displays the binomial pdf viewed both ways when $n = 4$: first (on the left) as a function of x when $\theta = .5$ and then (on the right) as a function of θ when $x = 1$. First, when $\theta = .5$, the pdf is evaluated for 5 possible values of x: 0, 1, 2, 3, 4 ,5. These are all the possible values of x. (When $n = 4$, these are all the possible values of x regardless of the value of θ, as long as it is a permissible value, i.e., it is between 0 and 1, which is often written $\theta \in (0, 1)$.) When $x = 1$ and the pdf is regarded as a function of θ there is a whole continuum of possible values of θ in (0,1). In the second part of the figure we set $x = 1$ and the pdf is evaluated for 99 values of θ, among all the possibilities for $\theta \in (0, 1)$. There is nothing of interest about the contrast between the picture on the left and the picture on the right *except* that the two representations are conceptually different.

When the pdf is considered as a function of the parameter θ rather than the values x of the random variable, it is called *the likelihood function*. We will denote it by $L(\theta)$. (Other notations are variations on this; all authors use some form of the letter "L.") The *maximum likelihood estimator (MLE)* is the value of θ that maximizes $L(\theta)$. We will denote it by $\hat{\theta}$. There is a potential for confusion, however, because the "hat" often denotes a generic estimator, which need not be the MLE. We will try to be careful about this, but in the literature, as here, the context should clarify what is meant by $\hat{\theta}$.

So far, we have discussed the pdf and likelihood based on a single (scalar) random variable. The concept generalizes immediately to vectors. In fact, one would typically have a vector of observed data $x = (x_1, \ldots, x_n)$ that has a joint pdf $f(x|\theta) = f(x_1, \ldots, x_n|\theta)$. In the subsequent parts of this chapter

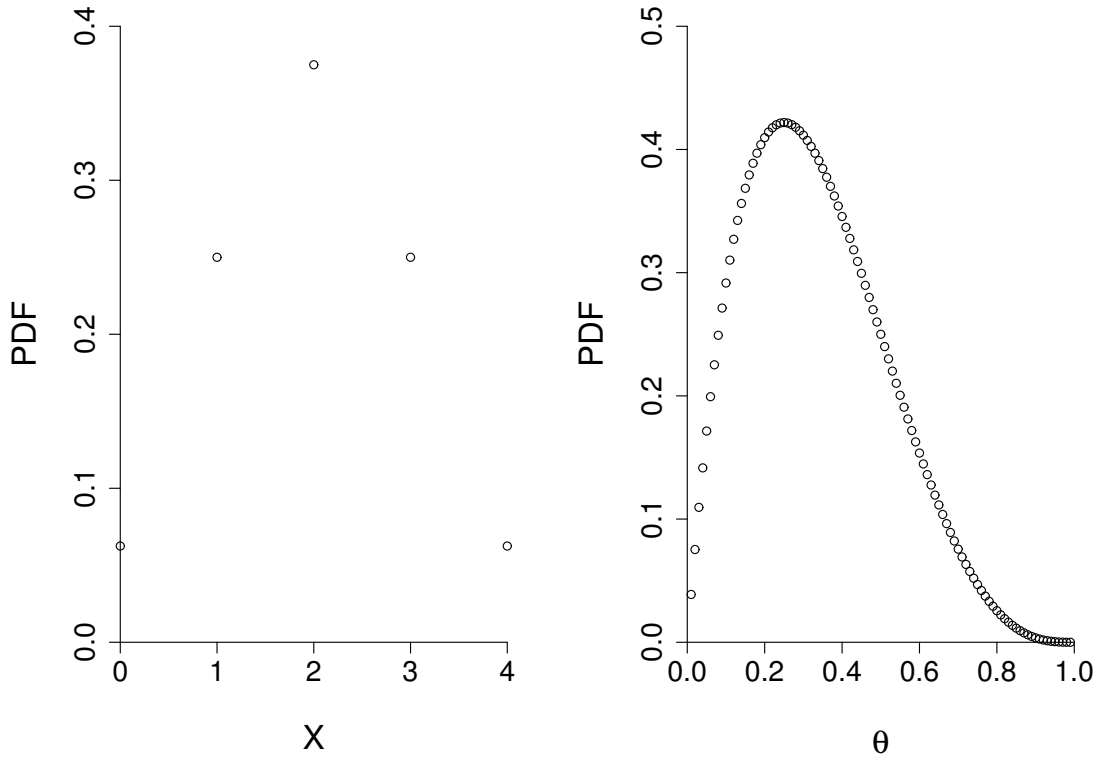


Figure 7.2: *Comparison of pdf $f(x|\theta)$ when viewed as a function of x with θ fixed at $\theta = .5$ (on left) or of θ with x fixed at $x = 1$ (on right). On the right-hand side, the pdf is evaluated for 99 equally-spaced values of θ from .01 to .99.*

we will take x to be a vector, often corresponding to a sample of data, and regard as simply a special case any application when it becomes a scalar.

Note that the value of θ maximizing $L(\theta)$ is the same as the value of θ maximizing $c \cdot L(\theta)$ for any positive constant c. We therefore always understand the likelihood function to be defined only up to a positive constant, and may write

$$L(\theta) \propto f(x|\theta)$$

choosing the constant for arithmetic convenience.

Illustration: Binomial likelihood We may write the binomial likelihood function as

$$L(\theta) = \theta^x (1-\theta)^{n-x}.$$

Here, in going from the pdf to the likelihood function we have omitted the factor $\binom{n}{x}$ because it does not involve θ. □

From the second part of Figure 7.2 it is apparent that when $x = 1$ the MLE is $\hat{\theta} = .25$, which is an instance of the formula $\hat{\theta} = x/n$. To find the maximum, more generally, some combination of analytic (calculus-based) and numerical methods may be used. In the simplest problems, analytic methods suffice. In either case, however, it is easiest to begin by taking logs, because the value maximizing $\log L(\theta)$ is the same as the value maximizing $L(\theta)$, and because the pdf typically has a product form which is thereby converted to a sum. Suitably enough, the log of the likelihood function is called the *loglikelihood function*. We denote it here by $\ell(\theta)$:

$$\ell(\theta) = \log L(\theta).$$

Note that in writing a formula for $\ell(\theta)$ we may omit any additive terms that do not involve θ, because these become multiplicative constants in $L(\theta)$ and do not affect the maximization.

Illustration: Binomial MLE. To derive the general form $\hat{\theta} = x/n$ for the MLE we begin with the loglikelihood function

$$\ell(\theta) = x \log \theta + (n-x) \log(1-\theta)$$

where we have omitted the term $\log \binom{n}{x}$ because it does not involve θ. To maximize this function we set its derivative equal to zero and solve:

$$0 = \ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

so that

$$x(1-\theta) = (n-x)\theta$$

which gives the solution

$$\hat{\theta} = \frac{x}{n}.$$

It is also easy to check that $\ell''(\hat{\theta}) < 0$, which verifies that $\hat{\theta}$ is a maximum. □

Illustration: Normal MLE. Suppose we have a sample $x_1, \ldots, x_n$ from a $N(\theta, \sigma^2)$ distribution, where σ is known and the problem is to estimate θ. The ith normal density has pdf

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i-\theta)^2}{2\sigma^2})$$

and the random variables $X_1, \ldots, X_n$ are independent, so the joint pdf is

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i-\theta)^2}{2\sigma^2}). \end{aligned}$$

From this, loglikelihood function is

$$\begin{aligned} \ell(\theta) &= -\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2} \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - 2x_i\theta + \theta^2 \\ &= -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta) + R \end{aligned}$$

where R is a term that does not involve θ. Because the loglikelihood function is defined only up to an additive constant, we have

$$\ell(\theta) = -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta). \tag{7.1}$$

Setting its derivative equal to 0 we obtain

$$0 = \frac{n}{2\sigma^2}(\theta - \bar{x})$$

so that $\hat{\theta} = \bar{x}$. □

7.3 Confidence Intervals

7.3.1 For scientific inference, estimates are useless without some notion of precision.

In Example 1.2.1 P.S. preferred the non-burning house about 80% of the time. However, this information by itself is not enough to say anything useful about her preferences: 4 out of 5 trials would also provide a preference for the non-burning house 80% of the time, as would 80 out of 100 trials. But 4 out of 5 is far different than 80 out of 100. With 100 trials we could say pretty accurately what her preference rate is, while with 4 out of 5 it would not be clear that this is different than guessing. In scientific contexts, an estimate is useless unless we have some idea how accurate it is. Although one need not always drag around a standard error or confidence interval, and it is common to speak in terms of estimates without stating uncertainty, this convention assumes the uncertainty to be small relative to the size of the effects under discussion. It is important to include a statement of uncertainty whenever the uncertainty is non-negligible. In our judgment, inclusion of uncertainty should be considered the rule rather than the exception. We keep returning to Example 1.2.1 precisely because 14/17 is intermediate between the obvious situations where one doesn't need uncertainty (80/100) and where the estimate is hopelessly uncertain (4/5). Even a trained statistician might have some trouble saying correctly where 14/17 falls in this continuum without doing some calculations. So let us look at $14/17 = .82$ and ask, "How much error is there in this estimate?"

At first glance it appears impossible to answer this question: if we knew θ then the error in estimating it with $\hat{\theta}$ would be $\hat{\theta} - \theta$; but we *don't know* θ, which is why we are trying to estimate it. Nonetheless, even though we can not say precisely how big the error is, we can use probability and say something about *the likely magnitude of error*. This is usually quantified with the *standard error*. The idea is simple enough. Every estimator $T = T(X_1, X_2, \ldots, X_n)$ exhibits variation. That is, if we were to examine T across many different samples we would get many different values. Because $X_1, \ldots, X_n$ are random variables having some probability distribution, T is a random variable. The simplest summary of the magnitude of the variation

of T is its standard deviation

$$\sigma_T = \sqrt{V(T)}.$$

This is almost, but not quite, the standard error of T. To get the standard error one additional step is taken. We illustrate it in the context of the blindsight example.

Example 1.2.1 (Continued) Let $Y \sim B(n, p)$ and note that the usual estimator of p is sample proportion $T = \hat{p} = Y/n$. Because $V(Y) = np(1-p)$ we have $V(T) = p(1-p)/n^2$. Thus, we have the formula

$$\sigma_T = \sqrt{\frac{p(1-p)}{n}}. \tag{7.2}$$

The formula in Equation (7.2) quantifies the variation we can associate with the observed proportion $\hat{p} = 14/17 = .824$. However, we can not compute a numerical value for σ_T from Equation (7.2) because we do not know what value of p to use. The obvious solution is to substitute $\hat{p}$ for p in Equation (7.2). When we do this we obtain the *standard error* for the binomial proportion

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7.3}$$

Applying this to the data from P.S. we get

$$SE = \sqrt{\frac{\frac{14}{17}(1-\frac{14}{17})}{17}} = .092.$$

We then typically write the estimate in the form $.824 \pm .092$, with the $\pm$ indicating that the likely variability in the estimate is .092. □

The general procedure for computing the standard error is, in essence, the same as in the binomial case. To emphasize the substitution of the estimated parameter for the unknown paramater we define the *standard error* of an estimator T to be of the form

$$SE(T) = \sqrt{\hat{V}(T)} \tag{7.4}$$

with the hat on V indicating that we have estimated the variance. In fact, definition (7.4) is very general in the sense that it does not specify *how* we

estimate the variance. As we will see in Chapters **??** and 9, several different methods are used to obtain variance estimates. We have used T in (7.4) to emphasize that it is a random variable, but in an alternative notation we use more often we may rewrite (7.4) as

$$SE(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}.$$

One note on terminology: the term "standard error" is sometimes used to refer to the standard error of the mean, as in Equation (7.10), which is a special case of (7.4).

It is very common practice to report an estimate together with its standard error in the form

$$\hat{\theta} \pm SE(\hat{\theta}).$$

This gives a simple, rough sense of how accurate the estimate is. A more refined statement comes from the use of a confidence interval. In general terms, a 95% *confidence interval (CI)* for a parameter θ is an interval of the form (L, U) (L for lower, U for upper), where $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ are random variables derived from the data and

$$P(L < \theta < U) = .95. \tag{7.5}$$

This rather abstract definition becomes clear by examining particular problems, as we do below. In words, Equation (7.5) says that if θ were the value of the unknown parameter, the probability that the interval would include this unknown value is 95%. The probability .95 is the *level of confidence* associated with the interval (L, U).

In many applications an estimator $\hat{\theta}$ follows an approximately normal distribution. This is a tremendous simplification, for it allows the 2/3–95% rule to be used. As we will see, this leads to a very general form for an *approximate* 95% confidence interval (L, U) by taking $L = \hat{\theta} - 2SE(\hat{\theta})$ and $U = \hat{\theta} + 2SE(\hat{\theta})$, that is,

$$\text{approx. } 95\% \text{ CI } = (\hat{\theta} - 2SE(\hat{\theta}), \hat{\theta} + 2SE(\hat{\theta})). \tag{7.6}$$

The ingeneously simple construction that drives confidence intervals is most easily understood in the case of estimating the mean of a normal distribution.

7.3.2 Estimation of a normal mean is a paradigm case.

Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, and that the value of σ is known. Here, for notational ease, we drop the subscript X from μ and σ. Note that μ may be estimated by the sample mean $\bar{X}$ with standard error $\mathrm{S}E(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Theorem If $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, with the value of σ known, then the interval $(\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))$ is a 95% CI for μ.

Proof: We must show that

$$P(\bar{X} - 2 \cdot SE(\bar{X}) \leq \mu \leq \bar{X} + 2 \cdot SE(\bar{X})) = .95. \tag{7.7}$$

We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

so that

$$P(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}) = .95. \tag{7.8}$$

We observe

$$\begin{aligned} \mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}} &\iff |\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}| \leq 2 \\ &\iff \bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Therefore, (7.8) gives (7.7). □

The beauty of confidence lies in the simple manipulations, given above, that allow us to reason from (7.8) to (7.7). We take a description of variation, $X \sim N(\mu, \frac{\sigma^2}{n})$, and convert it to a quantitative inference about the value of the unknown parameter μ.

7.3.3 For non-normal observations the Central Limit Theorem may be invoked.

Now suppose $X_1, \ldots, X_n$ form a sample from a distribution with mean μ and standard deviation σ, with the distribution not necessarily normal. For simplicity, suppose again that σ is known.

By the CLT we have

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{D} N(0, 1).$$

We now apply the same manipulations used in deriving (7.8). We have

$$P(|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}| \leq 2) \approx .95$$

and, in turn, this is equivalent to

$$P(\bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}) \approx .95. \tag{7.9}$$

Therefore, for n sufficiently large, Equation (7.9) provides an approximate 95% CI. Written slightly differently, an approximate 95% CI is given by $\bar{X} \pm 2 \cdot SE(\bar{X})$, where $SE(\bar{X}) = \sigma/\sqrt{n}$. The important point here is that we do not require the distribution of the data to be normal, yet we still get a quantitative inference based on asymptotic normality of the mean because of the CLT.

7.3.4 A large-sample confidence interval for μ is obtained using the standard error $s/\sqrt{n}$.

In Sections 7.3.2 and 7.3.3 we assumed σ was known. This was for purely pedagogical purposes. In practice, σ is almost always unknown and, as a consequence, we don't have a value to plug in when we want to calculate $SE = \sigma/\sqrt{n}$. The way to proceed, however, is pretty clear. As in the binomial standard error formula (7.3), we simply replace σ with an estimate, the obvious estimate being the sample standard deviation s. In the scenario

envisioned in Section 7.3.3, with σ unknown we replace it with s in $\sigma/\sqrt{n}$ to get the *standard error of the mean*,

$$SE(\bar{x}) = \frac{s}{\sqrt{n}} \tag{7.10}$$

and from this we obtain a more practical version of (7.9) for our approximate 95% CI. Because we state the result in terms of probability, we replace the observed value s with its random-variable counterpart S.

Result If $X_1, \ldots, X_n$ is a random sample from a distribution having mean μ and standard deviation σ, and n is sufficiently large, then an approximate 95% CI for μ is given by $\bar{x} \pm 2 \cdot SE(\bar{x})$, where $SE(\bar{x})$ is given by (7.10), i.e., for n sufficiently large,

$$P(\bar{X} - 2\frac{S}{\sqrt{n}} < \mu < \bar{X} + 2\frac{S}{\sqrt{n}}) \approx .95. \tag{7.11}$$

This result follows from manipulations similar to those used in deriving (7.7) and (7.9). In establishing (7.9) we applied the CLT. The following theorem modifies the CLT used in Section 7.3.3 by replacing σ with S.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a distribution having mean μ and standard deviation σ, let S_n be the sample standard deviation calculated from $X_1, \ldots, X_n$, and let $Y_n = \sqrt{n}(\bar{X} - \mu)/S_n$. Then, as $n \to \infty$,

$$Y_n \xrightarrow{D} N(0, 1).$$

Details: In order to prove the theorem we first need two lemmas. In stating these results we replace S with S_n to indicate that we are dealing with a sequence of standard deviations calculated from a sequence of random samples, as $n \to \infty$.

Lemma 1 Let $X_1, \ldots, X_n, \ldots$ be i.i.d. sequence and let S_n be the standard deviation calculated from $X_1, \ldots, X_n$. Then S_n converges to σ in probability.

Proof: Let $Y_i = (X_i - \mu)^2$, so that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Note that $E(Y_i) = \sigma_X^2$. By the law of large numbers we have that $\bar{Y}$ converges to σ_X^2. Because $n/(n-1) \to 1$, we also have that $\frac{n}{n-1}\bar{Y}$ converges to σ in probability. But $S_n = \frac{n}{n-1}\bar{Y}$. □

Lemma 2 (Slutsky's Theorem) If U_n converges to c in probability and V_n converges to Y in distribution, then $U_n V_n$ converges to cY in distribution.

Proof: The proof of this result, while straightforward, involves quite a bit of detailed manipulation. We omit it. (See Bickel and Doksum (2001), Theorem A.14.9. Bickel, Peter J. and Doksum, Kjell A. (2001) *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1*, Prentice-Hall.) □

Proof of Theorem: By the CLT $Z_n = \sqrt{n}(\bar{X} - \mu)/\sigma$ converges in distribution to $N(0, 1)$. Applying Lemma 1 we have that S_n converges to σ in probability or, equivalently, σ/S_n converges to 1 in probability. Writing $U_n = \sigma/S_n$ and $V_n = Z_n$, and noting that Y_n defined in the statement of the theorem satisfies $Y_n = U_n V_n$, we may apply Lemma 2 to obtain the desired convergence in distribution. □

Example 3.2.1 (continued) On page 154 we considered spike counts from a motor cortical neuron across 60 trials, each spike count being recorded during a 600 millisecond interval. The mean spike count across the 60 trials was 13.63 spikes. Converting the counts to firing rates (by dividing by .6 seconds), we get a mean of 22.72 spikes per second and a standard deviation of 7.17 spikes per second. This gives a standard error of

$$SE = \frac{7.17}{\sqrt{60}} = .93.$$

We might then report the firing rate of this neuron, under the particular experimental condition, to be 22.72 ($\pm$.93) spikes per second. An approximate 95% confidence interval for the firing rate is then $(20.8, 24.6)$ spikes per second. □

The result is tremendously important in practice. However, it leaves open the question of how large the sample must be in order for the approximation

to be good, i.e., for the probability of coverage (the probability the interval will cover μ) to be nearly .95. Because we have the exact result in (7.7), this approximation tends to be good for moderate-size samples when the data are nearly normal. It may not be very good in moderate-size samples with strongly non-normal data. This is why it is important to check normality. The small-sample case is more problematic. We mention it again in Section 7.3.10.

7.3.5 Standard errors lead immediately to confidence intervals.

We now return to the general form for an approximate 95% CI given by (7.6) and derive it. First we consider the special case of the binomial probability p. Recall that if $X_1, \ldots, X_n$ are Bernoulli trials with probability p, and if $Y = \sum_{i=1}^{n} X_i$, then $Y \sim B(n, p)$. We have $Y/n = \bar{X}$, $E(X_i) = p$ and $V(X_i) = p(1-p)$ so the CLT gives

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \xrightarrow{D} N(0, 1). \tag{7.12}$$

By the $\frac{2}{3}$–95% rule this implies

$$P(-2 \leq \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \leq 2) \approx .95$$

and, multiplying through the inequalities by $\sqrt{\frac{p(1-p)}{n}}$, we have

$$P(\bar{X} - 2 \cdot \sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + 2 \cdot \sqrt{\frac{p(1-p)}{n}}) \approx .95.$$

If we replace p by $\hat{p}$ we get

$$P(\bar{X} - 2 \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + 2 \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}) \approx .95 \tag{7.13}$$

which is (7.6) for the binomial case, where the standard error is given by (7.3). The replacement of p with $\hat{p}$ in the standard error formula is analogous to

the replacement of σ with s in Section 7.3.4. The binomial case is sufficiently important that we state it formally, rewriting (7.13) in terms of $\hat{p}$, where $\hat{p} = \bar{X}$ so that $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Result If $Y \sim B(n, p)$ then p may be estimated by $\hat{p} = Y/n$ with standard error $\mathrm{S}E(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. For large n, an approximate 95% CI is given by

$$\hat{p} \pm 2 \cdot SE(\hat{p}),$$

meaning that for n sufficiently large we have

$$P(\hat{p} - 2 \cdot SE(\hat{p}) \leq p \leq \hat{p} + 2 \cdot SE(\hat{p})) \approx .95. \qquad (7.14)$$

Details: To justify the replacement of p with $\hat{p}$ we first note that the LLN gives us

$$\bar{X} \xrightarrow{P} p.$$

Then, by Slutsky's Theorem (page 182), $\bar{X}(1 - \bar{X})$ converges to $p(1 - p)$ in probability and

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{\bar{X}(1 - \bar{X})}} \xrightarrow{D} N(0, 1)$$

which gives (7.13).

To generalize this argument we consider the problem of estimating a parameter vector θ in some probability model using an estimator $T_n = T(X_1, \ldots, X_n)$. We have written the subscript n on T to indicate that we are examining its behavior as $n \to \infty$. Two things drove the derivation of (7.14) above. First, the CLT was invoked to produce the approximate normality of $\bar{X}$ according to (7.12) and, second, in the standard deviation $\sqrt{\frac{p(1-p}{n}}$, p was replaced by $\hat{p}$ (which was justified by the convergence of $\bar{X}$ to p in probability). If we assume these two phenomena apply then we obtain (7.6) according to the following theorem.

Theorem If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{(T_n - \theta)}{\sigma_{T_n}} \xrightarrow{D} N(0,1)$$

and $\hat{\sigma}_{T_n}$ satisfies

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then we have

$$\frac{(T_n - \theta)}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0,1).$$

Proof: This follows by Slutsky's theorem, as in the binomial case. □

We now re-state the theorem as a "result," by putting it in a form that is less precise mathematically but more useful in practice.

Result If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{(T_n - \theta)}{\sigma_{T_n}} \xrightarrow{D} N(0,1) \tag{7.15}$$

and $\hat{\sigma}_{T_n}$ provides the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\text{approx. } 95\% \text{ CI } = (T_n - 2\hat{\sigma}_{T_n}, T_n + 2\hat{\sigma}_{T_n})$$

which may also be written, equivalently, in the form (7.6), i.e.,

$$\text{approx. } 95\% \text{ CI } = (\hat{\theta} - 2SE(\hat{\theta}), \hat{\theta} + 2SE(\hat{\theta})).$$

The method given by (7.6) is widely applicable because (i) lots of estimators are approximately normally distributed, as in the first assumption of the theorem, and (ii) there are good ways to get standard errors, as in the second assumption of the theorem. The useful "result" is imprecise because of the

approximation $\approx$. The precise statement is in the theorem. This degree of imprecision, and the unclear relevance of arguments that treat the sample size n as sufficiently large, or essentially infinite, are core components of the bond between theory and practice in data analysis.

> *A Detail:* An additional consequence of (7.15) returns us to the characterization, on page 176 of the standard error. After saying that the standard error represents the likely magnitude of error $T-\theta$ we then discussed standard error as estimating the standard deviation of T, which is not the same thing. It is in principle possible for the estimator T to be systematically wrong (being close to, say, $\theta + 10$ instead of θ) and yet have a small variance; in this case the standard error would not represent the likely magnitude of error. When (7.15) holds all is well: it says that $T - \theta$ is approximately normally distributed, with approximate standard deviation σ_{T_n}, so that σ_{T_n} is indeed the likely magnitude of error. This notion of standard error is justified because (7.15) holds in a variety of commonly-found cases.

7.3.6 Estimates and standard errors should be reported to two digits in the standard error.

We recommend rounding standard errors to two leading (nonzero) digits, and then rounding the estimate to match the standard error. For example, if we found an estimate to be 5.582 and the standard error to be .207 we would report the result as 5.58 $\pm$.21. Our reasoning is as follows. On the one hand, it is generally good to avoid too many digits both because numbers with many digits become hard to read, and also because extra digits may imply more accuracy than is present in the results. In this illustration, because the standard error is .21, the second digit in the estimate is already very uncertain: the 95% CI is (5.1,6.0) so we really don't know much about that second digit. We could report only a single digit in the standard error, but we prefer to report two because a standard error of .249 is quite a bit larger than a standard error of .151, yet to single-digit accuracy both would be rounded to .2. No rule is perfect, but it seems to us that reporting standard errors

to two digits, but not more, is a good idea. Thus, in Example 1.2.1 on page 177 we reported the estimate $\hat{p}$ of the propensity p to be .824 $\pm$.092, and in Example 3.2.1 on page 182 we reported the firing rate of the M1 neuron to be 22.72 $\pm$.93 spikes per second.

7.3.7 Appropriate sample sizes may be determined from desired size of standard error.

In Example 1.2.1 the results seemed conclusive, but in some situations we would like even stronger evidence. A natural question is then, How much data would we need to achieve a decisive result? By assuming preliminary data gives us a good idea of what to expect, we can answer this question. In the case of Example 1.2.1, we found $\hat{p} = .824$ with $SE = .092$. If we assume p is, in fact, somewhere around $\hat{p}$, the way we would obtain stronger evidence is by decreasing the standard error. In general terms we proceed in two steps. First, we determine how small we want the standard error to be. Writing our current standard error as SE_1 based on a sample size n_1 and our desired standard error as SE_2, we say that we want to choose a sample size n_2 so as to decrease the standard error by the factor SE_1/SE_2. Second, we apply the $\sqrt{n}$ law. Because the standard error is inversely proportional to $\sqrt{n}$ we may write

$$\frac{SE_1}{SE_2} = \sqrt{\frac{n_2}{n_1}}.$$

Solving for n_2 we get

$$n_2 = n_1 \left(\frac{SE_1}{SE_2} \right)^2. \tag{7.16}$$

If, for instance, we wanted to decrease the standard error by a factor of 2 we would have to multiply our current sample size by a factor of 4.

Using confidence intervals, the simple rule[2] in Equation (7.16) is about as far as we can go. An investigator may wonder about step one, the choice of SE_2. The selection of SE_2 must be determined by careful thinking about

[2]More complicated formulas exist, but they are usually unnecessary because the uncertainties in results from collecting more data are usually much larger than any extra precision one might gain from a more detailed calculation.

the scientific issues involved in the particular case at hand. The desired size of the standard error in Example 3.2.1, for instance, depends on the way the information about spike counts will be used as part of the overall project. In Example 3.2.1 a relatively large number of trials were collected because the experiment was part of a comparative study in which relatively small differences across conditions appeared possible—yet still would have been of interest. According to the standard error on page 182, the firing rate was determined within about ± 1 spike per second. If 15 trials had been used instead of 60, according to the $\sqrt{n}$ law and (7.16) we would expect an accuracy of only about ± 16 spikes per second, and for a mean rate of around 20 spikes per second this seems to be a rather large uncertainty unless the neural response was drastically changed under the alternative condition. On the other hand, such statistical consideration always must be balanced against experimental constraints.

7.3.8 Confidence assigns probability indirectly, making its interpretation subtle.

Here are two interpretations of the confidence interval found for the propensity p of P.S. to choose the non-burning house:

> *Interpretation A:* If p were the true value, then the probability that the interval given by (7.14) would contain p is approximately 95%. Based on the data from P.S., the CI is (.64,1.0).
>
> *Interpretation B:* Based on the data from P.S., the probability that (.64,1.0) contains p is approximately 95%.

It may seem that interpretation B is an immediate consequence of interpretation A. After all, once we apply interpretation A to all values of p, then, regardless of the data we observe, the CI will cover p with approximately 95% probability; we need only apply this to the data we actually did observe to get interpretation B. Unfortunately, to the shock and dismay of many students of statistical inference, this simple logic is fallacious. Interpretation B is a famously incorrect interpretation of a confidence interval. The correct interpretation of confidence, in interpretation A, can *not* be translated

into interpretation B because interpretation A involves the *random variables* L and U that specify the lower and upper endpoints of the CI; probability concerns random variables, not constants; and in interpretation B, .64 and 1.0 are constants, they are not random variables. Once the data have been observed, the probability formalism at the foundation of (7.14) no longer speaks. So it is *incorrect* to think that the confidence interval (.64,1.0) tells us there is a very large probability that p is in the range (.64,1.0). The math involved in deriving confidence intervals is clear, neat and clean. If we want to provide a linguistic interpretation of the confidence interval, however, we must revert to the somewhat clumsy and indirect interpretation A.

To highlight the meaning of CIs let us consider the blindsight example further.

Example 1.2.1 (continued) The first three columns of the table below gives possible CIs using (7.14) when $X \sim B(17, p)$. For example, when $X = 11$ we find $L = .415$ and $U = .879$ so that the CI becomes (.42,.88).

x	L	U	cover
7	.173	.650	N
8	.228	.713	N
9	.287	.772	N
10	.350	.827	Y
11	.415	.879	Y
12	.485	.927	Y
13	.559	.970	Y
14	.639	.008	Y
15	.726	.039	Y
16	.827	1.055	N
17	1	1	N

Now suppose the true value of p were .8. We would find that the CI would contain or "cover" p for some of the values of x but not others, as indicated in the fourth column of the table ("Y" for yes, the interval (L, U) covers .8, "N" for no it does not). The table shows that (L, U) covers .8 when $10 \leq x \leq 15$. To find the level of confidence associated with (L, U) we would compute $P(10 \leq X \leq 15)$ when $X \sim B(17, .8)$. We will return to this below. □

There is another way to look at confidence intervals. Suppose we draw N random samples, independently, and compute CIs (L, U) for each. Let $Y_i = 1$ if (L, U) contains p for the ith random sample and $Y_i = 0$ if not, so that $P(L \leq p \leq U) = P(Y_i = 1)$. Then $\bar{Y}$ is the fraction of random samples for which (L, U) contains p. By the LLN, $\bar{Y} \to P(Y_i = 1)$, that is, $\bar{Y} \to P(L \leq p \leq U)$. We may therefore consider the confidence level $P(L \leq p \leq U)$ to be the long-run limit of the fraction of confidence intervals that contain p.

Interpretation C: If we were to obtain CIs using (7.14) repeatedly, indefinitely many times, then, in the long run, approximately 95% of those CIs would contain p. Based on the data from P.S., the CI is (.64,1.0).

More generally, the level of confidence is usually considered to be the long-run frequency with which the CI covers the true value.

The big achievement of confidence intervals is the conversion of probability as a description of variation (the distribution $X \sim B(n, p)$) into a statement of knowledge. But this achievement comes at a cost: the statement of knowledge is very weak. We might prefer interpretation B, which is analogous to saying "I am 90% sure the capital of Louisiana is Baton Rouge," but confidence intervals do not carry with them this interpretation. An alternative approach, based on Bayes' Theorem, *does* provide an interpretation like interpretation B, but it has its own cost. See Section 7.3.9.

7.3.9 Bayes' Theorem may be used to assess uncertainty.

Recall Bayes' Theorem for random variables and vectors: for continuous random variables or vectors U and V we have

$$f_{U|V}(u|v) = \frac{f_{V|U}(v|u) f_U(u)}{\int f_{V|U}(v|u) f_U(u) du}. \tag{7.17}$$

Let us apply this to the problem of estimating the binomial parameter p. In this section we replace p by θ, so we suppose $X \sim B(n, \theta)$. To apply (7.17)

we take $U = \theta$ and $V = X$ to get

$$f_{\theta|x}(\theta|x) = \frac{f_{X|\theta}(x|\theta)f_\theta(\theta)}{\int f_{X|\theta}(x|\theta)f_\theta(\theta)d\theta}.$$

(We use θ for both capital and lower case theta.) Ordinarily we would take θ as a known constant. Here, however, we take θ to be a *random variable*. The interpretation is that we do not know the value of θ so we assign it a probability distribution. We take $f_\theta(\theta)$ to be the pdf representing our knowledge before seeing the data. It is the pdf corresponding to the *prior distribution*. Ordinarily, because θ is a known constant it is implicitly part of the binomial pdf, so we would write the binomial pdf as $f_X(x)$. Here, however, the binomial pdf must be determined *conditionally* on a value of θ, so it is written $f_{X|\theta}(x|\theta)$. The pdf that summarizes our knowledge *after observing the data* $X = x$ is $f_{\theta|X}(\theta|x)$. This is the pdf corresponding to the *posterior distribution*.

In order to do computations we must assign a specific probability distribution as the prior distribution. Assuming we know very little about the value of θ *a priori*, a natural choice is to use the uniform distribution, $\theta \sim U(0, 1)$, i.e., $f_\theta(\theta) = 1$. With this prior pdf we obtain

$$f(\theta|x) = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot 1}{\int \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot 1 d\theta}$$

which reduces to

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\int \theta^x(1-\theta)^{n-x}d\theta}. \tag{7.18}$$

This formula is a special case of a *Beta* distribution introduced briefly in Chapter **??**: in general, the $Beta(\alpha, \beta)$ density is

$$f(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}w^{\alpha-1}(1-w)^{\beta-1}. \tag{7.19}$$

Therefore, the posterior distribution of θ is $Beta(x+1, n-x+1)$ which has mean and standard deviation

$$\begin{aligned} \mu_{\theta|x} &= \frac{x+1}{n+2} \\ \sigma_{\theta|x} &= \sqrt{\frac{(x+1)(n-x+1)}{(n+2)^2(n+3)}}. \end{aligned}$$

Example 1.2.1 (continued) Let us apply this to the data from patient P.S. We find the posterior distribution based on $n = 17$ and $x = 14$ is $Beta(15, 4)$ and the posterior mean and standard deviation are $\mu_{\theta|x} = .79$ and $\sigma_{\theta|x} = .091$. Thus, roughly speaking, these data lead us to conclude that the frequency with which P.S. will prefer the non-burning house is approximately .79 and our uncertainty may be summarized by saying that the average amount by which this guess misses the truth is approximately .091. These numbers are similar to those obtained earlier, but here they have a different interpretation. Before giving this interpretation let us press on. We may obtain an interval having 95% posterior probability from the .025 and .975 percentiles of the $Beta(15, 4)$ distribution, which gives (.59,.94). That is, $P(\theta < .59|y) = P(\theta > .94|y) = .025$ so that $P(.59 < \theta < .94|x) = .95$. The posterior interval (.59,.94), sometimes called a *credible interval* to distinguish it from a confidence interval, is a succinct summary of what we know about θ based on the data. □

It is now legitimate to say what the posterior interval means, using words that are in essence just like interpretation B of Section 7.3.8.

Bayesian interpretation: Based on the data from P.S., together with the uniform prior, the probability that (.59,.94) contains θ is 95%.

The use of Bayes' Theorem has thus bought us a highly intuitive interpretation of the credible interval. Like confidence intervals, credible intervals convert probability as a description of variation (the distribution $X \sim B(n, p)$) into a statement of knowledge. In this case, unlike the indirect situation with credible intervals, the Bayesian statement is very much analogous to saying "I am 90% sure the capital of Louisiana is Baton Rouge."

The straightforward Bayesian interpretation is very appealing. We issue two notes of caution. First, as we said at the end of Section 7.3.8, Bayes' Theorem requires the additional assumption of a particular form for the prior distribution. For this problem it makes a good deal of sense to use the $U(0, 1)$ distribution for θ *a priori.*. In many settings, however, it is not clear what prior distribution should be used. Many proposals for rules to select prior distributions have been made over the years. The review by Kass and Wasserman (1996), for example, lists over 200 references. (Kass, R.E. and Wasserman, L. (1996) The selection of prior distributions by formal

rules, *J. Amer. Statist. Assoc.*, 91: 1343–1370.) In practice, in a particular data analytical context it may take considerable effort to determine how much the choice matters. Secondly, while confidence is undeniably less direct than posterior probability, we must keep in mind the fundamental distinction between the theoretical world of random variables and formal inferences, and the real world of data. There remains a degree of indirectness in the Bayesian statements as well, because they always say it is *as if* the data *were* to arise as random variables following the probability model (e.g., the binomial distribution). There is an inescapable divide between theoretical inferences and real-world conclusions; they are not quite the same thing, no matter what approach we take.

In practice, Bayesian and non-Bayesian methods are both applied, with the form of the problem and the predilections of the practitioner dictating the choice. There are many important theoretical results concerning posterior distributions. In particular, the approximate CIs given by (7.14) have a Bayesian justification, making interpretation B of Section 7.3.8 valid. We return to Bayesian methods in Chapter **??**.

7.3.10 In small samples it is customary to use the t distribution instead of the normal.

When the sample size is small, the approximation (7.9) may not be accurate. An alternative is to derive an "exact" confidence interval analagous to (7.7) that corrects for the substitution of s for σ. This leads to an adjustment of the multiplier put in front of the standard error. Recall from Chapter **??** that if $U \sim N(0,1)$ and $V \sim \chi^2_\nu$ independently then

$$W = \frac{U}{\sqrt{\frac{V}{n}}}$$

has a t distribution on ν degrees of freedom. The adjustment to the small-sample CI uses the t distribution: writing the .975 quantile as $t_{.975,\nu}$, i.e., $P(W \leq t_{.975,\nu}) = .975$, the value $t_{.975,\nu}$ multiplies the standard error instead of 2 in the formula (7.9), with $t_{.975,\nu}$ being somewhat larger than 2 (or, strictly speaking, $t_{.975,\nu}$ is somewhat larger than the more precise value 1.96 that most

books use instead of 2). The letter ν denotes the *degrees of freedom* of the t distribution. Here, $\nu = n - 1$. The distributional result that makes this work is the following.

Theorem If $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution, then $\bar{X}$ and S^2 are independent random variables with

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

and

$$\frac{S^2}{\sigma^2} \sim \chi^2_\nu$$

with $\nu = n - 1$.

Proof: We omit the proof of this theorem (which follows by manipulation of the joint pdf). □

Theorem If $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution, then a 95% CI is given by $\bar{x} \pm t_{.975,\nu} \cdot SE(\bar{x})$, where $\nu = n-1$ and $SE(\bar{x})$ is given by (7.10), meaning

$$P(\bar{X} - t_{.975,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{.975,n-1} \cdot \frac{S}{\sqrt{n}}) = .95. \quad (7.20)$$

Proof: Let us write

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{S^2}{\sigma^2}}}.$$

The previous theorem then gives the required t distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$. □

Formula (7.20) is the standard method used by most statistical software to provide a confidence interval for an unknown mean μ. When the sample size is large, say, $n \geq 12$, then $t_{.975,\nu} \approx 2$ and (7.20) agrees with (7.9). Customary terminology refers to the CI in (7.20) as based on t while the CI in (7.9) is

based on z. One would not need to bother with the distinction between these two formulas unless $n < 12$, except that as a matter of convention (found in many journals, for example), there tends to be a preference for procedures based on a t, such as (7.20). In other words, it is worth being aware that many people say they are reporting t-based intervals as in (7.20) even when n is large and they might just as well say they are reporting (7.9)—there is in that case no practical distinction between the two.

Example 3.2.1 (continued) Let us now consider the first 12 trials of counts from the motor cortical neuron, examined on page 182. We get a mean firing rate of 24.31 spikes per second, and a standard deviation of 5.20 spikes per second, giving a standard error of

$$SE = \frac{5.20}{\sqrt{12}} = 1.50$$

spikes per second. The t_ν-based CI uses $\nu = 12 - 1 = 11$ and we find $t_{.975,11} = 2.20$. For the 95% CI we take $L = 24.31 - 2.20(1.5) = 21.0$) and $U = 24.31 + 2.20(1.5) = 27.6$, giving us the CI (21.0,27.6) spikes per second. □

It is also worth emphasizing a fundamental difficulty with this approach: the cases in which (7.20) differs from (7.9) are those in which n is small. But in such situations it is quite hard to tell whether the sample is really close to being normal. Application of (7.20) based on small samples should be considered only rough guides to evaluation of uncertainty. We will return to this issue when we discuss nonparametric procedures.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 8

Estimation in Theory and Practice

In Section 7.2.1 we showed how the method of moments may be use to estimate the parameters of a $Gamma(\alpha, \beta)$ distribution, and we immediately stated that the method of maximum likelihood provides a better solution. How do we know this? In general, how should alternative methods of estimation be compared? In this chapter we lay out a series of principles that serve as guides to practice. The main ideas came from Fisher (1922); they were modified and made more precise by J. Neyman (1937), and have been refined and incorporated into textbooks on statistical theory ever since, beginning notably with Cramér (1945).

Let us assume we have a family of probability distributions that depends on a parameter θ, which must be estimated, and we have an estimator T. For now let assume that θ is a scalar. If we were to say that T is a good estimator of θ, what might we mean? In particular, what might we mean when we say that maximum likelihood produces a good estimator? Clearly, for T to be a good estimator it must be "close" to θ, but because T is a random variable the notion of closeness must be stated probabilistically. For

example, if we consider the mean $\bar{X}$ of a random sample $X_1, \ldots, X_n$ from a $N(\theta, 1)$ distribution, we might want to say that the mean $\bar{X}$ is close to θ when $|\bar{X} - \theta| < .1$. Because $\bar{X} \sim N(\theta, 1/n)$, even if n is large it is *possible* that $|\bar{X} - \theta| > .1$. We can not say that $|\bar{X} - \theta| < .1$. All we can say is the probabilty that $|\bar{X} - \theta| < .1$ is large, meaning close to 1 or, equivalently, the probability that $|\bar{X} - \theta| > .1$ is small, meaning close to 0.

For a general estimator T we can use the same approach and say that T is a good estimator of θ when it is *highly probable* that T is close to θ. Specifically, we introduce a tolerance ϵ, understanding that ϵ will be some small positive number, and then we require that $P(|\bar{X} - \theta| < \epsilon)$ is close to one or, equivalently, $P(|\bar{X} - \theta| > \epsilon)$ is close to zero. It is, in general, rather difficult to provide guarantees on the size of $P(|\bar{X} - \theta| > \epsilon)$ for fixed sample sizes. In most realistically complicated problems computer simulation studies must be used (as in Section 8.1.2) and they are based on specific cases so they do not provide general assurances. On the other hand, general results may be obtained asymptotically, letting the sample size grow indefinitely large. To take a concrete case, because the mean $\bar{X}$ of a random sample from a $N(\theta, 1)$ distribution follows a $N(\theta, 1/n)$ distribution, if we take $n = 10,000$, from the normal cdf we find $P(|\bar{X} - \theta| > .1) = 1.5 \cdot 10^{-23}$. Indeed, no matter how small we take ϵ we have $P(|\bar{X} - \theta| > \epsilon) \to 0$ as $n \to \infty$. This is simply a restatement of the law of large numbers,

$$\bar{X} \xrightarrow{P} \theta.$$

We discuss asymptotic results in Sections 8.2.1–8.3.1.

When we examine what happens as $n \to \infty$ it is helpful to write the generic estimator in the form $T_n = T(X_1, \ldots, X_n)$ to emphasize its dependence on n as we did in Section 7.3.5. One of the most important of the large-sample findings considers estimators that are *asymptotically normal*, as in Equation (7.15),

$$\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0, 1). \tag{8.1}$$

For such estimators, in large samples, the probabilistic closeness of T_n to θ depends entirely on σ_{T_n} and we seek estimators that make σ_{T_n} as small as possible. In Sections 8.2.2—8.3.1 we go over the remarkable discovery by Fisher that σ_{T_n} can be minimized, and the minimum is obtained by the MLE. There

has been quite a lot of theoretical work on the general subject of large-sample optimality, all of which leads to the conclusion that in well-behaved parametric problems, the method of maximum likelihood is essentially unbeatable. Fisher's insight seems to have been based on geometrical intuitions, which were elaborated in a mathematically rigorous framework by B. Efron in the 1970s and early 1980s. For details and references on the asymptotic arguments and their geometrical origins see Kass and Vos (1997). For a rigorous treatment in a more general context see van der Vaart (1998). (Kass, R.E., and Vos, P.W. (1997) *Geometrical Foundations of Asymptotic Inference*, Wiley; van der Vaart, A.W. (1997) *Asymptotic Statistics*, Cambridge.))

While asymptotic results are important, they have an inherent weakness: they apply when the sample size is large, but they do not say what "large" means in practice. In some cases $n = 20$ is more than adequate while in others $n = 20,000$ is not large enough. One approach to coping with this problem is to evaluate a measure of likely deviation for specific cases, with specified sample sizes. The most common assessment of deviation of T from θ is the *mean squared error (MSE)* defined by

$$MSE(T) = E((T-\theta)^2). \tag{8.2}$$

We discuss this in Section 8.1.

A rather different, but fundamental notion of what constitutes a good estimator is that it should capture as much as possible of the information about θ contained in the data. At first glance this may seem rather vague, but it has been formalized in fruitful ways. We discuss this in Section **??**. After reviewing the principles of estimation in Section 8.1 we go on to describe some of the practical considerations in applying ML estimation, in Section 8.4.

8.1 Mean Squared Error

The mean squared error criterion defined in (8.2) uses the squared magnitude of the deviation $T-\theta$ rather than its absolute value $|T-\theta|$ because it is easier to work with mathematically, and because it has a very nice decomposition given in Section 8.1.1. Intuitively, because $MSE(T)$ is an average of the values $(T-\theta)^2$, when $MSE(T)$ is small, large values of $(T-\theta)^2$ (and thus

also large values of $|T-\theta|$) must be highly improbable. In fact, even more is true: we have

$$P(|T-\theta|>\epsilon) < \frac{E((T-\theta)^2)}{\epsilon^2}. \tag{8.3}$$

Thus, we can make sure it is highly probable for T to be close to θ by instead making sure that $MSE(T)$ is small.

> *Details:* We can use Markov's inequality, which appeared as a lemma in Section 6.2.1, to guarantee that $P(|T-\theta|>\epsilon)$ will be small if $MSE(T)$ is small. First, we have
>
> $$P(|T-\theta|>\epsilon) = P((T-\theta)^2>\epsilon^2).$$
>
> Now, assuming $E((T-\theta)^2)<\infty$, Markov's inequality gives (8.3). □

In some cases $MSE(T)$ may be evaluated by analytical calculation, but in most practical situations computer simulation studies are used. We give two examples of such studies in Section 8.1.2.

8.1.1 Mean squared error is bias squared plus variance.

Two ways an estimator can perform poorly need to be distinguished. The first involves the systematic tendency for the estimator T to miss its target value θ. An estimator's *bias* is Bias(T) = E(T) − θ. When the bias is large, T will not be close to θ *on average*. The second is the variance $V(T)$. If $V(T)$ is large then T will rarely be close to θ. Figure 8.1 illustrates, by analogy with shooting at a bullseye target, the situations in which the bias is large, the variance is large, both are large (the worst case) and, finally, both are small (the best case).

Part of the appeal of mean squared error is that it combines bias and variance in a beautifully simple way.

Theorem Suppose $E((T-\theta)^2)<\infty$. Then

$$E((T-\theta)^2) = (E(T-\theta))^2 + V(T).$$

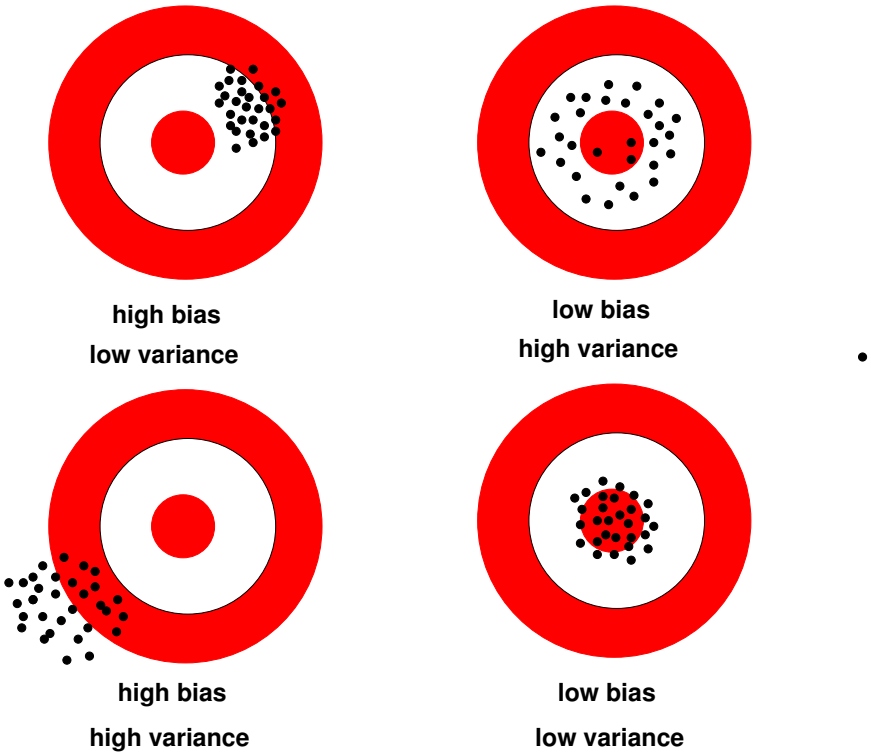


Figure 8.1: *Drawing of shots aimed at a target to illustrate the way estimates can miss their "target." They may be systematically biased, or they may have high variability, or both. The best situation, of course, is when there is little systematic bias and little variability.*

That is,

$$MSE(T) = \text{Bias(T)}^2 + \text{Variance(T)}.$$

Proof: Let us write $\mu_T = E(T)$ and $T - \theta = (T - \mu_T) + (\mu_T - \theta)$, and then square both sides to get

$$(T - \theta)^2 = (T - \mu_T)^2 + 2(T - \mu_T)(\mu_T - \theta) + (\mu_T - \theta)^2.$$

Now consider taking the expectation of the cross-product term on the right-hand side. The quantity $\mu_T - \theta$ is a constant (it is not a random variable), while because $E(T) = \mu_T$, we have

$E(T-\mu_T)=0$ and, therefore, $E(2(T-\mu_T)(\mu_T-\theta))=0$. Thus, we have

$$E((T-\theta)^2)=E((T-\mu_T)^2)+(E(\mu_T-\theta))^2$$

and, since $V(T)=E((T-\mu_T)^2)$, we have proven the theorem. □

Before we present an illustration of a MSE calculation, let us mention a property of the sample mean and sample variance. Assuming they are computed from a random sample $X_1,\ldots,X_n$, we have $E(\bar{X})=\mu_X$ which may be written

$$E(\bar{X})-\mu_X=0.$$

This says that the sample mean is *unbiased* as an estimator of the theoretical mean. When an estimator is unbiased we have $MSE(T)=V(T)$ so that consideration of its performance may be based on a study of its variance.

In addition to the sample mean being unbiased as an esimator of the theoretical mean, it also happens that the sample variance is unbiased as an estimator of the theoretical variance:

$$E(S^2)=\sigma_X^2 \tag{8.4}$$

so that the sample variance is unbiased as an estimator of the theoretical variance.

Details: We wish to evaluate

$$E(S^2)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right).$$

We write $X_i-\bar{X}=(X_i-\mu_X)+(\mu_X-\bar{X})$ and expand the square

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)^2\\ &= \sum_{i=1}^{n}(X_i-\mu_X)^2+\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X})+\sum_{i=1}^{n}(\mu_X-\bar{X})^2.\end{aligned}$$

We now rewrite the three terms in the last expression above. Because $E(X_i - \mu_X) = \sigma_X^2$, and the expectation of a sum is the sum of the expectations, the first term has expectation

$$E\left(\sum_{i=1}^{n}(X_i - \mu_X)^2\right) = n\sigma_X^2. \tag{8.5}$$

Next, the second term may be rewritten

$$\begin{aligned}\sum_{i=1}^{n} 2(X_i - \mu_X)(\mu_X - \bar{X}) &= 2(\mu_X - \bar{X})\sum_{i=1}^{n}(X_i - \mu_X) \\ &= -2(\bar{X} - \mu_X)\sum_{i=1}^{n}(X_i - \mu_X) \\ &= -2n(\bar{X} - \mu_X)^2,\end{aligned}$$

where the last equality uses $\sum_{i=1}^{n}(X_i - \mu_X) = n(\bar{X} - \mu_X)$, and then, because $E((\bar{X} - \mu_X)^2) = V(\bar{X}) = \sigma_X^2/n$, the expectation of the second term becomes

$$E\left(\sum_{i=1}^{n} 2(X_i - \mu_X)(\mu_X - \bar{X})\right) = -2\sigma_X^2. \tag{8.6}$$

Finally, because again, $E((\bar{X} - \mu_X)^2) = \sigma_X^2/n$, the expectation of the third term is

$$E\left(\sum_{i=1}^{n}(\mu_X - \bar{X})^2)\right) = \sigma_X^2 \tag{8.7}$$

and, combining (8.5), (8.6), and (8.7) we get

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = (n-1)\sigma_X^2$$

which gives (8.4). □

We will use the unbaisedness of the sample mean and sample variance in the following.

Illustration: Poisson Spike Counts We previously considered 60 spike counts from a motor cortical neuron, and found an approximate 95% CI for the resulting firing rate using the sample mean. The justification for that approximate CI involved the CLT, and the practical implication was that as long as the sample size is fairly large, and the distribution not too far from normal, the CI would have approximately .95 probability of covering the theoretical mean. In this case, the spike counts do, indeed, appear not too far from normal. Sometimes they are assumed to be Poisson distributed. This is questionable because careful examination of spike trains almost always indicates some deparature from the Poisson. On the other hand, the departure is sometimes not large enough to make a practical difference to results. In any case, for the sake of illustrating the MSE calculation, let us now *assume* the counts follow a Poisson distribution with mean λ. The sample mean $\bar{X}$ is a reasonable estimator of λ, but one might dream up alternatives. For example, a property of the Poisson distribution is that its variance is also equal to λ; therefore, the sample variance S^2 could also be used to estimate the population variance λ. This may seem odd, and potentially inferior, on intuitive grounds because the whole point is to estimate the mean firing rate, not the variance of the firing rate. On the other hand, once we take the Poisson model seriously the population mean and variance become equal and, from a statistical point of view, it is reasonable to ask whether it is better to estimate one rather than the other from their sample analogues. Our purpose here is to present a simple analysis that demonstrates the inferiority of the sample variance to the sample mean as an estimator of the Poisson mean λ. We are going through this exercise so that we can draw an analogy to it later on.

Now, because, as we mentioned immediately before beginning this illustration, $\bar{X}$ and S^2 are unbiased for the theoretical mean and variance they are, in this case, both unbiased as estimators of λ. As a consquence, $MSE(T) = V(T)$ for both $T = \bar{X}$ and $T = S^2$. Analytical calculation of the variance of these estimators (which we omit here) gives

$$\begin{aligned} V(\bar{X}) &= \frac{\lambda}{n} \\ V(S^2) &= \frac{\lambda}{n} + \frac{2\lambda^2}{n-1} \end{aligned}$$

where n is the number of counts (the number of trials). Therefore, the MSE

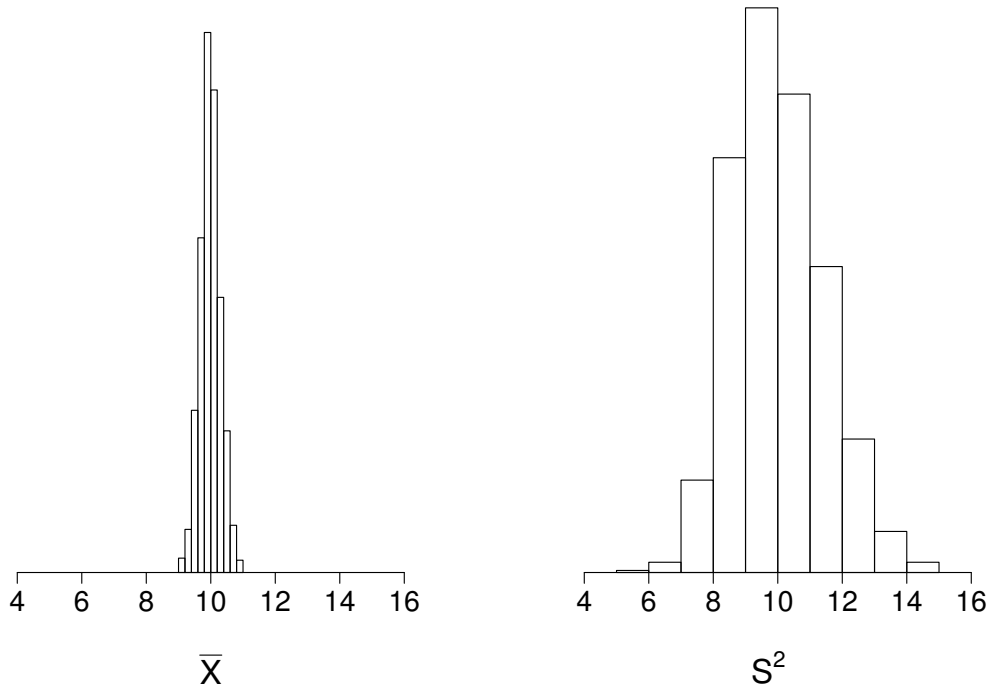


Figure 8.2: *Histograms displaying distributions of $\bar{X}$ (labeled $\bar{y}$ here) and S^2 based on 1000 randomly-generated samples of size $n = 100$ from a Poisson distribution with mean parameter $\mu = 10$. In these repeated samples both $\bar{X}$ and S^2 have distributions that are approximately normal (represented by the overlaid curves). Both distributions are centered at 10 (both estimators are unbiased) but the values of S^2 fluctuate much more than do the values of $\bar{X}$.*

of S^2 is always larger than that of $\bar{X}$ so that S^2 tends to be further from the correct value of λ than $\bar{X}$. For example, if we take $n = 100$ trials and $\lambda = 10$, we find $V(\bar{X}) = .10$ while $V(S^2) = 2.12$. The estimator S^2 has about 21 times the variability as $\bar{X}$, so that estimating λ using S^2 would require

about 2100 trials of data to gain the same accuracy as using $\bar{X}$ with 100 trials. Figure 8.2 shows a pair of histograms of $\bar{X}$ and S^2 values calculated from 1000 randomly-generated samples of size $n = 100$ when the true Poisson mean was $\lambda = 10$. □

This illustration nicely shows how one method of estimation can be very much better than another, but it is admittedly somewhat artificial: because the distribution of real spike counts may well depart from Poisson, a careful comparison of $\bar{X}$ versus S^2 should consider their behavior also under alternative assumptions. In this regard, the sample mean remains a reasonably good estimator of the population mean in large samples regardless of the probability distribution of the spike counts. The sample variance, on the other hand, does so only if the population variance is truly equal to the population mean; otherwise, as the sample size increases it will converge to the wrong value. This is likely to be an important consideration. However, even if one were convinced that counts truly followed a Poisson distribution, the analysis above would be compelling. It would be crazy to use S^2 instead of $\bar{X}$ in estimating λ.

Another thing to notice in Figure 8.2 is the approximately normal shape of the two histograms. Asymptotic normality of estimators is very common, and we have already relied on it in Section 7.3.5.

8.1.2 Mean squared error may be evaluated by computer simulation of pseudo-data.

In the Poisson spike count illustration on page 204 we were able to compute the MSE exactly. In more complicated situations this is often impossible. Instead we rely on either large-sample arguments, such as those in Section 8.2.2 or numerical simulations. The numerical method uses computer-generated *pseudo-data*, by which we mean numbers or vectors that are generated from known probability distributions in order to mimic the behavior of data. Because the distribution is known, there is a known correct value of θ to which T may be compared.

Suppose we wish to compute $MSE(T)$ in estimating θ under the assump-

tion that a random sample comes from a particular probability distribution having cdf $F(x)$. Assuming we know how to generate random samples from $F(x)$ on the computer, we may use this algorithm:

1. Take G to be a large integer (such as 1,000) and for $g = 1, \ldots, G$ do the following:

   (i) Generate a random sample $X_1^{(g)}, \ldots, X_n^{(g)}$ from $F(x)$.

   (ii) Compute $T^{(g)} = T(X_1^{(1)}, \ldots, X_n^{(g)})$, which is the value of the estimator T based on the gth random sample.

   (iii) Let $Y_g = (T^{(g)} - \theta)^2$.

2. Compute

$$\bar{Y} = \frac{1}{G} \sum_{g=1}^{G} Y_g. \tag{8.8}$$

By the LLN, we have that $\bar{Y}$ converges to the desired $MSE = E((T - \theta)^2)$ in probability. Thus, we take $\bar{Y}$ as our MSE.

This kind of computation is used in the following illustration. It involves the statistical efficiency of smoothing, a topic we take up in Chapter **??**. In presenting this now we omit details about the method.

Example 1.1.1 (continued) In Chapter 1 we discussed a study by Olson *et al.* (2000), in which neuronal spike trains were recorded from the Supplementary Eye Field under two different experimental conditions. As is usually the case in stimulus-response studies, the neuronal response—in this case, the firing rate—varied as a function time. For a particular neuron in one of the conditions, the PSTH in Figure 8.3 displays the way the firing rate changes across time. The data analytic challenge in the Olson *et al.* study was to characterize the distinctions between the firing rate functions under the two experimental conditions. One of the distinctions, evident in some of the plots, was that the maximal firing rate occurred somewhat later in one condition than in the other. How should this time of maximal firing rate be computed? One possibility is to use the PSTH, by finding the time bin for which the PSTH is maximized. Panel B of Figure 8.3 displays the resulting solution: according to the PSTH shown there, the maximal firing

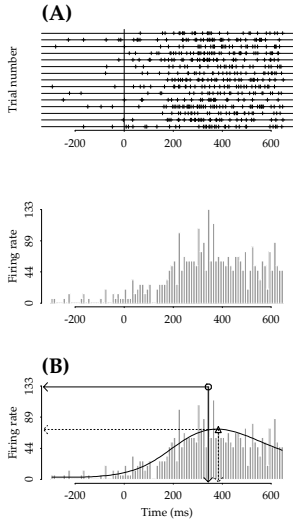


Figure 8.3: *Time of maximal firing rate. Part (A) displays a raster plot and Peri-Stimulus Time Histogram (PSTH). As explained in Chapter 1, the PSTH represents the firing rate as a function of time. Part (B) displays the time at which the maximal firing rate occurs, estimated (i) using the PSTH and (ii) using instead a smooth curve.*

rate of about 133 spikes per second occurs at a time marked by the arrow on the left along the time axis. However, this is clearly a noisy estimate. Slight variations in location of time bin, or width, would change this, as would consideration of new data from the same neuron. On the other hand, a second method based on first fitting a smooth curve to the PSTH and then finding its maximum, yields a different answer: the maximum firing rate of about 75 spikes per second occurs at a time indicated by the arrow on the right along the time axis. This value is less subject to fluctuations in the data.

We evaluate the benefit of smoothing the PSTH in Section 8.1.2. This example illustrates an additional, general point about estimation. In nearly all the problems we discuss in this chapter the parameter θ is a vector having a finite number of components. When we draw a smooth curve, as in Panel B of Figure 8.3, we are effectively estimating a *function*. Here we could write the firing rate as the function $f(t)$ and, if we allow $f(t)$ to be sufficiently general, the number of parameters we would need to represent it becomes infinite. At this point the problem becomes *non-parametric*, and a more elaborate theoretical framework is needed, which we do not pursue in this book. An alternative is to introduce a flexible parametric family of functions to represent $f(t)$ to which the standard theory of this chapter may be applied. We discuss this in some detail in Chapter **??**.

At the beginning of this chapter we returned to the SEF data from Chapter 1 and noted the potential benefit of smoothing the PSTH for the purpose of determining the time at which the maximal firing rate occurs. More generally we can inquire about the benefit of smoothing in estimating the firing rate at any particular point in time. Let us consider pseudo-data simulated, on the computer, from the true firing rate function shown in part A of Figure 8.4. When we simulate 16 trials (the number of trials in the SEF data shown in Figure 8.3) we get the PSTH shown in part B of Figure 8.4. We then determine a smooth firing-rate function, also shown in part B, as "estimated rate." (The method used here involves regression splines, which are discussed in Chapter **??**.) Note that the estimated rate is close to the true rate from the simulation, but it misses by a small amount due to the small number of trials we used in the simulation.

To quantify the deviation of both the PSTH and the smooth curve at

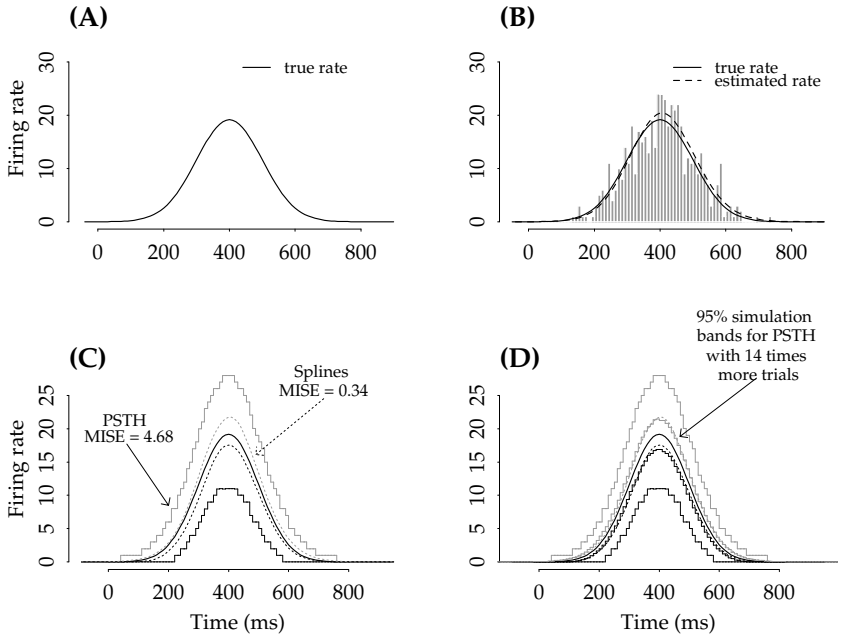


Figure 8.4: (A) True rate from which 16 trials are simulated; their PSTH is shown in (B), with true and estimated firing rates overlaid. (C) shows the true rate and 95% simulation bands obtained from smoothed and unsmoothed PSTHs. (D) shows the same curves as (C), as well as 95% simulation bands obtained from unsmoothed PSTHs with 16×14 trials instead of 16.

any one point in time t we could use MSE. Specifically, we would regard the true firing rate at time t as the value $\theta = \theta_t$ to be estimated, and we would compute $MSE(T) = MSE_t(T)$ when T is based on the PSTH and when T is based on the smooth curve. We have added the subscript t as a reminder that we have chosen a particular time point. If we then do this for every time value t we can find the total of all the mean squared errors by integrating across time. This defines what is *integrated mean squared error* or *mean integrated squared error (MISE)*,

$$MISE(T) = \int MSE_t(T)dt$$

where the integration is performed over the time interval of interest. The

integral may be evaluated easily simply by calculating the MSE along a fine grid of time values separated by some increment Δt

$$\int MSE_t(T)dt \approx \Delta t \sum_t MSE_t(T).$$

In order to compute the MSE at each time value t we used computer simulation: we generated data repeatedly, each time finding both the PSTH and the smooth curve. We simulated 1000 data sets, each involving 16 randomly-generated spike trains based on the true firing rate curve shown in Part A of Figure 8.4, and from these 1000 data sets we computed the $MISE$. We also computed 95% bands, within which fall 95% of the estimated curves. Part C of Figure 8.4 shows the two pairs of bands, now labeled with the two values of MISE: the spline-based estimate has a MISE of .34 (in spikes per second squared) while the PSTH has a MISE of 4.68, which is 14 times larger. This means that when the PSTH is used to estimate firing rate, 14 times as much data are needed to achieve the same level of accuracy. Similarly, the 95% bands for the PSTH are much further from the true firing-rate curve than the bands for the spline-based estimate. In part D of Figure 8.4 we add a pair of 95% bands obtained from the PSTH when 224 trials are used rather than 16 (because $224 = 14 \times 16$). This is another way of showing that the accuracy in estimating the firing rate using spline smoothing based on 16 trials is the same as the accuracy using the PSTH based on 224 trials. Clearly it is very much better to use smoothing when estimating the instantaneous firing rate. □

	r_{xy}	*Spearman*	New Method
$n = 15$	0.091	0.034	0.0072
$n = 30$	0.035	0.012	0.0013
$n = 60$	0.014	0.0052	0.00086
$n = 120$	0.0036	0.0026	0.00035

Table 8.1: *MSE for various sample sizes n Results show MSE in estimating ρ_{UV} using the uncorrected correlation r_{xy}, Spearman's method, and the new method of Behseta* et al.*. The simulation was carried out under what Behseta* et al. *called "scenario 2." In all cases, the new method has a much smaller MSE.*

Here is an interesting example in which MSE contributed valuable information. [**Sorry: I put much of this in Ch 12, but there's a new bit here involving MSE**]

Example 8.1.1 Correction for attenuation of the correlation in SEF selectivity indices When correlation is measured in the presence of noise, its value is decreased—the math showing this is given below. In single-neuron recording experiments, for example, the correlation of quantities based on firing rates in a pair of tasks may be assessed across neurons but, because the number of trials is limited, the measured quantities for each neuron will be noisy. This attenuates the correlation. A correction for such attenuation was proposed more than 100 years ago. Behseta *et al.* (2009) provided a new method and demonstrated through numerical simulation studies that it can perform much better than the older method. We outline one of the results here, and explain the methodology in Chapter **??**.

Suppose U and W are random variables having a positive correlation ρ_{UW} and ϵ and δ are independent "noise" random variables that corrupt the measurement of U and V producing $X = U + \epsilon$ and $Y = W + \delta$. A bit of calculation shows that

$$\rho_{XY} < \rho_{UW},$$

where ρ_{XY} is the correlation between X and Y.

Details: We assume that $V(\epsilon) > 0$ and $V(\delta) > 0$ and we begin by writing

$$\begin{aligned} Cov(X,Y) &= Cov(U+\epsilon, W+\delta) \\ &= Cov(U,W) + Cov(U,\delta) + Cov(W,\epsilon) + Cov(\epsilon,\delta). \end{aligned}$$

Because of independence the last 3 terms above are 0. Therefore, $Cov(X,Y) = Cov(U,W)$ and

$$\begin{aligned} \rho_{XY} &= Cor(U+\epsilon, W+\delta) \\ &= \frac{Cov(U,W)}{\sqrt{V(U+\epsilon)V(W+\delta)}} \\ &= \frac{Cov(U,W)}{\sqrt{(V(U)+V(\epsilon))(V(W)+V(\delta))}} \end{aligned}$$

$$\begin{aligned} &< \frac{Cov(U,W)}{\sqrt{Var(U)Var(W)}} \\ &= \rho_{UW}. \end{aligned}$$

□

In words, the presence of noise decreases the correlation. Thus, in many circumstances, a measured correlation will underestimate the strength of the actual correlation between two variables. However, if the likely magnitude of the noise is known it becomes possible to correct the estimate. The idea of introducing a "correction for attenuation" of the correlation goes back at least to Spearman (1904). In Table 8.1 we give MSE results of Behseta *et al.* for Spearman's method and an alternative labeled "new method." These results were presented to support the claim by Behseta *et al.* that the new method could perform much better than Spearman's.

Correcting for attenuation can be consequential. Behseta *et al.* reported analysis of data from an experiment on neural mechanisms of serial order performance. Monkeys were trained to perform eye movements in a given order signaled by a cue. For example, one cue carried the instruction: look up, then right, then left. Based on recordings of neural activity in frontal cortex (the supplementary eye field, SEF) during task performance, Behseta *et al.* reported that many neurons fire at different rates during different stages of the task, with some firing at the highest rate during the first, some during the second and some during the third stage. These rank-selective neurons might genuinely be sensitive to the monkey's stage in the sequence. Alternatively, they might be sensitive to some correlated factor. One such factor is expectation of reward. Reward (a drop of juice) was delivered only after all three movements had been completed. Thus as the stage of the trial progressed from one to three, the expectation of reward might have increased.

To see whether rank-selective neurons were sensitive to the size of the anticipated reward, the same monkeys were trained to perform a task in which a visual cue presented at the beginning of the trial signaled to the monkey whether he would receive one drop or three drops of juice after a fixed interval. The idea was that neuronal activity related to expectation of reward would be greater after the promise of three drops than after the promise of one. Spike counts from 54 neurons neurons were collected during the performance of both the serial order task and the variable reward task,

and indices rank selectivity in the serial order task and of selectivity for the size of the anticipated reward in the variable reward task were computed. The rank index was $I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. Similarly, the reward index was $I_{\text{reward}} = \frac{(f_b - f_s)}{(f_b + f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. The indices I_{rank} and I_{reward} turned out to be positively correlated, but that the effect was smaller than expected, with $r = 0.49$. The correlation between the rank and reward indices was expected to be larger because, from previous research, it was known that (a) the expectation of reward increases over the course of a serial order trial and (b) neuronal activity in the SEF is affected by the expectation of reward. Behseta *et al.* speculated that the correlation between the two indices had been attenuated by noise arising from trial-to-trial variations in neural activity. Correction for attenuation gave a dramatically increased correlation, with the new estimate of correlation becoming .83. The new estimate may be considered much more reliable than the original $r = .49$ because of the simulation studies of MSE that compared the estimators. □

A Detail: One issue that arises in numerical simulation is the accuracy of the computational results, because the value $\bar{Y}$ in (8.8) is itself an estimate of the MSE. However, if G is large, the standard error of $\bar{Y}$ will be small. Furthermore, because $\bar{Y}$ is a sample mean, we can apply the method of Section 7.3.4 and use $s/\sqrt{G}$ as its standard error, where $s^2 = \frac{1}{G-1}\sum_{g=1}^{G}(Y_g - \bar{Y})^2$. The standard error lets us determine whether G is adequately large. For instance, if we wish the MSE to be computed with accuracy δ, we can make sure $s/\sqrt{G} < \delta/2$. By the result in Section 7.3.4, an approximate 95% confidence interval would lie within the interval $(\theta - \delta, \theta + \delta)$.

8.1.3 In estimating a theoretical mean from observations having differing variances a weighted mean should be used, with weights inversely proportional to the variances.

In the illustration on Poisson spike counts, page 204, we used the MSE criterion to evaluate alternative estimators, based on an analytical expression. In that case both estimators were unbiased and the comparison was based on variance. Another illustration of this type arises when data are considered collectively across many similarly measured objects, such as neurons or people, with the observations from the different individuals contributing varying amounts of information; specifically, with the individual observations having different variances. In combining such discrepant observations, it is preferable not to use the sample mean, but instead to weight each observation according to the amount of information it contributes. Here we provide a theoretical analysis of this problem, and give the basic result.

Suppose we have two independent random variables X_i for $i = 1, 2$, with $E(X_1) = E(X_2) = \mu$ but $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, with the two variances possibly being different. After analyzing the two-observation case, we will present analogous results for n observations. Let us assume that σ_1 and σ_2 are known and ask how best to combine X_1 and X_2 linearly in order to estimate μ. We write a general weighted combination as

$$Y_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.9}$$

where $w_1 + w_2 = 1$. It turns out that the optimal special case is

$$\bar{X}_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.10}$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \tag{8.11}$$

for $i = 1, 2$.

Theorem Suppose X_1 and X_2 are independent random variables with $E(X_1) = E(X_2) = \mu$ and $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, and let Y_w be

defined as in (8.9). Then Y_w is unbiased, so that $MSE(Y_w) = V(Y_w)$, and this quantity is minimized among possible weighting pairs by taking $Y_w = \bar{X}_w$, i.e.,

$$V(\bar{X}_w) \leq V(Y_w)$$

or, equivalently,

$$MSE(\bar{X}_w) \leq MSE(Y_w)$$

with equality holding in both cases only if $Y_w = \bar{X}_w$.

Proof of Theorem: First, we have

$$\begin{aligned} E(Y_w) &= w_1 \cdot \mu + w_2 \cdot \mu \\ &= (w_1 + w_2)\mu \\ &= \mu. \end{aligned}$$

Thus, Y_w is unbiased and $MSE(Y_w) = V(Y_w)$. To derive the variance result we start with

$$V(w_1 \cdot X_1 + w_2 \cdot X_2) = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2.$$

Now we use $w_1 + w_2 = 1$ and replace w_2 with $1 - w_1$ to get

$$\begin{aligned} V(w_1 \cdot X_1 + w_2 \cdot X_2) &= w_1^2 \cdot \sigma_1^2 + (1 - w_1)^2 \cdot \sigma_2^2 \\ &= \sigma_1^2 w_1^2 + \sigma_2^2 - 2\sigma_2^2 w_1 + \sigma_2^2 w_1^2 \\ &= (\sigma_1^2 + \sigma_2^2) w_1^2 - 2\sigma_2^2 w_1 + \sigma_2^2. \end{aligned}$$

We now minimize this quantity by differentiating with respect to w_1, and setting the derivative equal to zero. We get

$$0 = 2(\sigma_1^2 + \sigma_2^2) w_1 - 2\sigma_2^2$$

and therefore

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Dividing the numerator and denominator of this fraction by $\sigma_1^2 \sigma_2^2$ gives

$$w_1 = \frac{\frac{\sigma_2^2}{\sigma_1^2 \sigma_2^2}}{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 \sigma_2^2}} = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

which is the desired result. □

As an example, suppose we had 100 independent observations $U_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, 100$, and grouped them unequally defining, say, $X_1 = \frac{1}{10}\sum_{i=1}^{10} U_i$ and $X_2 = \frac{1}{90}\sum_{i=11}^{100} U_i$. It would seem strange to use $\frac{1}{2}(X_1 + X_2)$ in this situation and the intuitive thing to do would be to use the weighted mean: here the weights are $w_1 = 10/100$ and $w_2 = 90/100$ (because $\sigma_1^2 = \sigma^2/10$ and $\sigma_2^2 = \sigma^2/90$) so we get $\bar{X}_w = \bar{U}$.

One way to interpret this is to say that using $\bar{X}$ instead of $\bar{X}_w$ is like throwing away a fraction of the data. For example, suppose X_1 and X_2 both represent means of counts from n trials. If σ_1 is half the size of σ_2 then, from the formula above, the ratio of variances is 1.56. This means that to achieve the same accuracy in the estimator, n would have to be 56% larger if we used the sample mean instead of the weighted mean. When σ_1 is one-third the size of σ_2 we would have to increase n by a factor of 2.78 (instead of 50 trials, say, we would need 139). In these cases we might say that the weighted mean is, respectively, 1.56 and 2.78 times more efficient than the ordinary sample mean. Here is the general result.

Theorem Suppose $X_1, \ldots, X_n$ are independent random variables with $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$ and $V(X_i) = \sigma_i^2$ for $i = 1, \ldots, n$, and define

$$Y_w = \sum_{i=1}^{n} w_i \cdot X_i$$

with $\sum_{i=1}^{n} w_i = 1$ and

$$\bar{X}_w = \sum_{i=1}^{n} w_i \cdot X_i \tag{8.12}$$

where, in (8.12),

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}.$$

Then

$$V(\bar{X}_w) \leq V(Y_w)$$

with equality holding if and only if $Y_w = \bar{X}_w$. Furthermore we have

$$SE(\bar{X}_w) = \sqrt{V(\bar{X}_w)} \tag{8.13}$$

where

$$V(\bar{X}_w) = \left(\sum_{i=1}^{n} \frac{1}{\sigma^2} \right)^{-1}.$$

Proof: The proof is analogous to that for the case $n = 2$. □

Example 8.1.2 Neural response to selective perturbation of a brain-machine interface In order to study learning-related changes in a network of neurons, Jarosiewicz *et al.* (2008) (PNAS, 105: 19486–19491) introduced a paradigm in which the output of a cortical network can be perturbed directly and the neural basis of the compensatory changes studied in detail. Using a brain-computer interface (BCI), dozens of simultaneously recorded neurons in the motor cortex of awake, behaving monkeys were used to control the movement of a cursor in a three-dimensional virtual-reality environment. This device creates a precise, well-defined mapping between the firing of the recorded neurons and an expressed behavior (cursor movement). In a series of experiments, they forced the animal to relearn the association between neural firing and cursor movement in a subset of neurons and assess how the network changes to compensate. Their main finding was that changes in neural activity reflect not only an alteration of behavioral strategy but also the relative contributions of individual neurons to the population error signal. As part of their study they compared firing rate modulation among neurons whose BCI signals had been artificially perturbed with that among neurons whose BCI signals remained as determined from their control responses. Because the uncertainties varied substantially across neurons, these comparisons among groups of neurons were carried out using weighted means. □

8.1.4 Decision theory uses mean squared error to represent risk.

8.2 Estimation in Large Samples

8.2.1 In large samples, an estimator should be very likely to be close to its estimand.

In the introduction to this chapter we offered the reminder that the sample mean satisfies $P(|\bar{X} - \theta| > \epsilon) \to 0$, which is the law of large numbers. Suppose T_n is an estimator of θ. If for every positive ϵ, as $n \to \infty$ we have

$$P(|T_n - \theta| > \epsilon) \to 0 \tag{8.14}$$

then T_n is said to be a *consistent* estimator of θ. This may also be written $T_n \xrightarrow{P} \theta$.

Note that, by (8.3), if $MSE(T_n) \to 0$ then T_n is consistent. Also, if T_n is asymptotically normal and $\sigma_{T_n} \to 0$ then T_n is consistent.

> *Details:* Multiplying the left-hand side of (8.1) by σ_{T_n} and applying Slutzky's theorem we have $T_n - \theta \xrightarrow{P} 0$, which is equivalent to $T_n \xrightarrow{P} \theta$. □

In words, to say that an estimator is consistent is to say that, for sufficiently large samples, it will be very likely to be close to the quantity it is estimating. This is clearly a desirable property. When T_n is asymptotically normal and $\sigma_{T_n} \to 0$ we will call T_n *consistent and asymptotically normal*.

8.2.2 In large samples, the precision with which a parameter may be estimated is bounded by Fisher information.

Let us now consider all estimators of θ that are consistent and asymptotically normal in the sense of Section 8.2.1. For such an estimator $T = T_n$ we may say that its distribution is approximately Normal:

$$T \dot{\sim} N(\theta, \sigma_T^2).$$

An approximate 95% CI for θ, based on T, would be $(T - 2\sigma_T, T + 2\sigma_T)$. If we had two such estimators T^A and T^B we would say that T^A is asymptotically more accurate than T^B if $\sigma_{T^A} < \sigma_{T^B}$. In general, we would prefer to use an estimator with a small σ_T, because it would be asymptotically more accurate—and, then, our CIs would be narrower. If we define the *precision* of an estimator to be $1/\sigma_T^2$, then we would prefer an estimator with the greatest precision. Fisher (1922) discovered that, in this sense, the precision of an estimator can not exceed a quantity known now as *Fisher information*.

For a family of distributions having pdf $f(x|\theta)$ the Fisher information is given by

$$I_F(\theta) = E\left(-\frac{d^2}{d\theta^2}\log f(x|\theta)\right)$$

and for a random sample drawn from this distribution the Fisher information is given by

$$I(\theta) = nI_F(\theta).$$

Because the expectation is used in defining $I(\theta)$, it is often called the *expected information* to distinguish it from the *observed information* which, as we discuss below, is used in data analysis.

Result Roughly speaking, for large samples, under certain general conditions, the precision $1/\sigma_T^2$ of a consistent and asymptotically normal estimator T can not exceed the Fisher information bound $I(\theta)$.

A detail: Rigorous formulations of this result involve the limiting precision. If we assume, in addition, that $\sigma_{T_n} \to 0$ at the rate $\sqrt{n}$

with $\lim \sqrt{n}\sigma_{T_n} = c_T(\theta)$ for some constant $c_T(\theta)$, then the limiting precision is $1/c_T(\theta)^2$ and we can obtain $1/c_T(\theta)^2 \leq I_F(\theta)$.

A carefully-stated theorem and proof involves too much mathematical development for us to go into it here. A particularly elegant framework for the theory involves exponential families; see Kass and Vos (1997).

A related result is the following. If T is an unbiased estimator of θ based on a random sample from the distribution with pdf $f(x|\theta)$, then

$$V(T) \geq (I(\theta))^{-1}. \tag{8.15}$$

This is usually called the *Cramér-Rao lower bound*. Although Equation (8.15) is of much less practical importance than the asymptotic result, authors often speak of the asymptotic Fisher information bound as a Cramér-Rao lower bound.

Fisher information also arises in theoretical neuroscience, particularly in discussion of neural decoding and optimal properties of tuning curves. (See Dayan and Abbott (2001). Dayan, P. and Abbott, L.F. (2001) *Theoretical Neuroscience*, MIT Press.)

8.2.3 Estimators that minimize large-sample variance are called efficient.

When an estimator attains the asymptotic precision bound in the result on page 8.2.2 it is said to be *efficient*. Fisher (1922) described efficient estimators by saying they contain the maximal amount of information supplied by the data about the value of a parameter—this is why $I(\theta)$ is called "Fisher information" and there are rigorous mathematical results that justify Fisher's use of these words. (See, for instance, Kass and Vos (1997).) Roughly speaking, the information in the data pertaining to the parameter value may be used well (or poorly) to make an estimator more (or less) accurate; in using as much information about the parameter as is possible, an efficient estimator uses the data most efficiently and reduces to a minimum the uncertainty attached to it. Other definitions of efficiency are used in statistical theory, but the one based on Fisher information remains most immediately relevant to data analysis.

8.3 Properties of ML Estimators

8.3.1 In large samples, ML estimation is optimal.

Although we do not here give careful statements of the optimality properties of ML estimation, it is important to state in rough terms the basic result.

Result Under certain general conditions, as $n \to \infty$ we have

$$\sqrt{I(\theta)}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.16}$$

That is, ML estimators are consistently asymptotically normal, and efficient.

In other words, when we consider what happens as $n \to \infty$, among all those "nice" estimators that are consistent and asymptotically normal, ML estimators are the best in the sense of having the smallest possible limiting standard deviation.

Results may also be derived in terms of MSE. Under certain conditions, an estimator T_n must satisfy

$$I(\theta) \cdot MSE(\hat{\theta}) \to c$$

where $c \geq 1$ and for the MLE $T = \hat{\theta}$) we have

$$I(\theta) \cdot MSE(\hat{\theta}) \to 1.$$

This is a different way of saying that, for large samples, ML estimation is as accurate as possible.

8.3.2 The standard error of the MLE is obtained from the second derivative of the loglikelihood function.

Although we have emphasized the theoretical importance of Equation (8.16), to be useful for data analysis it must be modified: the quantity $I(\theta)$ depends

on the unknown parameter θ, so we must replace $I(\theta)$ with an estimate of it. In other words, we must modify (8.16), to obtain a confidence interval. One possible such modification is fairly obvious, based on the way we modified initial asymptotic normality results in our discussion of confidence intervals in Section 7.3: we replace θ with the MLE $\hat{\theta}$. Under certain conditions we have

$$\sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.17}$$

Details: Because $\hat{\theta} \to \theta$ in probability (i.e., the MLE is consistent) we can again apply Slutsky's Theorem together with (8.16) to get (**??**).

It turns out that there is a more convenient version of the result. The difficulty with (8.17) is that in some problems it is hard to compute $I(\theta)$ analytically because of the required expectation. Instead, as a general rule, we replace $I(\theta)$ with the *observed information* given by

$$I_{OBS}(\hat{\theta}) = -\ell''(\hat{\theta}).$$

In other words, instead of the expected information evaluated at $\hat{\theta}$ in (8.17), we use the negative second derivative of the loglikelihood, evaluated at $\hat{\theta}$, without any expectation. (For the special class of models known as exponential families, which are used with the generalized linear models discussed in Chapter **??**, we have $I(\hat{\theta}) = I_{OBS}(\hat{\theta})$ (see, e.g., Kass and Vos, 1997) but this is not true in general.) Again, under certain conditions, we have

$$\sqrt{I_{OBS}(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.18}$$

Details: Note that

$$-\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

and that the expectation of the right-hand side is $I_F(\theta)$. From the LLN we therefore have

$$-\frac{1}{n}\ell''(\theta) \to I_F(\theta),$$

in probability, and it may also be shown that

$$\frac{I_{OBS}(\hat{\theta})}{I(\hat{\theta})} \to 1,$$

in probability, which, again by Slutzky's Theorem, gives (8.18). □.

Equation (8.18) provides large-sample standard errors and confidence intervals based on ML estimation.

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ satisfies (8.18), so that its standard error is given by

$$SE = \frac{1}{\sqrt{-\ell''(\hat{\theta})}} \tag{8.19}$$

and an approximate 95% CI for θ is given by $(\hat{\theta} - 2SE, \hat{\theta} + 2SE)$.

Additional insight about the observed information can be gained by looking at our paradigm case of the normal mean, discussed in Section 7.3.2. We find an interesting relationship. Differentiating the loglikelihood in Equation (7.1) twice we get

$$\ell''(\theta) = -\frac{n}{\sigma^2},$$

so that

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{-\ell''(\theta)}}. \tag{8.20}$$

In other words, $1/\sqrt{-\ell''(\theta)}$ gives the standard error of the mean in that case. Quite generally, for large samples, the likelihood function has an approximately normal form and there is a strong analogy with this paradagm case. Specifically, a second-order Taylor expansion of the loglikelihood function produces a normal likelihood having $1/\sqrt{-\ell''(\hat{\theta})^{-1}}$ as its standard deviation. This heuristic helps explain (8.19).

We now consider two simple examples.

Illustration: Exponential distribution Suppose $X_i \sim Exp(\lambda)$ for $i = 1, \ldots, n$, independently. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum x_i} \\ &= \lambda^n e^{-\lambda n \bar{x}} \end{aligned}$$

and the loglikelihood function is

$$\ell(\lambda) = n \log \lambda - n\lambda \bar{x}.$$

Differentiating this and setting equal to zero gives

$$0 = n(\frac{1}{\lambda} - \bar{x})$$

and solving this for λ yields the MLE

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

This is the estimator we discussed previously, when using Q-Q plots.

Continuing, we compute the observed information:

$$\begin{aligned} -\ell''(\hat{\lambda}) &= \frac{n}{\hat{\lambda}^2} \\ &= n\bar{x}^2 \end{aligned}$$

which gives us the large-sample standard error

$$SE(\hat{\lambda}) = \frac{1}{\bar{x}\sqrt{n}}.$$

□

Illustration: Binomial For a $B(n, p)$ random variable it is straightfoward to obtain the observed information

$$-\ell''(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}.$$

This gives

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which is the same as the SE found in Section 7.3.5. □

8.3.3 MLEs transform along with parameters.

It sometimes happens that we wish to consider an alternative parameterization of a pdf, say γ rather than θ, and then want find the MLE of γ. If $\gamma = g(\theta)$ for a transformation function g having nonzero derivative, then the MLE of the transformation equals the transformation of the MLE:

$$\hat{\gamma} = g(\hat{\theta}).$$

This is often called *invariance* or *equivariance.* The derivation of invariance of ML is perhaps most easily followed in a concrete example. The argument given next for the exponential distribution could be applied to any parametric family.

Illustration: Exponential distribution Suppose we parameterize the $Exp(\lambda)$ distribution in terms of the mean $\mu = 1/\lambda$ so that its pdf becomes

$$f(x) = \frac{1}{\mu} e^{-x/\mu}.$$

We previously found the MLE of λ based on a sample from $Exp(\lambda)$ is $\hat{\lambda} = 1/\bar{x}$. The invariance property of ML says that $\hat{\mu} = \bar{x}$.

To see why this works for the exponential distribution, let us use a subscript on the likelihood function to indicate its argument, $L_\lambda(\lambda)$ vs. $L_\mu(\mu)$. We find $L_\mu(\mu)$ by starting with

$$L_\lambda(\lambda) = \lambda^n e^{-\lambda n \bar{x}}$$

and writing

$$L_\mu(\mu) = L_\lambda(\frac{1}{\mu}) = \frac{1}{\mu^n} e^{-n\bar{x}/\mu}.$$

Thus, when we maximize $L_\mu(\mu)$ over μ, we are maximizing $L_\lambda(1/\mu)$ over μ which is the same thing as maximizing $L_\lambda(\lambda)$ over λ. We therefore must have $\hat{\mu} = 1/\hat{\lambda}$. More generally, the same argument shows that when $\gamma = g(\theta)$ we must have $\hat{\gamma} = g(\hat{\theta})$. □

Invariance is by no means a trivial property: some methods of estimation are *not* invariant to transformations of the parameter.

8.3.4 In large samples, ML estimation is approximately Bayesian.

In Chapter 7 we mentioned that Bayes' theorem may be used to provide a form of estimation based on the posterior distribution

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}.$$

One of the most important results in theoretical statistics is the approximate large-sample equivalence of inference based on ML and inference using Bayes' theorem.

Illustration: Binomial distribution Suppose $Y \sim B(n, \theta)$ with $n = 100$ and we observe $y = 60$. If we use the uniform prior, which is a $Beta(1, 1)$ distribution, we obtain a $Beta(61, 41)$ posterior. The observed proportion is the MLE $\hat{\theta} = x/n = .6$. The usual standard error then becomes $SE = \sqrt{\hat{\theta}(1-\hat{\theta})/n} = .049$. As shown in Figure 8.5 the normal distribution with mean $\hat{\theta}$ and standard deviation $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ is a remarkably good approximation to the posterior. □

The general result for the binomial may be stated as follows.

If $X \sim B(n, \theta)$ then θ may be estimated by $\hat{\theta} = X/n$ with standard error $SE = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$. An approximate 95% posterior probability interval is given by

$$\hat{\theta} \pm 2 \cdot SE,$$

meaning

$$P(\hat{\theta} - 2 \cdot SE \leq \theta \leq \hat{\theta} + 2 \cdot SE | X = x) \approx .95. \qquad (8.21)$$

For the data from subject P.S. in Example 1.2.1, which involves a small sample, we already noted that the usual approximate 95% confidence interval

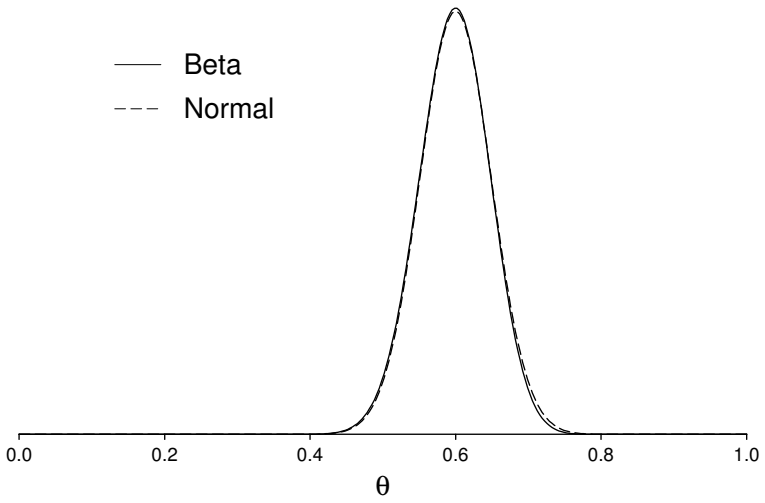


Figure 8.5: Normal approximation $N(.6, (.049)^2)$ to beta posterior $Beta(61, 41)$.

(.64, 1.0) differed by only a modest amount from the exact 95% posterior probability interval we obtained earlier, which was (.59,.94).

Result For large samples, under certain general conditions, the posterior distribution of θ satisfies is approximately normal with mean $\hat{\theta}$ and standard deviation given by the standard error formula (8.19).

8.3.5 Under normality, ML produces the weighted mean.

We now return to choosing the weights for a weighted mean, discussed in Section 8.1.3. Previously we found the weights that minimized MSE. A different way to solve the problem is to introduce a statistical model, and then apply the method of maximum likelihood. Let us do this.

To apply ML, we assume that X_1 and X_2 are both normally distributed. The loglikelihood is

$$\ell(\mu) = -\frac{(x_1-\mu)^2}{2\sigma_1^2} - \frac{(x_2-\mu)^2}{2\sigma_2^2}$$

and setting its derivative equal to zero gives

$$\begin{aligned} 0 &= -\frac{x_1-\mu}{\sigma_1^2} - \frac{x_2-\mu}{\sigma_2^2} \\ &= -\frac{x_1}{\sigma_1^2} - \frac{x_2}{\sigma_2^2} + \mu\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right). \end{aligned}$$

Therefore, multiplying through by $\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$, the MLE is

$$\hat{\mu} = w_1 \cdot X_1 + w_2 \cdot X_2,$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

for $i = 1, 2$. This is Equation (8.10).

8.4 Multiparameter Maximum Likelihood

The method of ML estimation was defined for the case of a scalar parameter θ in Section 7.2.2, together with Equations (8.18) and (8.19). More generally, when θ is a vector, the definitions of the likelihood function, log-likelihood function, and MLE remain unchanged. The observed information instead becomes a matrix, and the approximate normal distribution mentioned in conjunction with Equation (8.19) instead becomes an approximate *multivariate* normal distribution.

8.4.1 The MLE solves a set of partial differential equations.

In Section 7.2.2 we computed the MLE by solving the differential equation

$$0 = \ell'(\theta) \tag{8.22}$$

when θ was a scalar. To obtain the MLE of an m-dimensional vector parameter, we must solve precisely the same equation, except that now the derivative in Equation (8.22) is the vector

$$\ell'(\theta) = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_m} \end{pmatrix}$$

This means that Equation (8.22) is really a set of m equations, often called *the likelihood equations*, which need to be solved simultaneously.

Illustration: Normal MLE Let us return to finding the MLE for a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution. Previously we assumed σ was known, but now we consider the joint estimation of μ and σ. The log-likelihood function now must include a term previously omitted that involves σ. The joint pdf is

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

and the loglikelihood function is

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

The partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

Setting the first equation equal to 0 we obtain

$$\hat{\mu} = \bar{x}.$$

Setting the second equation equal to 0 and substituting $\hat{\mu} = \bar{x}$ gives

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

The MLE is thus slightly different than the usual sample standard deviation s, which is defined with the denominator $n - 1$ so that the sample variance becomes unbiased as an estimator of σ^2, as in (8.4). We have

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}} \cdot s.$$

Clearly the distinction is unimportant for substantial sample sizes. (We may obtain $\hat{\sigma} = s$ if we instead integrate out μ from the likelihood and then maximize the resulting function; this function is sometimes called an *integrated* or *marginal* likelihood, and in some situations maximizing the integrated likelihood yields a preferable estimator.) □

Illustration: Gamma MLE Let us rewrite the Gamma loglikelihood function:

$$\ell(\alpha, \beta) = n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i - n \log \Gamma(\alpha).$$

The partial derivatives are

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha} &= n \log \beta + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\partial \ell}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i\end{aligned}$$

where $\Gamma'(u)$ is the derivative of the function $\Gamma(u)$ (sometimes called the "digamma function"). Setting the second partial derivative equal to zero we obtain

$$\hat{\beta} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} x_i}.$$

When we set the first equation equal to zero and substitute this expression for $\hat{\beta}$, we get the nonlinear equation

$$n \log \hat{\alpha} - n \log \bar{x} + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

To obtain the MLE $(\hat{\alpha}, \hat{\beta})$ we may proceed iteratively: given a value $\hat{\beta}^{(j)}$ we can solve the first equation for $\hat{\alpha}^{(j+1)}$ and solve the second equation to obtain $\hat{\beta}^{(j+1)}$; we continue until the results converge. The second equation must be solved numerically, but it is not very difficult to use available software to do so. □

8.4.2 Least squares may be viewed as a special case of ML estimation.

In Example 1.2.2 we discussed data collected by Hursh (1939), indicating the linear relationship between a neuron's conduction velocity and its axonal diameter. We also briefly described the method of *least-squares regression*, based on the *linear regression model* (1.3), which is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Least-squares regression is discussed at length in Chapter 13. Here we show that the method of least squares may be considered a special case of ML estimation.

Least squares may be derived by assuming that the ϵ error variables in (**??**) are normally distributed, and that the problem is to estimate the parameter vector $\theta = (\beta_0, \beta_1)$. Specifically, we assume $\epsilon_i \sim N(0, \sigma^2)$, independently for all i. Calculation then shows that the ML estimate of θ is the least squares estimate. In other words, in the simple linear regression problem, ML based on the assumption of normal errors reproduces the least-squares solution.

Details: For a sample from a $N(\mu, \sigma^2)$ distribution we wrote down the loglikelihood function

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

and obtained the MLE $\hat{\mu} = \bar{x}$. Notice that, as a function of μ, the loglikelihood is maximized by minimizing the sum of squares $\sum_{i=1}^{n}(x_i - \mu)^2$. Thus, the MLE $\hat{\mu} = \bar{x}$ is also the least-squares estimator. The same is true much more generally.

For simple linear regression we may write a statistical model in the form

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently. The loglikelihood function then becomes

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

We can maximize $\ell(\beta_0, \beta_1, \sigma)$ by first defining $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ to be the maximum of $\ell(\beta_0, \beta_1, \sigma)$ over (β_0, β_1) for fixed σ, and then maximizing $\ell(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma), \sigma)$ over σ. However, from inspection of the formula above, for every σ the solution $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ (the maximum of $\ell(\beta_0, \beta_1, \sigma)$) is found by minimizing the sum of squares $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Therefore, the MLE $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ has the least-squares estimate as its first two components. □

8.4.3 The observed information is the negative of the matrix of second partial derivatives of the log-likelihood function, evaluated at $\hat{\theta}$.

In the multiparameter case the second derivative $\ell''(\theta)$ becomes a matrix,

$$\ell''(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} & \cdots & \frac{\partial^2 \ell}{\partial \theta_m^2} \end{pmatrix}.$$

(This second-derivative matrix is often called the *Hessian* of $\ell(\theta)$.) The *observed information matrix* is $-\ell''(\hat{\theta})$.

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ of the m-dimensional parameter θ is distributed approximately as an m-dimensional multivariate normal random vector with variance matrix

$$\hat{\Sigma} = -\ell''(\hat{\theta})^{-1}. \tag{8.23}$$

Example 8.4.1 Human perception of dim light Hecht *et al* (1942, *J. General Physiology*) investigated the threshold for visual perception by exposing human observers to very weak flashes of light in a darkened room. The response variable was an indication of whether or not light was observed ("yes" or "no"), and the explanatory variable was the intensity of the light (in units of average number of light quanta per flash). Several different intensities were used, and for each the experiment was repeated many times. (Similar experiments were performed by Hartline et al (1947) on the horseshoe crab *Limulus*, where the response was the proportion of trials on which an isolated optic nerve fiber fired.) The results for one series of trials in one subject are plotted in Figure 8.6.

The linear regression model (**??**) does not work very well in this example. As illustrated in the top part of Figure 8.6, the proportions vary between 0

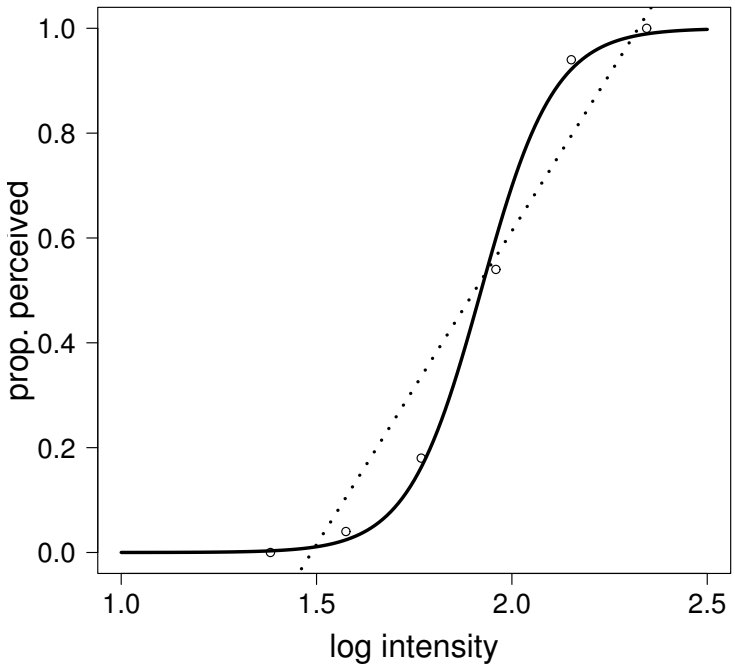


Figure 8.6: *Proportion of trials, out of 50, on which light flashes were perceived by subject S.S. as a function of log (base 10) intensity, together with fits. Dashed line is fit obtained by linear regression. Solid curve is fit obtained by logistic regression. Data from Hecht* et al. *(first series of trials).*

and 1 but a line $y = a + bx$ is unrestricted and thus can not represent the variation accurately, at least not for proportions that get close to 0 or 1. A solution is to replace the line $y = a + bx$ by a sigmoidal curve, which goes to zero as the explanatory variable x goes to $-\infty$ and increases to one as the $x \to \infty$, as illustrated in the bottom part of Figure 8.6. This fitted curve is based on the following statistical model: for the i-th value of light intensity we let Y_i be the number of light flashes on which the subject perceives light and then take

$$Y_i \sim B(n_i, p_i) \tag{8.24}$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.25}$$

This is known as the *logistic regression model*. There are many possible approaches to estimating the parameter vector $\theta = (\beta_0, \beta_1)$ but the usual solution is to apply maximum likelihood. The observed information matrix is then used to get standard errors of the coefficients. These calculations are performed by most statistical software packages. For the data in Figure 8.6 we obtained $\hat{\beta}_1 = 10.7 \pm 1.2$. Interpretation of this result is given in Chapter **??**. □

8.4.4 When using numerical methods to implement ML estimation, some care is needed.

There are three issues surrounding the application of numerical maximization to ML estimation. The first is that, while loglikelihood functions are usually well behaved near their maxima, they may be poorly behaved away from the maxima. In particular, a loglikelihood may well have multiple smaller peaks, and numerical methods often get stuck in a region away from the actual maximum. It is therefore essential to begin an iterative algorithm with a good preliminary estimate. Sometimes models may be altered and simplified in some way to get guesses at the parameter values. Sometimes the method of moments can be applied.

Illustration: Gamma distribution Recall that the method of moments estimator for the Gamma distribution is

$$\beta^* = \frac{\bar{x}}{s^2}$$

$$\alpha^* = \frac{\bar{x}^2}{s^2}.$$

In order to obtain the MLE of (α, β) we may use the iterative scheme described earlier, beginning with $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}) = (\alpha^*, \beta^*)$. □

With good initial values, the iterative maximization software usually only needs to run for a few iterations, after which the estimates don't change by more than a small fraction of the statistical uncertainty. (It may be shown, theoretically, that from any consistent estimator for which the MSE vanishes at the rate $1/n$, a single iteration of Newton's method for maximizing the loglikelihood function will produce an efficient estimator; see Lehmann (1983).)

Worries about initial values are substantially mitigated when the loglikelihood function is concave—because in that case it has a unique maximum. This occurs in exponential families (see Section **??**) for certain parameterizations and, as a result, for certain *generalized linear models*, which are discussed in Chapter **??**. One such generalized linear model is the logistic regression model of Example 8.4.1.

A second issue is that the second derivatives used in numerical maximization software are often themselves estimated numerically, and they are often estimated rather poorly (because they do not need to be estimated accurately to obtain the maximum). Thus, for the purpose of finding a variance matrix, one should either evaluate second derivatives separately (from an analytical formula, or from special-purpose software), or one should apply the parametric bootstrap (see Section 9.2).

The third issue is that parameterization can be important. Numerical maximization procedures tend to work well when the loglikelihood function is roughly quadratic, which means that the likelihood function is approximately normal. Transformations of parameters can improve this approximation. For example, it is often helpful to transform variance parameters by taking logs.

8.4.5 Maximum likelihood may produce bad estimates.

The method of ML is not universally applicable, nor does it guarantee good statistical results. The most serious concern with ML is that it is predicated on the description of the data according to a particular statistical model. If that model is seriously deficient, the MLE will be misleading. This underscores the essential role of model assessment, and the iterative nature of model building, emphasized in Chapter 1.

The provably good performance of ML estimation also applies only for large samples. What constitutes "large" is difficult to specify precisely, though attempts have been made occasionally. A key observation is that sample size must be judged relative to the number of parameters being estimated. In problems having large numbers of parameters and only modest sample sizes, we should expect neither ML estimates, nor their associated SEs, to be accurate. One standard approach to making progress in such situations is to build models that effectively reduce the number of parameters by restricting them in some way (often by introducing additional probability distributions). In some cases, however, ML must be abandoned. There is a large body of methods that are *nonparametric*, in the sense that they do not posit a statistical model with a finite number of parameters. There are many situations where nonparametric methods perform well, and save the difficulty and worry associated with careful model building.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 9

Propagation of Uncertainty and the Bootstrap

A fundamental goal of statistical analysis is to describe uncertainty. Statistical inference uses probabilistic descriptions of the variation in data to arrive at quantitative assessments of uncertainty about estimates of unknown parameters. The uncertainty given by a particular method, however, may not be in the form we need. More specifically, some statistical software may give us the covariance matrix of a random vector X, but we may wish to obtain the uncertainty about a function of that vector, $Y = f(X)$. Here is an example.

Illustration: Difference index for firing rates In single-unit electrophysiological studies, neural firing rates are often estimated under two experimental conditions. Let us label the conditions A and B, and suppose that for each neuron we have many trials of recordings under each of the conditions. Averaging across the trials gives sample mean firing rates, $\bar{X}_A$ and $\bar{X}_B$, which may be compared. However, comparisons are made across many neurons having quite different firing rates. For this reason, some sort of normalization is usually invoked. One commonly-used comparative measure

is the index

$$Y = \frac{\bar{X}_A - \bar{X}_B}{\bar{X}_A + \bar{X}_B}. \tag{9.1}$$

How should we obtain its standard error? It is easy enough to get standard errors $SE(\bar{X}_A)$ and $SE(\bar{X}_B)$. If s_A and n_A are the sample standard deviation and sample size under condition A, then we may take $SE(\bar{X}_A) = s_A/\sqrt{n_A}$, and similarly for condition B. We need a way of using the uncertainties $SE(\bar{X}_A)$ and $SE(\bar{X}_B)$ to get $SE(Y)$. □

To put this illustration in the more general framework we write $X_1 = \bar{X}_A$, $X_2 = \bar{X}_B$, $X = (X_1, X_2)$, and then

$$f(x) = \frac{x_1 - x_2}{x_1 + x_2}.$$

The problem of finding the standard error of Y defined by (9.1) then becomes a special case of the general problem of finding the standard error of $Y = f(X)$ when the uncertainty in X is known. This problem of transferring uncertainty about a random vector X to a random variable $Y = f(X)$ is the problem of propagation of uncertainty, or what was historically called "propagation of error" and, sometimes, "the delta method."

Here is another example.

Example 8.4.1 (continued) We previously displayed data from Hecht *et al* (1942), who investigated the threshold for visual perception by exposing human observers to very weak flashes of light in a darkened room. In the bottom part of Figure 8.6 we overlaid on the data a sigmoidal curve found from applying maximum likelihood to the logistic regression model

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \end{aligned}$$

We reported the values of the fitted coefficients and their standard errors.

Those data were from a single subject. What if we wanted to compare results across subjects? We would get a set of sigmoidal curves with somewhat different slopes, shifted to some extent to the left or right. One simple way to characterize the ability of a subject to perceive the dim light is the

intensity at which he or she will perceive it 50% of the time. This number easy to understand and it corresponds to the middle of the curve, thus being a nice single-number representation of the data. Let us label this value of the intensity x_{50}. Mathematically, to find it, we set $p = .5$ in the equation above, with the fitted coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ replacing β_0 and β_1. This gives

$$.5 = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{50})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{50})}$$

which we solve for x_{50}. Using the uncertainty in $(\hat{\beta}_0, \hat{\beta}_1)$ to obtain $SE(x_{50})$ is a problem in propagation of uncertainty. □

There are several varieties of propagation of uncertainty. The original method, historically, used mathematical analysis with $n \to \infty$ to derive an approximate $SE(Y)$ based on an approximate variance matrix for X. We discuss this in Section 9.1.1. This kind of analysis is relatively easy in many cases. It is often even easier to use a brute force computer simulation: if we can generate observations (on the computer) from the approximate distribution of X, we can also immediately obtain the approximate distribution of Y. We explain the idea, enumerating the steps, in Section 9.1.2. Realization of the simplicity and power of this computational approach was first clearly articulated by B. Efron when he introduced the bootstrap, but the bootstrap also involves an additional idea: the substitution of the empirical cdf $\hat{F}_n(x)$ for the unknown theoretical cdf $F(x)$. We discuss the bootstrap in Section 9.2. Propagation of uncertainty is also an essential part of modern Bayesian methods, which will appear in Chapter **??**.

9.1 Propagation of Uncertainty

We begin with the analytical approach to propagating uncertainty. Let us suppose we have a random variable or vector X, and a function $y = f(x)$, which we wish to apply to X. This will produce a random variable $Y = f(X)$. A good handful of special cases have been analyzed in the literature (mostly many, many years ago), which leads to some standard distributions we have already discussed. In practice, however, one often comes across cases that do not fit any specialized framework. Fortunately, there is a simple and powerful

method that may be applied in conjunction with a general theoretical result in order to get the approximate distribution of Y.

9.1.1 Functions of approximately normal random vectors are approximately normal.

Suppose, first, that X is a random variable having mean μ_X and standard deviation σ_X. The classical idea behind what is often called *the delta method* has three parts: first, the distribution of X is assumed to be concentrated around μ_X (so that σ_X is small); second, the function $f(x)$ is approximated linearly near μ_X; third, the approximating linear tranformation is applied to X to get the approximate distribution of Y. This works when (i) the distribution of X is, in fact, concentrated near μ_X, (ii) the function $f(x)$ is roughly linear in a range of highly probable values of X, such as $\mu_X \pm 2.5\sigma_X$, and (iii) the distribution of a new random variable having the form $a + bX$ is useful; this occurs when X is normally distributed (or approximately normally distributed) because then $a+bX$ itself is normal (or approximately normal). In fact, this approach is applied mainly when X is approximately normal.

Theorem Suppose that a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ satisfies

$$\frac{X_n - \mu}{\sigma_{X_n}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$, and that the function $f(x)$ is continuously differentiable with $f'(\mu) \neq 0$. Then

$$\frac{f(X_n) - f(\mu)}{\sigma_{Y_n}} \xrightarrow{D} N(0,1)$$

with $\sigma_{Y_n} = |f'(\mu)|\sigma_{X_n}$.

> *Proof:* We omit the proof, which is a consequence of Slutzky's theorem, page **??**, but give the essential idea below. □

We now explain why this result works. First, recall that when $f(x)$ is linear and X is normally distributed, $Y = f(X)$ is also normally distributed.

Specifically, if $Y = a + bX$ and $X \sim N(\mu_X, \sigma_X^2)$ then $Y \sim N(\mu_Y, \sigma_Y^2)$ with $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$. A pictorial display of this situation is given in Figure 9.1.

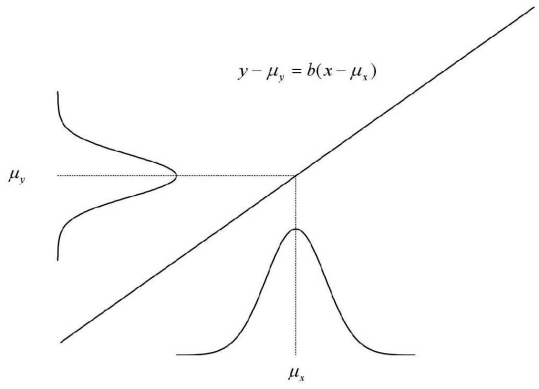


Figure 9.1: *The effect of the transformation $y = a + bx$ operating on a normally distributed random variable X having mean μ_X and standard deviation σ_X. The random variable $Y = a + bX$ is again normally distributed, with mean $\mu_Y = a + b\mu_X$ and standard deviation $\sigma_Y = |b|\sigma_X$. The normal distributions are displayed on the x and y axes; the linear transformation is displayed as a line, which passes through the point (μ_X, μ_Y) so that it may be written, equivalently, as $y - \mu_Y = b(x - \mu_X)$.*

Now, suppose that $f(x)$ is not linear, but let us assume that it is only mildly nonlinear within the "most probable" range of X. That is, $f(x)$ is mildly nonlinear within, say, $\mu_X \pm 2.5\sigma_X$, which is the range over which we are assuming X to be approximately normally distributed. Then we may approximate $f(x)$ with the best-fitting linear approximation at $x = \mu_X$:

$$f(x) \approx f(\mu_X) + f'(\mu_X)(x - \mu_X)$$

which is usually called a first-order *Taylor series* at $x = \mu_X$. That is, we have

$$f(x) \approx a + bx$$

with $a = f(\mu_X) - f'(\mu_X)\mu_X$ and $b = f'(\mu_X)$. Note that $a + b\mu_X = f(\mu_X)$. As a result, we have that $Y = f(X)$ is approximately normally distributed, with $\mu_Y \approx f(\mu_X)$ and $\sigma_Y \approx |f'(\mu_X)|\sigma_X$.

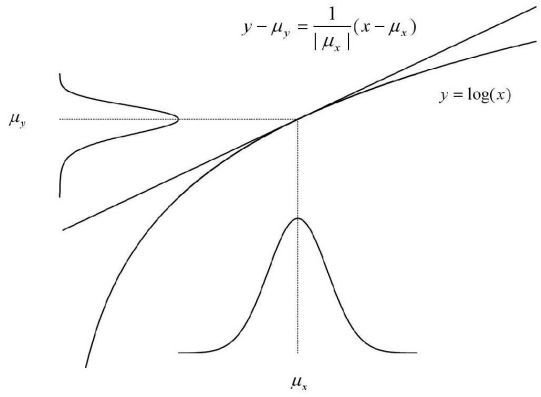


Figure 9.2: *The transformation* $y = \log(x)$ *operating on a normally distributed (or* approximately *normally distributed) random variable* X *having mean* μ_X *and standard deviation* σ_X *produces an* approximately *normally distributed random variable* Y *with mean and standard deviation* approximately *given by* $\mu_Y = \log(\mu_X)$ *and* $\sigma_Y = \sigma_X/|\mu_X|$. *The approximating line could also be written in the form* $y - \mu_Y \approx (x - \mu_X)/|\mu_X|$.

We now re-state this theorem in a less mathematically precise but more practical form. First, a detail for those who wish to follow the math fairly closely.

Result: Propagation of Uncertainty in the Scalar Case
If X is approximately $N(\mu_X, \sigma_X^2)$ and the function $f(x)$ is approximately linear with $f'(x) \neq 0$ near μ_X ("near" being defined probabilistically, in terms of σ_X), then
(1) $Y = f(X)$ is approximately normal, and
(2) the approximate normal mean and standard deviation are given by $\mu_Y \approx f(\mu_X)$ and $\sigma_Y \approx |f'(\mu_X)|\sigma_X$.

Note that both conclusions in this result are important: subsequently we will rely on the approximate normality in (1) using computer simulation in place of the analytical formula for the standard deviation appearing in (2). On the other hand, the formulas are sometimes valuable.

> *A detail:* Here is a technical point. In the statement of the theorem the numbers σ_{X_n} do not have to be the standard deviations of X_n. They can, instead, be some numbers that will serve as the approximate standard deviations. In practice, we often do not have the exact standard deviation but we do have a useful approximate value based on large-sample theory, as in Chapter **??**.

Illustration: Log transformation Suppose $f(x) = \log(x)$. Then $f'(x) = 1/x$, so that if X is approximately normal, with small σ_X, then Y is approximately normal with $\mu_Y \approx \log(\mu_X)$ and $\sigma_Y \approx \sigma_X/|\mu_X|$. The picture is given in Figure 9.2. Careful examination of the Figure reveals that the distribution of Y is not exactly normal (it is mildly skewed toward low values), but it is close. □

The illustration above, using the log transformation, serves mainly to show how the analytical calculation works in propagation of uncertainty. However, it also has a practical application to what are known as *variance-stabilizing transformations.*

Illustration: Variance-Stabilizing Transformations In common statistical procedures that compare observations across groups or experimental conditions, a simplifying assumption is that all the observations have the same variance. Similarly, one of the standard assumptions in linear regression is that the noise or error has the same variance for all observations. Sometimes, however, this is clearly violated. Suppose it is found, empirically, that the standard deviation is proportional to the mean. The illustration above shows that the log transformation removes this effect, and makes the variances homogeneous across observations.

Specifically, suppose we have a batch of random variables $X_1, \ldots, X_m$ for which σ_{X_i} is proportional to μ_{X_i}, with all $\mu_{X_i} > 0$. We may write this using the proportionality symbol as

$$\sigma_{X_i} \propto \mu_{X_i}. \tag{9.2}$$

For definiteness, let us assume the proportionality constant is c, so we have

$$\sigma_{X_i} = c\mu_{X_i}.$$

Now let $Y_i = \log(X_i)$. Then, by the analysis in the previous illustration, using $|\mu_{X_i}| = \mu_{X_i}$ because $\mu_{X_i} > 0$, we obtain

$$\sigma_{Y_i} \approx c.$$

In this context the log transformation is called *variance stabilizing*. Improving homogeneity of variances, making them more nearly equal, is one of the motivations for the log transformation in data analysis. □

Example 2.2.1 (continued) As part of their argument that it may be advantageous to transform high-field BOLD signal in fMRI data by taking logarithms, Lewis *et al.* (2005) provided plots of the standard deviation versus the mean for the BOLD signal and for the log-transformed BOLD signal. These plots are shown in Figures 9.3 and 9.4. The standard deviation is nearly proportional to the mean for the BOLD signal, but shows no relationship to the mean of the log-transformed BOLD signal. Because standard statistical procedures assume the standard deviation is more or less constant regardless of the mean, the authors suggested that taking logs might be a good idea.[1] □

[1] A subtlety here is that the assumptions of the analyses are not exactly of the form being examined by the authors; nonetheless, their results suggest that the issue be considered by any high-field fMRI data analyst.

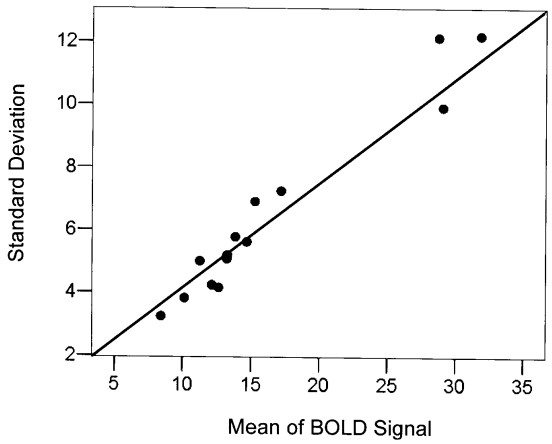


Figure 9.3: *Plot of standard deviation versus mean in BOLD signal across 15 subjects, from Lewis* et al. *(2005). The plot is nearly linear, so the standard deviation is very nearly proportional to the mean.*

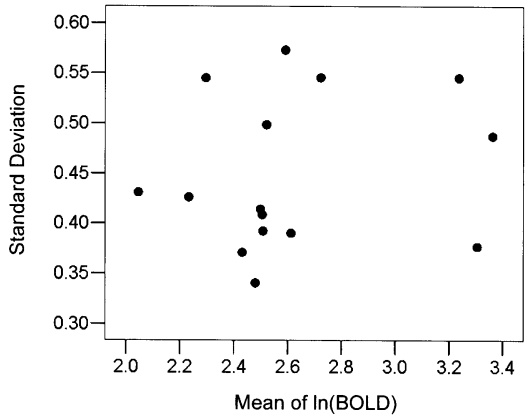


Figure 9.4: *Plot of standard deviation versus mean of log-transformed BOLD signal across 15 subjects, from Lewis* et al. *(2005). Here, in contrast to Figure 9.3, the standard deviation is approximately constant and shows no fixed relationship with mean.*

Example 9.1.1 Square-root transformation of spike counts in motor cortex When the variance of a bunch of spike counts is plotted against the mean it often happens that they are roughly proportional. That is, the spike counts $X_1, \ldots, X_m$ satisfy the proportionality relationship (9.3), at least approximately. (There are many references to this phenomenon; see Shadlen and Newsome, 1998, for some of them.) Let us rewrite this as in Equation

$$\sigma_{X_i} = c\mu_{X_i} \tag{9.3}$$

for some proportionality constant c. By examining the analysis in the foregoing illustrations of the log transformation it becomes apparent that a similar trick may be used here. From the propagation of uncertainty result $\sigma_Y \approx |f'(\mu_X)|\sigma_X$, together with (9.3) we have

$$\sigma_Y \approx |f'(\mu_X)|c\sqrt{\mu_X}. \tag{9.4}$$

In order to remove the effects in (9.3) we therefore should find $f(x)$ such that

$$f'(x) \propto 1/\sqrt{x} \tag{9.5}$$

because that will force the factors $|f'(\mu_X)|$ and $\sqrt{\mu_X}$ to cancel. The square-root function does the job: if $f(x) = \sqrt{x}$ then (9.5) is satisfied. For this reason, many authors have chosen to use square-root transformations of spike counts in their statistical analyses. In particular, Georgopoulos and Ashe (2000) (Georgopoulos, A.P. and Ashe, J. (2000) One motor cortex, two different views, *Nature Neuroscience*, 3: 963.) reported improvements when fitting spike counts to direction of movement by linear regression. For a similar reason, Yu et al. (2009) (Yu, B.M., Cunningham, J.P., Santhanam, G., Ryu, S.L., Shenoy, K.V., and Sahani, M. (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity, *J. Neurophysiology*, 102: 614–635.) used square-root transformations of spike counts in studying "neural trajectories" that summarize population activity in motor cortex during movement planning. □

We now extend the propagation of uncertainty argument to the vector case, which involves a multivariate first-order Taylor series expansion. The idea is to take a sequence of random vectors $X_1, X_2, \ldots$ that are approximately multivariate normal and apply the function $f(x)$ to each of them

and, as in the scalar case above, approximate $f(x)$ using a first-order Taylor series based on the derivative of $f(x)$. In this multidimensional case the derivative becomes the vector of partial derivatives. Specifically, for a vector x we let $f'(\mu)$ be the vector of partial derivatives (with respect to all components) of the real-valued function $f(x)$, evaluated at $x = \mu$. That is, the ith component of this derivative is

$$f'(\mu)_i = \frac{\partial f}{\partial x_i}|_{x=\mu}.$$

Theorem Let μ be an m-dimensional vector, and let $f(x)$ be a differentiable function for which $f'(\mu) \neq 0$. If $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of m-dimensional random vectors and Σ_n is a sequence of positive definite symmetric matrices such that for every nonzero m-dimensional vector w,

$$w^T \Sigma_n^{-1/2}(X_n - \mu) \xrightarrow{D} N(0,1),$$

then, writing $Y_n = f(X_n)$, we have

$$\frac{(Y_n - f(\mu))}{\sigma_Y} \xrightarrow{D} N(0,1) \tag{9.6}$$

where

$$\sigma_Y = \sqrt{f'(\mu)^T \Sigma_n f'(\mu)}.$$

Proof: Omitted. □

Here is the practical form of the method.

Result: Multivariate Propagation of Uncertainty If X is approximately multivariate normal, given by $N_m(\mu_X, \Sigma_X)$, and the function $f(x)$ is approximately linear with $f'(x) \neq 0$ near μ_X ("near" again being defined probabilistically), then
(1) $Y = f(X)$ is approximately normal, and
(2) the approximate normal mean and standard deviation are given by $\mu_Y \approx f(\mu_X)$ and

$$\sigma_Y \approx \sqrt{f'(\mu_X)^T \Sigma_X f'(\mu_X)}.$$

To see how we get this, consider the bivariate case. If we have $z = f(x, y)$ and we apply a first-order Taylor series expansion (a linear approximation) near a point (x_0, y_0), we get

$$z \approx f(x_0, y_0) + \frac{\partial f}{\partial x}|_{(x_0,y_0)}(x - x_0) + \frac{\partial f}{\partial y}|_{(x_0,y_0)}(y - y_0).$$

Analogously to what was done in the scalar case, we insert random variables X and Y and replace (x_0, y_0) with (μ_X, μ_y). With $Z = f(X, Y)$ we note that the first term in the variance $\sigma_Z^2 = V(Z)$ is $V(f(x_0, y_0)) = 0$ (because the variance of a constant is 0), and we then get

$$\begin{aligned}\sigma_Z^2 &= \left(\frac{\partial f}{\partial x}|_{(x,y)=(\mu_X,\mu_Y)}\right)^2 \cdot \sigma_X^2 + \left(\frac{\partial f}{\partial y}|_{(x,y)=(\mu_X,\mu_Y)}\right)^2 \cdot \sigma_Y^2 \\ &\quad +2 \cdot \frac{\partial f}{\partial x}|_{(x,y)=(\mu_X,\mu_Y)} \frac{\partial f}{\partial y}|_{(x,y)=(\mu_X,\mu_Y)} \rho\sigma_X\sigma_Y.\end{aligned}$$

The general multidimensional case is analogous.

Illustration: Difference index for firing rates (Continued) Let us return to the calculation of the variance of the index in (9.1). We define

$$f(x, y) = \frac{x - y}{x + y},$$

compute its partial derivatives, and apply the result under the assumption that $\bar{X}_A$ and $\bar{X}_B$ are independent. This is a reasonable assumption because $\bar{X}_A$ and $\bar{X}_B$ are based on separate recordings for different conditions. Calculations give the approximate standard error

$$SE(Z) = \sqrt{V(Z)} \tag{9.7}$$

with

$$V(Z) = \left(\frac{2\mu_B}{(\mu_A + \mu_B)^2}\right)^2 \sigma_A^2 + \left(\frac{2\mu_A}{(\mu_A + \mu_B)^2}\right)^2 \sigma_B^2. \tag{9.8}$$

Details: Recall the formula for the derivative of a ratio: if $h(x) = \frac{F(x)}{G(x)}$ then

$$h'(x) = \frac{F'(x)G(x) - F(x)G'(x)}{G(x)^2}.$$

Applying this, we obtain

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{x+y-(x-y)}{(x+y)^2} \\ &= \frac{-2y}{(x+y)^2}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial f}{\partial y} &= \frac{-(x+y)-(x-y)}{(x+y)^2} \\ &= \frac{-2x}{(x+y)^2}.\end{aligned}$$

Using independence of $\bar{X}_A$ and $\bar{X}_B$ the covariance term in the variance formula is zero and we therefore (after dropping the minus signs in the numerators of the partial derivatives because we are squaring those derivatives) obtain (9.8). □

Example 12.4.1 (continued) The rank index examined by Behseta *et al.* (2009) was $I_{\text{rank}} = \frac{(f_3-f_1)}{(f_3+f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials; and the reward index was $I_{\text{reward}} = \frac{(f_b-f_s)}{(f_b+f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. These had the form considered in the illustration above and, in computing their correction for attenuation of the correlation, Behseta *et al.* applied (9.7) with (9.8). □

9.1.2 Random draws from the distribution of the random variable X produce random draws from the distribution of the random variable $Y = f(X)$.

Calculations of the kind in the difference index illustration on 250 are just complicated enough that they may seem either painful or dangerous, in the sense that mistakes are possible. A remarkably effective way to propagate uncertainty, which may also reduce the chance of overlooking a math error, is to use numerical simulation. To understand the method, one must first be

sure to understand how to work with a probability distribution based on a transformation $y = f(x)$. Let us consider a simple example.

Illustration: Three possible values Suppose X can take the values 2,4, or 8 with probabilities .2,.5,.3, respectively, and we are interested in the transformation $y = \log_2(x)$. Then Y can take the values 1,2, or 3. To find the probability distribution of Y we simply note that

$$\begin{aligned} P(Y=1) &= P(\log_2(X)=1) = P(X=2) = .2 \\ P(Y=2) &= P(\log_2(X)=2) = P(X=4) = .5 \\ P(Y=3) &= P(\log_2(X)=3) = P(X=8) = .3. \end{aligned}$$

Thus, for example, if we wanted to find the mean of Y we would obtain

$$\begin{aligned} \mu_Y &= 1 \cdot P(Y=1) + 2 \cdot P(Y=2) + 3 \cdot P(Y=3) \\ &= 1 \cdot (.2) + 2 \cdot (.5) + 3 \cdot (.3). \\ &= 2.1. \end{aligned}$$

□

The calculation in the discrete case (as above) is very simple. In the continuous case, to get the pdf we would have to introduce a derivative factor $|\frac{dy}{dx}|$, as ordinary calculus requires when a variable is transformed. (This would be needed, for example, to derive the χ^2_1 pdf from the definition $Y = X^2$ with $X \sim N(0,1)$.) We will not pursue such calculations here. The point is that once we know the probabilities for X, we immediately obtain them for Y. In fact, in most data analysis situations one need not bother with these calculations anyway, because practical results may be obtained from computer simulation. Suppose we can, on the computer, generate observations ("draws") from the distribution of X, and let us denote a set of G such simulated observations by $U^{(1)}, U^{(2)}, \ldots, U^{(G)}$. If we define $W^{(1)} = f(U^{(1)}), W^{(2)} = f(U^{(2)}), \ldots, W^{(G)} = f(U^{(G)})$, we obtain a set of G draws from the distribution of Y.

Illustration: Three possible values (continued) In the discrete example above, suppose we wanted to find $P(Y = 1)$ without using the formula $P(Y = 1) = P(X = 2) = .2$. We could get an approximate answer by the following procedure:

1. For $j = 1$ to 10,000:

   Generate $U^{(g)}$ from the distribution of X.

   Compute $W^{(g)} = \log_2(U^{(g)})$.

2. Let N be the number of $W^{(g)}$ such that $W^{(g)} = 1$ and compute

$$P(Y = 1) \approx \frac{N}{10,000}.$$

To compute the mean of Y we could follow the same step 1, and then replace step 2 with

$$\mu_Y \approx \frac{1}{G} \sum_{j=1}^{G} W^{(g)}.$$

□

This computer-simulation procedure works for discrete and continuous random variables and random vectors.

Result: Simulation-Based Propagation of Uncertainty Suppose the random variable or random vector X has a probability distribution from which we are able to simulate observations, and we wish to find the distribution of a random variable $Y = f(X)$ defined by a real-valued function $f(x)$. Proceed as follows:

1. For $j = 1$ to G:

   Generate $U^{(g)}$ from the distribution of X.

   Compute $W^{(g)} = f(U^{(g)})$.

2. Step 1 gives us a sample $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$ from the distribution of Y. We can obtain whatever information we wish about the distribution of Y by taking G to be sufficiently large. In particular,

   (i) to get $P(a < Y < b)$ let N be the number of $W^{(g)}$ such that $a < W^{(g)} < b$ and compute

   $$P(a < Y < b) \approx \frac{N}{G}.$$

   (ii) To get σ_Y, compute the sample mean $\overline{W} = \frac{1}{G}\sum_{g=1}^{G} W^{(g)}$ and use the sample variance $\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2$ to get

   $$\sigma_Y \approx \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}. \tag{9.9}$$

The procedure is very general: it is applicable as long as it is possible to generate observations from the distribution of X. (The problem of creating algorithms that generate observations from a given distribution is itself a sub-specialty field of research; some additional comments about this may be found in Chapter **??**.) When we use simulation-based propagation of uncertainty together with the approximate normality of Y, due to the results in Section 9.1.1, we have a very powerful inference engine: we can apply them,

together, to obtain approximate 95% CIs in a wide variety of settings.

Result: Simulation-Based Propagation of Uncertainty in Estimation
Suppose X is an approximately multivariate normal estimator of θ having estimated covariance matrix $\hat{\Sigma}$, and we want to estimate $\phi = f(\theta)$ for some real-valued function $f(x)$. Let us take $Y = f(X)$ to be the estimator of ϕ. We will write the observed estimate of θ as $X = \hat{\theta}$ and the observed estimate of ϕ as $Y = \hat{\phi} = f(\hat{\theta})$. If the function $f(x)$ is approximately linear near $x = \hat{\theta}$ and $f'(\hat{\theta})$ is not the zero vector (i.e., not all of its partial derivatives are zero) then
(1) Y is approximately normally distributed, and
(2) the standard error obtained from (9.9) by simulation-based propagation of uncertainty

$$SE(\hat{\phi}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2} \tag{9.10}$$

furnishes approximate inferences. In particular, an approximate 95% CI is given by $(Y - 2SE(Y), Y + 2SE(Y))$.

Illustration: Exponentially-distributed ion channel opening durations We previously displayed data from Calquhoun and Sakmann (1985) which showed the distribution of ion channel opening durations to be consistent with exponentially-distributed durations. To illustrate propagation of uncertainty we created artificial data similar to the data displayed earlier: we simulated 200 values $x_1, x_2, \ldots, x_{200}$ from an $Exp(\lambda)$ distribution with $\lambda = 10$ milliseconds. We now use these data to ask, What is the probability that an ion channel will be open more than 20 milliseconds?

The probability we wish to obtain is $P(X > 20)$ where $X \sim Exp(\lambda)$. Let us write this as

$$\phi = P(X > 20) = 1 - F_\lambda(20)$$

where $F_\lambda(x)$ is the cdf of the $Exp(\lambda)$ distribution. One way to estimate this probability is to count the proportion of observations that are greater than 20. Because it is an observed proportion let us call this $\hat{p}$. Its standard error

is $\sqrt{\hat{p}(1-\hat{p})/n}$. That is an acceptable answer, even a good one (especially because the proportion is still a good estimator when the exponential distribution is not correct). However, let us also apply maximum likelihood and propagation of uncertainty, estimating ϕ as follows: (i) compute the MLE $\hat{\lambda}$, (ii) compute the MLE $\hat{\phi}$, (iii) compute the SE of $\hat{\lambda}$, (iv) propagate the uncertainty to get the standard error of $\hat{\phi}$.

There are two ways to propagate the uncertainty. We can apply the calculus-based formula, or we can use computer simulation. We begin both with the same first three steps (i), (ii), and (iii). Step (iv) is where the analytical and numerical methods differ.

Here are the steps in detail, but without numerical values.

(i) From our previous calculations (page **??**), the MLE of λ is $\hat{\lambda} = 1/\bar{x}$.

(ii) As in Section **??** we get the MLE of ϕ by plugging $\hat{\lambda}$ in for λ so that $\hat{\phi} = 1 - F_{\hat{\lambda}}(20)$.

(iii) From our previous calculations (page **??**), we have $SE(\hat{\lambda}) = \hat{\lambda}/\sqrt{n}$, where now $n = 200$.

(iv) *method 1, analytical:*

Writing $f(\lambda) = 1 - F_\lambda(20)$ we recall (from page **??**) that $F_\lambda(x) = 1 - \exp(-\lambda x)$ so that $f'(\lambda) = -20\exp(-\lambda 20)$. Plugging $\hat{\lambda}$ into the formula $SE(\hat{\phi}) = |f'(\hat{\lambda})|SE(\hat{\lambda})$. We get the first standard error

$$SE_1(\hat{\phi}) = |-20\exp(-\hat{\lambda}20)|SE(\hat{\lambda}).$$

(iv) *method 2, computer simulation:*

We simulate G values $U^{(g)}$ from a $N(\hat{\lambda}, SE(\hat{\lambda})^2)$ distribution; set $W^{(g)} = 1 - F_{U^{(g)}}(20)$; compute

$$s^2 = \frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2$$

and get the second standard error

$$SE_2(\hat{\phi}) = \sqrt{s^2}.$$

Here are all of the forgoing steps in the form of an algorithm that can be coded in software such as Matlab or R :

1. $n = 200$
2. $\lambda = .1$
3. $G = 1000$
4. create a data vector x by sampling n times from an $Exp(\lambda)$ distribution
5. $\hat{p} = \frac{1}{n}$(number of components i for which $x[i] > 20$)
6. $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
7. $\hat{\lambda} = 1/\bar{x}$
8. $SE(\hat{\lambda}) = \hat{\lambda}/\sqrt{n}$
9. $\hat{\phi} = 1 - F_{\hat{\lambda}}(20)$
10. $SE_1(\hat{\phi}) = |-20\exp(-\hat{\lambda}20)|SE(\hat{\lambda})$.
11. create a vector u of G draws from a $N(\hat{\lambda}, SE(\hat{\lambda})^2)$ distribution
12. create a vector w whose ith component is $w[i] = 1 - F_{u[i]}(20)$
13. $SE_2(\hat{\phi}) = \sqrt{var(w)}$, where $var(w)$ is the sample variance for the values in vector w.

Applying this algorithm we obtained $\hat{p} = .120$, $SE(\hat{p}) = .023$, $\hat{\lambda} = .103$, $\hat{\phi} = .128$, and finally $SE_1(\hat{\phi}) = .019$ and $SE_2(\hat{\phi}) = .020$. In this case, $\hat{p} \approx \hat{\phi}$ and also $SE(\hat{p}) \approx SE_1(\hat{\phi}) \approx SE_2(\hat{\phi})$. □

We now return to the example involving perception of dim light.

Example 8.4.1 (continued) We now show how to solve for x_{50}, and then propagate uncertainty about it. We may write the logistic regression model we have

$$p = \frac{\exp(u)}{1 + \exp(u)}$$

where $u = \beta_0 + \beta_1 x$. We can solve for u as follows:

$$u = \log \frac{p}{1-p}.$$

This may be checked by plugging the latter formula for u into the one above it to get $p = p$. If we set $p = .5$ we get $u = 0$. In other words, x_{50} must be the value of x for which

$$\beta_0 + \beta_1 x = 0.$$

Solving for x we get

$$\hat{x}_{50} = \frac{-\hat{\beta}_0}{\hat{\beta}_1}. \tag{9.11}$$

To get a standard error for x_{50} we propagate the uncertainty from the approximate variance matrix $\hat{\Sigma}$ for $(\hat{\beta}_0, \hat{\beta}_1)$, which is obtained as the inverse of the observed information matrix, $\hat{\Sigma} = I_{OBS}(\hat{\beta}_0, \hat{\beta}_1)^{-1}$. (The statistical software should furnish $\hat{\Sigma}$, though it may have to be coerced to do so.) We can then, as in the previous illustration, apply either the analytical method or the computer-simulation method. We again go through the steps.

Method 1, analytical: Let $x_{50} = g(\beta_0, \beta_1)$ according to Equation (9.11). Then its standard error is

$$SE = \sqrt{g'(\hat{\beta}_0, \hat{\beta}_1)^T \hat{\Sigma} g'(\hat{\beta}_0, \hat{\beta}_1)}$$

where the partial derivatives are

$$\begin{aligned} \frac{\partial g}{\partial \beta_0} \mid_{(\hat{\beta}_0, \hat{\beta}_1)} &= -\frac{1}{\hat{\beta}_1} \\ \frac{\partial g}{\partial \beta_1} \mid_{(\hat{\beta}_0, \hat{\beta}_1)} &= \frac{\hat{\beta}_0}{\hat{\beta}_1^2}. \end{aligned}$$

Plugging into the formulas above the values of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\Sigma}$, the $\log_{10}$ intensity at which subject S.S. (whose data were shown earlier, the scale on the x-axis having been $\log_{10}$(intensity)) would have perceived half the flashes is estimated to have been $\hat{x}_{50} = 1.921 \pm .019$. Note that the logistic regression model could be viewed here as a method of interpolating between the experimental values, while also providing a standard error of the interpolated quantity.

Method 2, computer simulation: An alternative method of propagating the uncertainty to find the standard error of x_{50} would be to generate many vectors $U^{(g)}$ from a bivariate $N(\hat{\beta}, \hat{\Sigma})$ distribution, where $\beta = (\beta_0, \beta_1)$, for each generated vector compute $W^{(g)} = \hat{x}_{50}^{(g)}$, and then find the standard deviation among the generated values $W^{(g)}$, as before. The algorithm is as follows:

1. For $g = 1, \ldots, G$

   simulate $U^{(g)} \sim N(\hat{\beta}, \hat{\Sigma})$

   Note that $U^{(g)}$ is a two-dimensional vector whose first component corresponds to β_0 and whose second component corresponds to β_1. Thus, we write $U^{(g)} = (U_1^{(g)}, U_2^{(g)})$.

   compute $W^{(g)} = \frac{-U_1^{(g)}}{U_2^{(g)}}$

2. Compute

$$SE(x_{50}) = \sqrt{\frac{1}{G-1}\sum (W^{(g)} - \overline{W}^{(g)})^2}.$$

Using simulation we again found the standard error of $\hat{x}_{50}$ to be $SE = .019$. □

One detail of the simulation procedures left unspecfied is the value of G to be used, i.e., the number of random variables or vectors $U^{(g)}$ to be generated on the computer. In general this is an empirical matter and an easy way to proceed is to pick a convenient value of G, such as $G = 1000$, and then run the entire procedure several times. Because new random variables will be generated each time the procedure is run, the several values of SE will be different. If they are all close to each other then it may be concluded that the SE is sufficiently accurate. If not, we can increase the size of G, keeping in mind that, according to Section 6.1.1, accuracy increases at the rate of $\sqrt{G}$.

9.2 The Bootstrap

9.2.1 The bootstrap is a general method of assessing uncertainty.

The *bootstrap* is a very simple way to obtain standard errors and confidence intervals. It has turned out to be one of the great inventions in the field of Statistics over the past 50 years. The bootstrap operates by *resampling* the data to generate *pseudo-data*, and then applies propagation of uncertainty, as in Section 9.1.

Here is the idea. Let us suppose $X_1, \ldots, X_n$ is a random sample from a distribution having distribution function $F_X(x)$. We write this as $X_i \sim F_X$, independently, for $i = 1, \ldots, n$. We wish to find the standard error of a statistic $T = T(X_1, \ldots, X_n)$. Notice, as we have said before, that T is obtained by applying some mapping to the random variables. Let us emphasize this still further by using the function $h(x_1, x_2, \ldots, x_n)$ to denote that mapping so that $T(X_1, X_2, \ldots, X_n) = h(X_1, X_2, \ldots, X_n)$. In the case of ML estimation, for instance, $h(x_1, x_2, \ldots, x_n)$ would be the function that gives the value of the MLE for a particular set of data $x_1, \ldots, x_n$. In some cases the function $h(x_1, x_2, \ldots, x_n)$ is explicit, as in ML estimation of the binomial propensity p, while in other cases it is implicit—the result of solving a differential equation. In either situation, however, $SE(T)$ is defined as the standard deviation of $T = h(X_1, X_2, \ldots, X_n)$ when the X_i random variables follow the distribution with cdf F_X. Now, if we were able to simulate observations from F_X on the computer, we could simulate G samples where G is a large number, proceeding as follows:

1. For $g = 1$ to G
   Generate a sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from F_X
   Compute $W^{(g)} = h(U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)})$
2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE_{sim}(T) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}.$$

This scheme would evaluate the estimator T not only on the original data $X_1, X_2, \ldots, X_n$, but also on all the sets of *pseudo-data* $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ for $g = 1, \ldots, G$. Each set of simulated values $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ may also be called a *pseudo-sample*. The squared value $SE_{sim}(T)^2$ is simply the sample variance of the $W^{(g)}$ random variables, and for large G it would become close to the variance $V(T)$ (because, in general, the sample variance converges to the theoretical variance, in probability, as in Section 7.3.4). Thus, for large G we would get $SE_{sim}(T) \approx SE(T)$. The only problem is that, in practice, we don't know the distribution F_U, so we don't know how to generate the pseudo-data. This situation is quite like the one we found in Section 7.3.4 where we could not compute $SE(\bar{X}) = \sigma_X/\sqrt{n}$ because we did not know σ_X. There, we solved the problem by substituting s for σ_X, which is often called a *plug-in* estimate, and this worked because $S^2 \to \sigma_X^2$ in probability. The idea of the bootstrap is, similarly, to replace F_X by an estimate of it and then to apply the algorithm above. If we have a parametric model and we use ML estimation to estimate the parameters, thereby obtaining a way to generate the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$, we get a *parametric bootstrap*. Otherwise, we replace F_X by the empirical cdf $\hat{F}_n$ and draw the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$ from $\hat{F}_n$. This is the *nonparametric bootstrap.*

9.2.2 The parametric bootstrap draws pseudo-data from an estimated parametric distribution.

Suppose we assume that a set of data $x_1, x_2, \ldots, x_n$ is a random sample from a distribution with pdf $f(x_i|\theta)$, and we estimate θ with the MLE $\hat{\theta}$. Let us assume for the moment that θ is a scalar. According to the scheme in Section 9.2.1, we may obtain the standard error of $\hat{\theta}$ as $SE_{sim}(\hat{\theta})$ by generating pseudo-samples $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution with pdf $f(x_i|\theta)$. Because we do not know the value of θ we plug in $\hat{\theta}$ and instead generate pseudo-samples from the distribution with pdf $f(x_i|\hat{\theta})$. This is a *parametric bootstrap*, and the resulting value of $SE_{sim}(\hat{\theta})$ is a *parametric bootstrap* standard error.

Parametric bootstrap estimate of standard error
We proceed as follows:

1. For $g = 1$ to G

   Generate a random sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution having pdf $f(x_i|\hat{\theta})$.

   Find the MLE $\hat{\theta}^{(g)}$ based on $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ and set $W^{(g)} = \hat{\theta}^{(g)}$.

2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE(\hat{\theta}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}.$$

When θ is a vector, the same method may be used to estimate the value $f(\theta)$ of any real-valued function $f(x)$. We modify the procedure as follows.

Parametric bootstrap when estimating $f(\theta)$
Suppose we want the standard error of $f(\hat{\theta})$. We proceed as follows:

1. For $g = 1$ to G

   Generate a random sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution having pdf $f(x_i|\hat{\theta})$.

   Find the MLE $\hat{\theta}^{(g)}$ based on $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ and set $W^{(g)} = f(\hat{\theta}^{(g)})$.

2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE(f(\hat{\theta})) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}.$$

Illustration: Exponentially-distributed ion channel opening durations (continued) Let us again find a standard error for the estimator $\hat{\phi}$ of

$$\phi = P(X > 20) = 1 - F_\lambda(20)$$

based on the MLE $\hat{\lambda}$ of λ. This time, we apply the parametric bootstrap. In this case, instead of repeatedly resample the 200 observations to get repeated sets of pseudo-data, we generate the 200 observations from an $Exp(\lambda)$ distribution with $\lambda = \hat{\lambda}$. Here is the algorithm:

1. Create a vector w of length G
2. For $g = 1$ to G,
   $U_1^{(g)}, U_2^{(g)}, \ldots, U_{200}^{(g)}$ from an $Exp(\hat{\lambda})$ distribution, then take u to be the vector whose ith component is $U_i^{(g)}$.
   $\hat{\lambda}^{(g)} = 1/mean(u)$
   $\hat{\phi}^{(g)} = 1 - F_{\hat{\lambda}^{(g)}}(20)$
   $w[g] = \hat{\phi}^{(g)}$
3. Compute the desired standard error
   $$SE_{parboot}(\hat{\phi}) = \sqrt{var(w)}$$
   where $var(w)$ is the sample variance for the values in vector w.

Carrying this out on the data we generated previously gave us $SE_{parboot} = .017$, which is close to the values .019 and .020 obtained earlier. □

Illustration: Gamma distribution We saw, above, that the MLE $(\hat{\alpha}, \hat{\beta})$ may be obtained by iteratively solving for α and β using nonlinear maximization software. Having done this, it is straightforward to generate pseudo-data from the $Gamma(\alpha, \beta)$ distribution; for each set of pseudo-data we compute the MLE. If we repeat this, say, $G = 1000$ times, can estimate any function of $\hat{\alpha}$ and $\hat{\beta}$ and find its standard error. □

Example 9.2.1 Action potential width and the preceding inter-spike interval As part of a study on the effects of seizure-induced neural activity,[2]

[2]Unpublished results from the lab of Dr. Alison Barth, obtained by Judy Xi.

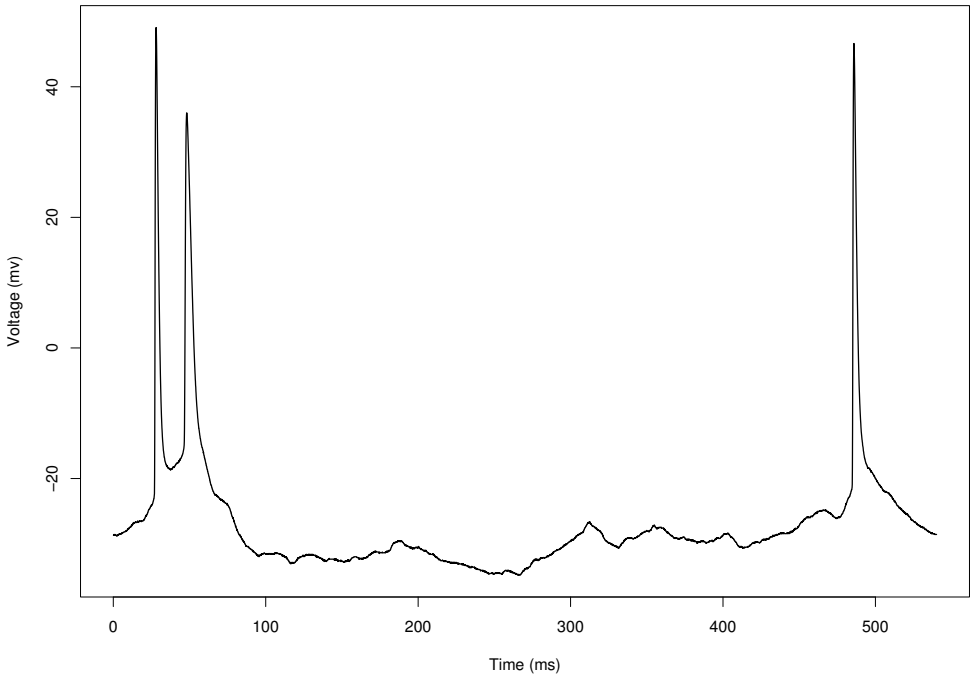


Figure 9.5: *When an action potential follows closely a previous action potential (with small ISI), the second action potential is broader than the first. When a long ISI intervenes, however, the second action potential is very similar to the first.*

spike trains were recorded from barrel cortex neurons in slice preparation. One of the interesting findings involved the relationship between the width of each action potential (spike) and its preceding ISI. As is well known, when a spike follows closely on a preceding spike, so that the ISI is relatively short, then the second spike will tend to be wider than the first. If, however, the ISI is sufficiently long, there will not be any effect of the first spike on the second, and the spike widths should be roughly equal. See Figure 9.6. How long is "sufficiently long?" This turns out to be dependent on previous neuronal activity.

Let Y be the spike width and x the preceding ISI length, and let us assume there is an ISI length τ such that, on average, Y is constant for

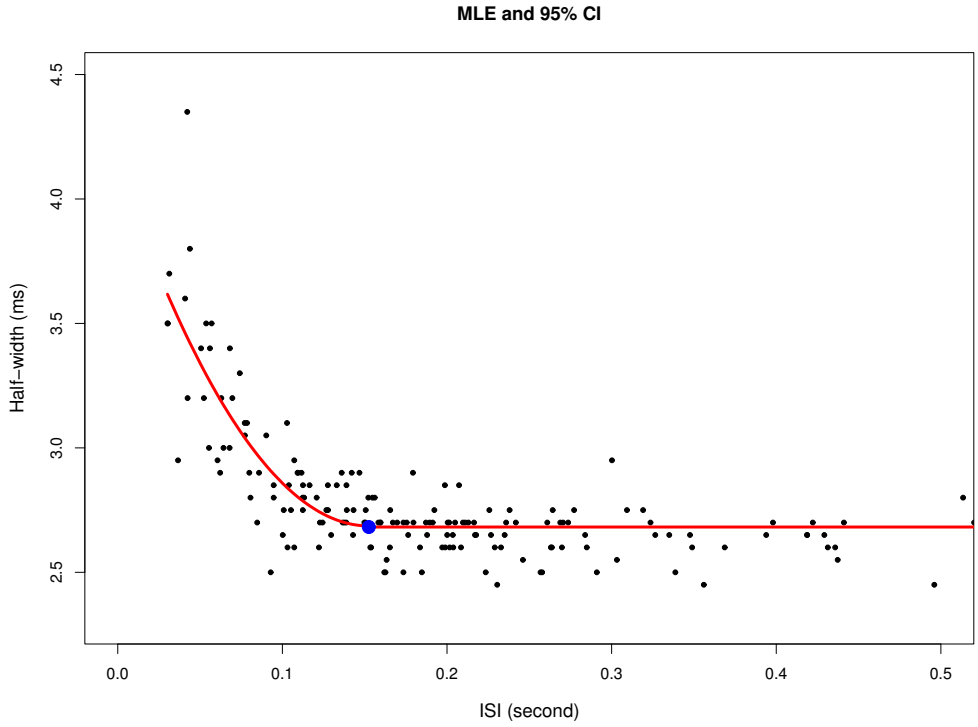


Figure 9.6: *Action potential width varies as function of previous ISI. The data are from many action potentials recorded for a single neuron. A fitted curve with a "change point" is also shown.*

all $x > \tau$. Among neurons taken from animals that had seizures, τ tended to be smaller than its value among control animals. Figure 9.6 displays some of the data, together with a fitted curve. The statistical model used for this curve assumes that, on average, Y decreases with x for $x < \tau$ but remains constant for $x > \tau$. In statistical jargon, τ is called a *change point*, because the relationship between Y and x changes at $x = \tau$. To estimate τ, ML estimation was used with the relationship between y and x assumed to be quadratic for $x < \tau$. The parametric bootstrap was then applied to obtain the $SE(\hat{\tau})$. The method was repeated for neurons from seizure and control animals to see whether there were systematic differences across the two treatment conditions. Figure 9.7 shows results for both groups. Note the very different standard errors across neurons. This suggests that in comparing the two groups it is advisable to use weighted means, as in

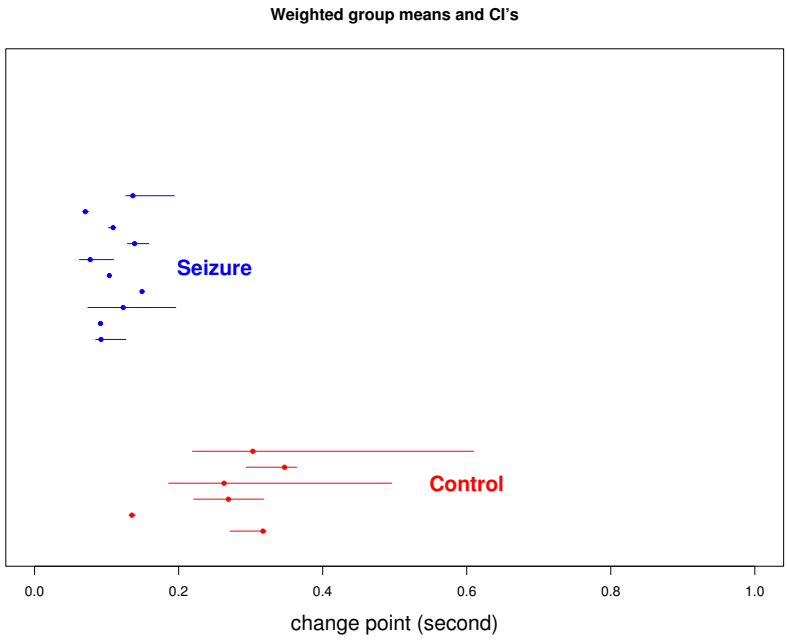


Figure 9.7: *Change points and SEs for neurons of both seizure and control groups. The results for the seizure group appear above those for the control group. The seizure group have change points that occur earlier and they tend to be less variable.*

Equation (8.12), together with standard errors given by Equation (8.13). The results were that the control group had mean change point of 190($\pm$32) milliseconds and the seizure group reset earlier, with mean change point 108($\pm$.012) milliseconds. □

9.2.3 The nonparametric bootstrap draws pseudo-data from the empirical cdf.

As we said in Section 9.2.1, the nonparametric bootstrap works by generating pseudo-samples from the empirical cdf $\hat{F}_n$ instead of F_X (in the scheme shown in that section).

How do we generate observations from the empirical cdf $\hat{F}_n$? This distribution assigns probability $\frac{1}{n}$ to each observation in the sample $x_1, x_2, \ldots, x_n$. Thus, in order to draw a single observation from the distribution $\hat{F}_n$, we randomly select one of the values $x_1, x_2, \ldots, x_n$ (with each value having probability $\frac{1}{n}$). In order to draw a set of pseudo-data, we simply repeat this process n times. Note that in doing so it is very likely to get repeats: we are sampling the values $x_1, x_2, \ldots, x_n$ each time; this is called *sampling with replacement* (because we "replace" each value after sampling it, before drawing again from all the values $x_1, x_2, \ldots, x_n$). bootstrap resampling is beautifully simple.

Nonparametric bootstrap estimate of standard error
We proceed as follows:

1. For $g = 1$ to G

   Generate a sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ by resampling, with replacement, the observations $x_1, \ldots, x_n$

   Compute $T^{(g)} = h(U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)})$

2. Compute $\bar{T} = \frac{1}{G}\sum_{i=1}^{G} T^{(g)}$ and then

$$SE(T) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(T^{(g)} - \bar{T})^2}.$$

This extends immediately to the case in which each X_i, and thus each $U_i^{(g)}$, is a random vector.

Why does the bootstrap work? It rests on the convergence of the empirical cdf to the theoretical cdf, discussed in Section **??**. Because $\hat{F}_n(x) \to F(x)$ in probability for every x, the data generated by $\hat{F}_n$ will be very similar to the data generated by F_X.

> *Details:* The rigorous theoretical justification for bootstrap-based confidence intervals actually requires a somewhat stronger statement, involving the *uniform* convergence of $\hat{F}_n$ to F, and we omit

further discussion of the technical foundation. For more information about the bootstrap, see Efron and Tibshirani (1993) and Davison and Hinkley (1997). (Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall. Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications*, Cambridge.) □

Illustration: Exponentially-distributed ion channel opening durations (continued) Let us again find a standard error for the estimator $\hat{\phi}$ of

$$\phi = P(X > 20) = 1 - F_\lambda(20)$$

based on the MLE $\hat{\lambda}$ of λ. This time, instead of assuming that the Exponential distribution describes the variation in the data, let us apply the nonparametric bootstrap. This assumes only that the ion channel opening durations are independently drawn from the same distribution. In this case we must repeatedly resample the 200 observations to get repeated sets of pseudo-data, and for each set of pseudo-data we must find the value of the estimator, i.e., for the gth set of pseudo-data we compute $\hat{\phi}^{(g)}$. We then calculate the mean and standard deviation among these values. Assuming we have the data in the vector x, here is the algorithm:

1. Create a vector w of length G

2. For $g = 1$ to G,

   Resample $x[1], x[2], \ldots, x[200]$ to get $U_1^{(g)}, U_2^{(g)}, \ldots, U_{200}^{(g)}$ then take u to be the vector whose ith component is $U_i^{(g)}$.

   $\hat{\lambda}^{(g)} = 1/mean(u)$

   $\hat{\phi}^{(g)} = 1 - F_{\hat{\lambda}^{(g)}}(20)$

   $w[g] = \hat{\phi}^{(g)}$

3. Compute the desired standard error

   $$SE_{boot}(\hat{\phi}) = \sqrt{var(w)}$$

   where $var(w)$ is the sample variance for the values in vector w.

Carrying this out on the data we generated previously gave us $SE_{boot} = .018$, which is very close to the values obtained by the parametric methods. □

Example: Difference index for firing rates: SEF neuronal activity under two conditions In the SEF example introduced in Chapter 1 there were two experimental conditions, and the problem was to compare the firing rates of a neuron under each of these conditions based on a limited number of trials. In a particular time interval we found mean firing rates of 48 spikes per second for the spatial condition versus 70 spikes per second for the pattern condition. As we have noted previously, because studies involve many neurons with varying firing rates, it is common to examine the difference index

$$Y = \frac{\bar{X}_A - \bar{X}_B}{\bar{X}_A + \bar{X}_B}.$$

In Section 9.1.2 we discussed generation of a standard error for T using propagation of uncertainty based on the asymptotic normality of $\bar{X}_A$ and $\bar{X}_B$. An alternative would be to apply the nonparametric bootstrap procedure given above. These would give very similar results, but let us make sure it is clear how the bootstrap would be applied. We would first draw a random samples of size 15 from the 15 firing rates under the spatial condition and another random sample of size 15 from the 15 firing rates under the pattern condition. For each we would compute the sample mean, which would give us $\bar{X}_A^{(1)}$ and $\bar{X}_B^{(1)}$. To these we would apply the difference index formula to get

$$Y^{(1)} = \frac{\bar{X}_A^{(1)} - \bar{X}_B^{(1)}}{\bar{X}_A^{(1)} + \bar{X}_B^{(1)}}.$$

We would then repeat this many times to get $Y^{(1)}, Y^{(2)}, \ldots, Y^{(G)}$ (where we would take something like $G = 1000$). Having done this, we would compute the mean and apply the formula for the standard error in step 2 of the algorithm. □

There are also more sophisticated ways to use the bootstrap than the scheme given above but the simple method often suffices to get a reasonable value for SE, and thus a reasonable approximate 95% CI. The nonparametric bootstrap has been studied extensively, and has been shown to work well in a variety of theoretical and empirical senses. (See Efron and Tibshirani (1993)

and Davison and Hinkley (1997).)

An important caveat is that arbitrary shuffles of the data do not necessarily produce bootstrap samples. The key assumption is *independent and identically distributed* sampling of $X_1, \ldots, X_n$, so that the key result $\hat{F}_n \to F_U$ applies. Many problems may be put in this form, but the nonparametric bootstrap only applies once they are.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 10

Models, Hypotheses, and Statistical Significance

The notion of *hypothesis* is fundamental to science. Typically it refers to an idea that might plausibly be true, and that is to be examined or "tested" with some experimental data. Sometimes, the expectation is that the data will conform to the hypothesis. In other situations, the hypothesis is introduced with the goal of refuting it. In either case, however, variation and experimental noise prevent a perfect determination of the veracity of the hypothesis: in reality, the hypothesis will at best predict only approximately the results of an experiment. But then, one might ask, in order to be judged favorably, how close to the data should a theoretical prediction be? Development of a systematic method of answering this question, the chi-squared *goodness-of-fit* test, was one of the great advances in the early part of the 20th century.

We describe chi-squared tests in Section 10.1. The idea is to use a statistical model to represent the theoretical predictions of the hypothesis. In this setting the model embodies the hypothesis, and we usually speak of assessing the fit of the model, as opposed to the accuracy of the hypothesis. The

statistical model assigns probabilities to possible data outcomes, and if the experimental data turn out to be very rare—according to the model—then the model is deemed a poor fit. Because the chi-squared procedure analyzes the discrepancy between model prediction and data outcome, it might better be called, as Tukey suggested, a "badness-of-fit" test. On the other hand, it is often applied as a way of checking that a model fits reasonably well—the expectation, or hope, being that it does.

When, instead, there is great interest in the possibility that the hypothesis may be wrong, we usually label it a *null hypothesis*, and if the data provide sufficient evidence against the null hypotheses we speak of *rejecting* it. Fisher introduced the general concept of *p-value*, with p standing for probability, to quantify the rarity of the data outcome under a null hypothesis. The notion is that when p is small, the data outcome is rare under the hypothesis, and thus casts doubt on the hypothesis. Fisher worked out specific procedures for obtaining p-values in many important problems, and his methodology became standard practice. We introduce p-values in the context of chi-squared tests, in Section 10.1.3, and we discuss the general framework and methodology in Section 10.3.

The null hypothesis and p-value are only part of the standard approach to testing hypotheses. An additional idea is to introduce a specific *alternative hypothesis*, which has the potential to replace the null. In the 1930s Jerzy Neyman and Egon Pearson provided a theoretical framework that explicitly included an alternative hypothesis. Specifically, Neyman and Pearson defined *type one error* (usually written *Type I*) as the probability of incorrectly rejecting the null hypothesis and *type two error* (*Type II*) as the probability of incorrectly rejecting the alternative hypothesis. The theory considers both kinds of errors, and analyzes statistical hypothesis tests according to the probabilities of making these errors. We go over the fundamental elements of the Neyman-Pearson framework in Section 10.4, and we also discuss several different points of view about the statistical assessment of hypotheses.

It is somewhat unconventional to present goodness-of-fit tests before other hypothesis tests. Our preference for this ordering is due to the smaller number of concepts and issues that arise in goodness-of-fit testing: from a pedagogical point of view, in this context it is easier to concentrate on the logic of p-values. We discuss other kinds of null hypotheses in Section 10.2.

10.1 Chi-Squared Statistics

We have described several studies where a theoretical model seemed to fit the data well, and was then used for scientific inference. For instance, the Hardy-Weinberg binomial model fit the nicotinic acetycholine receptor and ADHD data in Example 5.1.1, the Poisson distribution was used to fit quantal response in synaptic transmission data in Example 5.2.3, the normal distribution fit the background noise in MEG in Example 1.1.2, and the exponential and gamma distributions were used to fit ion channel opening duration data in Example 3.2.2. Previously we judged fit simply by looking at tables and graphs, informally. The chi-squared procedure provides a probabilistic quantification of the observed discrepancy between theoretical prediction and data.

The essence of goodness-of-fit assessement is as follows:

(i) We define a statistical model that assigns probabilities to potentially-observed outcomes,

(ii) we compute the discrepancy between the data values and the values obtained from the fitted model, and

(iii) assuming the data were generated by the hypothetical model, we determine whether the observed discrepancy would be considered rare; if observing such a large discrepancy constitutes a sufficiently rare event, then we consider this to be evidence that the model does *not* hold.

The discrepancy between observed data and fit is evaluated using a statistic, here a *chi-squared statistic*, and its rarity is judged by comparing the oberved value to a suitable probability distribution, here a chi-squared distribution, according to the p-value. The chi-squared statistic is used when each observation may be considered to arise as one of several possible categories.

10.1.1 The chi-squared statistic compares model-fitted values to observed values.

To assess the fit of a theoretical model to a set of data we begin with the obvious idea of examining the discrepancy between the model predictions and the data values.

Example 5.1.1 (continued) In Chapter 5 we displayed data from a study of genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. The table of frequencies (for a $T \rightarrow C$ exchange in one base in the gene sequence) among the 136 parents in the Kent *et al.* study is given again below:

	TT	CT	CC
Number	48	71	17
Frequency	.35	.52	.13
Hardy-Weinberg Probability	.38	.47	.15
Hardy-Weinberg Expected Number	51.7	63.9	20.4

We noted previously that the frequencies and Hardy-Weinberg probabilites are quite close. We have now added a fourth line in the table to indicate the predicted or "expected" number of each genotype. To judge the fit of the model we evaluate the discrepancy between the values in the first and last lines of this table. □

In Example 5.1.1 there are many possible ways to measure the discrepancy between the vector of observed values $(48, 71, 17)$ and the vector of theoretically-expected values $(51.7, 63.9, 20.4)$. The most common assessment is based on the chi-squared statistic. Let us denote observed values by O and theoretically-expected values by E, so that the first pair of O and E values are 48 and 51.7, the second pair are 71 and 63.9, and the third pair are 17 and 20.4. The chi-squared statistic is

$$\chi^2_{obs} = \sum \frac{(O-E)^2}{E} \tag{10.1}$$

where the sum is over all pairs of values, in this case the three pairs, and we have used the subscript on χ^2_{obs} to indicate that it is calculated from

the observed data. A large χ^2_{obs} indicates a failure of the model to fit the data. But how do we know when χ^2_{obs} should be considered large? The O values surely will, by chance fluctuation, deviate from the theoretical E values. The key is that when the theoretical model is valid the magnitude of this chance fluctuation becomes predictable. As motivation for χ^2_{obs} note that each O value is a count, counts are usually modeled as Poisson random variables, and for a Poisson random variable Y we have $V(Y) = E(Y)$. Therefore, if we wish to standardize each O value before combining all of them, a reasonable way to do it would be to subtract the expected value, which we here take to be E, and divide by the standard deviation which, if the observed value were Poisson would be the square root of the expectation, here $\sqrt{E}$. Each contribution $(O-E)^2/E$ may thus be considered the square of a standardized variable. It turns out that, for large samples, these standardized variables are approximately normally distributed. Recalling that the chi-squared distribution arises as a sum of squares of standard normal variables it then becomes at least plausible that a χ^2 distribution might be used to judge the magnitude of the chi-squared statistic. As we will mention again below, there is some subtlety to the argument, but it turns out to work.

To obtain the p-value for the chi-squared procedure we consider a random variable X having a χ^2_ν distribution and evaluate $p = P(X > \chi^2_{obs})$. We discuss the selection of ν in Section 10.1.2. If p is sufficiently small we consider the observed value to be rare. Typically, $p < .05$ is taken as modest evidence and $p < .01$ is taken as strong evidence that the model doesn't fit.

Example 5.1.1 (continued) For the ADHD data we get

$$\chi^2 = \frac{(48-51.7)^2}{51.7} + \frac{(71-63.9)^2}{63.9} + \frac{(17-20.4)^2}{20.4} = 1.62.$$

We compare this to a χ^2_1 distribution by taking X to be a random variable having a χ^2_1 distribution and evaluating $P(X > 1.62) = .20$, so that $p = .20$. (We use software such as Matlab to obtain the χ^2_1 probability.) This indicates a good fit of the Hardy-Weinberg model to these data. □

10.1.2 For multinomial data, the chi-squared statistic follows, approximately, a χ^2 distribution.

In Example 1.2.1 we introduced a binary random variable to analyze the variation across outcomes where each outcome was one of two possibilities, "burning house" or "non-burning house." In Example 5.1.1, we have a similar situation, except instead of two possible outcomes we have three: each of the 136 subjects contributed a genotype that was classified as TT, TC, or CC. To model the naturally occurring variation we work by analogy. Instead of the binary random variable for each trial in Example 1.2.1, we here introduce a three-dimensional random vector corresponding to each subject. Specifically, we take the vectors $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ to correspond to TT, TC, and CC, and then consider the ith subject's genotype to follow a three-dimensional random vector Y_i that takes the values $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ with probabilities p_1, p_2, or p_3, where $p_1+p_2+p_3=1$. Assuming the random variables Y_i are independent and summing them to get the triple set of counts $X=\sum_i Y_i$ we get a multinomial distribution $X \sim M(n,p_1,p_2,p_3)$. This is the fundamental assumption for the application of the chi-squared test on page 275. More generally, the theoretical starting point of every chi-squared test is the idea that the given set of counts may be considered an observation of a multinomial random variable. Here is a particularly straightforward example where the genetic model completely specifies the set of multinomial probabilities, leaving no free parameters.

Example 10.1.1 Allele frequencies in fruit flies Bridges and Moran are cited in Kempthorne (1957, p. 155) as investigating a cross breeding of vestigial (vg) by ebony (e) in certain fruit flies. According to equilibrium theory, the four possible results (denoted +,vg,e,vge) should be in the proportions 9:3:3:1. The four respective frequencies among 465 flies were 268, 94, 79, 24. The theoretical proportions are (.563,.188,.188,.0625) while the observed proportions were (.576,.202,.170,.0516). For instance, .576=268/465. In this case, we model the vector of numbers of phenotypes among 465 flies as a $M(n,p_1,p_2,p_3,p_4)$ distribution, where $n=465$ and p_1 is the probability that a given fly would be of type +, p_2 the probability the fly would be of type vg, etc. We would assume that the phenotypes are independent of each other across flies (so that knowing one fly's phenotype does not change another fly's phenotype probability distribution), and each as the same set of four

probabilities. Thus, under the model, the vector $(268, 94, 79, 24)$ is treated as if it were an observed value of the multinomial random vector. □

In applications of chi-squared methodology each O is a count associated with a particular data *category*. In Example 5.1.1, for instance, the categories were TT, CT, CC. The number of categories is important in determining the degrees of freedom ν. The value to use for ν depends on the problem. If we take the number of categories to be k and the number of estimated parameters to be m then ν is found from the formula

$$\nu = k - 1 - m \tag{10.2}$$

The degrees of freedom, often abbreviated *d.f.*, may be considered the number of free parameters. The idea and terminology of degrees of freedom come from mechanics: we count the number of dimensions in which the random variable is "free to move," often beginning with some apparent maximal number of dimensions and subtracting off constraints. The examples below should help clear this up, and there are general formulas for each type of problem. In Equation (10.2) we begin with a multinomial distribution that has k categories with probabilities $p_1, \ldots, p_k$. Because these sum to 1, there are only $k - 1$ free parameters. Then, after estimating m parameters for the null hypothetical model we are left with $\nu = k - 1 - m$ free parameters.

Just as the binomial may be approximated by a normal distribution for large n, so too may the multinomial be approximated by a multivariate normal for large n. This leads to the general result that the chi-squared statistic follows, approximately, a chi-squared distribution.

Result Suppose $X \sim M(n, p_1, p_2, \ldots, p_k)$ and we have a statistical model $p_1 = p_1(\theta), p_2 = p_2(\theta), \ldots p_k = p_k(\theta)$ based on an m-dimensionl parameter vector θ. Let $\hat{\theta}$ be the MLE and let Y_n be defined as in (10.1) by

$$Y_n = \sum_i \frac{\left(X_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})}.$$

Then, assuming suitable regularity conditions[1] on the statistical model, as $n \to \infty$ we have

$$Y_n \xrightarrow{D} \chi^2_\nu \tag{10.3}$$

where $\nu = k - 1 - m$.

In practice, the most important input to this theoretical result, which leads to the calculation of the p-value, is the assumption that the data may be represented by a multinomial random vector. As in the binomial case, the multinomial assumption will make sense when it is reasonable to assume the classification variables are independent across observations (across subjects in Example 5.1.1). Thus, as before, it is the judgment of independence that must be considered most carefully.

10.1.3 The rarity of a large chi-squared is judged by its p-value.

The conventional cut-offs for the p-value, reflecting modest and strong evidence, are .05 and .01. These two particular numbers were handed down from Fisher and are now imbedded in standard practice, but they are somewhat arbitrary and should be considered rough guides rather than finely tuned criteria. Articles in the literature often include statements in the form $p < .05$, with the result typically being called *statistically significant*, or $p < .01$, which may be labeled *highly significant*. However, it is not unusual to obtain a very small p-value (e.g., 10^{-4}), which is quite different than .01. Rather than saying $p < .01$, it is preferable to report the p-value, and it is also good

practice to say what statistic was computed, e.g., in Example 5.1.1 on page 275, one would report $p = .20$ for chi-squared on 1 degree of freedom.

Example 10.1.1 (continued) Returning to the allele frequencies example, the "observed values" O are 268, 94, 79, 24. The "expected values" E values must be calculated. If the ratios were 9:3:3:1, the corresponding proportions would be 9/16, 3/16, 3/16, 1/16. With 465 flies, we would therefore expect to see $\frac{9}{16} \cdot 465 = 261.6$, $\frac{3}{16} \cdot 465 = 87.2$, $\frac{3}{16} \cdot 465 = 87.2$, $\frac{1}{16} \cdot 465 = 29.1$. The O and E values are compared and summarized by the chi-squared statistic using (1):

$$\chi^2 = \frac{(268 - 261.6)^2}{261.6} + \cdots + \frac{(24 - 29.1)^2}{29.1} = 2.34.$$

Here there are 4 categories, so 3 degrees of freedom. We use the computer to find $p = P(X > 2.34) = 1 - P(X \leq 2.43)$ where X has a χ^2_3 distribution. We obtain $P(X \leq 2.43) = 0.4951$ and therefore $p = .50$. This p-value is large, much larger than the conventional values .05 and .01. Thus, data that deviate from expected values as much as these would not be rare and we conclude there is a good fit of the theoretical model to these data. □

Example 5.2.1 (continued) In the radioactive disintegration example, the statistical model is that the data follow a Poisson distribution. The O and E values are given in Table 10.1. The E values are obtained as $E_i = n\hat{p}_i$ with the $\hat{p}_i = p(\hat{\lambda})$ values in turn given by the Poisson pdf formula after the MLE $\hat{\lambda}$ is substituted for λ, i.e., we begin with $p_i = P(X = i) = e^{-\lambda}\lambda^i/i!$ and then substitute $\lambda = \hat{\lambda} = \bar{x}$. Thus, after computing $\hat{\lambda} = \bar{x} = 3.87$ we obtain $\hat{p}_i = e^{-\hat{\lambda}}\hat{\lambda}^i/i!$. The expected number of times we would observe one particle emitted is 2608 times the probability of getting one particle emitted, i.e., $2608 \cdot e^{-3.87}(3.87) = 210.523$.

Calculation of (10.1) gives $\chi^2_{obs} = 12.9$ and here there are $\nu = 11 - 1 - 1 = 9$ degrees of freedom: we start with $11 - 1 = 10$ degrees of freedom, because there are 11 categories, but we lose one degree of freedom from estimating λ. From the chi-squared cdf we find that when $X \sim \chi^2_{10}$, $P(X > 12.9) = .17$. Thus, $p = .17$ and there is no evidence of departure from the Poisson distribution despite the large sample size, which would have given an opportunity to detect even a small departure. □

A detail: A technical point arises in the example above from

Table 1

k	Observed Counts	Poisson Fitted Counts
0	57	54.399
1	203	210.523
2	383	407.361
3	525	525.496
4	532	508.418
5	408	393.515
6	273	253.817
7	139	140.325
8	45	67.882
9	27	29.189
≥ 10	16	17.075

Table 10.1: *Fit of Poisson distribution to the counts of α-particle emissions from a specimen during 2608 intervals. The frequency of counts $0, 1, 2, \ldots, 9, \geq 10$ appear beside those were based on the Poisson distribution.*

the observation that the number of categories here is actually somewhat arbitrary: we chose to use 11 categories, but could have chosen a different number. As a result, the large-sample distribution is not the claimed chi-squared, but a slightly different approximation (a pair of bounds) may be used for the p-value. In this case, the p-value would be somewhere between those obtained for 9 and 10 degrees of freedom. This would make the p-value a bit bigger than our reported $p = .17$. Many texts emphasize this technicality, but for models such as these, with a single parameter, it has little effect on the conclusions. □

10.1.4 Chi-squared may be used to test independence of two traits

Many studies seek to evaluate the association of two traits. In genetic epidemiology, for instance, it is useful to know whether a particular genotype

may be associated with a disease. When the occurrence of each trait is considered a random variable, the traits will fail to be associated if the two random variables are independent. Thus, the issue becomes one of evaluating the fit of a statistical model based on independence.

Example 10.1.2 Alzheimer's and APOE As part of a study of markers for late-onset Alzheimer's disease, Yu *et al.* (2007) (Yu, C.E., Seltman, H., Peskind, E.R., Galloway, N., Zhou, P.X., Rosenthal, E., Wijsman, E.M., Tsuang, D.W., Devlin, B., and Schellenberg, G.D. (2007) Comprehensive analysis of *APOE* and selected proximate markers for late-onset Alzheimer's disease: Pattern of linkage disequilibrium and disease/marker assocation, *Genomics*, 89: 655–665.) looked for the presence of the ε_4 allele of the apolipoprotein E gene (*APOE*), which had previously been associated with increased risk of Alzheimer's, among both Alzheimer's patients and controls. The following table summarizes some of the data they presented from 193 Alzheimer's patients (AD) and 232 controls:

	ε_4 absent	ε_4 present
AD	58	135
controls	162	70

At first glance it appears that the ε_4 allele is far more prevalent among the Alzheimer's patients than among the controls—and that this is probably not due to chance. This may be verified using a χ^2 test. □

Example 10.1.2 involves what is called a *two-by-two table* (written 2×2). In general, the probabilities for a 2×2 table may be represented as follows:

	1 absent	1 present	
2 absent	p_{11}	p_{12}	$p_{1\cdot}$
2 present	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	

Here the subscript ij corresponds to the (i, j) element in the table, meaning that p_{ij} is the probability in row i and column j. For example, p_{22} is the

probability that a random individual has both trait 1 and trait 2 (e.g., both ε_4 and AD). The probabilities along the margins of the table come from summing the probabilities along rows or columns. For example, $p_{\cdot 2} = p_{12}+p_{22}$ is the probability that the individual has trait 1 (e.g., ε_4) and $p_{2\cdot} = p_{21} + p_{22}$ is the probability that the individual has trait 2 (e.g., AD). Now, if independence holds, then the probability of having both trait 1 and trait 2 must equal the probability of having trait 1 times the probability of having trait 2, i.e., $p_{22} = p_{2\cdot}p_{\cdot 2}$. Filling out the rest of the table of probabilities the same way gives the independence model

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

for all i, j.

In order to apply χ^2 we need to compute the expected values, each of which is the number of individuals we would expect a particular entry of the table. In principle, the expected value for (i, j) entry in the table is $E = n \cdot p_{ij} = n \cdot p_{i\cdot}p_{\cdot j}$ for each of the four p_{ij}'s, but we don't know the values of $p_{i\cdot}$ and $p_{\cdot j}$. Here we resort to the standard "plug-in" method: we estimate these marginal probabilities from the data. For instance, in the Alzheimer's example there are a total of 425 individuals so we use $\hat{p}_{\cdot 2} = 205/425$ for the probability of having ε_4.

This estimation process causes the chi-squared distribution to lose degrees of freedom, as in the α-particle example. In general, if there are r rows and c columns we begin with $rc-1$ degrees of freedom: there are rc probabilities in the table but they must sum to 1, which means we lose 1 degree of freedom. We then lose another $r - 1$ degrees of freedom for estimating row marginal probabilities and $c - 1$ for estimating column marginal probabilities. This leaves $rc - 1 - (r - 1) - (c - 1) = rc - r - c + 1 = (r - 1)(c - 1)$ degrees of freedom.

Example: 10.1.2 (continued) In this example $r = 2$ and $c = 2$ so there is 1 degree of freedom. Entering the data into an appropriate statistical software package produces $\chi^2 = 65$ on 1 degree of freedom, and $p = 7\times10^{-16}$, which is truly tiny. Clearly there is an association here. □

Software used to get chi-squared results, as in Example 10.1.2 above, typically applies a variation of the chi-squared statistic that includes a "continuity correction." This adjusts the statistic slightly to make the continuous

chi-squared distribution match more closely the distribution of the discrete chi-squared statistic in small samples. It is also possible to use so-called "exact" methods, which avoid the χ^2 distribution altogether. While such methods are commonly applied, it is important to keep in mind that we are usually looking for clear and compelling results, either not significant or strongly significant, and borderline cases should be interpreted as such. That is, when the continuity correction—or the distinction between exact and approximate methods—is important to conclusions, this may signal a case in which a careful investigator will recognize the ambiguity of the data.

Example 10.1.2 (continued) The Alzheimer's and *APOE* data may be examined further to see if there is a difference between men and women. Here is the table for the AD patients:

	ε_4 absent	ε_4 present
women AD	32	70
men AD	26	65

The proportions appear to be about the same, and this time we get $\chi^2 = .071$ again on 1 degree of freedom, and $p = .79$, so there is no evidence of any discrepancy in ε_4 prevalence among the male and female AD patients. □

One final subtlety should be noted. The logic we have described here assumes that each subject in the study is drawn randomly from a population of potential subjects. That could be a good rough description of what happened in the Alzheimer's study: the incidence of Alzheimer's among relatively old subjects can be quite high. However, often a set of diseased patients is selected and then a set of controls is chosen separately. In epidemiology this is called a *case-control* study. It generates a different statistical model, but it turns out to give the same χ^2 test. (The cited study did not say which way the subjects were collected.) We will return to the issue of data collection strategies and their effects on scientific inference in Chapter **??**.

10.2 Null Hypotheses

10.2.1 Statistical models are often considered null hypotheses.

In talking about assessing fit we have used a "hypothesized model," i.e., the model being fit to the data. The standard terminology is to take such a model to be the "null" model, or the *null hypothesis*, often written as H_0. Sometimes the null hypothesis completely specifies the probability distribution, as in Example 10.1.1. In other cases it merely identifies a family of distributions, as in the α-particle emissions example (where there is still a free parameter λ), and in the Alzheimer's and $APOE$ example (where there remain two free parameters $p_{1\cdot}$ and $p_{\cdot 1}$). The "null" here indicates that such a hypothesis is often used with an eye toward collecting evidence against the hypothesis, the implicit understanding being that H_0 would eventually be replaced with something that could describe such data better.

10.2.2 Null hypotheses sometimes specify a particular value of a parameter within a statistical model.

Another possibility is that the null hypothesis specifies a particular value of a parameter within a family of distributions.

Example 1.2.1 (continued) In the investigation of blindsight in patient P.S., the possibility that P.S. was guessing corresponds to taking $p = .5$ in the binomial model. We write this as $X \sim B(17, p)$ with $H_0 : p = .5$. One way to test this is with χ^2. We take the observed values to be 14 and 3 (for the two categories "non-burning preferred" and "burning preferred") and take the expected values to be np_0 and $n(1-p_0)$, with $n = 17$ and $p_0 = .5$, or 8.5 and 8.5. Chi-squared is then $\chi^2_{obs} = ((14-8.5)^2/8.5)+(3-8.5)^2/8.5) = 7.12$. Here we have 2 categories and 0 estimated parameters, so d.f.=1. Comparing 7.12 to a χ^2_1 distribution gives a p-value of $p = .0076$, which[2] is strong

[2]In this example we use the notation p in two different ways: at first p stands for the probability that P.S. would choose the non-burning house, and then later it stands for the

evidence against H_0. □.

In Example 1.2.1 there is a simple null hypothesis and a chi-squared procedure to test it. Because the sample size there is small, however, the continuity correction mentioned on page 282 would change the p-value somewhat. We will obtain a more accurate p-value for Example 1.2.1 below.

10.2.3 Null hypotheses may also specify a constraint on two or more parameters.

In the blindsight example we had a single binomial and tested $H_0 : p = .5$. Now suppose we have two binomials, $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$ and we wish to test $H_0 : p_1 = p_2$. This is a special case of a widely-applied type of null hypothesis, namely one that corresponds to a constraint on some parameters in a statistical model. In the case of two binomials, $H_0 : p_1 = p_2$ may be assessed by comparing χ^2 to a χ^2_1 distribution: we begin with two free parameters p_1 and p_2 and lose a degree of freedom due to the constraint. In fact, this special case of χ^2 turns out to be mathematically equivalent to the test of independence examined above.

Example 10.1.2 (continued) We noted previously that the way the Alzheimer's data were collected would affect the way the statistical problem would be posed. If AD patients and controls were collected separately, then we would examine whether the probability of having the ε_4 genotype was the same in each population, i.e, we would have two binomials and would test $H_0 : p_1 = p_2$. To repeat, this test may be carried out using χ^2, exactly as done previously. □

In a similar way, data from two independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ may be used to test the hypothesis that the corresponding means μ_1 and μ_2 are equal, $H_0 : \mu_1 = \mu_2$. In the case of the SEF neuronal activity under two conditions discussed in Example 1.1.1 of Chapter 1, for example, there were 15 trials in both experimental conditions, generating mean firing rates of 48 spikes per second for the spatial condition and 70 spikes per second for the pattern condition across the time interval

p-value. We hope this is not confusing.

from 200 to 600 milliseconds after the onset of the cue. The null hypothesis $H_0 : \mu_1 = \mu_2$ would say that the two mean firing rates are equal. The standard statistical procedure for testing this hypothesis is called a t test, because it relies on the t distribution. We discuss this below. Here is another example.

Example 10.2.1 Test-enhanced learning Tests are used to assess whether students have learned subject-matter material. A line of research has emphasized the additional value of testing as a way to *enhance* learning (Karpicke and Roediger, 2008). (Karpicke, J.D., and Roediger, H.L. (2008) The critical importance of retrieval for learning, *Science*, 319: 966–968.) The idea is that when students are tested, they recall information and thereby reinforce memory of it. In one study, Roediger and Karpicke (2006) (Roediger, H.L., and Karpicke, J.D. (2006) Test-enhanced learning, *Psychological Science*, 17: 249–255.) had subjects read a short passage and then get tested on it after a delay period during which they would forget some of the material. Let us call this test the assessment test. After reading but before the assessment test there was an experimental manipulation: some subjects were asked to restudy the text, while other subjects were instead given a learning test, identical to the assessment test. These tests simply asked the subjects to write down everything they could remember about the passages. The tests were scored according to the number of "idea units" correctly recalled. A key part of the study focused on retention of the material following a delay period of 1 week, asking whether the learning-test group retained the material better than in the restudying group.

After finding strong evidence of a benefit from testing, the authors did a second experiment, using four study or testing sessions. In one condition, labelled SSSS, there were 4 study sessions, and in another, labelled SSST, there were 3 study sessions followed by a testing session. The assessment administered following a delay of 1 week had a maximal score of 30 idea units. Data from 60 subjects, 30 in each of the SSSS and SSST groups are displayed in Figure 10.1.

For the data displayed in Figure 10.1 the means were 11.9 and 16.7, with medians 12 and 16, and lower and upper quartiles (8.25, 15) and (11.25, 21). It appears that the SSST scores tend to be higher than the SSSS scores. To formalize the comparison, we consider the population mean scores under

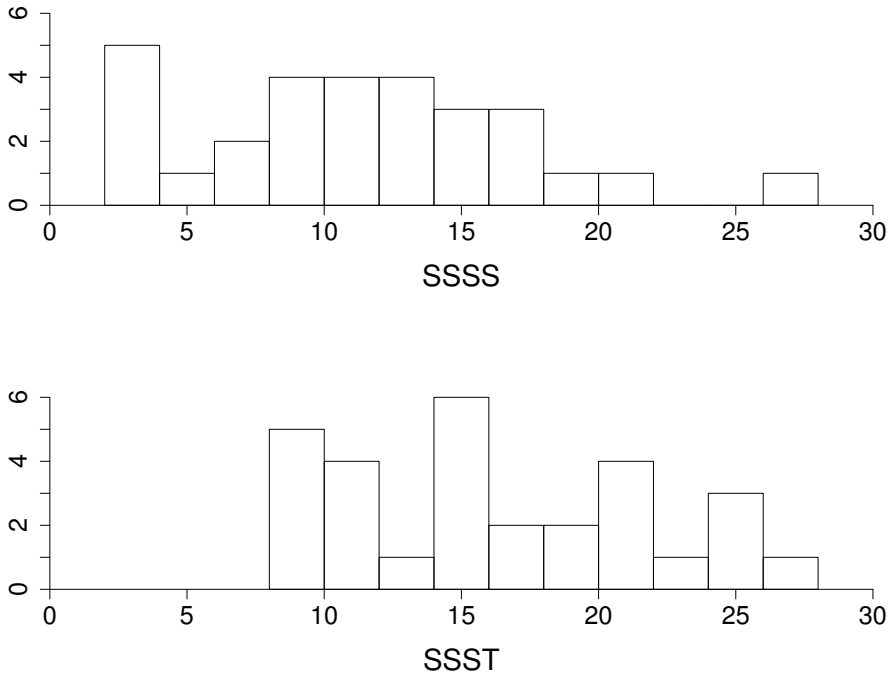


Figure 10.1: *Histograms of test-enhanced learning data. Data are assessment scores (number of recalled idea units, out of a maximum of 30) for 30 subjects under the SSSS condition (TOP) and the SSST condition (BOTTOM). Data courtesy of J.D. Karpicke.*

these two conditions. If we let X_{1i} be the score of the ith subject in the SSSS condition and X_{2i} be the score of the ith subject in the SSST condition and if μ_1 and μ_2 are the mean scores within these two conditions, we may take the null hypothesis to be $H_0 : \mu_1 = \mu_2$. The null hypothesis here says that the population mean scores in the SSSS and SSST conditions are identical. □

10.3 Testing Null Hypotheses

10.3.1 The hypothesis $H_0 : \mu = 0$ for a normal random variable is a paradigm case.

We have aleady noted that a null hypotheses may specify a particular value of a parameter. To establish intuition based on a widely-used form of test statistic, let us return to the prototypical situation in which we assume we have a sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution with σ is known. To test $H_0 : \mu = \mu_0$ we may form the ratio

$$z_{obs} = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$$

where $\bar{x}$ is the sample mean computed from the data and $SE(\bar{x}) = \sigma/\sqrt{n}$. If the magnitude $|z_{obs}|$ is sufficiently large we would say there is evidence against H_0. Specifically, we reason as follows. According to our assumption, the theoretical counterpart of z_{obs} is

$$Z = \frac{\bar{X} - \mu_0}{SE(\bar{X})} \tag{10.4}$$

where we have replaced $\bar{x}$ with $\bar{X}$ to indicate that we are now dealing with random variables rather than observed data. Because $\bar{X} \sim N(\mu, \sigma^2/n)$ we also have

$$Z \sim N(0, 1) \tag{10.5}$$

and we obtain a p-value from

$$p = P(|Z| \geq |z_{obs}|). \tag{10.6}$$

Because we rarely have a situation in which the data may be assumed normal with σ known, Equations (10.5) and (10.6) are not often applicable. However, the more general idea of comparing an estimator with a null hypothetical value after standardizing with the standard error is very useful. We present this in Section 10.3.2.

10.3.2 In large samples the hypothesis $H_0 : \theta = \theta_0$ may be tested using the ratio $(\hat{\theta} - \theta_0)/SE(\hat{\theta})$.

The uncertainty associated with an estimate is quantified by the estimate's standard error. In Example 1.2.1 we gave an approximate 95% confidence interval $(.64, 1.0)$ and noted that this was inconsistent with the probability of .5, which would correspond to guessing. But if we are mainly interested in whether the data are consistent with guessing, we could rephrase the problem using the observed discrepancy between $\frac{14}{17}$ and .5. The proportion $\hat{\theta} = \frac{14}{17}$ seems much too big to be consistent with guessing. So we may ask this question: If P.S. were guessing, how unlikely would it be that $\hat{\theta}$ would be as far from .5 as was $\frac{14}{17}$?

We will present several different procedures that provide slightly different answers to this question, all of which lead to the same conclusion. The one most closely related to the approximate confidence interval assesses the discrepancy between $\hat{\theta}$ and .5 in units of $SE(\hat{\theta})$. This relies on the approximate normality of the MLE $\hat{\theta}$.

Result: Suppose $X_1, \ldots, X_n$ has joint pdf $f(x_1, \ldots, x_n|\theta)$, with θ a scalar, and suppose further that T_n is an asymptotically normal estimator of θ with standard error $SE(T_n) = \hat{\sigma}_{T_n}$. Then the null hypothesis $H_0 : \theta = \theta_0$ may be tested by using the statistic

$$z_{obs} = \frac{T_n - \theta_0}{SE(T_n)}, \tag{10.7}$$

with large values of $|z_{obs}|$ indicating evidence against H_0. If the sample size is large, an approximate p-value may be obtained from

$$p = P(|Z| \geq |z_{obs}|) \tag{10.8}$$

where $Z \sim N(0, 1)$.

This result follows from the theorem in Section 7.3.5, which said that if $\hat{\sigma}_{T_n}$ is the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\frac{(T_n - \theta)}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0,1).$$

If $\theta = \theta_0$ then the random variable

$$Z = \frac{T_n - \theta_0}{SE(T_n)}$$

follows, approximately, for large n, a $N(0,1)$ distribution and the p value based on $Z \sim N(0,1)$ will be approximately correct. Because Z is a common notation for a $N(0,1)$ random variable, the value z_{obs} in (10.7) is often called a *z-score* and the procedure in (10.7) and (10.8) is a *z-test*.

Example 1.2.1 (continued) Suppose $X \sim B(n,\theta)$ and we wish to test $H_0 : \theta = \theta_0$. The usual formula for SE is $SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$. It is customary to find SE under the null hypothesis, $\theta_0 = .5$, i.e., we replace[3] $\hat{\theta}$ with $\theta_0 = .5$ in the calculation of SE. In the case of the data from P.S., we had $n = 17$ so we get $SE = \sqrt{(.5)(.5)/17} = .121$. This gives us Thus, $z_{obs} = (.824 - .5)/.121 = 2.68$. This gives us a p-value of .0074, which is nearly the same as the value .0076 obtained from the chi-squared analysis (see page 284). In fact, in this case, a little bit of manipulation shows that we have the arithmetic identity $z^2_{obs} = \chi^2_{obs}$. □

The identity above provides a way of understanding the chi-squared procedure. The definition of a χ^2_1 distribution is that it results from squaring a $N(0,1)$ random variable. When we replace the data with random variables we get the theoretical counterpart of the observed value z_{obs},

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})},$$

which has an approximate $N(0,1)$ distribution. Therefore, its square has an approximate χ^2_1 distribution, but its square is the theoretical counterpart of

[3]The logic of the procedure does not demand that we use θ_0 in place of $\hat{\theta}$. The justification of the large-sample significance test, the Theorem in Section 7.3.5 that says Z is approximately $N(0,1)$, and is not refined enough to distinguish between the two alternative choices for $SE(T_n)$ (both would satisfy the theorem). However, because we are doing the calculation under the assumption that $\theta = \theta_0$, it makes some sense to use the value $\theta = \theta_0$ in computing the standard error.

the observed value $z_{obs}^2 = \chi_{obs}^2$. In other words, the theoretical chi-squared statistic follows, approximately, a chi-squared distribution.

When θ is a vector essentially the same result as in (10.7) and (10.8) holds again for each component. That is, if θ_i is one component of θ and $T_{n,i}$ is the corresponding component of an asymptotically normal vector estimator T_n (which would be asymptotically multivariate normal), then we can test $H_0 : \theta_i = \theta_{i,0}$ by replacing T_n by $T_{n,i}$ and θ_i by $\theta_{i,0}$ in (10.7) and again using (10.8). For example, in simple linear regression we may have both an intercept and a slope, but we may wish to test the null hypothesis that the slope is zero—which would correspond to there being no linear relationship between the response and explanatory variables. We return to this in Chapter 13.

10.3.3 For two independent samples, the hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using the t-ratio.

Let us now apply the idea in Section 10.3.2 to the problem of testing $H_0 : \mu_1 = \mu_2$ based on two independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$. The obvious starting point is the difference between the sample means $\bar{X}_1 - \bar{X}_2$. The standard approach simply divides this quantity by its standard error.

Now, what is the standard error of $\bar{X}_1 - \bar{X}_2$? Because the two samples are independent we have

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{10.9}$$

where σ_1^2 and σ_2^2 are the respective variances of each X_{1i} and X_{2i}, within each of the two samples. The standard error will be the square-root of the variance in (10.9) after we plug in suitable estimates of σ_1 and σ_2. The most common procedure, the ordinary *t-test,* makes the assumption that $\sigma_1 = \sigma_2$, which greatly simplifies the theoretical results. We now label these standard deviations by σ (so that $\sigma = \sigma_1 = \sigma_2$). With this assumption, the two sample standard deviations s_1 and s_2 both estimate σ. We then pool the

data together by calculating

$$S^2_{pooled} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2\right)$$

which is taken as an estimator of σ^2 and gets plugged into (10.9) for σ_1 and σ_2. The test statistic becomes

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10.10}$$

and, assuming $\mu_1 = \mu_2$, as n_1 and n_2 become infinite T converges in distribution to $N(0, 1)$. This gives the following method (where the notation converts the capital T, X and S to lower case once T is applied to observed data).

Result: Suppose $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ are independent random samples from distributions having means μ_1 and μ_2 and standard deviations $\sigma_1 = \sigma_2$. The null hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

with large values of $|t_{obs}|$ indicating evidence against H_0. If the sample sizes are large, an approximate p-value may be obtained from

$$p = P(|Z| \geq |t_{obs}|)$$

where $Z \sim N(0, 1)$.

The result above is justified by the Central Limit Theorem. If, in addition, we are willing to assume normality of the distributions then we have an "exact" result, which applies in small samples.

Result: Suppose $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ are independent random samples from normal distributions having means μ_1 and μ_2 and standard deviations $\sigma_1 = \sigma_2$. The null hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{10.11}$$

with large values of $|t_{obs}|$ indicating evidence against H_0. A p-value may be obtained from

$$p = P(|T| \geq |t_{obs}|)$$

where $T \sim t_\nu$, with $\nu = n_1 + n_2 - 2$.

Example 10.2.1 (continued) In the test-enhanced learning study Roediger and Karpicke (2006) found strong evidence against H_0, the hypothesis the population mean scores in the learning-test group and the restudy groups were identical. Applying the two-sample t-test to the data displayed above we obtained $t_{obs} = -3.19$ on 58 degrees of freedom. Using the normal approximation this gives $p = .0014$ while using the t distribution we get $p = .0023$. Either way there is strong evidence against H_0, indicating strong evidence that the mean assessment score under the SSST condition is greater than the mean assessment score under the SSSS condition. □

In Example 10.2.1 the p-value is larger when the t distribution is used than when the normal distribution is used. This is generally the case, as the t distribution has thicker tails, so that it gives higher probability to values with large magnitudes. Standard practice is to report the t-based p-value.

10.3.4 Computer simulation may be used to find p-values.

We have gone over several examples of p-values but we have not actually spelled out what a p-value is supposed be. Recall the essential logic of p-values. If we consider a theoretical statistic Q, a possible observed value may

be written $Q = q$. When we apply the statistic to data we obtain an observed value which we write q_{obs}. We then suppose that the random variable Q is computed from random variables under the statistical model specified by the null hypothesis. The p-value is

$$p = P(Q \geq q_{obs}|H_0) \tag{10.12}$$

where we have used the conditioning notation to emphasize that[4] the probability is computed under the assumption that H_0 holds.

In many situations it is possible to generate *pseudo-data* under the null hypothesis. That is, the statistical model specified by the null hypothesis contains certain probability distributions, and it is often relatively easy to generate observations from these probability distributions on the computer. When this is feasible, the computer can be used to *simulate* a set of artificial data, according to the model, which will resemble the real data. Each such set of pseudo-data will, for example, have the same number of observations as the real data while having all the statistical properties specified by the model. We call such a set computer-simulated data "pseudo-data." The pseudo-data data will have the characteristics we assume in our theoretical world of statistical modeling. If we can create sets of pseudo-data repeatedly, a large number of times (each set of pseudo-data being different due to the randomness specified by the statistical model) then we can also compute the p value numerically.

The idea is to generate a large number G of pseudo-data sets (e.g., $G = 10,000$) and apply the statistic Q to each set of pseudo-data. This produces G computer-generated observations from the probability distribution of Q (under H_0). To find $p = P(Q \geq q_{obs})$ we then simply have to get the proportion of such generated observations (out of $G = 10,000$) for which Q is as large as q_{obs}. Let us use $Q^{(g)}$ to denote a value of Q computed from a set of pseudo-data, where $g = 1, 2, \ldots, G$. Here is the algorithm.

1. Generate G sets of pseudo-data labelled $g = 1, \ldots, G$ and for the gth

[4]This may be considered an abuse of the notation because we usually consider H_0 to be a fixed, non-random entity, so we are not really "conditioning" on it in the usual sense developed in Chapter 3. The exception occurs under the Bayesian interpretation given in Section 10.4.6, where H_0 is formally considered to be an event. In that scenario the probability in (10.18) does become a conditional probability.

set of pseudo-data compute $Q^{(g)}$.

2. Let N be the number of sets of pseudo-data for which $Q^{(g)} \geq q_{obs}$.

   The p-value is given by $p = \frac{N}{G}$.

Example 1.2.1 (continued) Let us take X to be random variable representing the number of non-burning house preferences. Under the null hypothesis we have $X \sim B(17, .5)$. As our test statistic we may use $Q = |X - 8.5|$, where 8.5 is the expected value of X and we are here judging small and large deviations from 8.5 to be equally important. We have $q_{obs} = 14 - 8.5 = 5.5$. We may then simulate 10,000 observations from a $B(17, .5)$ distribution and count the number N for which $Q \geq q_{obs}$. Doing this, we obtained $N = 126$ and $p \approx .013$. □

We used Example 1.2.1 to demonstrate the idea, but actually we do not need computer simulation in this case because binomial probabilities are readily available. We have that $Q \geq q_{obs}$ precisely when $x \geq 14$ or $x \leq 3$. Thus we simply find $p = P(X \leq 3) + P(X \geq 14)$. As it happens, for the $B(17, .5)$ distribution, by symmetry, $P(X \geq 14) = P(X \leq 3)$. The great virtue of p-values based on pseudo-data is that they can be easy to compute even in very complicated situations where direct calculation is impossible.

One issue is that the accuracy of such computer-generated p-values depends on the number of data sets generated. If we take G to be extremely large we can get a very accurate p-value, but in complicated problems the computing time may get too long. In most problems $G = 10,000$ is large enough to obtain reasonable accuracy.

> *Details:* In fact, we may compute the accuracy of such computer-generated p-values quite generally from the binomial standard error. If we generate G data sets, we have $N \sim B(G, p)$ where p is the desired p-value, which is estimated by $\hat{p} = N/G$. The standard error for this binomial proportion is $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/G}$. Thus, in the example above, the accuracy would be $SE = \sqrt{(.0126)(.9874)/10,000} = .0011$. Doubling this we get a 95% CI for p of $.013 \pm .002$. □

10.4 Interpretation and Properties of Tests

We now turn to some theoretical aspects of significance tests. In practice, new situations arise where no standard test is available. Researchers then invent significance tests, and sometimes they are not valid. What do we mean by this? The key property is Equation (10.18). For an evaluation of statistical significance to be correct, theoretically, (10.18) must be satisfied.

Let $F_Q(x)$ be the cdf of Q under the statistical model specified by H_0. We may then write an equivalent form of (10.18),

$$p = 1 - F_Q(q_{obs}). \tag{10.13}$$

This will help below. Sometimes (10.18) does not hold exactly, but it does hold approximately, as in the case of chi-squared tests. In Section 10.4.1 we derive a consequence that allows us to check whether (10.18) is approximately true and in Section 10.4.2 we go a step further with a more detailed calculation. Those two sections consider the behavior of a test when H_0 is true. In Section 10.4.4 we consider what happens when H_0 is false.

10.4.1 When the null hypothesis holds, the p-value should follow a uniform distribution.

Equation (10.18) defines the p-value using an observed value of the statistic q_{obs}. We may consider the theoretical properties of the p-value by examining its behavior when q_{obs} is computed from a random sample $X_1, \ldots, X_n$, which we may write as $q_{obs} = Q(X_1, \ldots, X_n)$. We then may also consider the p-value to be the observation of a random variable P, i.e., $p = P(X_1, \ldots, X_n)$. We have the following simple result.

Theorem Let $X_1, \ldots, X_n$ be a random sample from which $P = P(X_1, \ldots, X_n)$ is defined using (10.18). If H_0 holds the random variable P satisfies $P \sim U(0, 1)$.

Proof: Using (10.13) we have $P = 1 - F_Q(Q)$ and the theorem is an immediate consequence of the probability integral transform given in Section 3.2.5. □

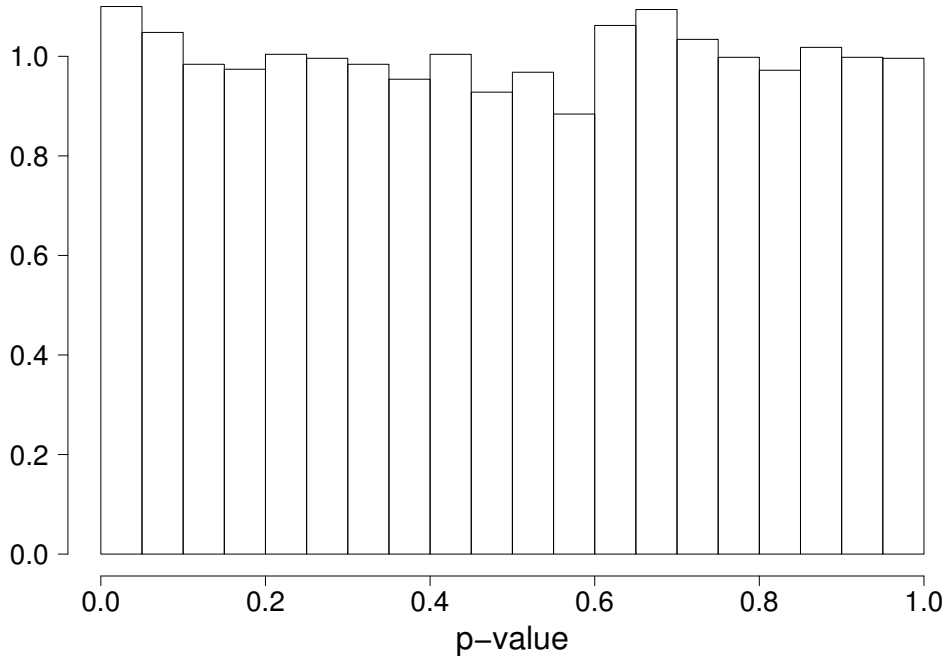


Figure 10.2: *Histogram of test-enhanced learning p-values under H_0. The p-values were computed by sampling at random the 60 data values under the SSSS and SSST conditions and arbitrarily putting them into two groups of 30 each, then running a t-test, as in the t-test on page 293.*

Example 10.2.1 (continued) To illustrate the uniformity of p-values guaranteed by the theorem, we generated samples of pseudo-data based on the real data used in the t-test on page 293. The idea was to begin with the 60 data values under the SSSS and SSST conditions and create 10,000 sets of pseudo-data like the real data except that for each set of pseudo-data H_0 was true. To force H_0 to hold we sampled the 60 data values and then arbitrarily put them into two groups of 30 values each, so that each of the two groups of pseudo-data would follow the same distribution.[5] We repeated this to get

[5]Specifically, both groups followed the distribution specified by the empirical cdf based on the 60 data values. This is an example of a *bootstrap distribution* and will lead to a *bootstrap test* discussed in Chapter **??**.

the 10,000 sets of pseudo-data, and then ran the t-test and computed the p-value for each set of pseudo-data. Figure 10.2 is a histogram of the resulting 10,000 p-values. The distribution is uniform. □

10.4.2 Statistical tests should have the correct probability of falsely rejecting H_0, at least approximately.

The criteria for deterimining statistical significance, usually taken to be .05 or .01, are called *significance levels.* Neyman and Pearson observed that research workers might routinely use $p < .05$ to claim that the null hypothesis is likely false. Indeed, this became standard practice. Neyman and Pearson then considered, formally, the behavior of such a procedure. They began by saying one might *reject* H_0 for sufficiently large values of the test statistic Q. If we let c be the cut-off value for which H_0 is rejected whenever $Q \geq c$, then c is called the *critical value* and

$$\alpha = P(Q \geq c)$$

is called the *level* of the test for the critical value c. Now, suppose for a particular test, such as the t-test on page 293 based on $Q = |T|$ and $q_{obs} = t_{obs}$ defined in (10.11), we choose $c = c_{.05}$ to be the crtical value producing $\alpha = .05$, i.e., $P(Q \geq c_{.05}) = .05$. Then, assuming $P(Q > c) < .05$ whenever $c > c_{.05}$, we have $p < .05$ precisely when $q_{obs} > c_{.05}$, and the probability of rejecting H_0 when it is actually true will be .05. In other words, by rejecting H_0 whenever $p < .05$, we will be making an incorrect decision 5% of the time.

> *A detail:* For continuous statistics like that in the t-test we can find $c_{.05}$ for which $P(Q \geq c_{.05}) = .05$ and $P(Q > c) < .05$ whenever $c > c_{.05}$, as we assumed in the preceding paragraph. In the discrete case, however, only particular values of probabilities actually occur, so there may not exist $c_{.05}$ for which $P(Q \geq c_{.05}) = .05$ and, furthermore, there will be values $a > b$ such that $P(Q > a) = P(Q > b)$. We ignore this technical point here. □

This kind of desirable calibration of p-values in terms of significance levels is satisfied when (10.18) holds. That is, for any α between 0 and 1, a test that rejects H_0 whenever $p < \alpha$ will have α as its significance level. Formula (10.18) holds for the t-test under the assumption of normality, but without the assumption of normality (10.18) is only approximately correct, as in the first version in Section 10.3.3. Similarly, (10.18) holds only approximately for the p-values computed from the chi-squared distribution based on the chi-squared statistics in Section 10.1. For approximate tests it is good to know how close the p-value is to being correct. Sometimes p-values may be obtained by computer simulation, as in Section 10.3.4, but this is not always possible. When a new statistical test is proposed to deal with a complicated or unusual situation, it may provide approximate p-values. In this case it is valuable to verify, by computer simulation, that the test has approximately the level $\alpha = .05$ when $p < .05$, and similarly for other levels such as $\alpha = 01$. For illustrative purposes we carried out the calculation in the case of the example on blindsight of patient P.S.

Example: Blindsight of P.S. Let us consider the use of χ^2 as we did on page 284. For a χ^2_1 distribution we have $c_{.05} = 3.84$, i.e., if $X \sim \chi^2_1$ then $P(X \geq 3.84) = .05$. For the case $n = 17$ and $p_0 = .5$ we may compute the value of $\alpha = P(Q \geq 3.84)$ where Q is the chi-squared statistic. This is easily done by computer simulation. We obtained $\alpha = .049$. Repeating this for $c_{.01} = 6.63$ we obtained $\alpha = .013$. For these standard cut-off values for p, and for this sample size, we conclude that the χ^2_1 distribution furnishes an accurate approximation. On the other hand, we should recall that the p-value we obtained for the data $x = 14$ was $p = .0076$ based on χ^2 while the exact p-value was $p = .0127$. □

10.4.3 A confidence interval for θ may be used to test $H_0 : \theta = \theta_0$.

Let us return to the "paradigm case" of Section 7.3.2 in which $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution with the value of σ known. In Section 7.3.2 we found a confidence interval for μ. Now let us consider, instead, the null hypthesis $H_0 : \mu = 0$. This hypothesis comes up frequently because many experiments generate a pair of observations for each subject,

one observation under each of two conditions, and the data may be reduced by taking the difference of the two observations. Thus, instead of n pairs we get n single-number differences X_i and the null hypothetical question becomes whether the mean of these differences is zero. In practice, the value of σ is unknown but here, as in Section 7.3.2, we assume it is known in order to make the derivation below a little more immediate.

As in Section 7.3.2 we have standard error $\mathrm{S}E(\bar{X}) = \sigma/\sqrt{n}$. In Section 7.3.2 we showed that the interval $(\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))$ is a 95% CI for μ, which means

$$P(\bar{X} - 2 \cdot SE(\bar{X}) \leq \mu \leq \bar{X} + 2 \cdot SE(\bar{X})) = .95.$$

There is a simple and intuitive way to test $H_0 : \mu = 0$. We simply check whether our 95% CI contains 0. If it does not, we have evidence against H_0.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, with the value of σ known, If $H_0 : \mu = 0$ holds, then we have

$$P(0 \notin (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) = .05.$$

Proof: For every μ we have

$$\begin{aligned} & P(\mu \notin (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) \\ = \; & 1 - P(\mu \in (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) \\ = \; & 1 - .95 = .05. \end{aligned}$$

The result follows by taking $\mu = 0$. □

This theorem says that the confidence interval for μ may be *inverted* to produce a test of $H_0 : \mu = 0$. We use the term "inverted" because instead of looking *within* the interval, as we do in the usual application of a confidence interval, in testing H_0 we are seeing whether it lies *outside* the confidence interval. When $\mu = 0$ lies outside the confidence interval we reject H_0 with significance level $\alpha = .05$, and can report $p < .05$.

The same logic may be used to state a version of the theorem in more general form.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a distribution that depends on a single parameter θ, and suppose $(\hat{\theta}-2\cdot SE(\hat{\theta}), \hat{\theta}+2\cdot SE(\hat{\theta}))$ is a 95% CI, i.e.,

$$P(\hat{\theta} - 2 \cdot SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 2 \cdot SE(\hat{\theta})) = .95.$$

If $H_0 : \theta = \theta_0$ holds, then we have

$$P(\theta_0 \notin (\hat{\theta} - 2 \cdot SE(\hat{\theta}), \hat{\theta} + 2 \cdot SE(\hat{\theta}))) = .05.$$

Proof: The argument is the same here as for the previous theorem. The details are omitted. □

This theorem says that whenever we have a 95% confidence interval for a parameter, we may invert it to get a test of a null hypothesis that takes the form $H_0 : \theta = \theta_0$. We stated the theorem to indicate generality, but actually the paradigm case of the normal sample with σ known furnishes one of the rare situations in which a standard confidence interval has exactly the correct coverage probability of .95. More commonly we rely on intervals that have *approximate* coverage probability .95. The method of using an approximate 95% confidence interval to test a hypothesis produces a signficance level of approximately .05. In practice, one makes sure that the null-hypothetical value is far outside the confidence interval, as in Example 1.2.1 in Chapter 1.

10.4.4 Statistical tests are evaluated in terms of their probability of correctly rejecting H_0.

In Section 10.4.2 we pointed out that a statistical test should have its significance levels match reasonably well its reported p-values, at least in the case of .05 and .01, and that this results in incorrect rejection of H_0 with the putative frequency (e.g., 5% or 1% of the time). But suppose we have two different ways of testing a hypothesis. How should we judge which way is better?

To answer this question, we may consider not only incorrect rejection of H_0 but also an incorrect decision not to reject. The two possible decisions may be identified as "reject H_0" and "accept H_0." There are then two types

of errors: incorrectly rejecting H_0 when it is in fact true, and incorrectly accepting H_0 when it is in fact false. These are called *type I* and *type II* errors. A good test would be one with small *type I* and *type II* errors. In order to evaluate the type II error we must introduce a particular non-null hypothesis. This is called the *alternative hypothesis* and is usually denoted H_A (or H_1). The *power* of the test is then the probability of correctly rejecting H_0 when H_A is true, i.e., it is one minus the type II error. The power is usually denoted by β. Thus, for a test based on large values of a statistic Q we have

$$\alpha = P(Q \geq c|H_0) \tag{10.14}$$

and

$$\beta = P(Q \geq c|H_A). \tag{10.15}$$

If we have two different tests that we want to compare, we may pick for each their respective critical values $c_{.05}$, and then ask, for a particular H_A, which test is more powerful. This is the general program laid out by Neyman and Pearson, and it is the standard way to evaluate competing statistical tests of hypotheses.

Illustration: Assessing Time-Varying Dependence Between Two Spike Trains Ventura, Cai, and Kass (2005) (*J. Neurophys.*) proposed a bootstrap method of testing the null hypothesis of independence between two spike trains. Their method not only tested independence but also found a window of time over which the two spike trains had increased joint activity. To compare the new method to an existing procedure that instead used contiguous time bins in the joint peri-stimulus time histogram, Ventura *et al.* computed power for a particular series of scenarios as the excess joint firing, above that predicted by independence, was increased. Figure 10.3 is a plot of power as a function of excess firing rate for the two methods. The purpose of such a plot is to demonstrate the superiority of a proposed method to an existing alternative. The plot in Figure 10.3 indicates especially large gains in power for 15-20% excess joint activity. □

Another use of power is to determine sample size. The idea is to choose an alternative H_A, considered to be plausible, and ask how big a sample size would be needed to achieve both a particular level α and a particular power β. The values $\alpha = .05$ and $\beta = .8$ are often used in medical applications, and planners of clinical trials typically must show to reviewers their calculation

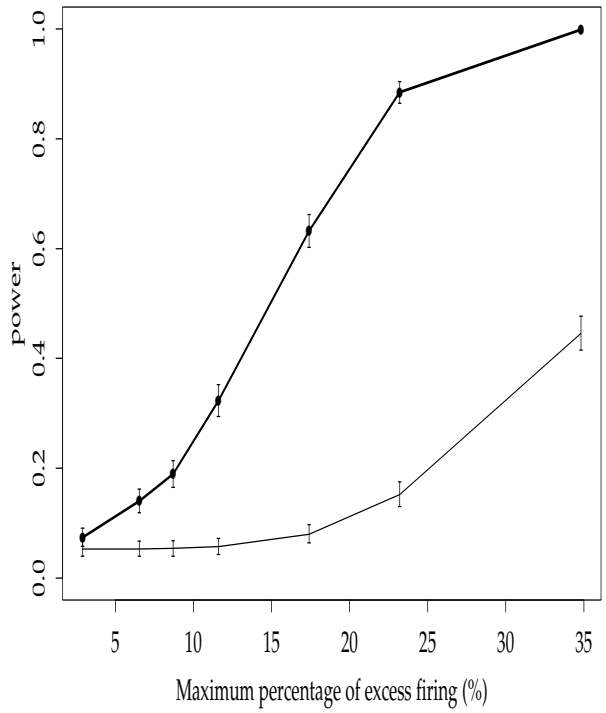


Figure 10.3: *Power of the bootstrap-based excursion test (thick black line) compared with the test using two contiguous bins in the JPSTH (thin line). Power is plotted against maximum percentage excess firing above that predicted by independence. Both tests have the same probability of rejecting* H_0 *when* H_0 *holds (type I error)* $\alpha = .05$*, indicated by the coincidence of the two power graphs when the percentage excess firing is zero.*

that the proposed sample size meets such specifications under reasonable assumptions.

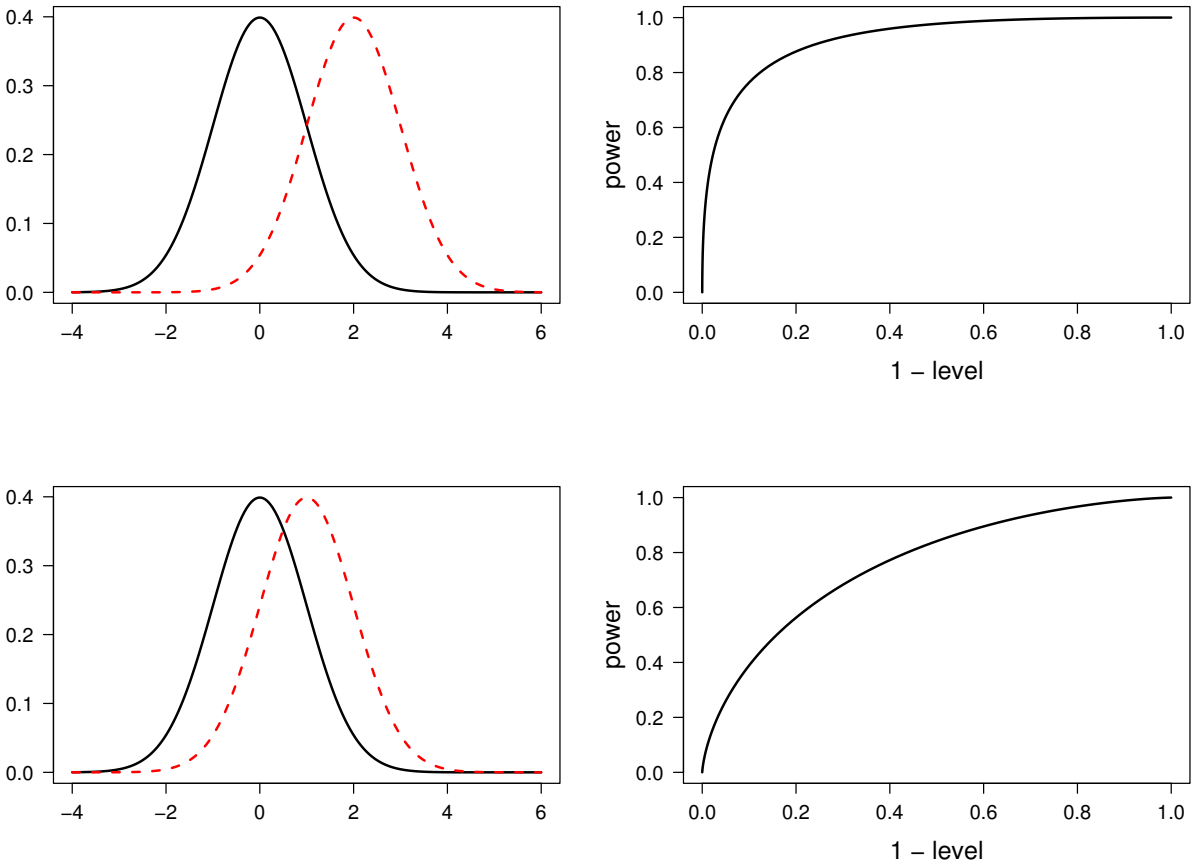


Figure 10.4: Two pairs of normal distributions and the resulting ROC curves. The left-hand side shows the pair of pdfs for $N(0,1)$ (solid) and $N(\delta,1)$ (dashed) and to the right are the corresponding ROC curves. Top: $\delta = 2$. Bottom: $\delta = 1$.

10.4.5 The performance of a statistical test may be displayed by the ROC curve.

According to (10.14) and (10.15), the level and power of a test based on large values of a statistic Q depend on the critical value c. Let us make this

dependence explicit by writing

$$\alpha(c) = P(Q \geq c|H_0) \tag{10.16}$$

and

$$\beta(c) = P(Q \geq c|H_A). \tag{10.17}$$

The choice of c is based on a trade-off of type I and type II errors: when c is increased, $\alpha(c)$ gets smaller so type I error decreases but $\beta(c)$ also gets smaller so type II error increases. The performance of a test may be examined by plotting $\beta(c)$ versus $\alpha(c)$ for a range of values of c. The customary plot is a slight variant of this, in which the x-axis is instead $1 - \alpha(c)$. The function $y = f(x)$ that traces values $(x, y) = (1 - \alpha(c), \beta(c))$ is called the *receiver-operating characteristic (ROC)* curve.

The simplest setting is the paradigm case of Section 10.3.1, where $\bar{X} \sim N(\mu, \sigma^2/n)$ and we wish to test $H_0 : \mu = \mu_0$. If H_0 holds, then the ratio Z defined in (10.4) satisifies $Z \sim N(0, 1)$ but if $H_A : \mu_1$ holds with $\mu_1 \neq \mu_0$, then $Z \sim N(\delta, 1)$ where $\delta = (\mu_1 - \mu_0)/SE(\bar{X})$. The ROC curves for $\delta = 2$ and $\delta = 1$ are shown in Figure 10.4. When $\delta = 1$ it is more difficult to discriminate between the two alternatives, so the power (β) is lower for a given value of the level (α) and the ROC curve is closer to the line $y = x$ (which is the ROC curve when $\delta = 0$). If we were instead to pick a very small value of δ the ROC curve would essentially fall on the line $y = x$, while if we picked a very large value of δ the ROC curve would hug the $y - axis$ near $x = 0$ and hug the asymptote $y = 1$ for increasing values of x. Thus, the higher the curve, the better its overall performance. Sometimes tests are compared by plotting their ROC curves. In addition, the area under the curve is often evaluated: it is 1 (the area of the 0-1 square) for a perfect test and .5 (the area within the square under the line $y = x$) for tests with no predictive ability at all (in the normal case corresponding to $\delta = 0$).

The ROC curve is also used in psychophysical analysis of perceptual detection of stimuli, called *signal detection theory*.

10.4.6 The p-value is *not* the probability that H_0 is true.

The p-value is commonly misinterpreted as the probability that the null hypothesis is true. This is quite wrong. A correct statement is necessarily rather cumbersome. Suppose we write a generic test statistic as Q and the value it takes when calculated from data as q_{obs}. In the case of the chi-squared tests we used $Q = X \sim \chi^2_\nu$ with $x_{obs} = \chi^2_{obs}$ and for the two-sided t test (10.11) $Q = |T|$ with $q_{obs} = |t_{obs}|$. The notation q_{obs} is used so that we can clearly distinguish the observed value from the theoretical random variable Q. The p-value is then

$$p = P(Q > q_{obs}|H_0) \tag{10.18}$$

where we have included the conditioning on H_0 to emphasize that the calculation of probability occurs under the assumption that H_0 holds. In words, p is the probability that one *would observe* a value of the test statistic as discrepant from the null hypothesis as the one observed from the data, *if the null hypothesis were true*. Or, again, in slightly different words: if the null hypothesis were true, the test statistic Q would have a probability distribution; the p-value is the resulting probability that Q would be as discrepant from the null hypothesis as the value q_{obs} actually observed. There is really no way to say this that is substantially simpler. The important point about the correct interpretation is its subjunctive nature: the p-value is a probability based on what *might* have happened if a random sample had been drawn under H_0.

Because the logic behind p-values is somewhat convoluted, they are very often misinterpreted to mean something much simpler and more direct, namely the probability that H_0 is true. There is no denying how nice it would be to have the probability that H_0 is true, based on the data. That probability may be obtained, instead, from Bayes' Theorem:

$$P(H_0|data) = \frac{P(data|H_0)P(H_0)}{P(data|H_0)P(H_0) + P(data|H_A)P(H_A)}.$$

From a practical point of view, however, the simplicity of this "Bayesian" result is deceptive. In data-analytic problems its application requires considerable care. For a detailed discussion see Kass and Raftery (1995) (Kass,

R.E. and Raftery, A. (1995) Bayes factors, *J. American Statistical Association*, 90: 773–795.). Nonetheless, with reasonable assumptions one may use Bayes' Theorem to get guidance on the interpretation of p-values. It turns out that $p = .05$ corresponds to values of $P(H_0|data)$ somewhere between roughly .5 and .7. In other words, a p-value of .05 is really only marginal evidence against H_0. Most importantly, $p = .05$ does *not* correspond to $P(H_0|data) = .05$. Additional discussion of these issues may be found in Edwards, Lindeman, Savage (1963), Kass and Raftery (1995), and Sellke, Bayarri, and Berger (2001).

10.4.7 The p-value is conceptually distinct from type one error.

We began by presenting p-values as a way of assessing evidence against a null hypothesis, and then reviewed the basic elements of the additional hypothesis testing framework based on evaluation of the performance of a test under both null and alternative hypotheses. The latter was introduced originally by Neyman and Pearson. Fisher disliked the Neyman-Pearson conception because he thought the alternative hypothesis was artificial and unnecessary—more than that, he thought it was counter-productive. In the Neyman-Pearson scheme there was no apparent role for p-values: in principle, one would pick a level α *a priori* and then determine whether $p < \alpha$ rather than reporting the p-value itself. Furthermore, the implication was that, in practice, the null hypothesis might routinely be accepted rather than rejected. This was the point that Fisher found most troubling. He said, "It is certain that the interest of statistical tests for scientific workers depends entirely [on] their use in rejecting hypotheses which are thereby judged to be incompatible with the observations." (R.A. Fisher (1935) Statistical tests, *Nature*, 136: 474.) From our current vantage point it is easy enough to step back from that early controversy. On the one hand, Fisher was correct that p-values and the rejection of statistical hypotheses would become a major activity of everyday science. On the other hand, the Neyman-Pearson conceptions have proven their worth in theoretical work. The modern point of view is thus a synthesis of Fisher's "significance testing" and the Neyman-Pearson "hypothesis testing." Like most others, we no longer feel any need to distinguish between these separate notions, which were once considered

incompatible, and we use the terms "signifance testing" and "hypothesis testing" interchangeably.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 11

General Methods for Testing Hypotheses

11.1 Likelihood Ratio Tests

Where do statistical tests come from? Sometimes they are based on intuition. A particular discrepancy measure may seem sensible as a way to capture the relevant departure from H_0. For instance, in the case of patient P.S. in Example 1.2.1 it would seem reasonable to use a test based somehow on $|\hat{p} - p_0|$. In Section 10.3.2 we have also suggested a general procedure, which is applicable to problems where H_0 involves only a single, scalar parameter, or a single component of a parameter vector, or a scalar function of a parameter vector. What about hypotheses that involve multiple parameters? Just as ML estimation is very widely applicable to parametric estimation problems, the *likelihood ratio test* is very widely applicable to parametric testing problems.

11.1.1 The likelihood ratio may be used to test $H_0 : \theta = \theta_0$.

The likelihood function assigns to alternative values of θ their plausability in light of the data $L(\theta)$. It can be used, analogously, when a particular value of θ is singled out in the form of a null hypothesis $H_0 : \theta = \theta_0$. That is, we consider the value $L(\theta_0)$ and assess whether it is nearly the same as the maximal value $L(\hat{\theta})$. Here, θ could be either a scalar or a vector. For a random sample $X_1, \ldots, X_n$, we may examine the likelihood ratio

$$LR = \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})}$$

and see how small it is. Because the MLE maximizes the likelihood function, we have $LR \leq 1$.

For a random sample $X_1, \ldots, X_n$ with joint pdf $f(x_1, \ldots, x_n|\theta)$, the likelihood ratio test of $H_0 : \theta = \theta_0$ evaluates the observed likelihood ratio statistic

$$LR_{obs} = \frac{f(x_1, \ldots, x_n|\theta_0)}{f(x_1, \ldots, x_n|\hat{\theta})}$$

and assigns the p-value

$$p = P(\frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})} < LR_{obs}) \tag{11.1}$$

computed under the assumption that $H_0 : \theta = \theta_0$ is satisfied, i.e., the assumption that $X_1, \ldots, X_n$ have pdf $f(x_1, \ldots, x_n|\theta_0)$.

Note that it is equivalent to examine the log of the likelihood ratio: in (11.1) we may take logs to get

$$p = P(\log \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})} < \log LR_{obs}).$$

As when maximizing a likelihood function, taking logs generally simplifies the expression. In addition, the log likelihood ratio is often multiplied by -1 so that larger values produce greater evidence against H_0, i.e., we compute

$$p = P(-\log \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})} \geq -\log LR_{obs}). \tag{11.2}$$

Example 1.2.1 (continued) Suppose $X \sim B(n,p)$ and we wish to test $H_0 : p = p_0$. In the case of the data from P.S., we would have $p_0 = .5$ and $\hat{p} = x/n$, with $n = 17$ and $x = 14$. The pdf is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

and the observed likelihood ratio statistic is

$$\begin{aligned} LR_{obs} &= \frac{p_0^x (1-p_0)^{n-x}}{\hat{p}^x (1-\hat{p})^{n-x}} \\ &= \frac{1}{2^n (\frac{x}{n})^x (1-\frac{x}{n})^{n-x}} \\ &= \frac{1}{2^n (\frac{14}{17})^{14} (1-\frac{14}{17})^3}. \end{aligned}$$

The negative log likelihood ratio becomes

$$\begin{aligned} -\log LR_{obs} &= n\log 2 + x\log\frac{x}{n} + (n-x)\log(1-\frac{x}{n}) \\ &= 17\log 2 + 14\log\frac{14}{17} + 3\log(1-\frac{14}{17}). \end{aligned}$$

□

This statistic is yet another way to test H_0 for this example. In general, for sufficiently large samples, the various methods we have presented so far will give equivalent results. The advantage of the likelihood ratio test is that it gives a specific method that can be applied in many, many problems and, furthermore, like ML estimation, it turns out to have very good properties in large samples.

11.1.2 *P*-values for the likelihood ratio test of $H_0 : \theta = \theta_0$ may be obtained from the χ^2 distribution or by simulation.

How do we find p-values for the likelihood ratio test? One way is to use the following very convenient result.

Result Under fairly general conditions, for large samples, if θ is m-dimensional then $-2\log LR$ is approximately distributed as χ^2_m, so that an approximate p-value may be obtained from the chi-squared distribution with m degrees of freedom.

Example 1.2.1 (continued) Continuing from the calculation above, we obtain

$$-2\log LR_{obs} = 2(17\log 2 + 14\log\frac{14}{17} + 3\log(1-\frac{14}{17})) = 7.72.$$

Here we have $m = 1$ degree of freedom for the chi-sqauared distribution. Writing $Y \sim \chi^2_1$ we find $P(Y \geq 7.72) = .0055$, i.e., we get $p = .0055$. This is only slightly different than the value $p = .0076$ obtained on page 284 from the χ^2 statistic. □

We have now used several alternative methods to test H_0 in Example 1.2.1. The chi-squared statistic and χ^2_1 distribution gave $p = .0076$. The likelihood ratio test and χ^2_1 distribution gave $p = .0055$. The exact calculation on page **??** gave $p = .013$. The discrepancies among these p-values are not very consequential for conclusions in this case. On the other hand, the numbers are different. This is due to the relatively small sample size. When conclusions depend on which test is used or the method of computing the p-value, the main message should be that the data are not decisive. When one must make a choice as to which p-value to report (in a publication), it is generally preferable to use an exact calculation of the p-value. The computation may be done by simulation. Specifically, under the assumption that H_0 holds, we generate a large number G of data sets and for each compute the test statistic—here, the likelihood ratio statistic—then find the proportion of such simulated test statistic that exceeds to observed value. We illustrate again using the blindsight example.

Example 1.2.1 (continued) For the responses of patient P.S. it is actually very easy to compute the exact p-value for the likelihood ratio. By symmetry about $p = .5$, it is apparent that $-2\log LR \geq -2\log LR_{obs}$ when $X \leq 3$ or $X \geq 14$. Thus, we would simply find $P(X \leq 3 \text{ or } X \geq 14)$ under the null-hypothetical assumption $X \sim B(17, .5)$. We computed this previously by simulation on page **??**, and we also noted on page **??** that simulation is unnecessary in this simple example. We found $p = .013$. Let us now write

out the steps in the simulation based on the likelihood ratio statistic, because these would be followed in more general contexts.

1. Define a function $LLR(x)$ that evaluates the loglikelihood ratio statistic. Here
$$LLR(x) = 17\log(2) + x\log(\frac{x}{17}) + (17-x)\log(\frac{17-x}{17}).$$
2. Make x a vector of G observations from the null distribution. Here we use $G = 100,000$ observations from $B(17, .5)$.
3. If there are possible values of the data that make the loglikelihood ratio become undefined (because the argument of a log would become zero), fix this. Here the log likelihood ratio is not defined when $x = 0$ or $x = 17$ so: if $x[g] = 0$ set $x[g] = 1$; if $x[g] = 17$ set $x[g] = 16$.
4. Set N equal to the number of values g for which $2LLR(x[g]) \geq 2LLR_{obs}$. This may be accomplished by creating a vector y of length G; if $2LLR(x[g]) \geq 2LLR_{obs}$ set $y[g] = 1$; otherwise set $y[g] = 0$; then $N = sum(y)$.

   Here $2LLR_{obs} = 7.72$.

   > *A detail:* The value 7.72 was actually rounded down slightly, so that we are computing $P(X \leq 3 \text{ or } X \geq 14)$ (rather than $P(X < 3 \text{ or } X > 14)$). We would rather compute $p = P(X \leq 3 \text{ or } X \geq 14)$ because it finds the probability of observing a value *at least as large as* LLR_{obs} instead of *larger than* LLR_{obs}, and is therefore more conservative in the sense of producing a larger p-value. □
5. Compute $p = \frac{N}{G}$.

□

11.1.3 The likelihood ratio test of $H_0 : (\omega, \theta) = (\omega, \theta_0)$ plugs in the MLE of ω, obtained under H_0.

We now consider the case in which the parameter vector may be decomposed into two sub-vectors ω and θ, having respective dimensions m_1 and m_2. For

example, in linear regression we would have a parameter vector (β_0, β_1) and we might decompose it as $\omega = \beta_0$ and $\theta = \beta_2$. We consider null hypotheses of the form $H_0 : \theta = \theta_0$ which now becomes a short-hand for $H_0 : (\omega, \theta) = (\omega, \theta_0)$. In linear regression, for example, we might consider whether there is a non-zero slope by introducing $H_0 : \beta_1 = 0$. This is short for $H_0 : (\beta_0, \beta_1) = (\beta_0, 0)$, which means simply that H_0 does not put any restriction on $\omega = \beta_0$. A wide variety of statistical models that are submodels of larger models may be written in this form. (See for example, Kass and Vos *The Geometry of Asymptotic Inference*, Theorem 2.3.2.)

To apply the likelihood ratio test, we must recognize that ω remains a free parameter under H_0. To evaluate the likelihood ratio we must pick a particular value of ω. To do so we maximize the likelihood, but under the null-hypothetical restriction $\theta = \theta_0$. That is, we maximize $L(\omega, \theta_0)$ over ω. Let us denote the solution by $\hat{\omega}_0$. Note that, in general $\hat{\omega}_0$ may not equal the global MLE $\hat{\omega}$ (though in some particular cases they will be equal).

For a sample $X_1, \ldots, X_n$ with joint pdf $f(x_1, \ldots, x_n|\omega, \theta)$, the likelihood ratio test of $H_0 : \theta = \theta_0$ evaluates the observed likelihood ratio statistic

$$LR_{obs} = \frac{f(x_1, \ldots, x_n|\hat{\omega}_0, \theta_0)}{f(x_1, \ldots, x_n|\hat{\omega}, \hat{\theta})}.$$

Computation of an exact p-value (by computer simulation) faces a substantial complication. In principle, to compute an explicit p-value, we would not only have to assume $\theta = \theta_0$ (which we do to satisfy H_0) but we would also have to assume some value for ω: to obtain

$$p = P(\frac{f(X_1, \ldots, X_n|\hat{\omega}_0, \theta_0)}{f(X_1, \ldots, X_n|\hat{\omega}, \hat{\theta})} \geq LR_{obs})$$

we must have an explicit probability distribution. Put differently, if we were to use computer simulation to find the exact p-value, we would have to know both the parameters ω, θ in order to do the simulation.

This problem is insoluble without introducing some further restriction or principle. (One leading idea is to find the "worst case" p-value (the largest) among all possible values of ω. This often remains intractable, except in

large samples.) Luckily, there are two good approximate solutions. Here is the first.

Result Under fairly general conditions, for large samples, if θ is a vector of length m then $-2 \log LR$ has an approximate $\chi^2_{m_2}$ distribution, so that an approximate p-value may be obtained from the chi-squared distribution with m_2 degrees of freedom.

The second method is to use $\omega = \hat{\omega}_0$ as a "plug-in" value, under which to compute the p-value. When this is done, i.e., when the p-value is computed by setting $(\omega, \theta) = (\hat{\omega}_0, \theta_0)$, generating many data sets, and then finding the proportion of them for which $LR < LR_{obs}$ the method is called a *parametric bootstrap*.

11.1.4 The likelihood ratio test reproduces, exactly or approximately, many commonly-used significance tests.

The likelihood ratio test may be used to derive the t test, and also other standard tests used in common situations. For testing independence of two traits, the likelihood ratio test is approximately equivalent to the χ^2 test of independence in large samples, meaning that in large samples it gives very nearly the same p-value as the χ^2 test of independence.

11.2 Permutation and Bootstrap Tests

11.2.1 Permutation tests consider all possible permutations of the data that would be consistent with the null hypothesis.

The idea behind permutation tests is illustrated by a famous example introduced by Fisher in his book *Design of Experiments*. There was, apparently, a lady who claimed to be able to tell the difference between tea with milk

added after the tea was poured, and tea with milk added before the tea was poured. Fisher asked how one might test this claim experimentally. His discussion emphasized the importance of *randomly* allocating the two treatments (milk second versus milk first) to many cups, without the subject's knowledge, and then asking for a judgment on each. He also considered the question of sample size, and the computation of a p-value. Fisher suggested using 8 cups of tea, 4 of which had the tea put in first and 4 had the milk put in first. There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ ways to select 4 tea-first cups among from the 8. Therefore, considering all these possible permutations, if the lady were randomly guessing, there would be a 1/70 chance she would correctly identify all cups of tea as either tea first or milk first. Thus, Fisher pointed out that, in that event that she correctly identified "before" or "after" for all 8 cups, there would be evidence against H_0 with $p = \frac{1}{70} = .014$. (Fisher also pointed out that with 6 cups there would be only 20 permutations and thus one would at best obtain $p = .05$; he considered this p-value too large to be useful.)

Example 10.2.1 (continued) We previously applied the two-sample t-test to the data displayed in Figure 10.1 obtained $t_{obs} = -3.19$ on 58 degrees of freedom, giving $p = .0023$. We now apply a permutation test analogous to that for the lady tasting tea.

In this data set there are two groups of 30 subjects. The permutation test considers all of the many ways that 60 subjects, with their learning results, could have been split into two groups of 30 and then asks, out of all those many ways of permuting the subjects, how many of them would have led to results as striking as the one actually observed? The number of ways of splitting 60 individuals into two groups of 30 is

$$\frac{60!}{30!30!} \approx 1.18 \times 10^{17}.$$

In other words, there are 10^{17} different samples of pseudo-data that would be obtained by permuting the group membership among the 60 subject values. The exact two-sample permutation test would, in principle, examine all of these 10^{17} samples and ask how many of them would produce a t-statistic at least as large in magnitude as $t_{obs} = -3.19$. This computation is possible, but it is just a bit complicated and we will skip it here. However, a variant on the idea is easy and will lead us naturally to the bootstrap procedure. Instead

of examining all 10^{17} permutations, we can *sample* from this distribution. (In R the function `sample(x)` produces a random permutation from the components of the vector x.) For example, a sample from the values 1,2,3,4,5 might be 1,5,3,2,4, which is a permutation of the original values. To get a relevant random permutation of the data we therefore sample the 60 data values and assign, the first 30 values to the first group (SSSS) and the last 30 values to the second group (SSST). We then compute the t-statistic for this permuted data set. If we repeat the procedure a large number of times (say, 10,000 times) we can thereby generate the distribution of the t-statistic under the permutations. □

To be clear, let us define the t-statistic as a function of data vectors x and y in several steps:

$$df = length(x) + length(y) - 2$$

$$v_{pooled}(x,y) = \frac{1}{df}\left((length(x)-1)var(x) + (length(y)-1)var(y)\right),$$

$$s_{pooled}(x,y) = \sqrt{v_{pooled}(x,y)}$$

and

$$t(x,y) = \frac{mean(x) - mean(y)}{s_{pooled}(x,y)\sqrt{\frac{1}{length(x)} + \frac{1}{length(y)}}}.$$

We then use the following algorithm.

1. For $i = 1$ to G:

   Generate $U_1^{(g)}, \ldots, U_{n_1+n_2}^{(g)}$ by permuting the components of the data vector $(x[1], \ldots, x[n_1], y[1], \ldots, y[n_2])$.

   Set $x^{(g)} = (U_1^{(g)}, \ldots, U_{n_1}^{(g)})$ and $y^{(g)} = (U_{n_1+1}^{(g)}, \ldots, U_{n_1+n_2}^{(g)})$.

   Compute $t^{(g)} = t(x^{(g)}, y^{(g)})$.

2. Set N equal to the number of values g for which $|t^{(g)}| \geq |t_{obs}|$.

3. Compute $p = \frac{N}{G}$.

Example 10.2.1 (continued) Applying the algorithm above with $G = 10,000$ we obtained $p = .0019$. Note that here the simulation standard error is $SE = \sqrt{(.0019)(.9981)/10,000} = .00044$. □

For large samples there is generally very little difference between the permutation distribution of the t statistic and the t distribution on which the usual t test is based. Results are likely to lead to different conclusions only for small samples.

One nice feature of the permutation test is that it is immediately applicable without using equal variances in the t-statistic. We simply re-define the t-statistic:

$$t(x, y) = \frac{mean(x) - mean(y)}{\sqrt{\frac{var(x)}{length(x)} + \frac{var(y)}{length(y)}}}.$$

Example 10.2.1 (continued) Applying the new version of the algorithm, without the assumption $\sigma_1 = \sigma_2 = \sigma$, we found $p = .0026$.

11.2.2 The Bootstrap samples with replacement.

Note that a permutation of a vector of data values x is actually a special case of sampling from that data set where (i) the sample size is equal to the length of x and (ii) the sampling is done *without replacement*, meaning that once a data value is selected it can not be selected again. An alternative type of sampling is *with replacement*. In this form, if n is the length of x, then 1 component of x is drawn at random repeatedly, with all components having equal probabilities of being drawn on all occasions, until a total n numbers are drawn. In this case, there may be repititions of values. For example, when the values 1,2,3,4,5 are sampled with replacement we might obtain 3,4,1,4,2. The bootstrap test is the same as this permutation sampling procedure except that the sampling is done with replacement. We therefore replace the permutation algorithm with the following variant.

1. For $i = 1$ to G:
   Generate $U_1^{(g)}, \ldots, U_{n1+n2}^{(g)}$ by sampling the components of the data vector $(x[1], \ldots, x[n_1], y[1], \ldots, y[n_2])$ with replacement.

Set $x^{(g)} = (U_1^{(g)}, \ldots, U_{n_1}^{(g)})$ and $y^{(g)} = (U_{n_1+1}^{(g)}, \ldots, U_{n_1+n_2}^{(g)})$.
Compute $t^{(g)} = t(x^{(g)}, y^{(g)})$.

2. Set N equal to the number of values g for which $|t^{(g)}| \geq t_{obs}$.

3. Compute $p = \frac{N}{G}$.

The only difference in the implementation (the R or Matlab code) would be that the line involving sampling without replacement is changed to sampling with replacement.

Example 10.2.1 (continued) Applying the bootstrap procedure without the assumption $\sigma_1 = \sigma_2$ we obtained $p = .0022$. □

This two-sample test provides a situation where it is especially straightforward to apply the bootstrap (or permutation test), because one merely merges the data across the two samples, resamples, and arbitrarily breaks each bootstrap sample into two groups of appropriate sizes. The idea is that the bootstrap draws pseudo-data from the a distribution that is approximately the same as the theoretical distribution that holds under H_0. In the two-sample problem, in theory, both groups of random variables are random samples from a distribution with cdf $F(x)$, which may be estimated by the empirical cdf $\hat{F}(x)$ found by merging the two sets of data. When we resample the merged data, we are drawing a new random sample of pseudo data from $\hat{F}(x)$ and, as in our previous use of the bootstrap for calculated SEs, we have $\hat{F}(x) \approx F(x)$.

In other situations the problem may be a bit more subtle, and it is important to keep in mind the theoretical underpinning of bootstrap methodology. As an example, let us consider testing the hypothesis $H_0 : \mu = \mu_0$ based on a sample $X_1, \ldots, X_n$ of spike counts on repeated trials. If the random variables X_i are assumed to be Poisson, then we may test the hypothesis quite easily by computing the p-value $p = P(\bar{X} \geq \bar{x}_{obs}|H_0)$ with $\bar{x}_{obs}$ being the observed sample mean. However, spike counts are often markedly non-Poisson. When the distribution of the X_i is unknown, we may apply the bootstrap. In this case, however, a bit of care is required in performing the bootstrap correctly. We can not merely sample the X_i's because their mean μ may not equal μ_0. Indeed, this is precisely the possibility we are examining when we perform

the statistical hypothesis test. Instead, we need to resample a version of the data that follow, approximately, the distribution the X_i variables *would have if their mean were* μ_0. Thus, we need to create variables that resemble the X_i's except that their mean is shifted to equal μ_0. To produce a version of X_i that has mean μ_0 rather than μ we would, in principle, want to subtract from X_i the quantity $(\mu - \mu_0)$. Because we do not know the value of μ we instead subtract $(\bar{X} - \mu_0)$ and consider the *translated* spike counts $x_1 - (\bar{x} - \mu_0), x_2 - (\bar{x} - \mu_0), \ldots, x_n - (\bar{x} - \mu_0)$. We then repeatedly resample these observations. That is, we draw observations $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ where each $U_i^{(g)}$ takes on the value of one of the observed $x_i - (\bar{x} - \mu_0)$'s, and all such values occur with probability $1/n$. Having thereby obtained a large number of samples we then compute the proportion of samples for which $|\bar{U}^{(g)} - \mu_0| \geq c$ which, again, estimates the desired $P(|\bar{X} - \mu_0| \geq c|H_0)$. This is an instance of the nonparametric bootstrap. It is nonparametric because it does not require the assumption of a specific parametric family of distributions (here, Poisson); instead, it requires the data values X_i to be *independent replications from the same probability distribution*. As before, it works because it implicitly uses the data $x_1, x_2, \ldots, x_n$ to estimate the probability distribution from which they are assumed to be drawn, and then computes the desired probability from this estimate.

11.3 Multiple Tests

11.3.1 When multiple independent data sets are used to test the same hypothesis, the p-values are easily combined.

11.3.2 When multiple hypotheses are considered, statistical significance should be adjusted.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 12

Linear Regression

One of the most important problems in data analysis, particularly when all its generalizations are included, is curve-fitting. In the univariate case, involving a single explanatory variable x, curve-fitting results in the summary of data pairs (x, y) as roughly following $y = f(x)$ for some suitable function $f(x)$. When f is a linear function of x, we may write it in the form $f(x) = \beta_0 + \beta_1 x$, and the statistical problem becomes one of estimating the coefficients β_0 and β_1 and, of course, representing the uncertainty in the estimates. There are many ways to fit lines to data, but the usual solution to the problem is in terms of the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, introduced in Chapter 1, together with their standard errors. The formulas (obtained by calculus) are

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{12.1}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{12.2}$$

The resulting fitted line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{12.3}$$

is usually called the (linear) *regression* line.

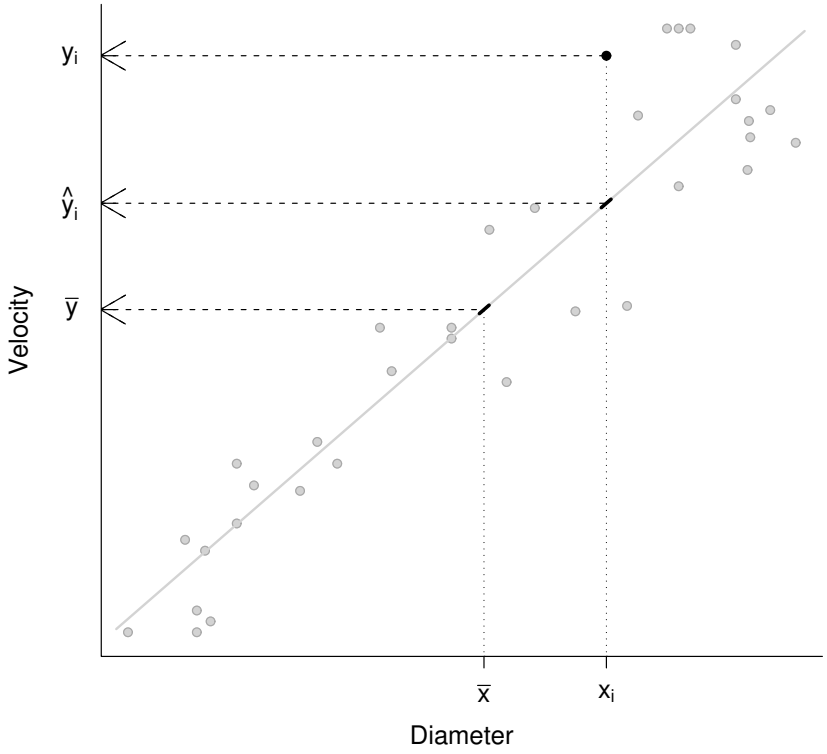


Figure 12.1: *Plot of the Hursh conduction velocity data set, with data points in gray except for a particular point* (x_i, y_i) *which is shown in black to identify the corresponding fitted value* $\hat{y}_i$*. The regression line also passes through the point* $(\bar{x}, \bar{y})$*, as indicated on the plot.*

There are numerous examples of linear regression in the neuroscience and psychology literature. For instance, Platt and Glimcher (1999, *Nature*) used linear regression to summarize the increasing trend in firing rate of intraparietal neurons with increasing expected gain in reward (volume of juice received) for successful completion of a task and Behrmann *et al.* (2002, *J. Cognitive Neuroscience*) used linear regression in examining the modulation of saccadic reaction time as a function of angle to target by eye, head, or trunk orientation among patients with hemispatial neglect. In Chapter 1 we used Example 1.2.2 on neural conduction velocity to illustrate linear regression.

Another plot of the neural conduction velocity data is provided again in Figure 12.1. The fitted values at each x_i are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{12.4}$$

If we plug (12.2) into (12.3) we get

$$y - \bar{y} = \hat{\beta}_1(x - \bar{x})$$

which shows that the regression line passes through the point $(\bar{x}, \bar{y})$, as may be seen in Figure 12.1.

12.1 The Linear Regression Model

12.1.1 Linear regression assumes linearity of $f(x)$ and independence of the noise contributions at the various observed x values.

Least-squares may be considered an algorithm for fitting a line to data. To get standard errors more structure is necessary: we need to have a statistical model. The starting point is to describe the deviations of the data from the line in terms of random variables. The fundamental assumption in linear regression is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{12.5}$$

for $i = 1, \ldots, n$, where ϵ_i is a random variable. This says that there is a *theoretical* linear relationship between x and y that holds except for "errors" ϵ_i. The ϵ_i's are assumed to be zero, on average, i.e., $E(\epsilon_i) = 0$ for all i, so that $E(Y_i) = \beta_0 + \beta_1 x_i$. Thus, the linear relationship $y = \beta_0 + \beta_1 x$ is assumed to hold "on average," that is, apart from errors that are on average zero; $\hat{\beta}_0$ and $\hat{\beta}_1$ are considered estimates of β_0 and β_1. The uncertainty arising from the ϵ_i's is used to quantify uncertainty in the least-squares estimates.

This linear model is the first crucial assumption. Another is that the errors ϵ_i are independent of each other. The independence assumption is often violated when there is a time trend in the observations (as when they

are recorded sequentially), in which case more elaborate methods are needed. Relationships may be examined in the presence of serial dependence using *time series* methods.

Important, though less potentially problematic, additional assumptions are that the variances of the ϵ_i's are all equal, so that the variability of the errors does not change with the value of x, and that the errors are normally distributed. These latter two assumptions are sufficient to guarantee that the 95% confidence intervals (see below) do have probability .95 of covering the coefficients. In sufficiently large samples these two assumptions become unnecessary, though in smaller samples they require attention, with substantial departures from the equality of variance assumption tending to be more problematic than departures from the normality assumption. The assumptions may be enumerated, in order of importance, as follows:

(i) the linear regression model (12.5) holds;

(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;

(iii) the errors ϵ_i are independent of each other;

(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and

(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

12.1.2 The relative contribution of the linear signal to the total response variation is summarized by R^2.

In Example 1.2.2 the least-squares line provides an excellent representation of the relationship between x and y, with the points clustering tightly around the line. In other cases there is much more "noise" relative to "signal," meaning that the relationship is not so close to being linear. There are two ways to summarize this overall tendency of the line to fit, or fail to fit the data: we use s and R^2.

The error standard deviation σ (see item (iv) in the assumptions in Section 12.1.1) represents the average size of the error, in the sense that it is an

average amount of deviation of each ϵ_i from zero. Thus, σ tells us how far off, on average, we would expect the line to be in predicting a value of y at any given x_i. It is estimated by $s = \sqrt{s^2}$ where

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

is usually called the "residual mean squared error." This definition of s makes it essentially the standard deviation of the residuals, except that $n-2$ is used in the denominator instead of $n-1$; here there are two parameters β_0 and β_1 being estimated so that two degrees of freedom are lost from n, rather than only one.

The other quantity, R^2, is interpreted as the fraction of the variability in Y that is attributable to the regression, as opposed to error. Its definition is

$$R^2 = 1 - \frac{SSE}{SST}$$

where $SSE = (n-2)s^2$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Here, SSE stands for "sum of squares due to error" and SST stands for "total sum of squares". The fraction SSE/SST is the proportion of the variability in Y that is attributable to error, and R^2 is what's left over, which is attributable to the regression line. The value of R^2 is between 0 and 1. It is 0 when there is no linear relationship and 1 when there is a perfect linear relationship. It is often interpreted as "the proportion of variability of Y that is explained by X." In different terminology, we may think of $SSR = SST - SSE$ as the *signal* variability (often called "the variability due to regression") and SSE as the *noise* variability. Then $R^2 = SSR/(SSR + SSE)$ becomes the relative proportion of signal-to-noise variability. (The ratio of signal-to-noise variabilities would be SSR/SSE.[1])

The decomposition of total variability (leading to SST) into the regression and error components (leading to SSR and SSE) begins with consideration of the values y_i, $\hat{y}_i$, and $\bar{y}$, as shown in Figure 12.1, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

[1]The signal-to-noise ratio is a term borrowed from engineering, where it refers to a ratio of the power for signal to the power for noise, and is usually reported in the log scale; it translates most immediately into a ratio of signal variance to noise variance, which is not quite what is being examined here, but is very closely related.

Writing $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$, we have

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2$$

but after plugging in the definition of $\hat{y}_i$ from (12.4) some algebra shows that the cross-product term vanishes and, defining $SSR = \sum (\hat{y}_i - \bar{y})^2$, we have

$$SST = SSR + SSE.$$

A demonstration of the vanishing of the cross-product term will be provided in our discussion of multiple regression.

The estimated standard deviation s has the units of Y and is therefore always immediately interpretable – at least to the extent that the Y measurements themselves are interpretable. But R^2 is dimensionless. Unfortunately, there are no universal rules of thumb as to what constitutes a large value: in some engineering applications one expects an R^2 of at least .99 while in many social science applications an R^2 of .40 would be considered quite substantial. One gets a feeling for the size of R^2 mainly by examining, and thinking about, many specific examples.

12.1.3 For large samples, if the model is correct, the least-squares estimate is likely to be accurate.

12.2 Checking Assumptions

12.2.1 Residual analysis is helpful because residuals should represent unstructured noise.

In examining single batches of data, in Chapter 2, we have seen how the data may be used not only to estimate unknown quantities (there, an unknown mean μ) but also to check assumptions (in particular, the assumption of normality). This is even more important in regression analysis and is accomplished by analyzing the residuals (the e_is). Sometimes the residuals are replaced by *standardized residuals*. The ith standardized residual is simply $e_i/SD(e_i)$, where $SD(e_i)$ is the standard deviation of e_i (as estimated

from the data). Dividing by the standard deviation puts the residuals on a familiar scale: since they are supposed to be normal, about 5% of the standardized residuals should be either larger than 2 or smaller than -2. Standardized residuals that are a lot larger than 2 in magnitude might be considered outliers.

> *A detail:* There are two different ways to standardize the residuals. We have here taken $SD(e_i)$ to be the estimated standard deviation of e_i. The formula for $SD(e_i)$ involves the x_i values. An alternative would be to compute the sample variance of the residuals
>
> $$s_e^2 = \frac{1}{n-1}\sum(e_i - \bar{e})^2$$
>
> and take its square root. The standardization using $SD(e_i)$, which allows the residual standard deviations to be different, is often called *studentization* (by analogy with Student's t ratio). The statistical software packages we are most familiar with use $SD(e_i)$ to standardize the residuals. □

Two kinds of plots are used. Residual versus fit plots are supposed to reveal (i) nonlinearity, (ii) inhomogeneity variances, or (iii) outliers. Plots having structure of the kind that would indicate these problems are shown in the Figure12.2. The first plot is typical of data with no systematic variation remaining after linear regression: the pattern is "random," specifically, it is consistent with errors that are independent and normally distributed, all having the same distribution. The second plot shows departure from linearity; the third indicates more variability for large fitted values than for smaller ones. The last plot has an outlier, indicating a point that is way off the fitted line.

Histograms and Q-Q plots of the residuals are also used to assess assumptions. These are supposed to (i) reveal outliers and (ii) check whether the the errors may be described, at least approximately, by normal distribution.

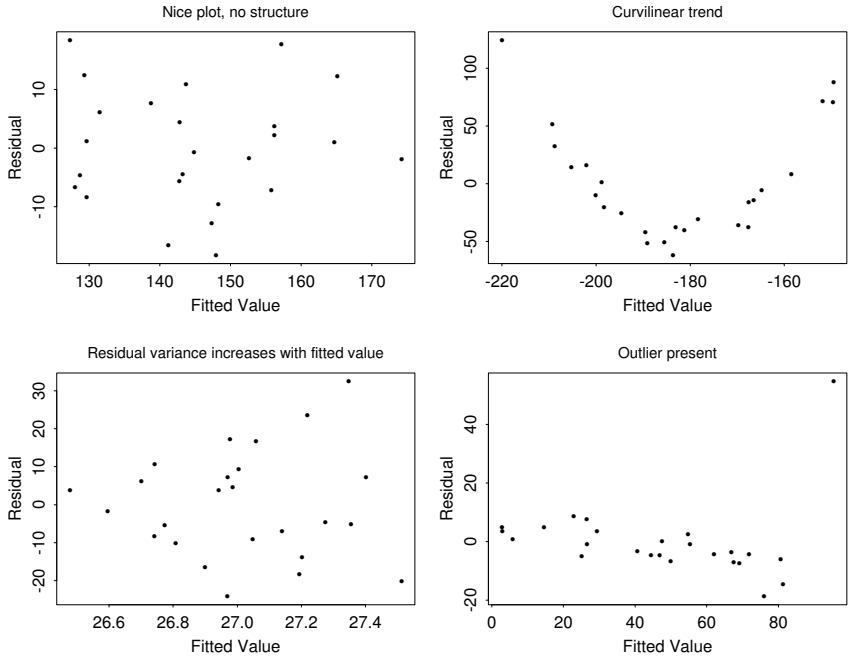


Figure 12.2: *Residual plots: the top left plot depicts unstructured noise while the latter three reveal structure, and thus deviations from the assumptions.*

12.2.2 Graphical examination of (x, y) data can yield crucial information.

Figure 12.3 shows a striking example in which four sets of data all have the same regression equation and R^2, but only in the first case (data set 1) would the regression line appropriately summarize the relationship. In the second case (data set 2) the relationship is clearly nonlinear, in the third case there is a big outlier and removing it dramatically changes the regression. In the fourth case the slope of the line is determined entirely by the height of the point to the right of the graph; therefore, since each point is subject to some random fluctuation, one would have to be very cautious in drawing conclusions. Taken together the main point of the example is that until one plots the data and/or the residuals it may not be clear how or whether the regression line summarizes the relationship.

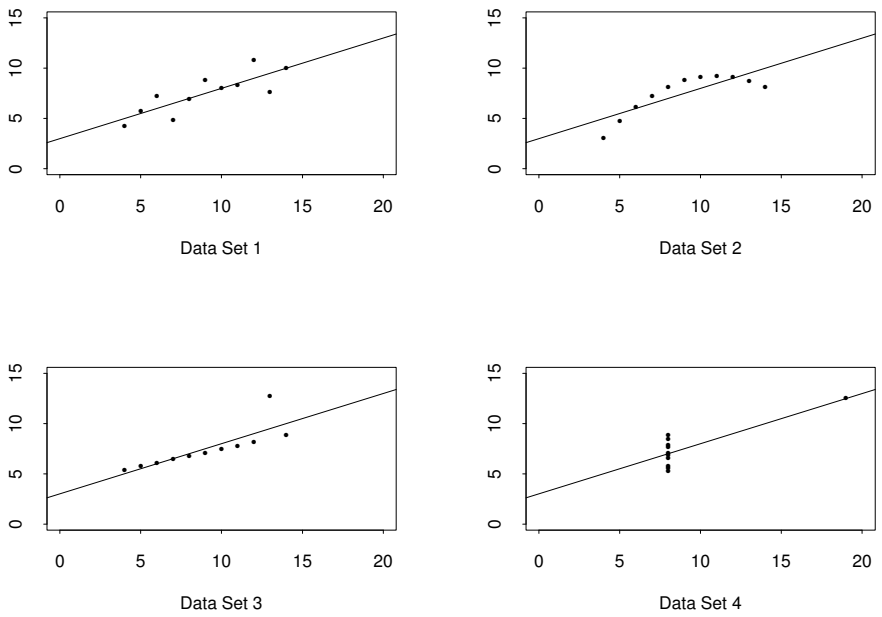


Figure 12.3: *Plots of four very different data sets all having the same fitted regression equation* $Y = 3 + .5x$ *and* $R^2 = .667$*. These were discussed in Anscombe, F.J. (1973), Graphs in statistical analysis,* American Statistician*, 27: 17-21.*

12.3 Evidence of a Linear Trend

12.3.1 Confidence intervals for slopes are based on SE, according to the general formula.

When reporting least-squares estimates, standard errors should also be supplied. Thus, in Example 1.2.2 we found that, on average, action potential velocity increases by $6.07 \pm .14$ meters per second for every micron increase in diameter of a neuron. To get a confidence interval we again use the general form

$$estimate \pm 2 \cdot SE$$

as an approximate 95% CI. Thus, for the conduction velocity example we have that an approximate 95% CI for the slope of the regression line is $6.07 \pm 2(.14)$ or $(5.79, 6.35)$.

An alternative in small samples, which is very much analogous to the small sample procedure in estimating a population mean, is to substitute for 2 the value $t_{.025,\nu}$, where now $\nu = n - 2$ because we have estimated two parameters (intercept and slope) and thus have lost two degrees of freedom. Thus, we would use the formula $\hat{\beta}_1 \pm t_{.025,n-2} \cdot SE(\hat{\beta}_1)$ as our 95% CI.

12.3.2 Evidence in favor of a linear trend can be obtained from a t-test concerning the slope.

In the conduction velocity example it is obvious that there is a linear trend in the data. This kind of increasing or decreasing tendency is sometimes a central issue in an analysis. For example, it is of interest to know whether the temperature in North America, and elsewhere, has had a tendency to increase over the years 1880-1980; this is the so-called phenomenon of "global warming." Here, detailed study would show that temperature has not increased linearly over this period of time, but there is a tendency toward increase that may be summarized with a line. Similarly, it was not crucial to the results of Platt and Glimcher (1999) that the relationship between firing rate and volume of juice received was exactly linear; the most important thing was to establish there was *some* upward tendency. In problems such as this, it makes sense to assume that y is roughly linear in x but to consider the possibility that in fact the slope of the line is zero—meaning that y is actually constant, on average, as x changes, that is, that y is really not related to x at all. We formalize this possibility as the null hypothesis $H_0 : \beta_1 = 0$.

When the 95% CI $\hat{\beta}_1 \pm t_{.025,n-2} \cdot SE(\hat{\beta}_1)$ provides an interval that omits the value 0, as in the conduction velocity case where the approximate CI is $(5.79, 6.35)$, we may regard this as evidence against $H_0 : \beta_1 = 0$. We may, instead test this hypothesis by computing a p-value based on the t-ratio:

$$t\text{-ratio} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.$$

For large samples, under H_0, this statistic has a $N(0,1)$ distribution. For small samples, if the assumptions of linear regression are satisfied, under H_0 the t-ratio has a t distribution on $\nu = n - 2$ degrees of freedom (which for large n is approximately normal). Thus, generally speaking, when the t-ratio is much larger than 2 the p-value will be small (much less than .05, perhaps less than .01) and there will be clear evidence against $H_0 : \beta_1 = 0$ and in favor of the existence of a linear trend. Again, it is conventional to consider $p < .05$ to indicate mild evidence against H_0 and $p < .01$ to indicate strong evidence against H_0. And, again, it is important to remember that departures from H_0 that are so small as to be irrelevant are very likely to be identified as significant in large samples. In the case of the conduction velocity example we obtained $p < 10^{-15}$; if it were necessary to report this, we would remember that such small p-values are sensitive to distributional form and we might instead say simply that the result is clearly statistically significant and report $p << .0001$.

12.3.3 The fitted relationship may not be accurate outside the range of the observed data.

An interesting related issue arises in Example 1.2.2. There, the fitted line does not go through the origin $(0, 0)$. In fact, according to the fitted line, when the diameter of the nerve is 0, the conduction velocity becomes negative! Should we try to fix this?

It is possible to force the line through $(0, 0)$, by omitting the intercept in the fitting process. Regression software typically provides an option for leaving out the intercept. However, for this data set, and in many others, omission of the intercept may be unwise. The reason is that the relationship may well be nonlinear near the origin, and there are no data to determine the fitted relationship in that region. Instead, we would view the fitted relationship as accurate only for (roughly) the range of values of diameters actually examined in the data.

12.4 Correlation and Regression

Sometimes the "explanatory variable" x is observed, rather than fixed by the experimenter. In this case the pair (x, y) is observed and we may model this by considering a pair of random variables X and Y and their *joint* distribution. Recall that the *correlation coefficient* ρ is a measure of linear association between X and Y. As we discussed in Chapter 4, the best linear predictor $\beta_0 + \beta_1 X$ of Y satisfies we have

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \cdot \rho. \tag{12.6}$$

as in Equation (**??**). Also, the theoretical regression of Y on X is defined to be $E(Y|X = x)$, which is a function of x, and it may happen that this function is linear:

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

In Chapter 4 we noted that the regression is, in fact, linear when (X, Y) has a bivariate normal distribution and then (12.6) holds. This linearity, and its interpretation, was illustrated in Figure 4.3. However, the right-hand plot in Figure 4.3 concerns data, rather than a theoretical distribution, and there is an analogous formula and interpretation using the *sample correlation*, which is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{s_x s_y}$$

where s_x and s_y are the sample standard deviations among the x_is and y_is. It may be shown that the sample correlation r is the MLE of ρ.

The sample correlation is related to the relative proportion of signal-to-noise variability R^2 by $R^2 = r^2$. Important properties are the following:

- $-1 \leq r \leq 1$ with $r = 1$ when the points fall exactly on a line with positive slope and $r = -1$ when the points fall exactly on a line with negative slope;
- the value of r does not depend on the units in which the two variables are measured;
- just as ρ measures linear association between random variables X and Y, so too may r be considered a measure of *linear* association.

There are no general guidelines as to what constitutes a "large" value of the correlation coefficient. Interpretation depends on the application.

12.4.1 The correlation coefficient is determined by the regression coefficient and the standard deviations of x and y.

Equation (12.6) gives the relationship of the theoretical slope β_1 to the theoretical correlation coefficient ρ. For data pairs (x_i, y_i) we have the analogous formula

$$\hat{\beta}_1 = \frac{s_Y}{s_X} \cdot r.$$

As a consequence, if x and y have about the same variability, the fitted regression slope becomes approximately equal to the sample correlation. In some contexts it is useful to standardize x and y by dividing each variable by its standard deviation. When that is done, the regression slope will equal the sample correlation.

12.4.2 Association is not causation.

There are two very important points to keep in mind. First, correlation is a measure of *linear* association; for instance, a correlation can be zero even if x perfectly, but nonlinearly, predicts y. Second, association is not causation: there are numerous examples of two variables having a high correlation while no one would seriously suggest that high values of one causes high values of the other. For example, one author looked at data from many different countries and pointed out that the number of telephones per capita had a strong correlation with the death rate due to heart disease. (Brownlee, KA (1965) *Statistical Theory and Methodology in Science and Engineering*, Wiley.) In such situations there are confounding factors that, presumably, have an effect on both variables and thus create a "spurious" correlation. Only in well-performed experiments, often using randomization[2], can one be

[2]Randomization refers to the random assignment of treatments to subjects, and to the process of randomly ordering treatment conditions; we discuss this further in Chapter **??**.

confident there are no confounding factors.

12.4.3 When noise is added to two variables, their correlation diminishes.

When correlation is measured in the presence of noise, its value is decreased—the math showing this is given below. In single-neuron recording experiments, for example, the correlation of quantities based on firing rates in a pair of tasks may be assessed across neurons but, because the number of trials is limited, the measured quantities for each neuron will be noisy. This attenuates the correlation.

Theorem Suppose U and W are random variables having a positive correlation ρ_{UW} and ϵ and δ are independent random variables that are also independent of U and V. Defining $X = U + \epsilon$ and $Y = W + \delta$ we have

$$\rho_{XY} < \rho_{UW},$$

where ρ_{XY} is the correlation between X and Y.

Proof details: We assume that $V(\epsilon) > 0$ and $V(\delta) > 0$ and we begin by writing

$$\begin{aligned} Cov(X,Y) &= Cov(U+\epsilon, W+\delta) \\ &= Cov(U,W) + Cov(U,\delta) + Cov(W,\epsilon) + Cov(\epsilon,\delta). \end{aligned}$$

Because of independence the last 3 terms above are 0. Therefore, $Cov(X,Y) = Cov(U,W)$ and we then have

$$\begin{aligned} \rho_{XY} &= Cor(U+\epsilon, W+\delta) \\ &= \frac{Cov(U,W)}{\sqrt{V(U+\epsilon)V(W+\delta)}} \\ &= \frac{Cov(U,W)}{\sqrt{(V(U)+V(\epsilon))(V(W)+V(\delta))}} \\ &< \frac{Cov(U,W)}{\sqrt{Var(U)Var(W)}} \\ &= \rho_{UW}. \end{aligned}$$

□

In words, when measurements are corrupted by noise their correlation decreases.

Example 12.4.1 Correction for attenuation of the correlation in SEF selectivity indices The theorem above indicates that when measurements are subject to substantial noise a measured correlation will underestimate the strength of the actual correlation between two variables. However, if the likely magnitude of the noise is known it becomes possible to correct the estimate. Correcting for attenuation can be consequential. Behseta, Berdyyeva, Olson, and Kass (2009, *J. Neurophysiol.*) reported analysis of data from an experiment on neural mechanisms of serial order performance. Monkeys were trained to perform eye movements in a given order signaled by a cue. For example, one cue carried the instruction: look up, then right, then left. Based on recordings of neural activity in frontal cortex (the supplementary eye field, SEF) during task performance, Behseta *et al.* reported that many neurons fire at different rates during different stages of the task, with some firing at the highest rate during the first, some during the second and some during the third stage. These rank-selective neurons might genuinely be sensitive to the monkey's stage in the sequence. Alternatively, they might be sensitive to some correlated factor. One such factor is expectation of reward. Reward (a drop of juice) was delivered only after all three movements had been completed. Thus as the stage of the trial progressed from one to three, the expectation of reward might have increased.

To see whether rank-selective neurons were sensitive to the size of the anticipated reward, the same monkeys were trained to perform a task in which a visual cue presented at the beginning of the trial signaled to the monkey whether he would receive one drop or three drops of juice after a fixed interval. The idea was that neuronal activity related to expectation of reward would be greater after the promise of three drops than after the promise of one. Spike counts from 54 neurons neurons were collected during the performance of both the serial order task and the variable reward task, and indices rank selectivity in the serial order task and of selectivity for the size of the anticipated reward in the variable reward task were computed. The rank index was $I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean

being taken across trials. Similarly, the reward index was $I_{\text{reward}} = \frac{(f_b - f_s)}{(f_b + f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. The indices I_{rank} and I_{reward} turned out to be positively correlated, but that the effect was smaller than expected, with $r = 0.49$. The correlation between the rank and reward indices was expected to be larger because, from previous research, it was known that (a) the expectation of reward increases over the course of a serial order trial and (b) neuronal activity in the SEF is affected by the expectation of reward. Behseta *et al.* speculated that the correlation between the two indices had been attenuated by noise arising from trial-to-trial variations in neural activity. Correction for attenuation gave a dramatically increased correlation, with the new estimate of correlation becoming .83. Results given by Behseta *et al.* showed that the new estimate may be considered much more reliable than the original $r = .49$. □

12.5 Multiple Linear Regression

The simple linear regression model (12.5) states that the response variable Y arises when a linear function of a single predictive variable x is subjectd to additive noise ϵ. The idea is easily extended to two or more predictive variables. Let us write the ith observation of the jth predictive variable as x_{ji}. Then, for p predictive variables the linear regression model becomes

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \tag{12.7}$$

where the ϵ_i's have the same assumptions as in (12.5).

Just as $y = \beta_0 + \beta_1 x_1$ describes a line, the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ describes a plane. When only a single variable x_1 is involved, the coefficient β_1 is the slope: $\beta_1 = \Delta y / \Delta x$. For example, if we increase x by $\Delta x = 2$ then we increase y by $\Delta y = 2\beta_1$. In the case of the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, if we increase x_1 by $\Delta x_1 = 2$ and ask what happens to y, the answer will depend on how we change x_2. However, if we hold x_2 fixed while we increase x_1 by $\Delta x_1 = 2$ then we will increase y by $\Delta y = 2\beta_1$. In general, when there are two variables, β_1 is interpreted as the change in y for a one-unit change in x_1 *when* x_2 *is held fixed*. Thus, linear regression is often used as a way

of assesssing what *might* happen if we *were* to hold one variable fixed while allowing a different variable to fluctuate.

Example 12.5.1 Neural correlates of developmental change in working memory from fMRI Many studies have documented the way visuo-spatial working memory (VSWM) changes during development. Kwon, Reiss, and Menon (2002; Kwon, H., Reiss, A.L., and Menon, V. (2002) Neural basis of protracted developmental changes in visuo-spatial working memory, *Proc. Nat. Acad. Science*, 99: 13336–13341.) used fMRI to examine neural correlates of these changes. These authors studied 34 children and young adults, ranging in age from 7 to 22. Each subject was given a VSWM task while being imaged. The task consisted of 12 alternating 36-second working memory (WM) and control epochs during which subjects viewed items on a screen. During both the WM and control versions of the task the subjects viewed the letter "O" once every 2 seconds at one of nine distinct locations on the screen. In the WM task the subjects responded when the current location was the same as it was when the symbol was presented two stimuli back. This required the subjects to engage their working memory. In the control condition the subjects responded when the "O" was in the center of the screen.

One of the y variables used in this study was the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex. They were interested in the relationship of this variable with age (x_1). However, it is possible that Y would increase due to better performance of the task, and that this would increase with age. Therefore, in principle, the authors wanted to "hold fixed" the performance of task while age varied. This is, of course, impossible. What they did instead was to introduce two measures of task performance: the subjects' accuracy in performing the task (x_2) and their mean reaction time (x_3). □

12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables.

To demonstrate multiple regression in action we consider a simple example.

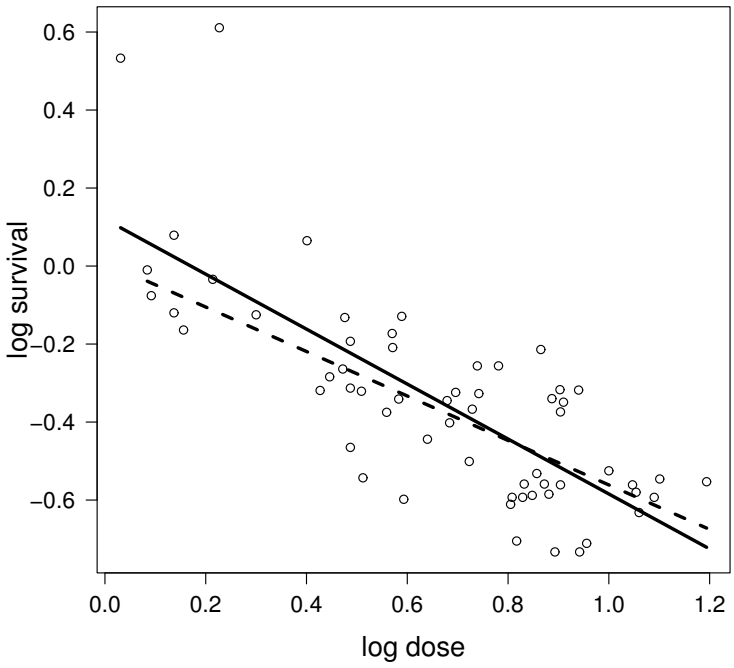


Figure 12.4: *Plot of log survival time (*$\log(w/1000)$ *where* w *was minutes survived) versus log dose (1.5 plus log milligrams) of sodium asenate in silkwork larvae; data from Bliss (1936). Lines are fits based on linear regression: solid line used the original data shown in plot; dashed line after removing the two high values of survival at low dose.*

Example 12.5.2 Toxicity as a function of dose and weight Many studies of toxicity, including neurotoxicity (Makris *et al.*, 2009), (Makris SL,

Raffaele K, Allen S, Bowers WJ, Hass U, Alleva E, et al. (2009) A Retrospective Performance Assessment of the Developmental Neurotoxicity Study in Support of OECD Test Guideline 426. *Environ. Health Perspect.* 117:17-25.) a drug or other agent is given to an animal and toxicity is examined as a function of dose and animal weight. A relatively early example was the study of sodium arsenate (arsenic) in silkworm larvae (Bliss, 1936) (Bliss, C.I. (1936) The size factor in the action of arsenic upon silkworm larvae, *J. Exp. Biol.* 13: 95-110.). We reanalyzed data reported there. The response variable (y) was $\log(w/1000)$ where w was minutes survived, and the two predictive variables were log weight, in log grams, and log dose, given in 1.5 plus log milligrams. A plot of survival versus dose is given in Figure 12.4. Because there were two potential outliers that might affect the slope of the line fitted to the plotted data we have provided in the plot the fitted regression lines with and without those two data pairs. The results we discuss were based on the complete set of data.

The linear regression of log survival on log dose gave the fitted line

$$\text{log survival} = .120(\pm.056) - .704(\pm.078)\text{log dose}$$

which says that survival decreased roughly .704(±.078) log 1000 minutes for every log milligram increase in dose. The regression was very highly significant ($p = 10^{-12}$), consistently with the obvious downward trend.

The linear regression of log survival on both log dose and log weight gave the fitted line

$$\text{log survival} = .120(\pm.056) - .704(\pm.078)\text{log dose} + 1.07(\pm.16)\text{log weight}.$$

In this case, including weight in the regression does not change very much the relationship between dose and survival: the slope is nearly the same in both cases. □

12.5.2 Response variation may be decomposed into signal and noise sums of squares.

As in simple linear regression we may decompose the total sum of squares into two pieces:

$$SST = SSR + SSE$$

and we define

$$R^2 = \frac{SSR}{SST}$$

as a measure of the strength of the linear relationship between y and the multiple explanatory variables. We may again use the sum of squares of the residuals to estimate the noise variation σ^2, here using

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2,$$

where $r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi})$ is the ith residual.

In multiple regression the F statistic is used to test the hypothesis of no linear relationship between the x variables and y, i.e., H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. Let us define the mean squared errors

$$MSR = \frac{SSR}{df_{regr}}$$

and

$$MSE = \frac{SSE}{df_{error}}$$

Here, with p explanatory variables, there are p degrees of freedom for regression and there are $n-1-p$ degrees of freedom for error (so that there are $n-1$ degrees of freedom in total). We then have

$$F = \frac{MSR}{MSE}.$$

In words, F is the ratio of the mean squared errors for regression and error, which are obtained by dividing the respective sums of squares by the appropriate degrees of freedom. We also have

$$s^2 = MSE = \frac{SSE}{df_{error}}.$$

If H_0 is true then

$$SSR = \sum(\hat{y}_i - \bar{y})^2 \sim \chi^2_{\nu_1}$$

where $\nu_1 = p$ is the degrees of freedom for regression,

$$SSE = \sum(y_i - \hat{y}_i)^2 \sim \chi^2_{\nu_2}$$

where $\nu_2 = n-1-p$ is the degrees of freedom for error, and it may be shown that these are independent. Therefore, the ratio

$$F = \frac{SSR/\nu_1}{SSE/\nu_2}$$

is a ratio of independent chi-squared random variables divided by their degrees of freedom, which, by definition (given in Chapter 5, has an F_{ν_1,ν_2} distribution. We provide a geometrical interpretation of the sum of squares decomposition below.

In simple linear regression, where there is only one explanatory variable, $\nu_1 = 1$ and F is equal to the square of the t-ratio. Because the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution, it follows that the t-test and the F-test of H_0: $\beta_1 = 0$ are identical. In multiple regression, hypotheses may also be tested about the individual coefficients, e.g., H_0 : $\beta_2 = 0$, using t tests.

Example 12.5.2 (continued) Returning to the toxicity data, here are the results for the regression of log survival on log dose:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12005    0.05665   2.119   0.0384 *
log dose    -0.70445    0.07773  -9.063 1.04e-12 ***

Residual standard error: 0.1736 on 58 degrees of freedom
Multiple R-Squared: 0.5861,     Adjusted R-squared: 0.579
F-statistic: 82.13 on 1 and 58 DF,  p-value: 1.045e-12
```

Here are the results for the regression of log survival on both log dose and log weight:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14002    0.05696  -2.458   0.0170 *
log dose    -0.73433    0.05831 -12.595  < 2e-16 ***
log weight   1.07247    0.15699   6.831 6.08e-09 ***
```

```
Residual standard error: 0.1298 on 57 degrees of freedom
Multiple R-Squared: 0.7724,     Adjusted R-squared: 0.7644
F-statistic: 96.74 on 2 and 57 DF,  p-value: < 2.2e-16
```

We see that the strength of the linear relationship increases greatly after weight is added to the regression: R^2 goes from .58 to .77 and the noise magnitude measured by s similarly decreases from .17 to .13.

We would interpret the t ratios as follows: given that log weight is in included in the regression model, there is very strong evidence ($t = -12.6$, $p << 10^{-5}$) that log survival has a decreasing linear trend with log dose; similarly, given that log dose is in the model, there is very strong evidence ($t = 6.8$, $p << 10^{-5}$) that survival has an increasing linearly trend with log weight. □

Example: Neural correlates of developmental change in working memory from fMRI (continued) Recall that in one of their analyses Kwon *et al.* defined Y to be the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex and considered its linear relationship with age (X_1), accuracy (X_2) and reaction time (X_3). They then performed multiple linear regression and found $R^2 = .53$ with $\beta_1 = .75(\pm.20)$, $p < .001$, $\beta_2 = -.21(\pm.19)$, $p = .28$, and $\beta_3 = -.15(\pm(.17)$, $p = .37$. They interpreted this as showing that the right PFC tends to become much more strongly activated in the VSWM task as the subjects' age increases, and that this is not due solely to improvement in performance of the task. □

The distributional results are based on normality of the errors. On the other hand, for sufficiently large samples they will be approximately correct even if the errors are non-normal. However, the independence assumption is again crucial. The standard errors and other distributional results generally may be trusted for reasonably large samples when the errors are independent, but they require correction otherwise. Generally, assumptions are examined using residual plots, as in simple linear regression.

12.5.3 Multiple regression may be formulated concisely using matrices.

Mathematical manipulations in multiple regression could get very complicated. A great simplification is to collect multiple equations together and write them as single equations in matrix form. We start by writing the n random variables Y_i as an $n \times 1$ random vector

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and then similarly write

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

The linear regression model may then be written in the form

$$Y = X\beta + \epsilon \tag{12.8}$$

where it is quickly checked that both left-hand side and right-hand side are $n \times 1$ vectors. The usual assumptions may also be stated in matrix form. For example, we have

$$\epsilon \sim N_n(0, \sigma^2 \cdot I_n)$$

which says that ϵ has a multivariate normal distribution of dimension n, with mean equal to the zero vector and variance matrix equal to σ^2 times the n-dimensional identity matrix, i.e.,

$$V(\epsilon) = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \ddots & 0 \\ 0 & & \cdots & 0 & 1 \end{pmatrix}.$$

Equation (12.8), together with the assumptions, is often called the *general linear model*. It accommodates not only multiple regression but also a large variety of Analysis of Variance models.[3]

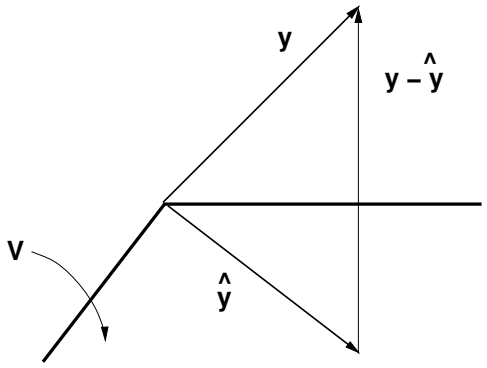


Figure 12.5: *Orthogonal projection of the vector y onto the vector subspace V resulting in the vector $\hat{y}$ in V.*

In matrix form the least-squares problem is to choose β so as to minimize $||y - X\beta||^2$, where $||w||$ is used to indicate the length of the vector w. The solution is found by solving the equations

$$X^T X\beta = X^T y \tag{12.9}$$

numerically (by numerically stable methods) and the solution may be written in the form[4]

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{12.10}$$

[3]Sometimes when someone refers to the general linear model they may allow the variance matrix to be different, or they may allow for non-normal errors.

[4]The equations are *not* solved merely by inverting the matrix $X^T X$; this can lead to grossly incorrect answers due to seemingly innocuous round-off error.

Formula (12.10) may be obtained by a simple geometrical argument. We begin by thinking of y as a vector in n-dimensional space and we consider the subspace V consisting of all linear combinations of the columns of X (we say that V is linear subspace *spanned* by the columns of X). The problem of least squares, then, is to minimize the distance between y and V. The solution to this minimization problem is the fitted vector $\hat{y} = X\hat{\beta}$. See Figure 12.5. This geometry also gives us the Pythagorean relationship

$$||y||^2 = ||\hat{y}||^2 + ||y - \hat{y}||^2$$

which is the basis for[5] the ANOVA decomposition $SST = SSR + SSE$.

Details: Euclidean geometry says that $\hat{y}$ must be obtained by orthogonal projection of y onto the subspace spanned by the columns of X and, as a result, the residual $y - \hat{y}$ must be orthogonal to the subspace spanned by the columns of X, which means that $y - \hat{y}$ must be orthogonal to $X\beta$ for every β. This, in turn, may be written in the following form: for all β,

$$< X\beta, y - \hat{y} >= 0 \tag{12.11}$$

where $< u, v >= u^T v$ is the inner product of u and v. Substituting $\hat{y} = X\hat{\beta}$ we have

$$< X\beta, y - X\hat{\beta} >= 0$$

for all β, and rewriting this we find that

$$\beta^T X^T y = \beta^T X^T X\hat{\beta}$$

for all β, which gives us Equation (12.9). Equation (12.9) is sometimes called the set of *normal equations* (presumably using "normal" in the sense of "orthogonal"; and plural because (12.9) is a vector equation and therefore a set of scalar equations). Because (12.11) holds for all β, it holds in particular for $\beta = \hat{\beta}$, i.e.,

$$< \hat{y}, y - \hat{y} >= 0$$

[5] Actually, a little more care is needed here because, in order to make this conform precisely to the usual ANOVA decomposition we must consider the intercept explicitly. But the idea remains the same.

as illustrated in Figure 12.5. □

By a little bit of algebra it may be shown that if W is a $p \times 1$ random vector with variance matrix $V(W) = \Sigma$ and A is a $k \times p$ matrix, then the variance matrix of AW is

$$V(AW) = A\Sigma A^T.$$

This is a generalization of the formula $\sigma^2_{aX} = a^2\sigma_X$ for a random variable X. Using this variance matrix formula we obtain

$$\begin{aligned} V(\hat{\beta}) &= ((X^TX)^{-1}X^T)\sigma^2 I_n((X^TX)^{-1}X^T)^T \\ &= \sigma^2 \cdot (X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2 \cdot (X^TX)^{-1}. \end{aligned}$$

This matrix would, for instance, give us the variance of $\hat{\beta}_k$ as its k-th diagonal element. To use it with data, however, we must substitute s for σ. We thus have the estimated variance matrix

$$\hat{V}(\hat{\beta}) = s^2 \cdot (X^TX)^{-1}$$

and the standard errors are given by

$$SE(\hat{\beta}_k) = \sqrt{s^2 \cdot (X^TX)^{-1}_{kk}}.$$

12.5.4 The linear regression model applies to polynomial regression and cosine regression.

In many data sets the relationship of y and x is mildly nonlinear, and a quadratic in x may offer better results than a line. Even though a quadratic is nonlinear, a neat trick allows us to fit quadratic regression via multiple linear regression. The trick is to set $x_1 = x$ and to define a new variable $x_2 = x^2$. Then, when y is regressed on both x_1 and x_2 this amounts to fitting a general quadratic of the form $y = a + bx + cx^2$, where now $a = \beta_0$, $b = \beta_1$ and $c = \beta_2$.

In quadratic regression generally, there are several possibilities. First, there may be evidence of a linear association between y and x (from the

simple linear regression), but the relationship appears nonlinear and there is also evidence of a linear association between y and both x and x^2 combined. This latter evidence would come from the combined regression output of (i) a statistically significant F-ratio and (ii) a significant t-ratio for the coefficient of x^2. This case is illustrated below. Note that it is possible for the coefficient of x in the combined regression to be non-significant. This should not necessarily be taken to mean that there is no linear component to the relationship: it is generally preferable to use the general form $y = a + bx + cx^2$, which requires the bx term and thus the x variable. Actually, it is possible for the coefficients of *both* x and x^2 to be non-significant while the F-ratio is significant; this occurs when the two variables are themselves so highly correlated that neither adds anything to the regression when the other is already used.

As a second possibility, there may be evidence of a linear association between y and x (from the simple linear regression), but there is no evidence of a quadratic relationship. The latter would be apparent from (i) an OK (not curved) residual plot in the simple linear regression and (ii) a non-significant t-ratio for the coefficient of x^2. The third possibility is that there may be no evidence of a relationship between y and *either* x by itself or X combined with x^2. This would be evident from an insignificant t-ratio in the simple linear regression and an insignificant F-ratio in the combined regression.

Let us now turn to an example.

Example ?? (continued) We previously examined spike train data recorded from a barrel cortex neuron in slice preparation, which was part of a study on the effects of seizure-induced neural activity. Figure 12.6 (repeated here) displayed the decreasing width of action potentials with increasing length of the interspike interval. Figure 12.7 shows a plot of many action potential widths against preceding interspike interval (ISI), where the data have been selected to include only ISIs of length less than 120 milliseconds. In the plot, the downward trend begins to level off hear 100 milliseconds, and a quadratic curve fitted by linear regression is able to capture the leveling off reasonably well within this range of ISI values. In this case the linear and quadratic regression coefficients were both highly significant ($p = 6 \times 10^{-6}$ and $p = .0017$, respectively, with the overall F-statistic giving $p = 8 \times 10^{-14}$) and $R^2 = .61$. □

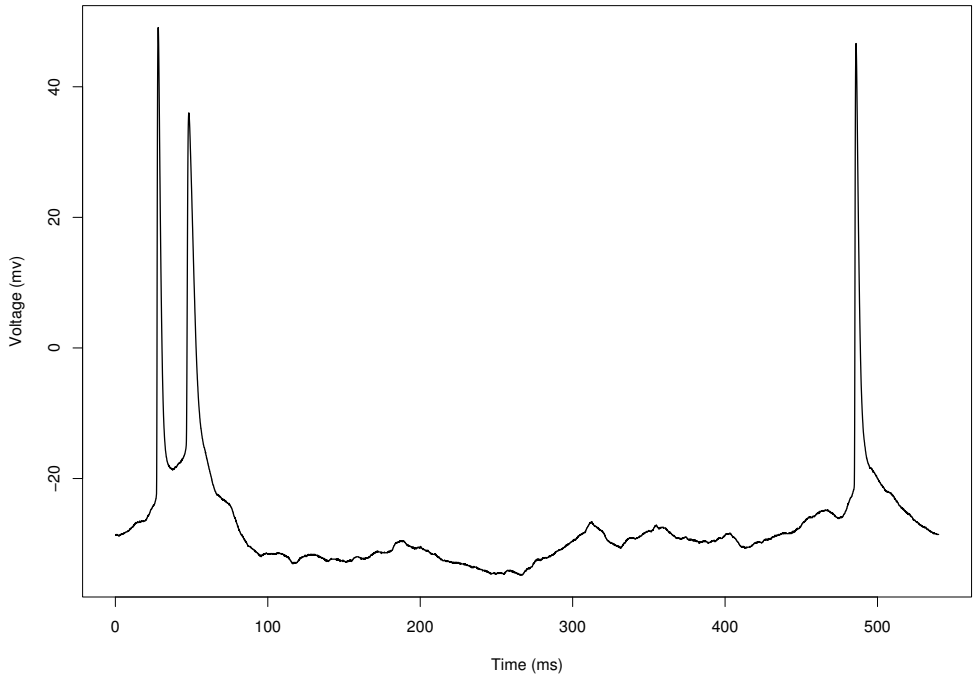


Figure 12.6: *When an action potential follows closely a previous action potential (with small ISI), the second action potential is broader than the first. When a long ISI intervenes, however, the second action potential is very similar to the first.*

It is sometimes useful to remember that x and x^2 are often highly correlated. In Example **??**, above, the correlation between ISI and its square was $r = .98$. An easy way to reduce correlation is to subtract the mean of the x variable before squaring, i.e., take $x_1 = x$ and $x_2 = (x - \bar{x})^2$. With x_1 and x_2 defined in this way for $x = ISI$ in the example above we obtained $r = -.08$.

A second nonlinear function that may be fitted with linear regression is the cosine.

Example 12.5.3 Directional Tuning in Motor Cortex In a well-known

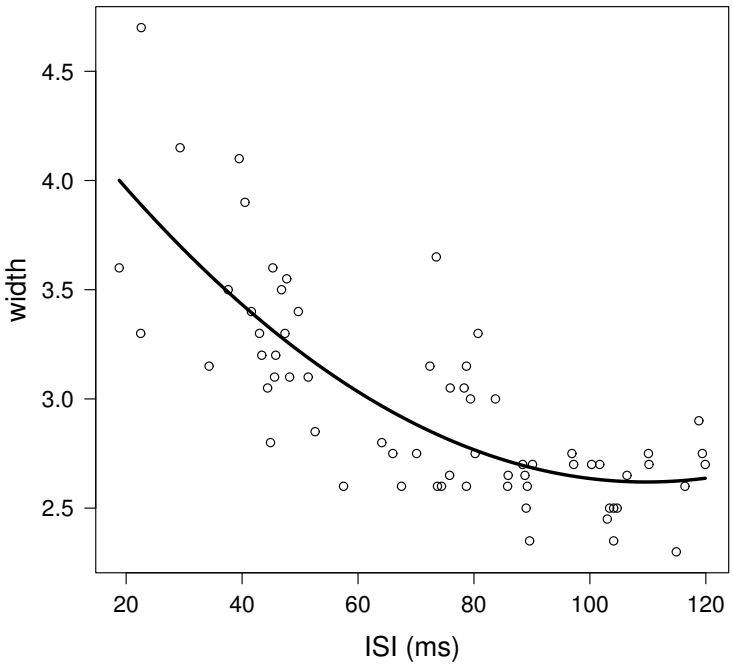


Figure 12.7: *Plot of action potential width against length of previous ISI, together with quadratic fitted by linear regression.*

set of experiments, Georgopolous, Schwartz and colleagues showed that motor cortex neurons are directionally "tuned." Figure 12.8 shows a set of PSTHs for a "center-out" reaching task: the monkey reached to one of eight points on a circular image, and this neuron was much more active for reaches in some directions than for others. The bottom part of Figure 12.8 shows a cosine function that has been fitted to the mean firing rate as a function of the angle around the circle, which indicates the direction of reach. For example (and as is also shown in the PSTH diagrams), reaches at angles near 180 degrees from the x-axis produced high firing rates, while those at angles close to 0 degrees (movement to the right) produced much lower firing rates. The angle at which the maximum firing rate occurs is called the "preferred direction" of the cell. It is obtained from the cosine function.

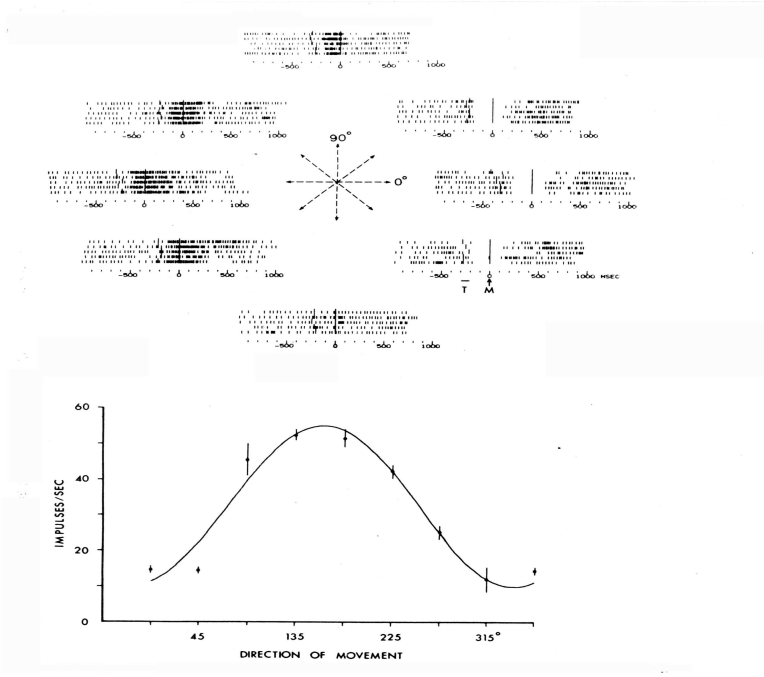


Figure 12.8: *Directional tuning of motor cortex neurons (from Georgopolous et al., 1982). Top displays the PSTH for each of eight reaching directions. Bottom displays corresponding mean firing rates.*

To fit a cosine to a set of spike counts, multiple linear regression is used. Let $v = (v_1, v_2)$ be the vector specifying the direction of movement and let $d = (d_1, d_2)$ be the preferred direction for the neuron. Both v and d are unit vectors. Assuming cosine tuning, the firing depends only on $\cos\theta$, where θ is the angle between v and d. We have

$$\cos\theta = v \cdot d = v_1 d_1 + v_2 d_2.$$

Letting $\mu(v)$ be the mean spike rate in a given interval of time when the movement is in direction v, if we let the minimal firing rate be B_{min} and the maximal firing rate be B_{max}, then cosine tuning may be written as the requirement that

$$\mu(v) = B_{min} + \frac{B_{max} - B_{min}}{2} + \frac{B_{max} - B_{min}}{2}\cos\theta.$$

(Recall that the minimal value of the cosine is -1, and its maximal value is 1.) If we now define $\beta_1 = \frac{B_{max}-B_{min}}{2}d_1$, $\beta_2 = \frac{B_{max}-B_{min}}{2}d_2$, and $\beta_0 = B_{min} + \frac{B_{max}-B_{min}}{2}$ we obtain the linear form

$$\mu(v) = \beta_0 + \beta_1 v_1 + \beta_2 v_2. \tag{12.12}$$

Taking $Y_i(v)$ to be the firing rate for the ith trial in direction v, we assume

$$Y_i(v) = \mu(v) + \epsilon_i(v). \tag{12.13}$$

Together, Equations (12.13) and (12.12) define a two-variable multiple linear regression model from which the tuning parameters may be obtained. □

12.5.5 Effects of correlated explanatory variables can not be interpreed separately.

When two explanatory variables that are substantially correlated the results must be interpreted with care.

Illustration: Quadratic regression To demonstrate the interpretive subtlety when explanatory variables are correlated we set $x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ and then defined

$$y_i = x_i + u_i$$

where $u_i \sim N(0, 4)$. Regressing the variable y on both x and x^2 we obtained the following results:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.36954    2.48590  -0.953    0.372
x            1.86379    1.03822   1.795    0.116
x2          -0.06750    0.09198  -0.734    0.487

Residual standard error: 2.114 on 7 degrees of freedom
Multiple R-Squared: 0.7724
F-statistic: 11.88 on 2 and 7 DF,  p-value: 0.005623
```

This regression might at first glance appear to provide no evidence that y was linearly related to either x nor x^2. However, regressing y on either x

or x^2 alone produces a highly significant linear regression. Furthermore, as displayed above, the F-statistic from the regression on both variables together is highly significant. These potentially puzzling results come from the high correlation of explanatory variables: the correlation between x and x^2 is $r = .975$. Keep in mind that the t-statistic for x^2 in the regression results displayed above reflects the contribution of x^2 *after* the variable x has been used to explain y and likewise the t-statistic for x reflects the contribution of x after the variable x^2 has been used to explain y. □

12.5.6 In multiple linear regression interaction effects are often important.

We saw earlier that it is possible to fit a quadratic in a variable x using linear regression by defining a new variable x^2 and then performing multiple linear regression on x and x^2 simultaneously. Now suppose we have variables x_1 and x_2. The general quadratic in these two variables would have the form

$$y = a + bx_1 + cx_2 + dx_1^2 + ex_1x_2 + fx_2^2.$$

Thus, we may again use multiple linear regression to fit a quadratic in these two variables if, in addition to defining new variables x_1^2 and x_2^2 we also define the new variable $x_1 \cdot x_2$. This latter variable is often called the *interaction* between x_1 and x_2. To see its effect consider the simpler equation

$$y = a + bx_1 + cx_2 + dx_1x_2. \tag{12.14}$$

Here, for instance, we have $\Delta y/\Delta x_1 = b + dx_2$. That is, the slope for the linear relationship between y and x_1 depends on the value of x_2 (and similarly the slope for x_2 depends on x_1). When $d = 0$ and we graph y vs. x_1 for two different values of x_2 we get two parallel lines, but when $d \neq 0$ the two lines are no longer parallel.

Interaction effects are especially important in analysis of variance models, which we discuss in Chapter **??**.

12.5.7 Regression models with many explanatory variables can often be simplified.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 13

Analysis of Variance

Many experiments examine the effects of multiple experimental conditions. When each measured response from a subject is a single-number, the data are usually analyzed with *analysis of variance (ANOVA)*. The name has a certain logic because, as we will see, the technique rests on a breakdown of sums of squares (assessing variation), but the null hypothesis typically takes the theoretical means to be equal among the experimental conditions, specifying no treatment effect, so that one might just as well think of the methodology as an investigation of means. The general ideas developed in Chapters 10 and 11 carry over to ANOVA. One additional, very important notion involves the structure of the experiment. This is spelled out in Section 13.1

13.1 One-Way and Two-Way ANOVA

ANOVA can take many forms, depending on the design of the experiment and the resulting structure of the data. We consider here only the two simplest

kinds of ANOVA and introduce them with a pair of examples.

Example 13.1.1 Stimulation and development of motor control Zelazo, Zelazo, and Kolb (1972, *Science*, 176:314-315) conducted a study to see whether stimulation of infants during the first eight weeks of life could make them walk earlier. The stimulation involved a simulation of walking in which a parent held the baby in a manner that would make it respond reflexively with walking-type leg movements. The data in Table 13.1 are ages in months at which 24 infants were judged to begin walking.[1] Each 1-week-old infant was assigned to one of four groups, namely, an experimental group (active-exercise) and three control groups (passive-exercise, no-exercise, 8-week control).[2]

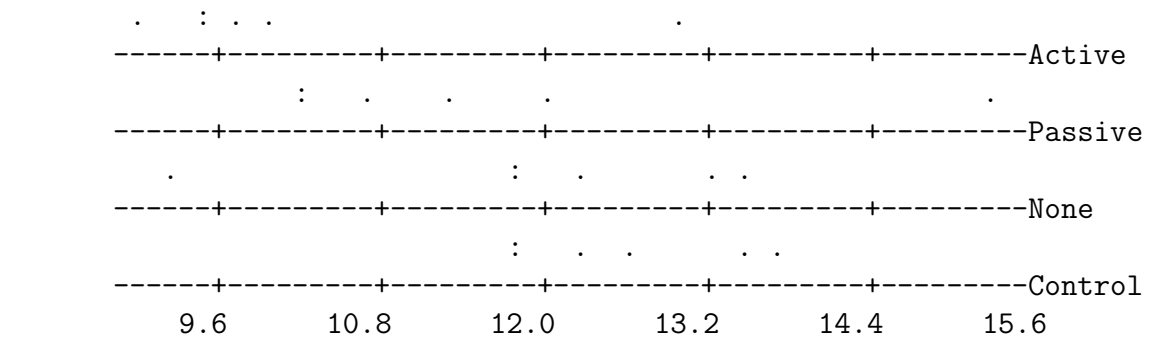
Active-exercise Group	Passive-exercise Group	No-exercise Group	8-week Control Group
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	12.35

Table 13.1: *Data from motor control experiment of Zelazo* et al. *(1972).*

The issue is whether the active-exercise group walked earlier than the controls. Here is a "dotplot" of the data:

[1]For pedagogical simplicity, we wanted the number of subjects per group to be equal. This is not required for ANOVA; it merely makes things a bit easier to discuss. In the original data there were only 5 subjects in the 8-week control group. We therefore added the 12.35 value to the 8-week control group.

[2]Infants in the active-exercise group received stimulation of the walking and placing reflexes during four 3-minute sessions that were held each day from the beginning of the second week until the end of the eighth week. The infants in the passive-exercise group received equal amounts of gross motor and social stimulation as those who received active-exercise, but unlike the active-exercise group, these infants had neither the walking nor placing reflex exercised. Infants in the no-exercise group did not receive any special training, but were tested along with the active-exercise and passive-exercise subjects. The 8-week control group was tested only when they were 8 weeks of age; this group served as a control for the possible helpful effects of repeated examination.



It may be seen that the Active-group infants had somewhat earlier reported ages of walking than those in the three control groups. However, there is quite a bit of variability, with one of the 5 infants in the Active group being relatively late (13.25) and one in the None group being quite early (9.0). Thus, it's hard to tell whether there is a consistent pattern. This is a case in which a formal procedure is helpful. □

Notice the layout of the data in the example above: it makes sense to display them in columns, with each column identified with a different treatment. The next example is different.

Example 13.1.2 Finger tapping in response to stimulants Scott and Chen (1944; *J. Pharmacol. Exptl. Therap.*, 82: 89-97) conducted an experiment on finger tapping in response to orally-administered stimulants. Four subjects were each given three different treatments and then their finger-tapping rates were analyzed. The treatments were caffeine (Ca), 1-ethyltheobromine (Th); this is the stimulant in chocolate, and it is very similar to caffeine, and a placebo (Pl). The tapping rates (rate minus 440, with "rate" not defined but possibly taps per minute) are shown in Table 13.2.

In this case, once again, we would be interested in comparing the three treatments. The mean tapping rates for Pl, Th, and Ca are 22, 39, and 41. Is this evidence that theobromine and caffeine lead to increased tapping rates? □

	Subject No.			
DRUG	1	2	3	4
Pl	11	56	15	6
Th	26	83	34	13
Ca	20	71	41	32

Table 13.2: *Data from finger tapping experiment of Scott and Chen (1944).*

An important distinction between the two experiments above is that in the finger tapping experiment in Example 13.1.2 each subject received *all* of the treatments. Thus, there were only four subjects in the experiment, not twelve, which led to 12 data values. In the motor control experiment of Example 13.1.1, each subject received only one treatment. The two situations require related but different statistical methods. Table 13.1 is sometimes called a *one-way* table and is treated by *one-way ANOVA* while Table 13.2 is called a *two-way* table and is treated by *two-way ANOVA*. The data in Table 13.2 are also often called *blocked,* with the subjects being the blocks. (The "block" terminology comes from the origin of ANOVA in agricultural field trials, where it referred to a block of land in a field.)

13.1.1 ANOVA is based on a linear model.

The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where Y_{ij} is the j-th observation in the i-th group, $\mu + \alpha_i$ is the mean for the i-th group and ϵ_{ij} is the error for the j-th observation in the i-th group (the discrepancy between Y_{ij} and $\mu + \alpha_i$). Here, μ is the overall mean (the "grand mean") and α_i is the increment added to that overall mean in obtaining the mean for the i-th group. Here we assume there are a total of I groups, so that $i = 1, 2, \ldots, I$.

The additional assumptions, beyond the additive-error assumption inherent in the model formulations above, are that the errors are independent, have the zero expectation, have equal variances ("homogeneous" variances),

and are normally distributed. Note that these are the same assumptions as in regression. As a result, residual analysis may be used in very much the same way as in regression. Indeed, mathematically, analysis of variance may be considered a special case of linear regression. We return to this shortly.

The purpose of this model is to provide a basis for formal comparison of the group means $\mu + \alpha_i$. That is, we ask whether there is evidence that the means are different and, if so, we can estimate how different they are. Formally, we want to test the null hypothesis that the groups means are equal:

$$\mu + \alpha_1 = \mu + \alpha_2 = \cdots = \mu + \alpha_I.$$

The usual way the hypothesis is stated is as follows:

$$\mathrm{H}_0 : \alpha_i = 0$$

for all i, which implies that the group means are equal. It also satisfies the condition that the grand mean μ remains the expectation of Y_{ij} under H_0.

13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis.

Recall that in defining R^2 we started with the total sum of squares $SST = \sum(y_i - \bar{y})^2$, which represents the total amount of variation in y, and then considered the residual or "error" sum of squares $SSE = \sum(y_i - \hat{y}_i)^2$. We take R^2 to be 1 minus the ratio of the latter to the former. In Analysis of Variance we decompose the variability in the data as represented by the total sum of squares again into two pieces, a sum of squares due to the different treatments, which we here call "groups," and a sum of squares due to error (or residuals). The idea is to compute a measure of the *average* amount of variability due to the groups, and an *average* amount of variability due to error, then compare these. Under the null hypothesis that the group means are equal, there should be no systematic variability due to groups, so that the variability we see in our "average variability due to groups" is simply the result of background variability in the measurements themselves, that is, the

error variability. In other words, the average variability due to groups should be about the same size as the average variability due to error.

The F-ratio compares these two masures of average variability and when the ratio is much larger than 1 there is evidence against H_0, in favor of there being differences among the groups. Again, the p-value assesses whether the value of F is sufficiently large to be considered unusually rare. Formally, we take the total sum of squares to be

$$SST = \sum (y_{ij} - \bar{y}_{..})^2$$

where the subscript .. indicates that the mean is being taken over all the values of y (averaging across both rows and columns) In the infant exercise example we average across all 24 values. We also define the error (residual) sum of squares to be

$$SSE = \sum (y_{ij} - \bar{y}_{i.})^2$$

where the subscript $i.$ indicates that the mean is being taken *within* the i-th group. In the infant exercise example there would be 4 means $\bar{y}_{i.}$ for $i = 1, 2, 3, 4$ and each would be an average across all 6 values in the appropriate column. The group sum of squares is then

$$SS_{group} = SST - SSE.$$

We next obtain averages of the group and error sums of squares by dividing by their respective degrees of freedom, $df_{group} = I - 1$ and $df_{error} = n - I$, where n is the total number of observations. These averages are called *mean squares*:

$$\begin{aligned} MS_{group} &= SS_{group}/df_{group} \\ MSE &= SSE/df_{error} \end{aligned}$$

Finally, we obtain from these the F-ratio

$$F = MS_{group}/MSE.$$

Under the null hypothesis, this ratio follows an F distribution, which is used to compute the p-value.

Note that in a certain sense "analysis of variance" is a misnomer. We are really analyzing several means, and determining whether there's evidence

that they are different. However, the basic tool for doing so is a comparison of sums of squares, that is, a comparison of different sources of variability, which explains the terminology.

Example 13.1.1 (continued) Here is the basic ANOVA breakdown.

```
ANALYSIS OF VARIANCE
SOURCE     DF        SS        MS        F        p
Groups      3     15.74      5.25     2.40    0.098
Error      20     43.69      2.18
Total      23     59.43
```

and this should be examined together with the individual group means:

```
GROUP        N      MEAN     STDEV
  1          6    10.125     1.447
  2          6    11.375     1.896
  3          6    11.708     1.520
  4          6    12.350     0.860

 POOLED STDEV =    1.478
```

Here, $F = 2.40$ on 3 and 20 d.f. with $p = .098$ indicating no evidence of any differences among the means. Although from the sample means (and the individual CI's) it is appears, from this analysis, that the first group has a mean somewhat smaller than that for the control groups there is enough variability in the data that this is consistent with chance fluctuation. □

13.1.3 When there are only two groups, the ANOVA F-test reduces to a t-test.

In the special case of only two groups with two means μ_1 and μ_2, the null hypothesis $H_0 : \mu_1 = \mu_2$ may be tested with a t-test. This turns out to be

equivalent to the ANOVA F test and, in fact, the square of the t-statistic is equal to the F-statistic.

Example 13.1.1 (continued) The means displayed previously were 10.1, 11.4, 11.7, 12.35. From the pooled estimate of σ, $s = 1.48$ we get the standard error of each mean $SE = s/\sqrt{6} = .60$. Comparing the active exercise group mean with the eight-week control we have a difference of $12.35 - 10.1 = 2.25$. Using the pooled estimate s, this difference has a standard error of $SE(\bar{X}_4 - \bar{X}_1) = s\sqrt{\frac{1}{6} + \frac{1}{6}} = .853$ and the ratio is $2.25/.853 = 2.6$ and with 20 degrees of freedom (corresponding to the ANOVA MSE because $s = \sqrt{MSE}$), we get $p = .017$. An altnative analysis compares the active exercise group with the other three groups, all of which could be considered controls. This is apparently the way the investigators thought of them. In this case, we could pool them together and thereby end up with two groups: active versus controls, the latter now having 18 observations. We may then use the "two-sample t" analysis:

```
TWOSAMPLE T FOR active VS controls
           N      MEAN     STDEV   SE MEAN
active     6     10.12      1.45      0.59
controls  18     11.81      1.46      0.34

95 PCT CI FOR MU active - MU controls: ( -3.11,  -0.26)

TTEST MU active = MU controls (VS NE): T= -2.46  P=0.022  DF=  22

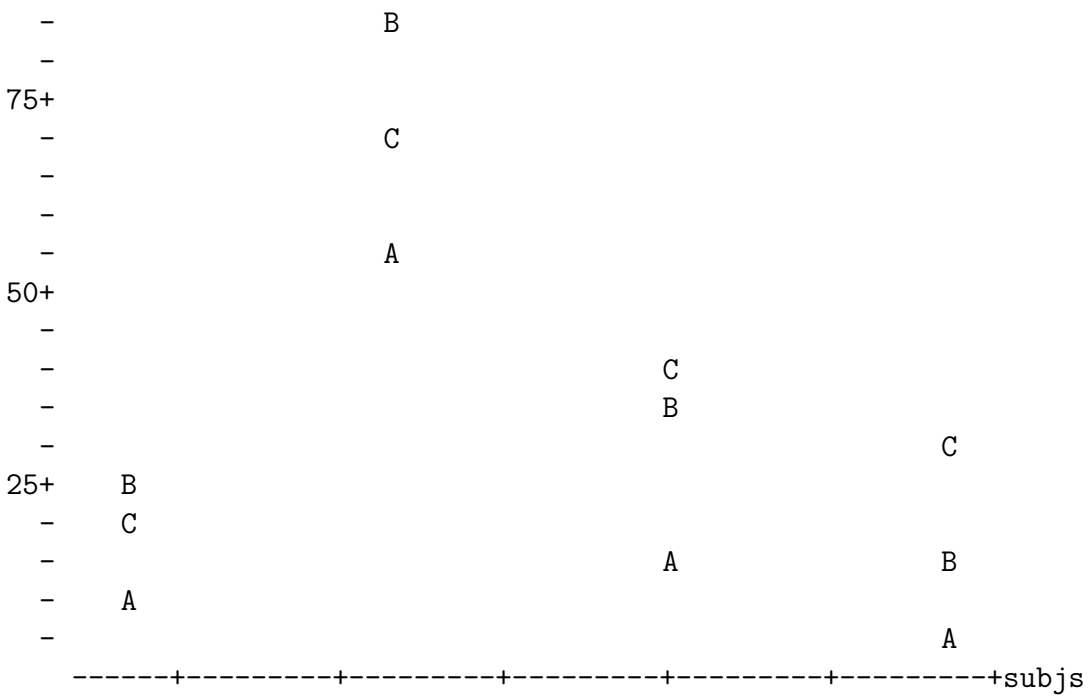
POOLED STDEV =       1.46
```

Here, we obtain (i) a test of the null hypothesis that the means for these two groups are equal, which we may write as $H_0 : \mu_{active} - \mu_{controls} = 0$, and (ii) a 95% CI for the difference between the means $\mu_{active} - \mu_{controls}$. Based on the results above, we find the observed difference to be statistically significant with $t = 2.46$ on 22 degrees of freedom leading to $p = .022$, and our 95% CI for the decrease in mean age of walking for the active group compared with controls to be (.26,3.11) months.

The conclusions from this analysis are quite different from that in the previous subsection, based on the F-test. We return to this example later. □

13.1.4 Two-way ANOVA assesses the effects of one factor while adjusting for the other factor.

We use the finger tapping example to illustrate two-way ANOVA. Once again, the first step is to plot the data. A very useful way to do this is to label each point according to treatment, and to use the x-axis to represent subjects, as shown below.



Example 13.1.2 (continued) In this plot, the tapping rates for the four subjects are plotted separately, and the three experimental treatments are identified as `A, B,` or `C`, with `A` corresponding to placebo, `B` to theobromine, and `C` to caffeine. Here, we can see clearly that the subjects have very different

tapping rates, but for all four of them the placebo rate is noticeably lower than that obtained with theobromine or caffeine, while neither thoebromine nor caffeine consistently raises the rate faster than that for the other. □

The two-way model is $Y_{ij} = \mu+\alpha_i+\beta_j+\epsilon_{ij}$, where Y_{ij} is the observation for the i-th treatment on the j-th "block," often, as in the tapping example, the j-th subject, $\mu+\alpha_i+\beta_j$ is its mean, and ϵ_{ij} is the error for the i-th treatment and j-th block. Here, α_i is the increment added to the overall mean μ in obtaining the mean for the i-th treatment while β_j is the increment added to overall mean in obtaining the mean for the j-th block.

For both one-way and two-way models the null hypothesis that is tested with ANOVA is $H_0 : \alpha_i = 0$ for all i. In the two-way case it is possible to formulate the hypothesis that all the β_j's are zero, as well. This is not usually a main interest in blocked experiments, but statistics packages print out F-statistics and p-values for both hypotheses, so it's important to keep them straight.

Example 13.1.2 (continued) In the ANOVA for the finger tapping data there are two "factors" to be considered, treatments and subjects.

Source	DF	SS	MS	F	P
subjs	3	5478.0	1826.0	33.00	0.000
treats	2	872.0	436.0	7.88	0.021
Error	6	332.0	55.3		
Total	11	6682.0			

Here, $F = 7.88$ on 2 and 6 d.f. with $p = .021$ indicates some evidence that the treatment means are different. Notice that there is also an F-ratio for subjects, which in fact is much larger and has a considerably smaller p-value: in this example, there is a very substantial difference among the subjects. In particular, the second subject has a much higher tapping rate than the others. The variability among subjects might be important to the conclusions one would wish to draw.

We may say something about the means, as well. For the three groups the mean tapping rates are, respectively, 22, 39, and 41. Standard errors are

found by plugging in an estimate s of σ and again applying $SE = s/\sqrt{n}$. We have $s = \sqrt{MS_{Error}} = \sqrt{55.3} = 7.44$. Since there are 4 observations per treatment group, we use $n = 4$ and get 22 ± 3.7, 39 ± 3.7 and 41 ± 3.7. Clearly, the caffeine and theobromine groups have tapping rates substantially above that for the placebo group. □

13.1.5 Additional analyses, involving multiple comparisons, may require adjustments to p-values.

Before talking about additional ANOVA analyses, let us consider a hypothetical setting analogous that of the blindsight patient P.S., but where, in choosing between the "burning" and "non-burning" house pictures, the subject was actually guessing in a completely random way with equal probability of picking each of the two pictures. In this case we would consider the number of times the subject chose the non-burning house out of an experiment involving 17 trials to be a Binomial $B(17, .5)$ random variable. Now, imagine we repeated this experiment for 500 subjects, all of whom were guessing in the same random way. Most of these subjects would pick the non-burning house a number of times not too far different from the expectation $np = 8.5$. However, it would be very likely that at least one of them would respond by choosing the non-burning house at least 14 times. Specifically, if $X_i \sim B(17, .5)$ for $i = 1, \ldots, 500$, independently, then

$$P(X_i \geq 14 \text{ for at least one } i) = 1 - \prod P(X_i \leq 13) = 1 - .9936^{500} = .96.$$

Now, imagine picking such a subject and reporting his or her findings. In other words, imagine saying "We have a statistically significant result: there is solid evidence that this subject has blindsight." That would clearly be preposterous; we would have cheated and it would be bad science. Methods of *multiple comparisons* are supposed to guard against similar mistakes, where the situation is less extreme and somewhat more subtle but nonetheless similar in principle. These methods work by using additional probability calculations like the one above, where we obtained the probability that at least one subject would choose the non-burning house at least 14 times.

Let us put the issue in the context of ANOVA by going back to the 1-way ANOVA example on development of ability to walk. One of the ways

we looked at the data was to compare the active exercise group with the eight-week control group, testing the null hypothesis $H_0 : \mu_1 = \mu_4$ with a t-ratio. If we consider the t-ratio to be a random variable T, the significance test has the property that, if there were actually no effect, so that H_0 were true, then $P(|T| > t_\nu(.975))) = .05$, where $t_\nu(.975)$ is the .975 quantile of the t distribution on ν degrees of freedom. This is a precise way of saying that results as surprising as those giving $p = .05$ would occur 5% of the time. Now suppose we also tested $H_0 : \mu_1 = \mu_3$ under the condition that both of these null hypotheses were true, i.e., $\mu_1 = \mu_3 = \mu_4$. Let us label the first t statistic T_1 and the second T_2. Now, if we consider either T_1 or T_2 separately, we would continue to have $P(|T_j| > t_\nu(.975))) = .05$, for $j = 1, 2$. However, if we conducted *both* tests we would be examining *both* results. In practice, it would be an interesting finding if either one of them, or both, were significant. But, continuing to assume $\mu_1 = \mu_3 = \mu_4$, the chance of either T_1 *or* T_2 *or* both being significant is considerably larger than the probability that T_1 is significant. In other words, it is considerably larger than .05. And in that example, because there were 4 treatment groups in all, there were additional opportunities for testing hypotheses. It thus becomes clear that in using t-tests (or related procedures) there will have to be some adjustment of p-values.

Such considerations have led to many, many methods of multiple-comparisons procedures. Perhaps the simplest is the *Bonferroni correction*, in which one counts the total number of comparisons that might have been contemplated and multiplies the p-value by this number. For example, if 3 comparisons could have been made and the initial p-value is .01, then the Bonferroni-adjusted p-value becomes .03. This is justified by the *Bonferroni inequality*, which says that for sets $A_1, A_2, \ldots, A_n$ we have

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n). \tag{13.1}$$

Details: Recall that for two events A and B we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{13.2}$$

This implies

$$P(A \cup B) \leq P(A) + P(B). \tag{13.3}$$

Now consider three events C, D, E. Applying the formula (13.2) with $A = C \cup D$ and $B = E$ we get

$$P(C \cup D \cup E) = P(C \cup D) + P(E) - P((C \cup D) \cap E)$$

and applying (13.2) to the right-hand side with $A = C$ and $B = D$ we obtain

$$P(C \cup D \cup E) = P(C) + P(D) - P(C \cap D) + P(E) - P((C \cup D) \cap E)$$

which gives

$$P(C \cup D \cup E) \leq P(C) + P(D) + P(E). \tag{13.4}$$

The inequalities (13.3) and (13.4) are examples of the Bonferroni inequality. We can continue the same argument to obtain (13.1). □

If we apply inequality (13.1) to the case in which A_i is the event that a particular significance test would yield, say, $p = c$ then the probability of getting at least one significant result is no more than nc. Thus, if we want the overall significance level to be, say, less than .05 we would have to take $c = .05/n$. In the case of n two-sided t-tests, this means we would use the cutoff value $t_\nu(1 - .025/n)$ and declare a particular test significant if $|T_{obs}| > t_\nu(1 - .025/n)$.

The Bonferroni adjustment is easy to apply, but it is usually quite conservative in the sense that it makes it difficult to find statistically significant tests. This has led to development of many other procedures.

In many cases a very workable way to procede, however, is to begin by relying on the F test. If one obtains a significant F-statistic there is evidence for a difference among the means, and it therefore makes sense to go ahead and examine whichever means happen to look interesting, without worrying much about the process of selecting them. In other words, a widely-advocated method (sometimes called the *protected least-significant difference*) is to require a significant F statistic and then to report results from the many t tests, or any of them that seem to be of interest.

Example 13.1.1 (continued) Where does all this leave us in this example? We may summarise by saying that there is some evidence, but not

strong evidence, that the active group mean age of walking is a bit younger than that for the control groups. The marginal nature of this evidence becomes clear when we ignore the special feature that the latter three groups are all controls and look for differences among all four groups: we find no evidence for this, according to the F-test. Given that it may be difficult to determine exactly when a given child walks, and it is not clear that the parents made this determination in the absence of knowledge about what to expect based on the experimental hypothesis, some skepticism would seem appropriate.[3] □

13.2 ANOVA as Regression

13.2.1 The general linear model includes both regression and ANOVA models.

Linear regression may be used to solve problems of Analysis of Variance. These concern the comparison of means among several groups. Let us consider two simple examples. Suppose X is the $n \times 1$ vector of 1s

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

We then compute $X^T X = n$ and $X^T Y = \sum y_i$ and find

$$(X^T X)^{-1} X^T y = \bar{y}.$$

Therefore, the sample mean may be found by applying regression with this very special version of the X matrix.

[3]On the other hand, the paper by Zelazo *et al.* presented an additional measure where the results were more striking. On this subject, see Adolph (2002). (Adolph, K.E (2002), Babies steps make giant strides toward a science of development *Infant Behavior and Development*, 25: 86–90.)

Next, consider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$ and define

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{2m} \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}$$

where the first column contains m rows of 1s followed by m rows of 0s and the second column contains m rows of 0s followed by m rows of 1s. We compute

$$X^T X = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

$$X^T y = \begin{pmatrix} \sum y_{1i} \\ \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the first and second group means.

In a similar manner we may use linear regression to compute several group means, and the Analysis of Variance table from this regression becomes the same as the Analysis of Variance table used in, well, Analysis of Variance.

Before leaving the subject of indicator variables, let us make the further point that there are typically many reasonable choices of the way to code

the columns of the X matrix. For example, if we reconsider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$, we could take

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}$$

In this case, X is no longer made up of indicator variables, but its columns span the same space as that spanned by the indicator variables given above. That is, a vector v is a linear combination of the columns of X using this formulation if and only if it is a linear combination of the columns of X using the indicator variables (though the coefficients of the linear combinations will be different in the two cases). We now compute

$$X^T X = \begin{pmatrix} 2m & 0 \\ 0 & 2m \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} \sum y_{1i} + \sum y_{2i} \\ \sum y_{1i} - \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} \bar{y} \\ (\bar{y}_1 - \bar{y}_2)/2 \end{pmatrix}$$

where $\bar{y}$ is the overall mean. The second component $(\bar{y}_1 - \bar{y}_2)/2$ is often called a *contrast*, because it is "contrasting" the means of the groups. Generally speaking, a contrast vector (leading to a contrast estimate) is one whose components add to zero. In ANOVA settings, where there are multiple groups, it is often of interest to define an X matrix made up of contrast vectors, together with the vector whose components are all 1.[4]

[4]It is also convenient to require the vectors to be orthogonal to one another, in which ase they are called *orthogonal contrasts*. For orthogonal contrasts, each estimate is independent of the others. This is a topic discussed in many books on regression analysis and/or experimental design.

13.2.2 In multi-way ANOVA, interactions are often of interest.

Interaction effects are very important in many ANOVA settings. Here, we consider interactions in regression, and comment at the end how they occur, as special cases, in ANOVA.

Now suppose Equation (12.14) holds and we let x_2 take on values of either 1 or 0 (as in Section 13.2.1). We find that the slope of the linear relationship between y and x_1 is $b + d$ when $x_2 = 1$ and it is b when $x_2 = 0$. Thus, we have a way of describing two lines in one equation. Let us now put this sort of reasoning to work in a statistical setting.

Example 2.1.1 (continued) In the experiment on saccadic reaction time, suppose Y is saccadic reaction time, X_1 is the distance from eye fixation to target, and X_2 is 0 if the fixation point is straight ahead of the subject and 1 if it is 30 degrees to the right. Assuming that the head and trunk are oriented straight ahead, X_2 indicates whether or not the angle of the eyes is 30 degrees to the right, as opposed to straight ahead.

If have data in the two eye-orientation positions for several different saccadic distances we may now perform a linear regression of Y on X_1 and X_2. Let us consider three models. First, if we fit

$$Y_i = \beta_0 + \beta_2 x_2 + \epsilon_i$$

we note that

$$Y_i = \beta_0 + \epsilon_i$$

for the eyes-straight condition while

$$Y_i = \beta_0 + \beta_2 + \epsilon_i$$

for the eyes-right condition. Thus, β_2 corresponds to the increase in reaction time, on average, when the eyes are to the right. Carrying out calculations like those in Section 13.2.1 we will obtain the mean for the eyes-straight condition as $\hat{\beta}_0$ and that for the eyes-right condition as $\hat{\beta}_0 + \hat{\beta}_2$. The quantity $\hat{\beta}_2$, together with its standard error (and t-ratio) could be used to make inferences. This would be of interest, but it is potentially problematic if the

saccadic distances for the two conditions are not properly balanced to be comparable.

A second model would adjust for saccadic distance. If we fit

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

we find that

$$Y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$$

for the eyes-straight condition while

$$Y_i = \beta_0 + \beta_2 + \beta_1 x_1 + \epsilon_i$$

for the eyes-right condition. Here, we would be saying that the saccadic reaction time varies with target distance in the same linear way for both conditions, but those for the eyes-right condition are slower on average by a fixed amount β_2. We could again use $\hat{\beta}_2$ together with its standard error to make inferences.

However, what if reaction times do not vary in the same linear way for both conditions? The third model allows for this possibility. So let us imagine fitting

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$$

where $x_3 = x_1 x_2$. Now the reaction time varies with target distance differently in the two condition. For the eyes-straight condition the slope is β_1 while in the eyes-right condition it is $\beta_1 + \beta_3$. That is, when the patient is in the eyes-right condition he or she would move more slowly to the target, on average, by an amount β_3 per unit of distance. In addition, the patient would also be slower by the fixed amount β_2 on average. Behrmann *et al.* found that hemi-spatial neglect patients have substantially slower saccades in the eyes-right condition (and the effect was highly statistically significant). The increment here labelled as β_2 was not, however, significantly different from zero. □

It is also possible to consider interactions when both X_1 and X_2 are indicator variables. This becomes a two-way ANOVA model, and interactions often have interesting interpretations.

Example 2.1.1 (continued) Suppose X_2 is the indicator for eyes right, as before, but X_1 is an indicator for whether or not a distractor is present.

Then the absence of an interaction might say that the patient had slower reactions for eyes right, and also slower with the distractor present, but that the difference in reaction times between eyes right and eyes straight remained the same regardless of whether the distractor was present. As an alternative, one plausible form of an interaction would be if the reaction times between eyes right and eyes straight became larger when the distractor was present than when it was absent. □

Interactions are widely analyzed in ANOVA models. Our point here is simply that they may be understood as special cases of interactions in multiple linear regression. The case in which one variable is quantitative and one or more variables are indicator variables is usually called *analysis of covariance* or ANCOVA. This is because the full regression analysis effectively replaces the ANOVA (which would ignore the quantitative variable); in so doing, it "adjusts" the ANOVA for the "covariation" in the quantitative variable.

13.3 Nonparametric Methods

13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks.

The ANOVA methods we have so far discussed are *parametric* statistical methods, meaning methods that apply to probability models characterized by a small number of parameters. An alternative strategy is to apply methods that are designed to work reasonably well both when familiar parametric models are valid and when they are not. These are called *nonparametric* methods, and we have already presented permutation tests and bootstrap tests for the case of two independent samples. An alternative is to use methods based on the *ranks* of the data. That is, to replace each data value by its rank within the whole data set. Rank-based procedures remove the assumption of a distributional form. Please note, however, that all these procedures continue to make the big assumptions, additivity and independence of the errors.

Example 13.3.1 Alcohol metabolism among men and women Women seem to have a lower tolerance for alcohol than men, and are more prone to develop alcohol-related diseases. When men and women of the same size and history of drinking consume equal amounts of alcohol, the alcohol in the bloodstream of the women tends to be higher. In research by Frezza, et al. (1990, *New England Journal of Medicine*, 322: 95-99), the "first-pass" metabolism of alcohol in the stomach was studied. The data shown below come from 18 women and 14 men who volunteered to be studied. Each subject was given two doses of .3 grams ethanol per kilogram of body weight, one orally and one intraveneously on two different days. The difference in concentrations of alcohol in the blood (at some fixed time after administration), between the intraveneous dose and the oral dose, provides a measure of first-pass metabolism in the blood; this defines the response variable in the table, with units in mmols per liter per hour. If first-pass metabolism were more effective in men than women, the difference in levels following intraveneous and oral administration would tend to be higher among men.

Before we move on to the analysis, there is a more basic concern in this example: we must wonder about the extent to which the volunteers represent the population as a whole, and whether the particular men and women in the study might for some reason self-select in a manner that was related to their alcohol metabolism. We will ignore this possibility here, and instead confront the more technical problem that the values for the alcoholic men have a very wide range compared to the other values, and the assumptions of ANOVA may not be satisfied.

Alcoholic Women	Non-alcoholic Women	Alcoholic Men	Non-alcoholic Men
0.6	0.4	1.5	0.3
0.6	0.1	1.9	2.5
1.5	0.2	2.7	2.7
	0.3	3.0	3.0
	0.3	3.7	4.0
	0.4		4.5
	1.0		6.1
	1.1		9.5
	1.2		12.3
	1.3		
	1.6		
	1.8		
	2.0		
	2.5		
	2.9		

The alcohol metabolism data are printed out in the following table, with each *rank* listed at the end. The rank goes from 1 up to 32, with the smallest value getting the rank 1 and the largest value getting the rank 32. (Ranks ending in .5 represent "ties," i.e., cases in which some data value appears twice.) The women in the study have a 1 in the `females` column.

case	difference	female	rank
1	0.6	1	8.5
2	0.6	1	8.5
3	1.5	1	14.5
4	0.4	1	6.5
5	0.1	1	1.0
6	0.2	1	2.0
7	0.3	1	4.0
8	0.3	1	4.0
9	0.4	1	6.5
10	1.0	1	10.0
11	1.1	1	11.0

12	1.2	1	12.0
13	1.3	1	13.0
14	1.6	1	16.0
15	1.8	1	17.0
16	2.0	1	19.0
17	2.5	1	20.5
18	2.9	1	24.0
19	1.5	0	14.5
20	1.9	0	18.0
21	2.7	0	22.5
22	3.0	0	25.5
23	3.7	0	27.0
24	0.3	0	4.0
25	2.5	0	20.5
26	2.7	0	22.5
27	3.0	0	25.5
28	4.0	0	28.0
29	4.5	0	29.0
30	6.1	0	30.0
31	9.5	0	31.0
32	12.3	0	32.0

□

Rank-sum methods compare the ranks of the two groups. That is, if one group has values of its ranks that are sufficiently much larger than those of the other group, there will be evidence that the means of the two groups are different. More specifically, we may find the sum of the ranks from one of the groups and see whether it is either much larger or much smaller than would be expected if, in fact, the two groups followed the same distribution. Based on the null hypothesis that the probability distributions for the two groups are the same, we can get a p-value. This is the *rank-sum test*. The test statistic W is simply the sum of the ranks from one of the two groups. It is sometimes called the Wilcoxin rank-sum test, and it is also often called the Mann-Whitney test.

To be specific about the procedure, suppose the alcohol metabolism data consisted only of the following four observations:

```
case  difference    females
   1    0.6         1
  18    2.9         1
  19    1.5         0
  32   12.3         0
```

Here, we would rank the data as 1, 3, 2, 4 (0.6 is the smallest, 2.9 is the third smallest, 1.5 is the second smallest, and 12.3 is the fourth smallest). Then we would add up the values of the ranks for the females to get the statistic $W = 1 + 3 = 4$.

Example 13.3.1 (continued) Here is some software output for the Mann-Whitney rank-sum test on the alcohol metabolism data:

```
males      N =  14     Median =       3.000
females    N =  18     Median =       1.050
W = 330.0
Test is significant at 0.0002
```

We find that $W = 330$ and the test is significant with $p = .0002$. This may be compared with the usual t-based method:

```
          N      MEAN     STDEV   SE MEAN
males    14      4.12      3.23      0.86
females  18     1.100     0.824      0.19

TTEST: T= 3.41  P=0.0042  DF=  14
```

In this case, we get similar results and are reassured that the assumption of normality is not crucial.

In fact, we might choose to transform the data by taking logs, and the results turn out to be that the mean log first-pass metabolism is higher than

that for the females with $p = .0002$. We could then add that nonparametric analysis with the rank-sum test provides corroboration with an approximate $p = .0002$ once again. Overall, we still must consider the strength this evidence in light of the subjects being volunteers. We would want to know more about how plausible it might be that these volunteers are different than randomly selected men and women. We return to this point below but the size of the effect and its p-value might make us expect the results to be fairly likely to hold in the general population. □

An analogous procedure for several groups is called the *Kruskal-Wallis test*. It may be used in place of the usual F-statistic from an ANOVA.

Example 13.3.1 (continued) When all four groups are used, and the data are transformed by logs, we find $p = .003$ from the ANOVA F-test. Using the Kruskal-Wallis test we get $p = .002$, which again corroborates the procedure. In fact, the residual analysis for the log-transformed data looks pretty good and we would find no reason to worry about the assumption of normality. □

Although in this example the nonparametric methods merely add a small degree of reassurance, in some cases the data may be grossly non-normal and transformations may not help, so it is worth knowing that alternatives to the standard normality-based results do exist.

13.3.2 Permutation and bootstrap tests may be used to test ANOVA hypotheses.

In Section 11.2 we described how permutation and bootstrap tests may be used as an alternative to the t-distribution for computing a p-value in order to test $H_0 : \mu_1 = \mu_2$. The essential method was to (i) merge the data, then (ii) repeatedly draw pseudo-data random samples of size n_1 and n_2, (iii) to each pair of samples apply the t-statistic, and finally (iv) see what proportion of the pseudo-data give t-statistic values greater than that observed in the real data. When the sampling is done without replacement the method is a permutation test, and with replacement it becomes a bootstrap test.

For one-way ANOVA the procedure is exactly analogous. For instance,

with 3 conditions we would follow step (i) then in (ii) draw samples of sizes n_1, n_2, n_3, in (iii) we would get the F-statistic, and likewise in (iv) see what proportion of the pseudo-data F values exceed the F obtained for the real data.

Two-way ANOVA is a bit more complicated, but the concept is the same.

13.3.3 When there are many levels of a factor it may be useful to model its effects as random.

13.4 Randomization, Causation, and Observational Studies

To introduce randomization, causation and observational studies we consider an example.

Example 13.4.1 IQ and breast milk In a study reported in the British medical journal *Lancet* (1992, 339: 261-64), IQ test scores were obtained from 300 children who had been premature infants and initially fed milk by a tube. The children were 8 years old when they took the IQ test. The milk they were fed by tube was either breast milk or prepared formula, or some combination of the two. Of interest was the relationship between IQ test scores and the proportion of milk the infants received that was breast milk. The amount of breast milk a baby drank was determined by whether or not the mother wished to feed the infant by breast milk, and how much milk the mother was able to express. □

In Example 13.4.1, immediately we must be aware of possible *confounding factors*. The decision to administer the treatment, i.e., use breast milk or not, was the mother's; whatever might determine that decision *and also be related to susequent IQ* would affect the observed relationship between IQ and consumption of breast milk. If, for example, mothers who chose to breast feed were also more likely to provide intellectual stimulation to their

young children then the decision to breast feed could appear to raise IQ even though it was the increased stimulation that had the greater impact. The study would be free of these concerns if babies instead received a randomly-determined percentage of breast milk, but few mothers would give up this decision in order to be part of a scientific investigation.

Example 13.4.1 (continued) In an attempt to control confounding factors (and to reduce variability and make the comparisons more sensitive), the researchers performed a regression that included characteristics of both the mothers and the babies: social class (ordered from 1 to 5 with 5 being highest), mother's education (ordered from 1 to 5 with 5 being highest), whether or not the child was a female (1 if female, 0 if male), the number of days of ventilation of the baby after birth, and whether or not the baby received any breast milk (1 if yes, 0 if no). The results of the regression are as follows:

Explanatory Variable	Estimated Coefficient	p-Value
Social class	-3.5	.0004
Mother's education	2.0	.01
Female or not	4.2	.01
Days of ventilation	-2.6	.02
Received breast milk or not	8.3	$< .0001$

Let us begin by interpreting the main finding. If we hold fixed social class, mother's education, sex of the baby, and days of ventilation, there is a highly significant effect of whether or not the baby received breast milk, with breast milk increasing subsequent IQ, on average, by 8.3 points. This is quite a large effect. If it were felt appropriate to generalize from these data to the population at large, this effect would certainly be something the pediatric professions would pay attention to.

Should we believe that early consumption of breast milk would tend to increase IQ in the general population? □

13.4.1 Randomization eliminates effects of confounding factors.

To analyze the possibility of confounding factors it is useful to introduce some terminology and list some basic points.

1. An *experiment* imposes a treatment on some person or thing that will be observed; measurements on that person or thing are called the *response variable*. On the other hand, *Observational studies* examine relationships between response variables and potential explanatory variables, which could become treatments, but there is no active administration of a treatment.

2. In experiments and observational studies, we are typically interested in effects of some explanatory variable or treatment on a response variable.

3. A *confounding factor* (or *confounding variable*) is one that affects both the response variable and an explanatory variable; its effects on the response can not be distinguished from the effects of the explanatory variable of interest on the response.

4. The particular subjects or objects being experimented upon may have special characteristics that make them different than those about which one may wish to draw conclusions.

5. Response variables, and explanatory variables, may not accurately capture what they are purported to be measuring.

6. In many situations, carefully designed experiments can avoid these difficulties. *Randomization*, meaning the random allocation of the treatment to the experimental unit (the person or thing that will be measured to provide a response variable), provides a way of avoiding confounding variables; *double-blind* experiments can avoid hidden biases in the response measurements; strict adherence to the experimental *protocol* can also help avoid mismeasured variables.

13.4.2 Observational studies often produce substantial evidence.

It is important to appreciate the power of well-designed experiments, and especially the role of randomization: they can support causal explanations for associations between response and explanatory variables. More specifically, based on a well-designed experiment, it may be possible to say that, up to some degree of statistical uncertainty (represented by a standard error or confidence interval), a response will on average increase or decrease by a particular amount when an explanatory variable changes its value by some number of units (including being present rather than absent, as is the case for typical treatments). However, there are situations in which it is impossible to randomly assign subjects to treatments. For example, one can not tell people whether they will be in "smoking" or "non-smoking" groups. Still, very convincing evidence can accumulate from observational studies—as in fact has happened in the case of smoking. Several observed patterns may increase the plausibility of an explanatory variable as a cause of a response variable:[5]

- The explanatory variable or treatment precedes observation of the response (and in terms of timing can thus act as a cause).
- Large effects are observed; this makes it less likely that the association is due to a confounding variable. (One often-cited example is that mortality due to scrotum cancer among chimney sweeps was about 200 times above the population levels early in the 20th century.)
- A quantitative "dose-response" relationship is observed, in which an increase in an exposure to the explanatory variable increases (or decreases) the observed response, as opposed to simply an observation of an effect when a treatment is applied versus not applied.
- There is biological evidence to support a theory that could explain the putative causal relationship.

[5] A widely-cited source for many of these ideas is A.B. Hill, *Principles of Medical Statistics*, ninth edition, Oxford University Press, 1971.

- There are no anomalous results that seem difficult to explain. (Anomalous results may signal the presence of confounding variables.)

- Similar results are obtained from differing experimental conditions (because a confounding variable may be unlikely to be present in each of the different studies).

Example 13.4.1 (continued) Now, let us reexamine the IQ and breast milk results with these principles in mind. First, the study is prospective, in the sense that children received some percentage of breast milk and then were followed over time to see what IQ score they got many years later. Second, the estimated effect is reasonably large—8 IQ points is about half of a standard deviation in the population as a whole. Third, there is some biological relevance: pediatricians recommend that mothers breast-feed their babies for nutritional reasons; on the other hand, there probably isn't accompanying evidence from some other species.

Concerning the dose-response relationship, in the regression reported above the breast milk variable merely indicates whether or not the infant received breast milk; but the authors reported a similar regression using instead *percentage* breast milk where the regression coefficient was .09, which says that holding the same variables fixed, for every 10% increase in breast milk the subsequent IQ would go up on average by nearly a full point. This last result is important: by removing the decision of whether or not to use breast milk as an explanatory variable, the confounding variables associated with that decision are no longer a concern.[6] Now we must shift to the question of whether some confounding variables may affect both the amount of milk a mother can express and the subsequent IQ of the child. If not, we would be regarding the percentage breast milk actually delivered as if it were a randomly-determined percentage. One possible confounding variable would be the health of the mother during pregancy: mothers who are unable to express much milk might conceivably have been providing worse nutrition to the fetus.

[6]We are here assuming that the reported regression is not being driven primarily by inclusion of lots of babies with zero percent breast milk, but rather holds among the non-zero percentage babies.

As far as anomalous results are concerned, here are two: first, given the other variables, subsequent IQ decreases as social class increases, which is surprising. Second, given the other variables, female babies have higher subsequent IQs. Is there a reason for this? (Here we need to know more about the specific test that was administered and whether sex effects are usually seen.) Unless these issues are resolved, they raise doubts.[7]

Overall, there is clearly a substantial association between increased administration of breast milk and increased IQ, when social class (measured in the way the authors did), mother's education, and days of ventilation are held fixed. It remains possible that some confounding variables affect breast-milk expression and IQ. We would consider this study to be very interesting and find that it provides some evidence but, based on what we currently know about the details, we would not consider this strong evidence. □

Example 13.3.1 (continued) Returning to the alcohol metabolism example, let us now consider the possibility of confounding due to the use of volunteers in the study. The chief concern is whether volunteers are different than the rest of the population with respect to alcohol metabolism. This is at least plausible, though in order to affect the study, the volunteer men and women would have to be different. For example, if the women who volunteered tended to have trouble with alcohol metabolism (perhaps they thought the study sounded interesting because they knew they couldn't "handle the liquor") but men just wanted the money, then the differential effect would tend to be larger in this sample than in the population. Is this kind of hypothetical scenario reasonable, or really a stretch of the imagination? Your answer to this question determines how much faith you will put in the results. □

[7]We do not have the full results when percentage breast milk is used, so we don't know whether these assocations diminish or change sign in that case.

©2010 SPRINGER SCIENCE+BUSINESS MEDIA, LLC. All rights reserved. No part of this work may be reproduced in any form without the written permission of SPRINGER SCIENCE+BUSINESS MEDIA, LLC.

Chapter 14

Generalized Linear and Nonlinear Regression

Multiple linear regression is a powerful method of exploring relationship between a response Y and a set of potential explanatory variables $x_1, \ldots, x_p$, but it has an obvious flaw: it assumes the predictive relationship is, on average, linear. In addition, its fitting criterion, least squares, effectively assumes that the noise contributions (the deviations from the underlying linear model) among the various data points are homogeneous (they have the same variability) and they follow, roughly, a normal distribution. During the latter part of the 20th century a great deal of attention was directed toward the development of generalized regression methods that could be applied to nonlinear relationships, with non-constant and non-normal noise variation. In this chapter, and in Chapter 16 we discuss several of the most common techniques that come under the heading "modern regression."

The essential modeling idea behind modern regression is very simple. If we consider a single explanatory variable x, we may write the ordinary linear

regression model as

$$Y_i = \mu_i + \epsilon_i \quad (14.1)$$
$$\mu_i = \beta_0 + \beta_1 x_i \quad (14.2)$$

where $\epsilon_i \sim N(0, \sigma^2)$. The point here, is that we are separating the linear, deterministic part of the model in Equation (14.2) from the probabilistic part in (14.1), which of course represents deviations from the systematic relationship in (14.1) as additive noise. Modern regression models have the more general form

$$Y_i \sim p(y_i|\theta_i) \quad (14.3)$$
$$\theta_i = f(x_i) \quad (14.4)$$

where $p(y|\theta)$ is some family of distributions that depend on a parameter θ, which is related to x according to a function $f(x)$. Here, not only is $f(x)$ in (14.4) allowed to be nonlinear, but also the probabilistic representation of noise in (14.3) is more general than in (14.1). In Sections 14.1.2 and 14.2, respectively, we take the response distributions in (14.3) to be binomial and Poisson but in (14.4) we retain the linear dependence on x for suitable parameters θ_i. In Section 14.3 we describe the use of nonlinear functions $f(x) = f(x;\theta)$ that remain determined by a specified vector of parameters θ (such as $f(x;\theta) = \theta_1 \exp(-\theta_2 x)$); and in Section 14.4 we discuss the formal framework known as *generalized linear models* that encompasses methods based on normal, binomial, and Poisson distributions, along with several others.

14.1 Logistic Regression

14.1.1 Logistic regression may be used to analyze responses that are binomial.

There are many situations where outcomes are binary. For instance, behavioral responses are sometimes either correct or incorrect and we may wish to consider the probability of correct response as a function of some explanatory variable or variables, or across experimental conditions.

Example 2.1.1 (continued) On page 371 of Section 13.2.2 we discussed interactions in the context of the study by Behrmann *et al.* (2002) on hemispatial neglect, where the response was saccadic reaction time and one of the explanatory variables was angle of the starting fixation point of the eyes away from "straight ahead." A second response variable of interest was saccadic error, i.e., whether the patient failed to execute the saccade within a given time window. Errors may be coded as 0 and successful execution as 1. Behrmann *et al.* (2002) used logistic regression to analyze the error rate as a function of the same explanatory variables. They found, for example, that the probability of error increased as eyes fixated further to the right. □

We next reconsider Example 8.4.1.

Example 8.4.1 (continued) In Figure 8.6 we displayed a sigmoidal curve fitted to the classic psychophysical data of Hecht *et al.* on perception of dim light. The method was to apply maximum likelihood estimation to the logistic regression model in (8.24) and (8.25). This[1] is known as *logistic regression.* □

[1]It may be of interest to note that the analysis of Hecht *et al.* was different, but related. They wished to obtain the minimum number of quanta, n, that would produce perception. Because quanta are considered to follow a Poisson distribution, in the notation we used above, they took $W \sim P(\lambda)$ and $c = n$, with λ, the mean number of quanta falling on the retina, being proportional to the intensity. This latter statement may be rewritten in the form $\log \lambda = \beta_0 + x$, with x again being the log intensity. Then $Y = 1$ (light is perceived) if $W \geq n$ which occurs with probability $p = 1 - P(W \leq n-1) = 1 - F(n-1|\lambda)$, where F is the Poisson cdf. This is yet another latent-variable model for the proportional data. It could be fitted by finding the MLE of β_0, though Hecht *et al.* apparently did the fitting by eye. Hecht *et al.* then determined the value of n that provided the best fit. They concluded that a very small number of quanta sufficed to produce perception. (Hecht, S., Schlaer, S. and Pirenne, M. (1942) Energy, quanta and vision. *J. General Physiology*, 25: 819–40; but see also Teich MC, Prucnal PR, Vannucci G, Breton ME, McGill WJ (1982) Multiplication noise in the human visual system at threshold. 3. The role of non-Poisson quantum fluctuations, *Biol. Cybernetics* 44:157-65.)

14.1.2 The logistic regression model assumes binomial responses and converts the binomial probability p to log odds $\log \frac{p}{1-p}$.

In linear regression we have $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$, which implies $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. There are three problems in applying ordinary linear regression with dichotomous or proportional responses: (i) a line won't be constrained to (0,1), (ii) the variances are not equal, and (iii) the responses are not normal (unless we have proportions among large samples). The first problem, illustrated in Figure 8.6, is that the linear regression may not make sense beyond a limited range of x values: if $y = a + bx$ and $b > 0$ then y must become infinitely large, or small, as x does. In many data sets with dichotomous or proportional responses there is a clear sigmoidal shape to the relationship with x. The second problem was discussed in the simpler context of estimating a mean, in Chapter 8. There we derived the best set of weights to be used for that problem, and showed that an esimator that omits weights can be very much more variable, effectively throwing away a substantial portion of the data. Much more generally it is also possible to solve problem (ii) by using weighted least squares, and such solutions apply—often quite nicely—to the logistic regression setting. We will say a few words about this below. The third problem can make distributional results (standard errors and p-values) suspect. Logistic regression fixes all three problems.

In logistic regression we solve problem (i) with the log-odds transformation. Recall that when p is a probability the associated *odds* are $p/(1-p)$. The number p lies in the range (0,1) while the associated odds is in the range $(0, \infty)$. If we then take logs, the number $\log(p/(1-p))$ will lie in the range $(-\infty, \infty)$, which corresponds to what we need for in infinite straight line. Therefore, instead of taking the expected value of Y to be linear in x ($E(Y_i) = \beta_0 + \beta_1 x_i$) we note that when $Y_i \sim B(n_i, p_i)$ we have $E(Y_i/n_i) = p_i$ and use $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i$. Using

$$z = \log(\frac{w}{1-w}) \Longleftrightarrow w = \frac{\exp(z)}{1+\exp(z)}$$

with p substituted for w and $\beta_0 + \beta_1 x_i$ substituted for z, the logistic regression model may then be written in the form for binomial data becomes

$$Y_i \sim B(n_i, p_i)$$

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i.$$

The log-odds (or "logit") transformation is helpful in interpreting the results. The log odds (of a response) are linear in x. Thus, β_1 is the change in the log odds for a unit change in x.

The log odds scale itself is a bit awkward to think about, though if the base of the logarithm is changed from e to 2 (or 10) it becomes easier. It is often useful to transform back to the odds scale, where an increase of 1 unit in x is associated with an increase in the odds (that $Y = 1$) by a factor of $\exp(\beta_1)$. If we wish to interpret the change in probabilities, we must pick a particular probability p and conclude that a unit increase in x is associated with an increase from p to $expit(logit(p)+\beta_1)$, where $logit(z) = \log(z/(1-z))$ and $expit(w) = \exp(w)/(1+\exp(w))$. This is illustrated below.

The logistic regression model extends immediately to multiple explanatory variables: for m variables $x_1, \ldots, x_m$ we write

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_m x_{mi}.$$

14.1.3 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x.

As we indicated in Chapter 8, it is not hard to write down the likelihood function for logistic reqression. The responses Y_i are independent observations from $B(n_i, p_i)$ distributions, so each pdf has the form $\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}$ and the likelihood function is

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i-y_i} \\ p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} \end{aligned}$$

where the second equation is substituted into the first. Standard statistical software may be used to maximize this likelihood. The standard errors are

obtained from the observed information matrix.

The likelihood ration test may be used to test hypotheses involving the coefficients. Let us write the log likelihood ratio in the form

$$-2\log LR = -2[\log(\hat{L}_1) - \log(\hat{L}_2)]$$

where $\hat{L}_i$ is the maximum value of the likelihood under model i. Here we assume that model 1 is obtained from model 2 when some coefficient (under model 2) vanishes. As discussed in Chapter 11 for large samples, $-2\log LR$ follows the χ^2 distribution with degrees of freedom equal to the number of additional parameters in model 2.

In some software, the results are given in terms of "deviance." The *deviance* for a given model is $-2\log(\hat{L})$. The *null deviance* is the deviance for the "intercept-only" model. Thus, the usual test of $H_0 : \beta_1 = 0$ compares the deviance for the logistic regression model to the deviance for the null model.

Example 8.4.1 (continued) The output from least-squares software looks like this:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7792     0.3032  -5.867  0.00421 **
intensity     1.1962     0.1603   7.464  0.00172 **


Residual standard error: 0.129 on 4 degrees of freedom
Multiple R-Squared: 0.933,
F-statistic: 55.72 on 1 and 4 DF,  p-value: 0.001721
```

while that for logistic regression looks like this:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -20.54       2.39  -8.594   <2e-16 ***
intensity      10.69       1.24   8.620   <2e-16 ***
```

```
    Null deviance: 257.2584  on 5  degrees of freedom
Residual deviance:   2.9215  on 4  degrees of freedom
```

Note that the difference in deviance is

$$\text{null deviance - residual deviance} = 257.3 - 2.9 = 256.4$$

which should be compared to the chi-squared distribution on 1 degree of freedom. It is very highly significant.

We interpret the fitted model as saying that, on average, for every increase of intensity by a factor of 10 (1 unit on the scale of the explanatory variable) there is a 10.7 ± 1.2 increase in the log odds of a response. To get an approximate 95% CI for the factor by which the odds increase we exponentiate, $\exp(10.7 \pm 2(1.2))$, i.e., (4023,489000). A more interpretable intensity change, perhaps, would be doubling. An increase in intensity by a factor of 2 corresponds to .30 units on the scale of the explanatory variable (because $\log_{10}(2) = .301$). For an increase of intensity by a factor of 2 the log odds thus increase by $3.21 \pm .72$. This gives an approximate 95% CI for the factor by which the odds increase, when the intensity doubles, of (12.1,50.9). When the probability of perception is .5, a doubling of the intensity would increase the probability to (.80,.92), with 95% confidence. □

Finally, we note that polynomial terms in x may be handled as in linear regression.

Example 8.4.1 (continued) To consider whether an additional, nonlinear component might contribute usefully to the linear logistic regression model, we may square the intensity and try including it in a two-variable logistic regression model. In this case it is interesting to note that intensity and its square are highly correlated. To reduce the correlation it helps to subtract the mean before squaring. Thus, we define intsq = $(\text{intensity})^2$ and int2 = $(\text{intensity} - \text{mean(intensity)})^2$ and perform logistic regression, which gives the following as output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)   -4.344     15.756  -0.276    0.783
intensity     -6.628     16.972  -0.391    0.696
intsq          4.593      4.557   1.008    0.313

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -20.295      2.329  -8.716   <2e-16 ***
intensity     10.491      1.216   8.630   <2e-16 ***
int2           4.593      4.557   1.008    0.313
```

Using either of these two logistic regression summaries we would conclude the quadratic term does not improve the fit. The first might be a bit confusing, however, because of the nonsignificant p-values. As we noted in our discussion of linear regression, this is a fairly common occurrence with highly correlated explanatory variables, as x and x^2 often are. Recall that each nonsignificant p-value leads to the conclusion that its corresponding variable contributes little *in addition to* the other variable. Since we already found a very highly significant logistic linear relationship, we would conclude that the quadratic doesn't improve the fit. Again, though, the interpretation appears cleaner in the second formulation. □

14.1.4 The logit transformation is one among many that may be used for binomial responses, but it is the most commonly applied.

The *expit* function $\exp(x)/(1+\exp(x))$, defined in Section 14.1.2, is one of many possible sigmoidal curves and thus logistic regression is only one of many possible models for binary or proportion data. In fact, expit(x) has an asymptote at 0 as $x \to -\infty$ and at 1 as $x \to \infty$, and is increasing, so it is a cumulative distribution function. The distribution having expit(x) as its cdf is called the *logistic distribution*, but the cdf of any continuous distribution could be used instead. One important alternative to logistic regression is the Probit regression model, which substitutes the normal cdf in place of the *expit*: specifically, the probit model is

$$Y_i \sim B(n_i, p_i)$$

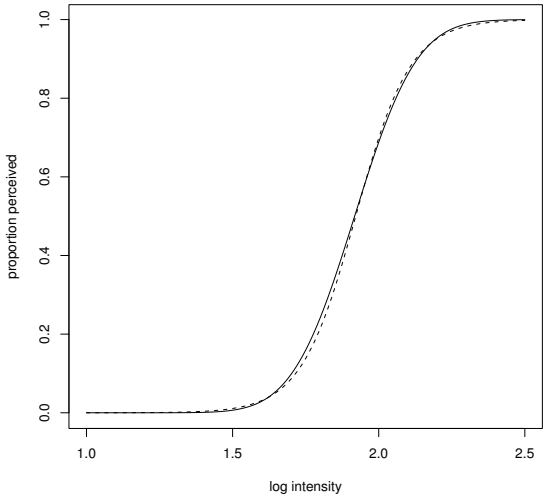


Figure 14.1: The fitted curve from probit regression (dashed line) is shown together with the fitted curve from logistic regression. The fits are very close to each other.

$$\Phi^{-1}(p_i) \quad = \quad \beta_0 + \beta_1 x_i$$

where $\Phi(z) = P(Z \leq z)$, with $Z \sim N(0,1)$. The fitted curve is then obtained from $y = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x)$.

Example: Threshold for visual perception (continued) Figure 14.1 displays the fitted curves from probit and logistic regression for the threshold data. The two models produce nearly identical fitted curves. □

As with the threshold data, the fitted curves from probit and logistic regression are generally very close to each other. This is because the graph of the logistic cdf (the *expit* function) is quite close to the graph of the normal cdf. Two things are special about the logistic regression model. First, it gives a nice interpretation of the coefficients in terms of log odds. Second, in the logistic regression model (but not the Probit or other versions) the loglikelihood function is necessarily concave (as long as there are at least two distinct values of x). This means that there is a unique MLE, which can be

obtained from an arbitrary starting value in the iterative algorithm. Logistic regression is the standard method for analyzing dichotomous or proportional data, though in some contexts probit regression remains quite popular.[2]

An interesting interpretation of binary phenomena involves the introduction of *latent variables*, meaning random variables that become part of the statistical model but are never observed. Let us discuss this in terms of perception, and let us imagine that the binary experience of perception, as "perceived" or "not perceived" is controlled by an underlying continuous random variable, which we label W. We may think of W as summarizing the transduction process (from light striking the retina to firing rate among multiple ganglion cells), so that perception occurs whenever $W > c$ for some constant c. Neither the precise meaning of W, nor the units of c need concern us. Let us take W to be normally distributed and, because the units are arbitrary, we take its standard deviation to be 1. Finally, we take this latent transduction variable, on average, to be a linear function of the log intensity of light x and we write this in the form $\mu_W = c + \beta_0 + \beta_1 x$. We now have the probit regression model: $Y = 1$ when $W > c$ but, defining $-Z = W - \mu_W$ (so that $-Z \sim N(0,1)$ and $Z \sim N(0,1)$),

$$W > c \Longleftrightarrow W - \mu_W > c - \mu_W \Longleftrightarrow -Z > c - \mu_W \Longleftrightarrow Z < \mu_W - c.$$

In other words, $Y = 1$ when $Z < \beta_0 + \beta_1 x$, which occurs with probability $p = \Phi(\beta_0 + \beta_1 x)$.

This latent-variable interpretation helps transfer the intuition of linear regression models over to the binary case, and provides an appealing way to think about many phenomena. Note that logistic regression is obtained by taking W to have a logistic distribution.[3]

[2]We have not discussed residual analysis here. It may be performed using "deviance residuals" (or other forms of residuals, including Pearson chi-squared residuals).

[3]Probit regression was introduced by Bliss in 1934, but the latent variable idea and normal cdf-transformation was part of Fechner's thinking about psychophysics in 1860; logistic regression was apparently discussed first by Fisher and Yates in 1938. See Agresti *Categorial Data Analysis* for much more extensive discussion of the methods described briefly here.

14.2 Poisson Regression

14.2.1 The usual Poisson regression model transforms the mean λ to $\log \lambda$.

The simplest distribution for counts is Poisson. Poisson regression is useful when we have counts depending on one or more explanatory variables.

Example 14.2.1 Directional sensitivity of an SEF neuron Olson *et al.* (2000, *J. Neurophys.*) reported data collected from many individually-recorded neurons in the supplementary eye field (SEF). In this experiment, a monkey was to trained to translate one of four possible icons displayed at the fixation point into an instruction of a location to which he was to move his eyes: either left, up, right, or down. SEF neurons tend to be directionally sensitive. To establish direction sensitivity, Olson *et al.* examined the number of spikes occurring 600 to 750 ms after presentation of the cue. Here are the data for the various trials:

```
left
9 6 9 9 6 6 8 5 7 9 4 8 8 3 6
up
2 0 6 4 4 0 0 0 5 2 1 0 3 0
right
4 8 2 2 4 0 3 4 1 1 0 3 4 0 2
down
1 5 1 2 0 4 4 4 4 3 6 1 1 1
```

Is this neuron directionally sensitive?

By eye it appears that the firing rate is higher for the "left" condition than for the other conditions. There are various versions of ANOVA that may be used to check this. Analysis of these neurons revealed that while the spike counts deviated from that predicted by a Poisson distribution, the deviation was small (Ventura *et al.*, 2001). Here we will use the data to illustrate a version of ANOVA based on Poisson regression. □

The problem of fitting counts is analogous to, though less extreme than, that of fitting proportions. For proportions, the (0,1) range could make linear regression clearly inappropriate. Counts have a range of $(0, \infty)$. Because the ordinary regression line is not constrained, it will eventually go negative. The simple solution is to use a log transformation of the underlying mean. The usual Poisson regression model is

$$Y_i \sim P(\lambda_i) \tag{14.5}$$
$$\lambda_i = \exp(\beta_0 + \beta_1 x_i). \tag{14.6}$$

To interpret the model we use the log transformation:

$$\log \lambda_i = \beta_0 + \beta_1 x_i.$$

For example, in the SEF data $\log \lambda_i$ is the spike count and x_i is the experimental condition (up, down, left, right) for the ith trial. The advantage of viewing ANOVA as a special case of regression is apparent: we immediately generalize Poisson ANOVA by applying our generalization of linear regression to the Poisson regression model above.

14.2.2 In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends.

As in logistic regression we use ML estimation and the likelihood ratio test ("analysis of deviance").

Example 14.2.1 (continued) We perform Poisson regression using indicator variables to achieve an ANOVA-like model. First, let us look at the output from ordinary ANOVA:

```
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
cond       3  228.8215 76.27384 18.75505 1.836788e-08
Residuals 54  219.6095  4.06684
```

Here is ANOVA output performed instead using indicator variables and least-squares regression:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   3.4929   0.2650    13.1828   0.0000
      cond1   2.1119   0.3747     5.6362   0.0000
      cond2  -0.7405   0.2138    -3.4629   0.0011
      cond3  -0.5214   0.1547    -3.3701   0.0014

Residual standard error: 2.017 on 54 degrees of freedom
Multiple R-Squared: 0.5103
F-statistic:
18.76 on 3 and 54 degrees of freedom, the p-value is 1.837e-08
```

Note that the F and p-value results for least-squares regression are identical with those from ANOVA (as they must be if there is no computational error).

The Poisson regression output is similar in structure:

```
Coefficients:
                 Value Std. Error    t value
(Intercept)  1.1212194 0.07897981  14.196279
      cond1  0.4774091 0.09581966   4.982371
      cond2 -0.1732444 0.06277747  -2.759659
      cond3 -0.1548059 0.05231702  -2.958998

    Null Deviance: 149.7523 on 57 degrees of freedom
Residual Deviance: 92.48045 on 54 degrees of freedom
```

As in logistic regression we look at difference in deviance:

$$\text{null.deviance - residual deviance} = 149.8 - 92.5 = 57.3$$

which should be compared to the chi-squared distribution on 3 degrees of freedom. It is very highly significant. Roughly speaking the results are the same as with ordinary linear regression (standard ANOVA), but the details are different and this is an indication that in some situations the conclusions could be different. □

14.3 Nonlinear Regression

14.3.1 Nonlinear regression models may be fitted by least squares.

In Chapter 13 we pointed out that polynomials in x could be fitted to (x, y) data using linear regression. This involved the "trick" of starting with an initial definition of x, relabeling it as x_1 and then defining the new variable $x_2 = x_1^2$, and so on for higher-order polynomials. The resulting expectation of Y,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

followed the form required in the linear regression model. In particular, although the relationship of Y and x, on average, was nonlinear, the *coefficients* entered linearly into the model and therefore—as in any linear regression model—the likelihood equations could be solved easily by linear algebra.

There are, however, many nonlinear relationships where this sort of manipulation does not apply. For example, if

$$E(Y) = \theta_1 e^{-\theta_2 x}$$

it is not possible to redefine the x variable so that the form becomes linear in the parameters. Instead, we have the *nonlinear regression model*,

$$Y_i = f(x_i; \theta) + \epsilon_i \tag{14.7}$$

$$f(x_i; \theta) = \theta_1 e^{\theta_2 x}. \tag{14.8}$$

Here, the usual assumption is $\epsilon_i \sim N(0, \sigma^2)$, independently (though, again, normality is not crucial).

Models of the form (14.7) may still be fit by least-squares and, in fact, least squares remains a special case of ML estimation. What is different is that the equations defining the least-squares solution (the likelihood equations) are no longer solved by a single linear algebraic step. Instead, they must be solved iteratively. The problem is thus usually called *nonlinear least squares*.

Nonlinear least squares is especially common in pharmacokinetic studies.

Example 14.3.1 Magnesium block of NMDA receptors NMDA receptors, which are ubiquitous in the vertebrate central nervous system, may be blocked by Magnesium ions (Mg^{2+}). To investigate the quantitative dependence of NMDA-receptor currents on the concentration of Mg^{2+}, Qian, Buller, and Johnson (2005, *J. Physiology*) measured currents at various concentrations, then summarized the data using the equation

$$\frac{I}{I_0} = \frac{1}{1 + (\frac{[Mg^{2+}]}{IC_{50}})^{n_H}}$$

where the measurements are the current I and the Magnesium concentration $[Mg^{2+}]$, I_0 being the current in the absence of Mg^{2+}. The free parameters are the "Hille constant" n_H and the 50% inhibition concentration IC_{50} (when $[Mg^{2+}] = IC_{50}$ we get $I/I_0 = .5$). The authors examined IC_{50} across voltages, and across receptor subunit types. □

Example 14.3.2 Non-cosine directional tuning of motor cortical neurons Amerikian and Georgopoulos (2002, *Neurosci. Research*) investigated cosine and non-cosine directional tuning among motor cortical neurons. Let $Y_i(\theta)$ be the number of times a neuron fires during a specified interval of time of length T (such as 200 milliseconds) in direction θ during the ith repetition of the experiment (the ith "trial"). The simplest model assumes the count $Y_i(\theta)$ is normally distributed with mean $T \cdot \lambda(\theta)$ and with

$$\begin{aligned} \lambda(\theta) &= \mu + \alpha \cos(\theta - \theta_0) \\ &= \mu + \beta_1 \cos(\theta) + \beta_2 \sin(\theta). \end{aligned} \tag{14.9}$$

Linear regression may be used to fit the resulting *cosine tuning function*. The parameter θ_0 is commonly referred to as the neuron's "preferred direction." The second form again illustrates the possibility of converting an apparently nonlinear model to a linear model; the latter form is easy to fit and propagation of error may be used to make inferences about $\theta_0 = \arctan(\beta_2/\beta_1)$. To investigate departures from cosine tuning, Amirikian and Georgopoulos used a class of functions, having an exponentiated form, which are not amenable to reconfiguration in a linear model. For example, the functions shown fit to data from two neurons in Figure 14.2 are

$$\lambda(\theta) = \mu + \beta \exp(\kappa \cos(\theta - \tau + \eta \cos(\theta - \tau))) \tag{14.10}$$

for the first neuron, and

$$\lambda(\theta) = \mu + \beta_1 \exp(\kappa_1 \cos(\theta - \tau_1)) + \beta_2 \exp(\kappa_2 \cos(\theta - \tau_2)). \qquad (14.11)$$

for the second neuron. The results given here come from Kaufman, Ventura, and Kass (2005, *Statistics in Medicine*), who also considered nonparametric methods, discussed Chapter 16. The function in (14.10) includes parameters corresponding roughly to the baseline firing rate, the amplitude, width, and location of the mode, and the skewness about the mode. The function in (14.11) includes parameters corresponding to two modes, one of which is constrained to be in the positive direction and the other in the negative direction. This is of use in fitting the data for the Neuron 2 in Figure 14.2. For both neurons the data indicate mild but noticeable departures from cosine tuning. □

14.3.2 In solving nonlinear least-squares problems, good starting values are important, and it can be helpful to reparameterize.

As in maximization of any likelihood, use of the numerical procedures require a bit of care. Two important issues are, once again, the choice of initial values, and of parameterization. Both of these may be illustrated with the exponential model (14.8).

Illustration: Exponential regression We consider here some data fabricated to follow the exponential model (14.8). Specifically, we took $x = 1, 2, \ldots, 15$ and generated Y values using (14.8), with $\epsilon_i \sim N(0, \sigma^2)$. The parameters were $\theta_1 = 1$, $\theta_2 = .25$, $\sigma = .04$. A plot of the data, together with the nonlinear least-squares fit, is shown in Figure 14.3.

In this case, we first reparameterized from θ to ω using $\omega_1 = \log(\theta_1)$ and $\omega_2 = \theta_2$ so that the expected values have the form

$$E(Y) = \exp(\omega_1 + \omega_2 x).$$

The loglikelihood is typically a bit closer to being quadratic as a function of ω than as a function of θ. Taking logs of both sides of this expectation

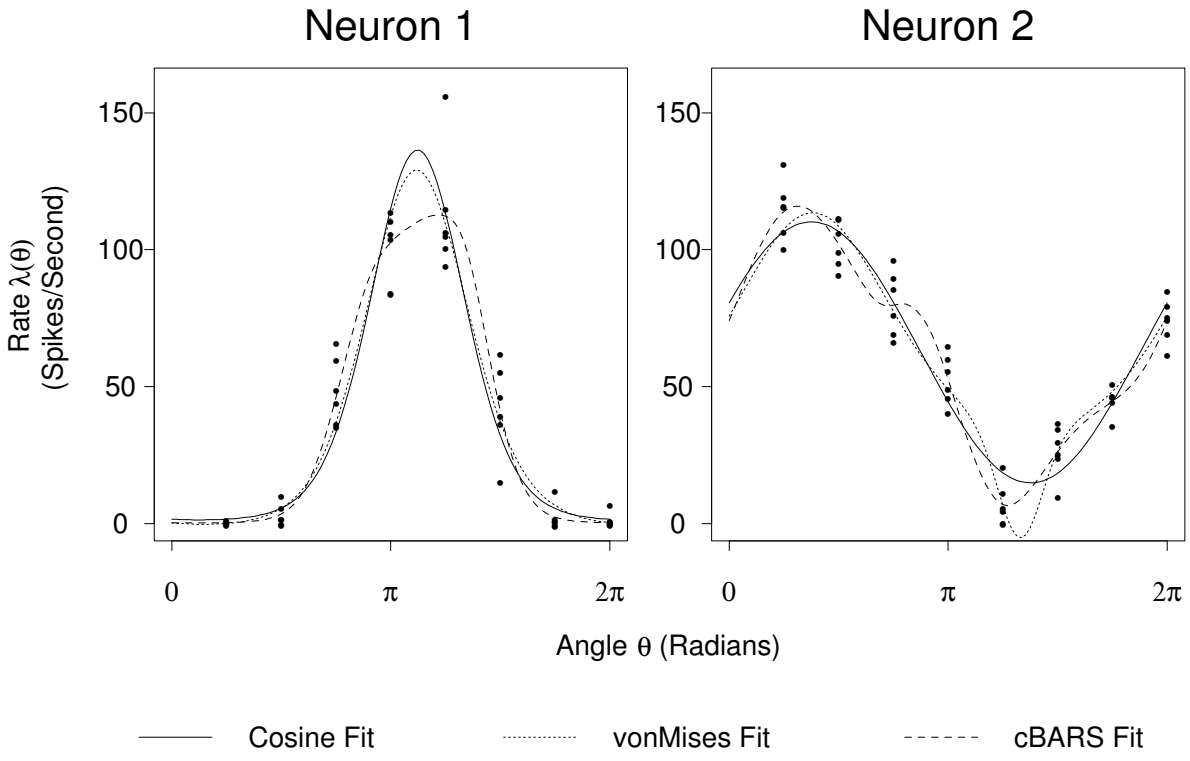


Figure 14.2: *Fits to activity of two neurons in primate motor cortex (from Kaufman, Ventura, and Kass, 2005,* Statistics in Medicine*). Each datapoint represents the observed firing rate of a neuron in the motor cortex of a monkey during one repetition of a wrist movement to a particular target. The cosine fits use the cosine function in Equation (14.9) and the von Mises fits use more complicated parametric forms given by Equation (14.10), for Neuron 1, and Equation (14.11) for Neuron 2. The cosine and von Mises parametric fits use Poisson maximum likelihood for Neuron 1 and least squares for Neuron 2. Also shown is the fit from a nonparametric regression method called cBARS, described by Kaufman, Ventura, and Kass.*

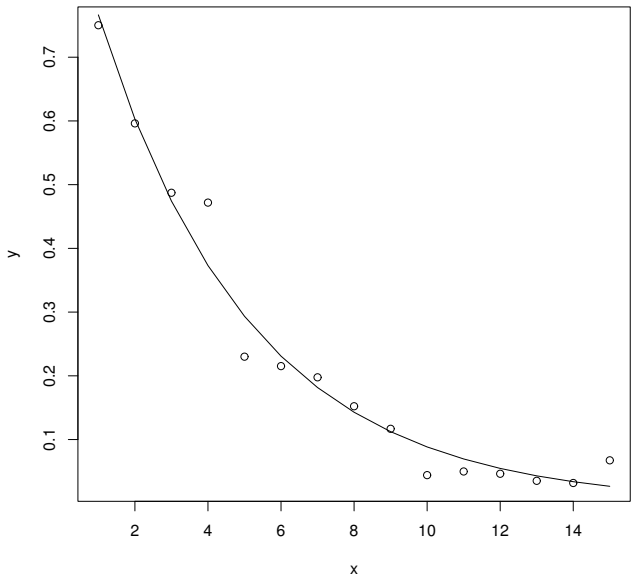


Figure 14.3: *Data simulated from exponential regression model together with nonlinear least-squares fit.*

equation gives

$$\log E(Y) = \omega_1 + \omega_2 x.$$

This suggests we may define $U_i = \log(Y_i)$ and apply the linear model,

$$U_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{14.12}$$

The resulting fitted values $\hat{\beta}_0$ $\hat{\beta}_1$ make good starting values for the iterative procedure used to obtain ω_1 and ω_2. □

It is important to recognize the distinction between the exponential model in (14.7) and (14.8) and the linearized version (14.12). Either could be used to fit data, but they make different assumptions about the way the noise contributes. In many examples, the fits based on (14.7) and (14.12) would be very close, but sometimes the resulting inferences would be different. It is an empirical question which model does a better job of describing the data. The point here, however, is that if the exponential form is preferred, the log-linear form may still be used to obtain starting values for the parameters. The linearization method of obtaining starting values is frequently used in fitting nonlinear models. (See Bates and Watts *Nonlinear Regression Analysis and its Applications* for further discussion.)

14.3.3 Nonlinear models may also be used with binomial or Poisson responses.

The term "nonlinear regression" usually refers to models of the form (14.7). However, similar models may be used with binomial or Poisson responses, and may be fit using ML. Indeed, the results for Neuron 1 in Figure 14.2 come from a Poisson model, as mentioned again in Section 14.4.

Example 14.3.3 Onset latency in a basal ganglia neuron. An unfortunate symptom of Parkinson's disease (PD) is muscular rigidity. This has been associated with increased gain and inappropriate timing of the long latency component of the stretch reflex, which is a muscular response to sudden perturbations of limb position. One of the important components of the stretch reflex is mediated by a trans-cortical reflex, probably via cortico-spinal neurons in primary motor cortex that are sensitive to kinesthetic input. To investigate the neural correlates of degradation in stretch reflex, Dr.

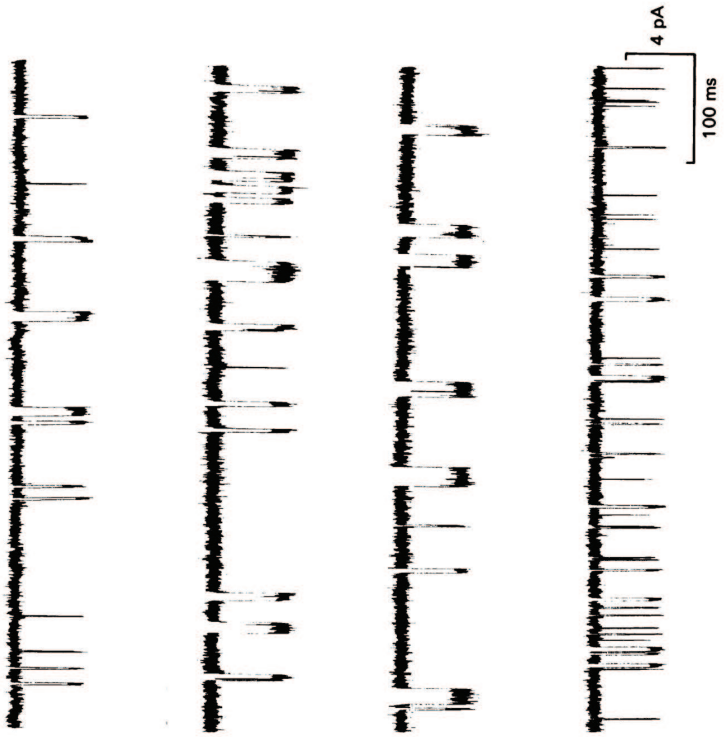


Figure 14.4: *Initiation of firing in a neuron from the basal ganglia: change-point and bootstrap confidence intervals when a quadratic model is used for the post-change-point firing rate. Two forms of approximate 95% confidence intervals are shown. The first is the usual estimate $\pm 2SE$ interval. The second is the interval formed by the .025 and .975 quantiles among the bootstrap samples. The latter typically performs somewhat better, in the sense of having coverage probability closer to .95.*

Robert Turner and colleagues at the University of Pittsburgh have recorded neurons in primary motor cortex of monkeys before and after experimental production of PD-like symptoms. One part of this line of work aims at characterizing neuronal response latency following a limb perturbation. Figure 14.4 displays a PSTH from one neuron prior to drug-induced PD symptoms. The statistical problem is to identify the time at which the neuron begins to increase its firing rate, with the goal being to compare these latencies in the

population of neurons before and after induction of PD.

To solve this problem we have used a *change-point model* similar to that used in the example on action potential width in Chapter 8. In this case, we assume the counts within the PSTH time bins—after pooling the data across trials—follow Poisson distributions. Let Y_t be the pooled spike count in the bin centered at time t and let $\mu(t)$ be its mean. The change-point model assumes the mean counts are constant up until time $t = \tau$, at which time they increase. For simplicity, we assume the count increases as a quadratic. This gives us the Poisson change-point model

$$Y_t \sim P(\mu(t))$$

with

$$\mu(t) = \begin{cases} \beta_0 & \text{if } t \leq \tau \\ \beta_0 + \beta_1(t-\tau)^2 & \text{if } t > \tau \end{cases}$$

The value τ is the change point. For any fixed τ the change-point model becomes simply a Poisson regression model. Specifically, for a given τ we define

$$x = \begin{cases} 0 & \text{if } t \leq \tau \\ (t-\tau)^2 & \text{if } t > \tau \end{cases}$$

We then apply Poisson regression with the regression variable x.

However, the parameter τ is unknown and is, in fact, the object of interest. We may maximize the likelihood function iteratively over τ. That is, in R or Matlab we set up a loop within which, for a fixed τ, we perform Poisson regression and obtain the value of the loglikelihood. We then iterate until we maximize the loglikelihood across values of τ. This gives us the MLE of τ. We may then obtain a SE for τ by applying a parametric bootstrap. Results are given in Figure 14.4. □

14.4 Generalized Linear Models

14.4.1 Generalized linear models extend regression methods to response distributions from exponential families.

We began this chapter by saying that modern regression models have the form given by (14.3) and (14.4), which for convenience we repeat:

$$\begin{aligned} Y_i &\sim p(y_i|\theta_i) \\ \theta_i &= f(x_i). \end{aligned}$$

The simple logistic regression model may be put into this form by writing

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \theta_i &= \beta_0 + x_i\beta_1 \end{aligned}$$

where

$$\theta_i = \log \frac{p_i}{1-p_i}$$

or, more succinctly,

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log \frac{p_i}{1-p_i} &= \beta_0 + x_i\beta_1. \end{aligned}$$

Similarly, the simple Poisson regression model may be written

$$\begin{aligned} Y_i &\sim P(\lambda_i) \\ \log \lambda_i &= \beta_0 + x_i\beta_1. \end{aligned}$$

Logistic and Poisson regression are special cases of *generalized linear models*. These generalize linear regression by allowing the response variable to follow a distribution from a certain class known as *exponential families*. They also use a *link* function that links the expected value (the mean) μ_i of the data with the linear model $\beta_0 + \beta_1 x_i$. For example, the usual link functions for binomial and Poisson data are the log odds and the log, respectively, as shown above.

Exponential families have pdfs of the form

$$f(y|\eta(\theta)) = h(y)\exp(\eta(\theta)T(y) - B(\theta)). \tag{14.13}$$

For instance, in the Poisson case $Y \sim P(\lambda)$, the pdf (from Chapter 5) is

$$P(Y = y) = \frac{1}{y!}\lambda^y e^{-\lambda}.$$

We can rewrite this in the form

$$\frac{1}{y!}\lambda^y e^{-\lambda} = \frac{1}{y!}\exp(y\log\lambda - \lambda)$$

If we let $\theta = \lambda$, $\eta(\theta) = \log\lambda$, $B(\lambda) = \lambda$, $T(y) = y$ and $h(y) = 1/y!$ we obtain (14.13). Now, with $\mu = \lambda$, if we define the link function to be

$$g(\mu) = \log\mu \tag{14.14}$$

the simple Poisson regression model becomes

$$g(\mu) = \beta_0 + \beta_1 x_i.$$

Here, the log provides the link in the sense that it is the function by which the mean is transformed before being equated to the linear model.

We may rewrite (14.13) in the form

$$f(y|\eta) = h(y)\exp(\eta T(y) - A(\eta))$$

in which case $\eta = \eta(\theta)$ is called the *natural parameter* (or *canonical parameter*). In the Poisson case the natural parameter is $\log\lambda$. The logarithmic link function is thus often called *the canonical link*. In the binomial case the log odds function becomes the canonical link. The extension to the multiparameter case, in which η and $T(y)$ are vectors, is immediate:

$$f(y|\eta) = h(y)\exp(\eta^T T(y) - A(\eta)).$$

Assuming that Y_i comes from an exponential family, and writing $\mu_i = E(Y_i)$, we obtain a generalized linear model by writing

$$g(\mu_i) = \beta_0 + \beta_1 x_i \tag{14.15}$$

as in (14.14), where now $g(\mu)$ may be any link function. Common distributions forming exponential families include binomial, multinomial, Poisson, normal, inverse Gaussian, gamma, and beta. The introduction of generalized linear models allowed regression methods to be extended immediately to all of these families. The unification of mathematical form meant that implementation of maximum likelihood, and likelihood ratio tests, could use the same algorithms with only minor changes in each particular case. Furthermore, for the canonical link it turns out (under relatively mild conditions on the x and y variables[4]) that the loglikelihood function is concave so that the MLE is unique. This guarantees that the maximum of the loglikelihood function will be found by the function maximizer (using *Newton's method*, i.e., iterative quadratic approximation) beginning with any starting value, and furthermore convergence will tend to be fast. Generalized linear models are part of most statistical software.

In addition to the canonical link, several other link functions are usually available in software. For example, it is usually possible to perform binomial regression using the probit link instead of the log odds, or logit link. Similarly, a Poisson regression could be performed using the identity link so that

$$\log \lambda_i = \beta_0 + \beta_1 x_i$$

is replaced by

$$\lambda_i = \beta_0 + \beta_1 x_i.$$

Occasionally, the identity link provides a better description of the data than the canonical link. This was the case in the non-cosine directional tuning example mentioned below.

Exponential families have special structure that make them easy to handle for theoretical purposes.[5] Their most important property in applications to generalized linear models is that, under relatively mild restrictions on the x and y values, with the canonical link the loglikelihood function is concave

[4]The regularity conditions insure non-degeneracy. For example, if there is only one x variable, it must take on at least 2 distinct values so that a line may be fitted. The y observations also must correspond to values that are possible according to the model; in dealing with proportions, for instance, the observed proportions can not all be zero.

[5]See Barndorff-Nielsen, O.E. (1978) *Information and Exponential Families in Statistical Theory*, Wiley.

and the β paramer vector has a unique MLE. This means that the algorithms used to fit generalized linear models with the canonical link are very robust.

The terminology "generalized linear model" should not to be confused with "the general linear model," which is the matrix form of regression and includes ANOVA. Also, the "linear" part of the terminology is misleading because the framework really includes *nonlinear* and *nonparametric* models, as well. Specifically, while linear models with the canonical link have especially nice properties, more generally in Equation (14.4) $f(x_i)$ does not need to be linear.

Example 14.3.2 (continued) The data in Figure 14.2 coming from Neuron 1 exhibited roughly Poisson variation. The fits shown there were based on the model $Y_i \sim P(\lambda(\theta))$ with $\lambda(\theta)$ given by Equation (14.10). This is a case of Poisson nonlinear regression using the identity link. □

14.4.2 Generalized linear models often involve multiple explanatory variables.

At the end of Section 14.1.2 we mentioned that the logistic regression model extends to multiple explanatory variables, and in Section 14.1.3 two-variable logistic regression was applied in Example 8.4.1. In Sections 14.2.1 and 14.2.2 a three-variable Poisson regression was used to produce Poisson ANOVA results for Example 14.2.1.

Multiple logistic regression may be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log \frac{p_i}{1-p_i} &= x_i^T \beta \end{aligned}$$

where β is the coefficient vector and x_i is the vector of values of the several explanatory variables corresponding the ith unit under study. Similarly, a multiple-variable generalized linear model may be written

$$\begin{aligned} Y_i &\sim p(y_i|\mu_i) \\ g(\mu_i) &= x_i^T \beta \end{aligned}$$

where $\mu_i = E(Y_i)$ and $g(\mu)$ is again the link function. Here is another example.

Example 14.4.1 A Poisson regression model for a hippocampal place cell Neurons in rodent hippocampus have spatially specific firing properties, whereby the spiking intensity is highest when the animal is at a specific location in an environment, and falls off as the animal moves further away from that point. Such receptive fields are called *place fields*, and neurons that have such firing properties are called *place cells*. Panel A of Figure 14.5 shows an example of the spiking activity of one such place cell, as a rat executes a free-foraging task in a circular environment. The rat's path through this environment is shown in blue, and the location of the animal at spike times is overlain in red. It is clear that the firing intensity is highest slightly to the southwest of the center of the environment, and decreases when the rat moves away from this point.

One very simple way to describe this hippocampal neural activity is to use a Poisson generalized linear model for spike counts in successive time bins while the rat forages, and to assume that the spike count depends on location in the environment. Let us take $Y_t \sim P(\lambda_t)$, with t signifying time, and then define

$$\lambda_t = \exp\left\{\alpha - \frac{1}{2}\begin{pmatrix} x(t)-\mu_x & y(t) & -\mu_y \end{pmatrix}\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}^{-1}\begin{pmatrix} x(t)-\mu_x \\ y(t)-\mu_y \end{pmatrix}\right\}. \tag{14.16}$$

The explanatory variables in this model are $x(t)$ and $y(t)$, the animal's x and y-position. The model parameters are $(\alpha, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$, where (μ_x, μ_y) is the center of the place field, $\exp\alpha$ is the maximum firing intensity at that point, and σ_x^2, σ_y^2, and σ_{xy} express how the intensity drops off away from the center. Note that it is the shape of the place field that is Gaussian, not the distribution of the spiking activity, which is again an inhomogeneous Poisson process. Panel B of Figure 14.5 displays a fit of the place field to the data in panel A. We will discuss models of this sort when we discuss point processes in Chapter **??**. □

Figure 14.5: *Spiking activity of a rat Hippocampal place cell during a free-foraging task in a circular environment. (A) Visualization of animal's path and locations of spikes. (B) Gaussian place field model for this neuron, with parameters fit by the method of maximum likelihood.*

Copyright, 2010, R.E. Kass, E.N. Brown, and U. Eden
REPRODUCTION OR CIRCULATION REQUIRES PERMISSION OF THE AUTHORS

Chapter 16

Nonparametric Regression

Nonlinear regression models, of the form (14.7), are useful whenever there is a specific function $f(x;\theta)$, as in (14.8), that does a good job of fitting the data. Such functions are often derived from differential equations, as in Example 14.3.1 on magnesium block of NMDA receptors, or from modifications of more elementary forms, as in Example 14.3.2 on non-cosine directional tuning of motor cortical neurons. In many problems, however, no such functional form is apparent, and it becomes highly desirable to be able to fit more general functions $f(x)$ to the data. This is the subject of *nonparametric regression*. The terminology "nonparametric" refers to the absence of a specified parametric form, such as in (14.8).

There are two general approaches to nonparametric regression. The first attempts to represent a function $f(x)$ in terms of a set of more primitive functions, such as polynomials, which are often called *basis functions*. The methods following the second approach estimate $f(x)$ by weighting the data (x_i, y_i) according to the proximity of x_i to x, a process called *local fitting*. We take up these two topics in Sections 16.2 and 16.3. in this chapter we focus mainly on functions of a single variable x, and the fitted values $\hat{y}_i = \hat{f}(x_i)$ produce fitted points $(x_i, \hat{y}_i)$ which collectively become a *smoothed* version

of the original data points. (See Example 16.1.1 below.) Thus, the nonparametric regression algorithm that is applied to the data is often called a *smoother*.

16.1 Smoothers

As always, we are concerned with the use of statistical models both to generate estimates of scientifically interesting quantities and to provide measures of uncertainty. For both purposes we need to begin by defining the quantities we want to know about. In linear regression and generalized linear models, and in their nonlinear counterparts, these are usually coefficients or simple functions of them such as x_{50} in Example 8.4.1 of Chapter 9, where we discussed propagation of uncertainty. With nonparametric regression the trick is to phrase inferential problems in terms of the function values themselves, which avoids any reference to a specific functional form. In fact, x_{50} in Example 8.4.1 could be considered an example of this, because even if some other function (some nonlinear, nonparametric function) were used to link log odds of perception with light intensity, that function would necessarily define a value x_{50} of the intensity at which the probability of perception would be 50%. Here is another example.

Example 16.1.1 Peak minus trough differences in response of an IT neuron Some neurons in the inferotemporal cortex (IT) of the macaque monkey respond to visual stimuli by firing action potentials in a series of sharply defined bursts. Rollenhagen and Olson (2005, *J Neurophysiology*) found that displaying an object image in the presence of a different, already-visible "flanker" image could enhance the strength of the oscillatory bursts. Figure 16.1 displays data (in the form of PSTHs) from an IT neuron under two conditions: in the first, a black patterned object was displayed as the stimulus for 600 milliseconds; in the second condition, prior to the display of the stimulus a pair of blue rectangles appeared (as a flanker image) and these remained illuminated while the patterned-object stimulus was displayed. Overlaid on the PSTHs are fits obtained by the nonparametric regression method BARS, which will be explained briefly in Section 16.2.6. In part B of Figure 16.1 the BARS fits are displayed together, to highlight the

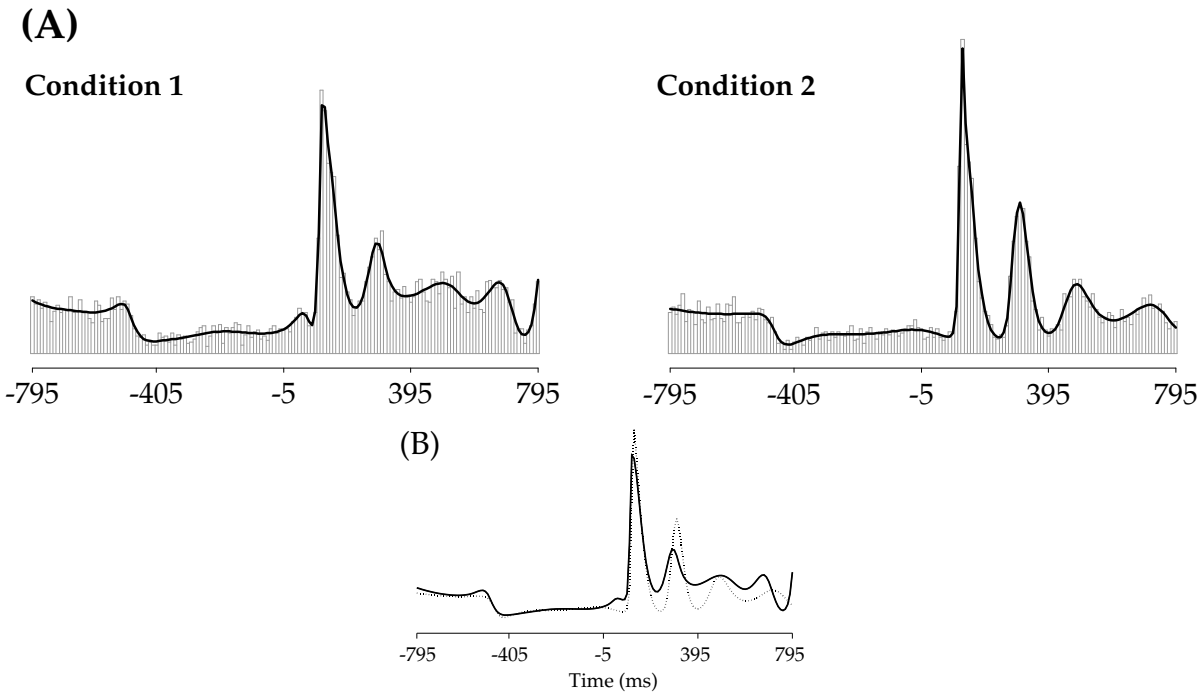


Figure 16.1: *(A) PSTHs and BARS fits for an IT neuron recorded by Rollenhagen and Olson (2005) under two conditions. (B) The two BARS fits are overlaid for ease of comparison. See text for explanation.*

differential response. One way to quantify the comparison is to estimate the drop in firing rate from its peak (the maximal firing rate) to the trough immediately following the peak in each condition. Let us call these peak minus trough differences, under the two conditions, ϕ^1 and ϕ^2. BARS was used to propagate the error (see DiMatteo, Genovese, and Kass, 2001, *Biometrika*). The results, for this neuron, were $\hat{\phi}^1 = 131.8(\pm 4.4)$, $\hat{\phi}^2 = 181.8(\pm 20.4)$ spikes per second, and $\hat{\phi}^1 - \hat{\phi}^2 = 50.0(\pm 20.8)$ spikes per second (where parenthetical values are SEs). □

A variety of nonparametric regression methods have been proposed. Some are linear and some nonlinear in a sense spelled out in Section 16.1.1.

16.1.1 Linear smoothers are fast.

We say that a nonparametric regression method results from a *linear smoother* if the fitted function values $\hat{f}(x_i)$ are obtained by linear operations on the data vector $y = (y_1, \ldots, y_n)^T$, that is, if we can write

$$(\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T = Hy \tag{16.1}$$

for a suitable matrix H. In other words, according to (16.1), for these linear smoothers, each fitted value is a linear combination of the data values y_i. The only nonlinear smoothing method we mention is that used in Example 16.1.1, BARS.

Linear smoothers are particularly easy to implement but, in practice, their big advantage is their speed. For large data sets, where speed becomes important, the relatively small number of arithmetic operations involved in (16.1) can make it possible to computer an answer.

16.1.2 For linear smoothers, the fitted function values are obtained via a "hat matrix," and it is easy to apply propagation of uncertainty.

The matrix H in (16.1) is called the *hat matrix*, because it produces estimates denoted with "hats". For example, in linear regression we have

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(see Chapter 13) so that

$$\begin{aligned} (\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T &= X\hat{\beta} \\ &= X(X^T X)^{-1} X^T y \end{aligned}$$

and the hat matrix is $H = X(X^T X)^{-1} X^T$. In the case of linear regression we are able to propagate uncertainty using the distribution of $\hat{\beta}$ (as we did, similarly, for logistic regression in Chapter **??**), but we could instead propagate the uncertainty from the distributions of the fitted values $X\hat{\beta}$: we simply

need the variance

$$\begin{aligned} V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) &= HV(Y)H^T \\ &= \sigma^2 HH^T. \end{aligned} \tag{16.2}$$

In the case of linear regression this simplifies because (as is easily checked) $H^T = H$ and $HH^T = H$ so that

$$V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) = \sigma^2 H.$$

For linear smoothers more generally, $H \neq HH^T$ but, in the case of data for which $V(Y_i) = \sigma^2$ with the Y_is being independent of each other, the variance formula (16.2) continues to hold, and it remains easy to apply propagation of uncertainty. In other words, even though we do not have an estimated parameter vector, such as $\hat{\beta}$, from which to compute quantities of interest and their SEs, we can often compute quantities of interest directly from the fitted values, as in the peak minus trough example above, and can then obtain SEs from the variance formula (16.2) together with the large-sample result that the fitted values are approximately normally distributed. Similarly, when linear smoothing methods extend to logistic or Poisson regression it again remains easy to propagate uncertainty.

16.2 Splines

16.2.1 Splines may be used to represent complicated functions.

Suppose $f(x)$ is a continuous function on an interval $[a, b]$. A famous theorem in mathematical analysis (the Weierstrass Approximation Theorem) says that $f(x)$ may be approximated arbitrarily well by a polynomial of sufficiently high order. One might therefore think that polynomials might provide flexibility in fitting functions to data. It turns out that they tend to perform rather badly, however. As illustrated in Figure 16.2, even a twentieth-order polynomial can fail to represent adequately a relatively well-behaved function in the presence of minimal noise.

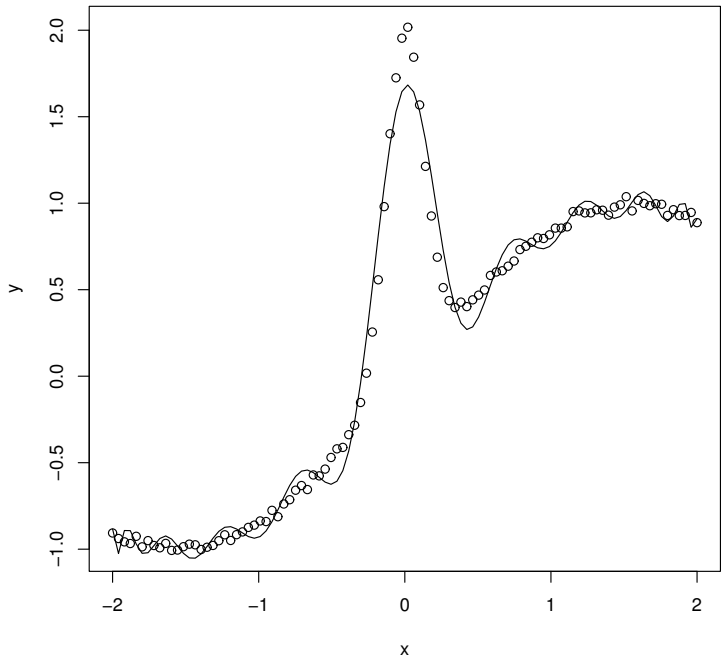


Figure 16.2: *Data simulated from function* $f(x) = \sin(x) + 2\exp(-30x^2)$ *together with twentieth-order polynomial fit (shown as line). Note that the polynomial is over-fitting (under-smoothing) in the relatively smooth regions of* $f(x)$*, and under-fitting (over-smoothing) in the peak. (In the data shown here, the noise standard deviation is 1/50 times the standard deviation of the function values.)*

The problem in Figure 16.2 is that the function $f(x)$ is not very close to being a low-order polynomial; in particular, it has a quite different form near $x = 0$ than it does as the magnitude of x increases. A possible solution here, and in other problems, is to glue together several pieces of polynomials. If the pieces are joined in such a way that the resulting function remains smooth, then it is called a *spline*. We will discuss cubic splines. Specifically, a cubic spline on an interval $[a, b]$ is a function $f(x)$ made up of cubic polynomials joined at a set of *knots*, $\xi_1, \xi_2, \ldots, \xi_p$, with $a < \xi_1 < \xi_2 < \cdots < \xi_p < b$, such

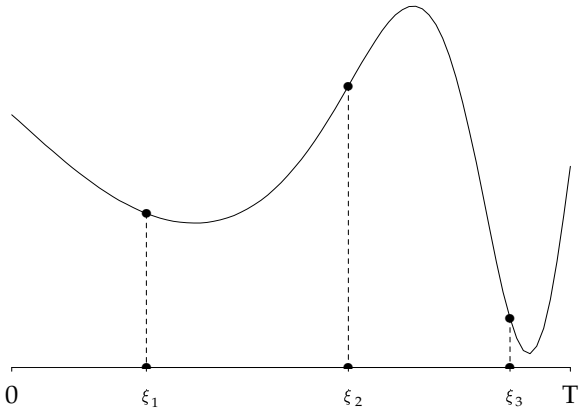


Figure 16.3: *A cubic spline with three knots, on an interval* $[0, T]$. *The function* $f(x)$ *depicted here is made up of distinct cubic polynomials (cubic polynomials with different coefficients) on each sub-interval* $[0, \xi_1]$, $[\xi_1, \xi_2]$, $[\xi_2, \xi_3]$, $[\xi_3, T]$.

that $f(x)$ is continuous and its first two derivatives $f'(x)$, and $f''(x)$ are also continuous. This restriction of continuity applies at the knots; in between the knots, each cubic polynomial is already continuous with continuous derivatives. A cubic spline is shown in Figure 16.3, and the result of fitting a cubic spline to the data of Figure 16.2 is shown in Figure 16.4.

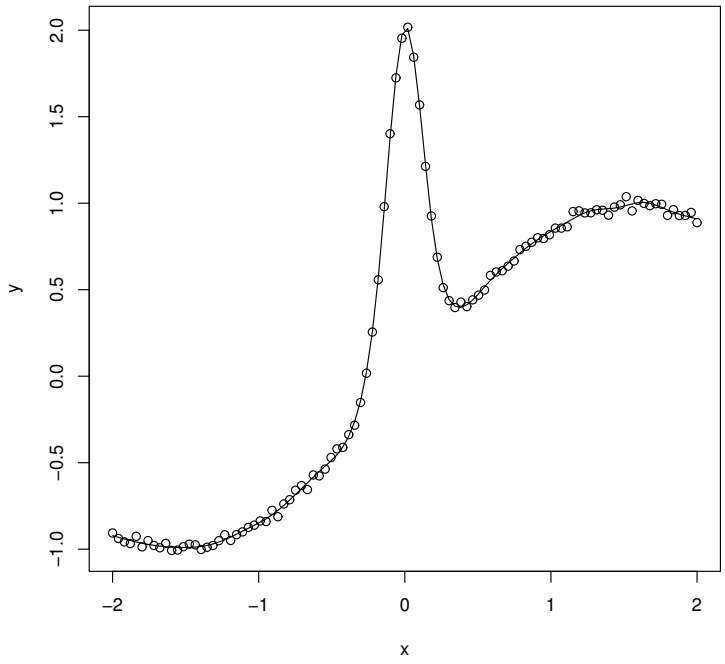


Figure 16.4: *A cubic spline fit to the data from Figure 16.2. The spline has knots* $(\xi_1, \xi_2, \ldots, \xi_7) = (-1.8, -.4, -.2, 0, .2, .4, 1.8)$.

16.2.2 Splines may be fit to data using linear models.

It is easy to define a cubic spline having knots at $\xi_1, \xi_2, \ldots, \xi_p$. Let $(x - \xi_j)_+$ be equal to $x - \xi_j$ for $x \geq \xi_j$ and 0 otherwise. Then the function

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ &+ \beta_4 (x - \xi_1)_+^3 + \beta_5 (x - \xi_2)_+^3 + \cdots + \beta_{p+3} (x - \xi_p)_+^3 \end{aligned} \quad (16.3)$$

is twice continuously differentiable, and is a cubic polynomial on each segment $[\xi_j, \xi_{j+1}]$. Furthermore, with $f(x)$ defined by (16.3),

$$Y_i = f(x_i) + \epsilon_i$$

becomes a standard linear regression model (assuming $\epsilon_i \sim N(0, \sigma^2)$, independently), so that standard software may be used to obtain spline-based curve fitting. Specifically, we define $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_4 = (x-\xi_1)+^3$, $\ldots$, $x_{p+3} = (x - \xi_p)_+^3$ and then regress Y on $x_1, x_2, \ldots, x_{p+3}$. To be concrete, let take an extremely simple special case. Suppose we have 7 data values $y_1, \ldots, y_7$ observed at 7 x values, $(x_1, \ldots, x_6) = (-3, -2, -1, 0, 1, 2, 3)$ and we want to fit a spline with knots at $\xi_1 = -1$ and $\xi_2 = 1$. Then we define $y = (y_1, \ldots, y_7)^T$, $x_1 = (-3, -2, -1, 0, 1, 2, 3)^T$, $x_2 = (9, 4, 1, 0, 1, 4, 9)^T$, $x_3 = (-27, -8, -1, 0, 1, 8, 27)^T$, $x_4 = (0, 0, 0, 1, 8, 27, 64)^T$, $x_5 = (0, 0, 0, 0, 0, 1, 8)^T$, and we regress y on x_1, x_2, x_3, x_4, x_5.

An important caveat in applying (16.3), however, is that the variables $x_1, x_2, \ldots, x_{p+3}$ will be highly correlated. The possibility of polynomial x variables being correlated was considered in Section 12.5.4 and again in Section 14.1.3. Here there are two good solutions to this problem. The first is to *orthogonalize* the x variables. The trick of subtracting the mean, used in the earlier sections, is a special case of orthogonalization. The general method is to first replace x with $x_1^* = x - \bar{x}$; then regress $(x_1^*)^2$ on x_1^* and replace x^2 with x_2^* defined to be the residual from that regression; then regress $(x_1^*)^3$ on x_1^* and x_2^* and replace x^3 with x_3^* defined to be the residual from that regression; etc., continuing through the remainder of the regression variables to get a new set of variables $x_1^*, x_2^*, \ldots, x_{p+3}^*$ which are used instead of $x_1, x_2, \ldots, x_{p+3}$. The second, more commonly-applied alternative is to use a different version of splines, knowns as *B-splines*. These have similar properties but do not produce high correlation among the regression variables. A variant of B-splines, known as *natural splines*, assumes the function is linear for x outside a specified range—which is often taken to be the range of the data (i.e., the function is linear for $x < x_{min}$ and $x > x_{max}$ where x_{min} and x_{max} are the smallest and largest values of x in the data). Natural splines are often recommended.

16.2.3 Splines are also easy to use in binomial or Poisson regression models.

Note that (16.3) may also be applied in logistic regression or Poisson regression. When splines are used in regression models, they are often called

regression splines.

Example 1.1.1 (continued) In Chapter 1 we discussed the problem of describing a neural response to a stimulus under two different experimental conditions in the context of recordings made from the SEF. In Chapter 8 we returned to the example to describe the value of smoothing the PSTH, using Figure 8.3 to illustrate. We did not, however, say specifically how the smoothing was done. We obtained the smooth curve in Part B of Figure 8.3 by fitting a Poisson regression spline. Specifically, spike counts Y were pooled across trials in 10 millisecond bins centered at times $x = -295, -285, -275, \ldots, 635, 645$ relative to appearance of the cue at time $x = 0$. Then the statistical model was

$$\begin{aligned} Y_i &\sim P(\lambda_i) \\ \log \lambda_i &= f(x_i) \end{aligned}$$

with $f(x)$ being a regression spline having knots at $-200, 200$. The fitted values $\hat{f}(x_i)$ were obtained using generalized linear model software and x_{max} was the value of x_i at which maximum among the $\hat{f}(x_i)$ values occurred. (Interpolation could be have been used to get a more refined maximum, but this was not considered necessary.) In Figure 8.3, the arrow indicating the maximum of the fitted curve was plotted at $x = x_{max}$. It is straightforward to obtain a SE for $x = x_{max}$ by propagation of uncertainty. □

16.2.4 With regression splines, the number and location of knots controls the smoothness of the fit.

Splines are very easy to use because the problem of spline fitting may be formulated in terms of a linear model. This, however, assumes that the knot set $\xi_1, \xi_2, \ldots, \xi_p$ has been determined. The choice of knots can be consequential: with more knots, the spline has greater flexibility, but also provides less smoothness. In addition, the placement of knots can be important. Figure 16.5 displays three alternative spline fits. The first two use splines with 5 and 15 knots having locations that are equally-spaced according to the quantiles of x so, for example, 5 knots would be placed at the $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}$ quantiles. Spacing the knots according to the quantiles of x allows more knots to be placed where there are more data values. The third spline uses

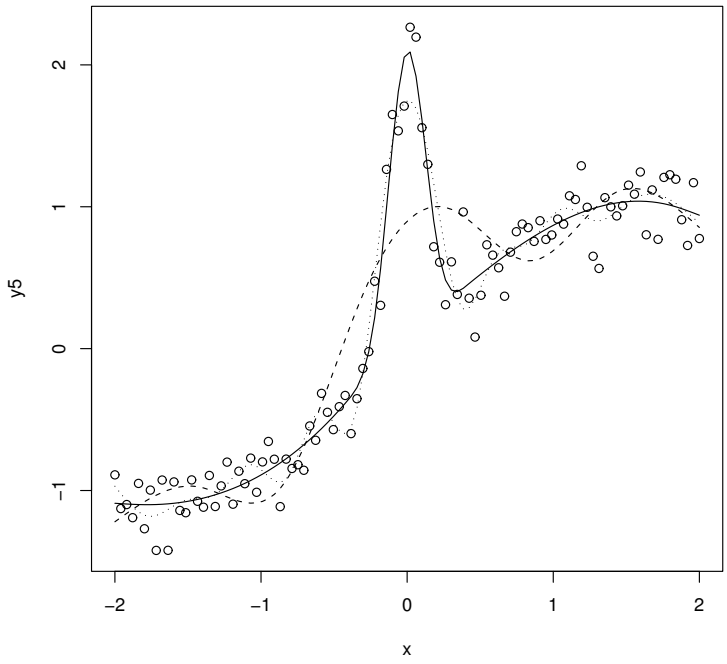


Figure 16.5: *Three cubic spline fits to data generated from the same test function as Figure 16.2, but with more noise. Splines with 5 and 15 knots are shown (dashed and dotted lines), with knot locations selected by default in R. The spline with 5 knots provides more smoothing than the spline with 15 knots and, as a result, does a poorer job of capturing the peak in the function. The spline shown in the solid line has 7 knots chosen to be* $\xi = (-1.8, -.4, -.2, 0, .2, .4, 1.8)$.

7 knots chosen by eye. The spline with 7 knots fits well because 5 knots are placed in the middle of the range, where the function variation is large, while only 2 are placed on the flanks where the variation is small.

16.2.5 Smoothing splines are splines with knots at each x_i, but with reduced coefficients obtained by penalized ML.

The problem of choosing knots may be solved in various ways, and in many situations it is adequate to select knots based on preliminary examination of the data and/or some knowledge of the way the function $f(x)$ is likely to behave. This is admitedly somewhat arbitrary, and two kinds of alternatives have been proposed that are more automated.

The first approach is to use a large number of knots, but to reduce, or "shrink," the values of the coefficients. One intuition here is that using a large number of knots in a regression spline would allow it to follow the function well, but would make it very wiggly; reducing the size of the coefficients will tend to smooth out the wiggles. A second intuition is obtained by replacing the least-squares problem of minimizing the sum of squares

$$SS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

with the *penalized least squares* problem of minimizing the penalized sum of squares

$$PSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

where λ is a constant. Here, the squared second derivative is a *roughness penalty*: wherever $(f''(x))^2$ is large, the function is fluctuating substantially, and the integral of this quantity is a measure of the total fluctuation, or roughness. Thus, the value of the coefficient vector β^* that minimizes PSS will achieve some compromise between fitting the y_i values and keeping the function smooth. As λ increases, the resulting fit becomes increasingly smooth, and in the limit $\lambda \to \infty$ it becomes a line. It turns out that the solution to the penalized least squares problem is a cubic spline with knots at every value of x_i, but with coefficients that are smaller in magnitude than those of the regression spline with knots at every x_i (which would correspond to taking $\lambda = 0$). This solution is called *a smoothing spline*.

Smoothing spline technology has a strong theoretical foundation, and is among the most widely-used methods for nonparametric regression. There

is also much well-developed software for smoothing splines. In the case of binomial or Poisson regression, the smoothing spline will maximize a penalized likelihood.

There remains the problem of choosing λ. Various alternative choices of λ may be tried. A more automatic method of choosing λ uses *cross-validation* (or *generalized cross-validation*), which will be discussed below. It is also possible to use a version of ML, often called *generalized maximum likelihood*. A smoothing spline fit to the data of Figure 16.2 is visually indistinguishable from the spline fit in Figure 16.4.

16.2.6 A method called BARS chooses knot sets automatically, according to a Bayesian criterion.

One defect of smoothing spline technology, and many other nonparametric methods, is that it assumes the degree of smoothness of $f(x)$ remains about the same across its domain, i.e., throughout the range of x values. An alternative is to devise a method that selects good knot sets based on the data. One of the most successful such procedures is called BARS (DiMatteo, Genovese, and Kass, 2001, *Biometrika*). In Figure 1.7 of Example 1.2.4 BARS was applied to data from an electrooculogram, which produces voltage traces that are similar to many others, including EEG, ECoG, and LFP. There, BARS was able to retain the high-frequency signal (the sudden drop and sudden increase in voltage associated with an eye blink) while filtering high-frequency noise. In Figure 16.1 of Section 16.1 we displayed BARS fits to two peristimulus time histograms. Briefly, BARS uses a Bayesian framework, and produces a posterior probability distribution on knot sets; knot sets are then generated by simulation from the posterior distribution; a fitted curve is easily obtained from this distribution; and propagation of uncertainty is used to provide standard errors or intervals for quantities of interest. Figure 16.6 compares the BARS and smoothing spline fits to the data from Figure 16.5.

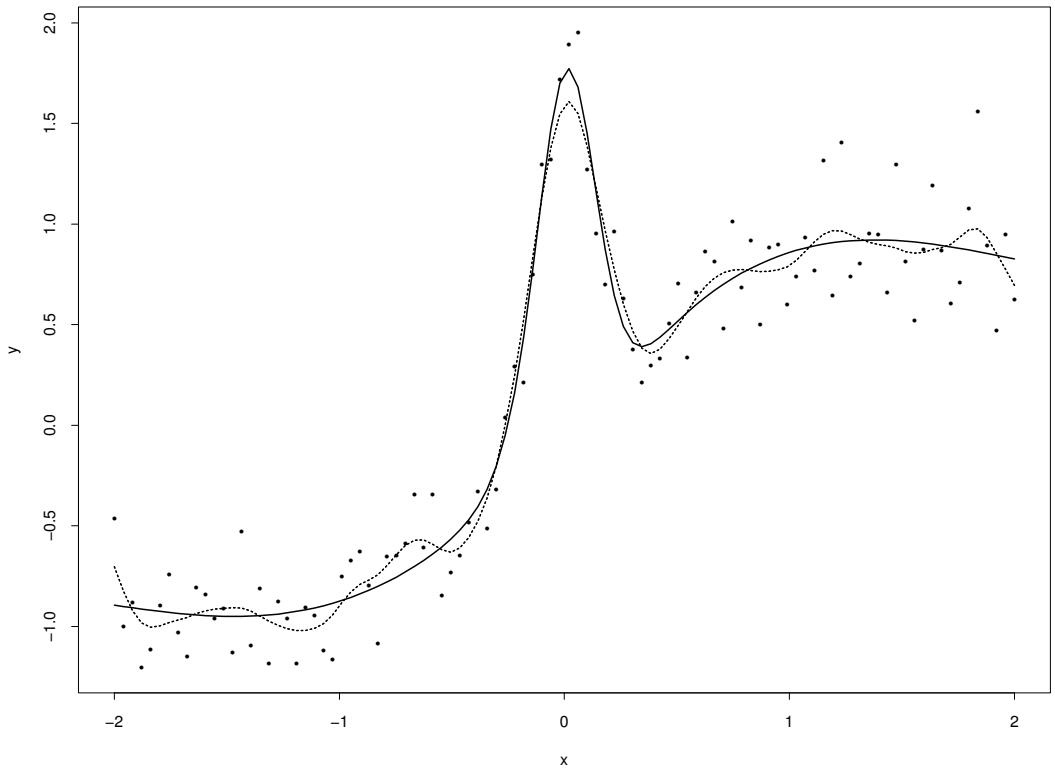


Figure 16.6: *Data from the test function of Figure 16.2, but with more noise, as in Figure 16.5, together with smoothing spline fit (dotted line) and BARS fit (solid line).*

16.2.7 Cross-validation may be used to estimate MSE and control smoothness.

16.3 Local Fitting

16.3.1 Functions may be estimated by local fitting procedures.

The second general approach to nonparametric regression is to use local fitting. Recall that in ordinary linear regression, the regression line is the

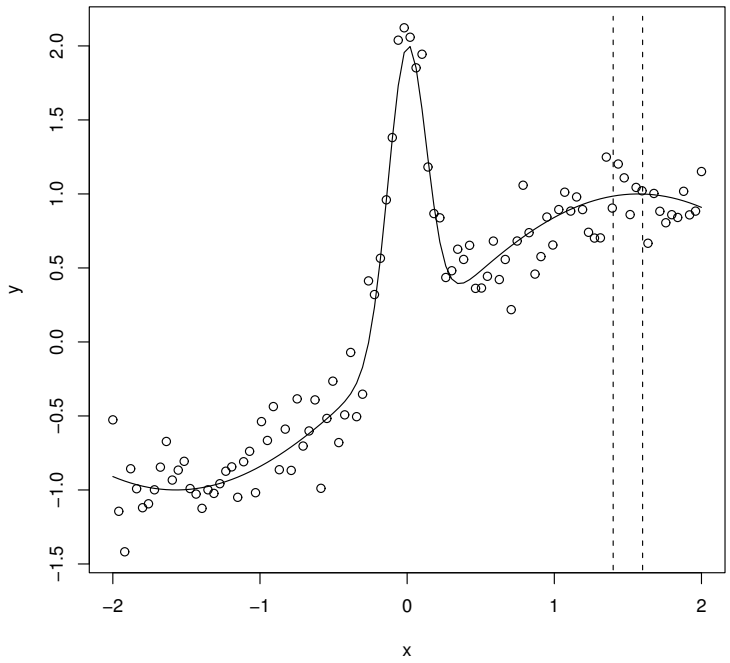


Figure 16.7: *Data simulated from function* $f(x) = \sin(x) + 2\exp(-30x^2)$ *(shown as dark line). The idea of local fitting begins with the notion that, just as in linear regression, for large data sets, the regression curve* $f(x)$ *at* $x = 1.5$ *should average the* y*-values among the points within the dashed lines. However, for smaller data sets, like that shown here, the region within the dashed lines contains relatively few points.*

expectation of Y as a function of x: we have $E(Y_i) = \beta_0 + \beta_1 x_i$ and could extend this to some newly-observed value of x by writing

$$E(Y|x) = \beta_0 + \beta_1 x. \tag{16.4}$$

In (16.4) we mean to include the case in which the data collection process makes it more reasonable to think of x as non-random. However, we have written $E(Y|x)$ to indicate functional dependence on x because if X is also a random variable then the regression line becomes the conditional expectation of Y given $X = x$. This was discussed in Section 4.2.4.

Now, just as the expectation of a random variable is generally estimated by a sample mean, so the conditional expectation in (16.4) may be estimated as the mean of y_i values for which $X = x_i$, at least approximately. This is indicated in Figure 4.3. When we generalize (16.4) to

$$E(Y|x) = f(x) \tag{16.5}$$

we may, in principle, also estimate $f(x)$ by averaging y_i values for $X = x_i$, approximately, as illustrated in Figure 16.7. For large data sets the average gives an answer very close to the expectation. An immediate issue, however, is how to choose the size of the *window* (between the dashed lines in Figure 16.7). Furthermore, in estimating $f(x)$ even with moderate-size data sets, it is possible to improve on the arithmetic mean among y_i values corresponding to x_i near x. For instance, in Figure 16.7, there are not many values of x_i that are very close to any particular x. The idea of local fitting is to consider x_i values that are somewhat more distant from x, but to *weight* the various x_i values according to their proximity to x.

Two different ways to accomplish local fitting are distinguished by the names *kernel regression* and *local polynomial regression*.

16.3.2 Kernel regression estimates $f(x)$ with a weighted mean defined by a pdf.

Recall that, given a set of positive numbers $w_1, \ldots, w_n$, the weighted mean of $y_1, \ldots, y_n$ is

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

where w_i becomes the weight attached to the ith value. In kernel regression, each value $f(x)$ is estimated as a weighted mean of the observations y_i, with the weights increasing as x_i gets closer to x. The weights are defined by

$$w_i = K(\frac{x - x_i}{h}) \tag{16.6}$$

for a suitable function $K(u)$, which is called a *kernel*. The constant h is usually called the *bandwidth* (by analogy with similar methods in spectral analysis). The most commonly-used kernel is the $N(0, 1)$ pdf, in which case h effectively plays the role of a standard deviation, i.e., we have $w_i \propto K_h(x-x_i)$ where $K_h(u)$ is the $N(0, h^2)$ pdf. That is, $K(\frac{x-x_i}{h})$ is proportional to a normal pdf centered at zero having standard deviation h. This puts very nearly zero weight on y_i values for which $|x - x_i| > 3h$. Because many applications of smoothing arise in signal processing, some of the terminology is taken from that domain. In particular, when a normal kernel is used, it is often called a *normal filter* or *Gaussian filter*.

More generally, any pdf could be used as a kernel. The formula for the kernel-regression fit is

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}. \tag{16.7}$$

Example ?? (continued): Previously we provided some results from a study of action potential width as a function of the preceding ISI, and Figure 9.6 displayed a plot of some data from one neuron recorded from rat barrel cortex in a slice preparation. A portion of the data are shown again here, in Figure 16.8. Only the data points for which ISI was less than 200 milliseconds are displayed, and the analysis here only considered this truncated data set. Kernel regression, with a normal kernel, produced the fitted relationship shown by the dashed line in the figure. The bandwidth used was 30 milliseconds. □

The choice of bandwidth h in kernel regression is important, and corresponds roughly to the choice of number of knots with regression splines: when h is small, the estimate tends to follow the data closely, but is very rough, while when h is large the estimate becomes smooth but may ignore places

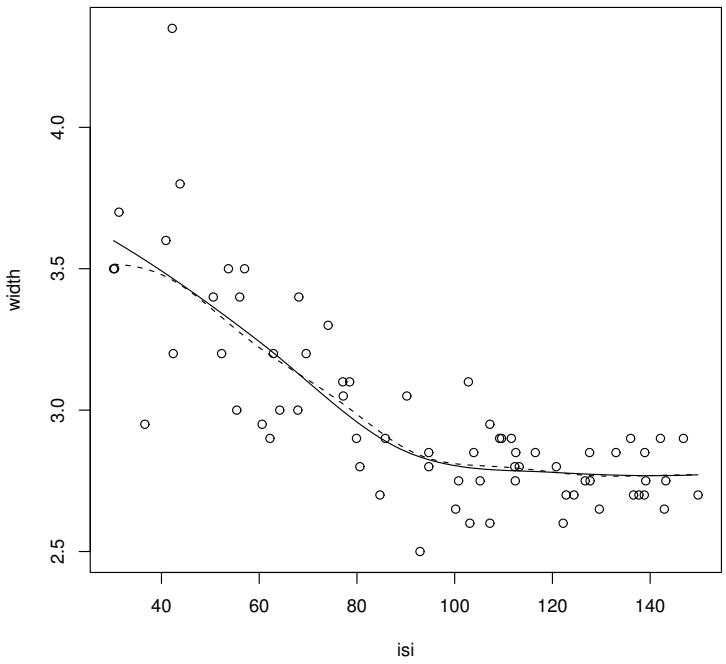


Figure 16.8: *Data showing the relationship of spike width to preceding ISI length for a neuron recorded in slice preparation. A kernel regression estimator is superimposed on the plot (dashed line) together with a local linear fit (solid line).*

where the function seems to vary. Bandwidth selection involves a "bias versus variance" trade-off: small h reduces bias (and increases variance) while large h reduces variance (but increases bias).

16.3.3 Local polynomial regression solves a weighted least squares problem with weights defined by a kernel.

A second idea in local fitting $f(x)$ is to solve a weighted least-squares problem defined at x by suitable weights $w_i = w_i(x)$. In particular, *local linear regression* at x minimizes

$$WSS(x) = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1(x - x_i))^2 \tag{16.8}$$

where the weights w_i are defined in terms of a kernel, as in (16.6). A normal pdf may be used as the kernel, but an alternative is

$$K(u) = (1 - |u|^3)^3$$

for $|u| < 1$ and $K(u) = 0$ otherwise. The latter form of the kernel is used as the default in software that comes under the name *locfit*, available for both R and Matlab. Extensive study of this methodology has shown that local linear regression is effective in many situations.

Example ?? (continued): In Figure 16.8 we displayed a plot of some actional potential width data together with a nonparametric regression fit based on a normal kernel (or Gaussian filter). A local linear fit is also shown in Figure 16.8. In this example the local linear fit is nearly identical with the kernel regression fit. □

As with kernel regression, in local polynomial regression there remains a choice of bandwidth. For further information about bandwidth selection, and local fitting in general, see Loader (1999, *Local Regression and Likelihood*).

An important feature of local polynomial regression, and of the *locfit* software, is that it may be extended to non-normal families such as binomial and Poisson. The idea is very simple. In place of the locally weighted sum of squares in (16.8) we can, for any value of the explanatory variable x_i, maximize a locally weighted loglikelihood having the form

$$WLL(x) = \sum_{i=1}^{n} w_i \ell(\beta_0 - \beta_1 x_i).$$

More specifically, in the case of binomial local linear fitting, with $Y_i \sim B(n_i, p_i)$, we have

$$\begin{aligned} WLL(x) &= \sum_{i=1}^{n} w_i(y_i \log p_i + (n - y_i)\log(1 - p_i)) \\ \log \frac{p_i}{1 - p_i} &= \beta_0 + \beta_1 x_i. \end{aligned}$$

Maximizing this loglikelihood for each successive x_i produces the fit at x_i.

16.3.4 Simple theoretical considerations lead to bandwidth recommendations for linear smoothers.

Recall that $MSE = \text{BIAS}^2 + \text{Variance}$. A minimal requirement of an estimator, in large samples, is that its bias and variance vanish (as $n \to \infty$). Consider estimation of $f(x)$ at the single point x. A linear smoother is, at x, a linear combination of the data response values y_i, so that the estimator may be written in the form

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i$$

where $w_i(x)$ emphasizes that the weights are determined for each x. We want

$$E(\hat{f}(x)) \to f(x) \tag{16.9}$$

and

$$V(\hat{f}(x)) \to 0. \tag{16.10}$$

Because $E\hat{f}(x) = \sum w_i(x) f(x_t)$, the bias vanishes, as stated in (16.9), if the weights $w_i(x)$ become concentrated near x and the function $f(x)$ is smooth. For the weights to become concentrated it is sufficient that $\sum (i-x)^2 w_i(x) \to 0$. Assuming $V(Y_i) = \sigma^2$ (or, at least, that the variances do not vary rapidly), the variance vanishes if $\sum w_i(x)^2 \to 0$. These sorts of conditions on the weights, to guarantee (16.9) and (16.10), need to be assumed by any large-sample theoretical justification of a linear smoothing method. An explicit expression for the MSE of kernel estimators was given by Gasser and Müller

(1984; Estimating regression functions and their derivatives by the kernel method, *Scandavian J. Statist.*, 11: 171-185). This allows a theoretical bias versus variance trade-off, i.e., a formula for bandwidth selection as a function of n.

16.4 Density Estimation

Suppose we have a sample $U_1, \ldots, U_n$ from a distribution having pdf $f_U(u)$. If $f_U(u)$ is specified by a parameter vector θ (so that $f_U(u) = f_U(u|\theta)$) we may apply ML to estimate θ and thereby determine $f_U(u)$. Sometimes, however, we do not wish to assume a particular parametric form, yet we still want to obtain an estimate of the pdf. This presents the problem of nonparametric *density estimation*.

16.4.1 Kernels may be used to estimate a pdf.

One of the most popular ways to estimate a density is to apply a kernel, in the form we give below. It is possible to view the problem of density estimation as a special case of the problem of nonparametric regression, and in particular to derive a kernel density estimate from (16.7). We provide some discussion of this in the next subsection. Here we consider a somewhat simpler motivation for the procedure.

Recall that, for small h,

$$f_U(u) \approx \frac{P(u-h<U<u+h)}{2h}.$$

Then a direct estimate of $f_U(u)$ is

$$\hat{f}_U(u) \approx \frac{\text{no. obsn's falling in (u-h, u+h)}}{2nh}. \tag{16.11}$$

This estimate can be written in terms of the kernel $K(z) = \frac{1}{2}$ for $|z| < 1$ and 0 otherwise: we have

$$\hat{f}_U(u) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{u-u_i}{h}). \tag{16.12}$$

This direct (or "näive") estimate is also essentially a histogram with bins centered at the observations: if we normalize an ordinary histogram to give it the form of a pdf we get

$$\hat{f}_{U,hist}(u) = \frac{\text{no. obsn's in same bin as u}}{2nh}.$$

Both the histogram and the estimate in (16.11) suffer from being rectangular, and thus unable to produce a smooth curve as an estimate of the pdf. If we instead replace the kernel $K(z) = \frac{1}{2}$ for $|z| < 1$ with a smooth kernel, such as the normal pdf, we will get a smooth density estimate. In this general form the result of applying (16.12) produces what is known as a *kernel density estimate*. Kernel density estimation may be considered a way of getting a smooth density to replace the histogram. The normal (Gaussian) kernel is often used, though other choices are generally available in density estimation software.

As in kernel regression, the bandwidth parameter h is important. As we discussed in Chapter **??**, choice of bin width is similarly important when using a histogram. For small h the estimate will tend to follow the data, but will be wiggly, while for large h the estimate will be smooth, but may not respond quickly to bunching of points that should indicate an increase in probability density. A variety of methods have been proposed for automatic selection of h, but many analysts choose h based on examination of the data, and experience with similar data (often picking a round number for h partly as in indication of the arbitrariness in the choice).

Example ?? (continued): We now examine only the ISI component of the data considered earlier, including all ISIs under 1000 milliseconds. A Gaussian kernel density estimate is shown in Figure 16.9 superimposed on an ISI histogram. □

16.4.2 Other nonparametric regression methods may be used to estimate a pdf.

It is possible to view density estimation as a problem in binary nonparametric regression: we consider a very fine grid of values of u and define a

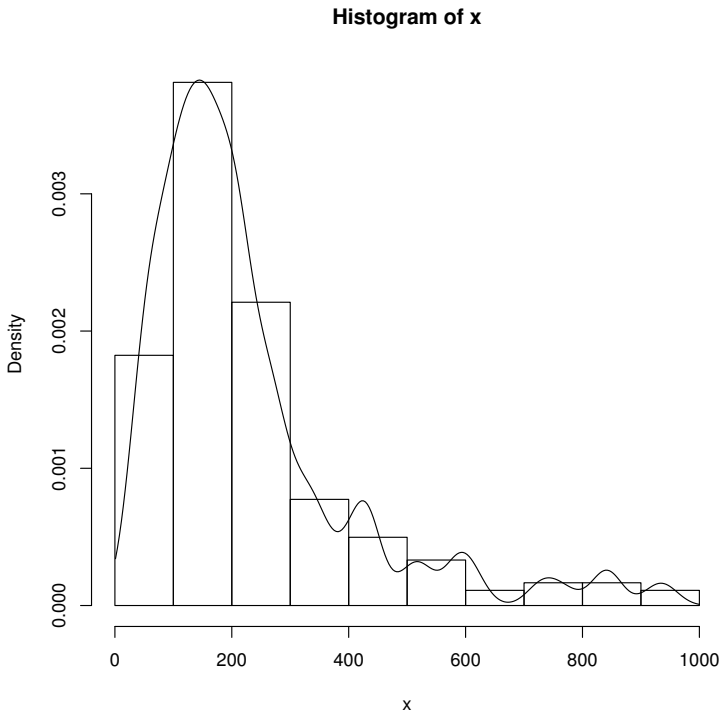


Figure 16.9: *A Gaussian kernel density estimator superimposed on an ISI histogram for the ISI data of Figure 16.8. Here the histogram bin width was chosen using the "oversmoothed" rule from Scott (1992,* Multivariate Density Estimation*, p. 55), which produced 10 bins of width 100 milliseconds; the bandwidth of the Gaussian kernel was set at 100 milliseconds.*

variable that is 1 whenever a grid interval contains an observation, and 0 otherwise; estimating the expectation of these binary random variables amounts to estimating the pdf of U.

> *Details:* Let us suppose we wish to obtain $\hat{f}_U(u)$ at some grid of u values, as we would in order to plot $\hat{f}_U(u)$, and let us write the grid as $x_1, x_2, \ldots, x_m$, so that the pairs we would plot would be $(x_j, \hat{f}_U(x_j))$, for $j = 1, \ldots, m$. We are using the notation x_j to distinguish the grid points from the random variable observations u_i. For the purpose of plotting this pdf we would, typically—

as in plotting any function—choose m to be a fairly large value (such as 200), so that the plotted graph would not appear jagged. For convenience, let us take $\Delta x = x_j - x_{j-1}$, assuming the grid points to be equally spaced. Then taking a large m is equivalent to making Δx small. Let us assume that the grid is chosen to be sufficiently fine that there is at most 1 observation u_i in any given interval (x_{j-1}, x_j). (We may take $x_0 = x_1 - \Delta x$.) Viewing this procedure probabilistically, we can set up our grid prior to observing $U_1, \ldots, U_n$ and take it to be sufficiently fine that the probability of obtaining more than one observation in any given interval is negligible. The probability that an observation U_i will fall in interval (x_{j-1}, x_j) is approximately $f_U(x_j)\Delta x$. (We could improve the approximation somewhat by instead taking it to be $f_U(\frac{x_j+x_{j-1}}{2})\Delta x$, but will ignore this distinction here, as we are assuming Δx is small, so that $f_U(x_j) \approx f_U(\frac{x_j+x_{j-1}}{2})$.) Now let $Y_j = 1$ if the interval (x_{j-1}, x_j) contains an observation U_i (for some i) and 0 otherwise. Then Y_j, for $j = 1, \ldots, m$, forms a sequence of binomial random variables with

$$E(Y_j) \approx n f_U(x_j)\Delta x. \tag{16.13}$$

Because Y_j varies with j, it varies also with x_j and we may think of this expectation as a conditional expectation $E(Y_j|x_j)$; and because nonparametric regression methods estimate such conditional expectations, we may apply a kernel method to the estimation of the left-hand side of (16.13) in order to obtain an estimate of $f_U(u)$, which appears on the right-hand side. Specifically, writing $x = x_j$ and applying (16.7), we have

$$\begin{aligned} n\hat{f}_U(x)\Delta x &= \frac{\sum_{i=1}^n K(\frac{x-x_i}{h})y_i}{\sum_{i=1}^n K(\frac{x-x_i}{h})} \\ &= \frac{\Delta x \sum_{i=1}^n K(\frac{x-x_i}{h})y_i}{\sum_{i=1}^n K(\frac{x-x_i}{h})\Delta x}. \end{aligned} \tag{16.14}$$

For large m, the sum appearing in the denominator is approximately equal to an integral and, because $K(z)$ (where z is used to stand for the generic argument of the kernel) is itself a pdf, it is easy to show that the integral is nh. In the numerator, we note

that $y_j = 0$ except when there is an observation u_i in $(x_{j-1} - x_j)$, in which case $x_j \approx u_i$. Plugging these into (16.14), canceling Δx, and replacing x with u then gives (16.12). □

Many alternatives to kernel density estimation have been studied, and some of these can provide better estimates in certain situations. Indeed, as the derivation of (16.12) given above indicates, any method of nonparametric regression for binary data may be used: after the regression estimate is normalized so that it integrates to 1 it may be considered a density estimate. The virtue of kernel density estimation is that it is fast, easy, and reasonably effective. When some imprecision in the estimate is tolerable, it is often a method of choice.