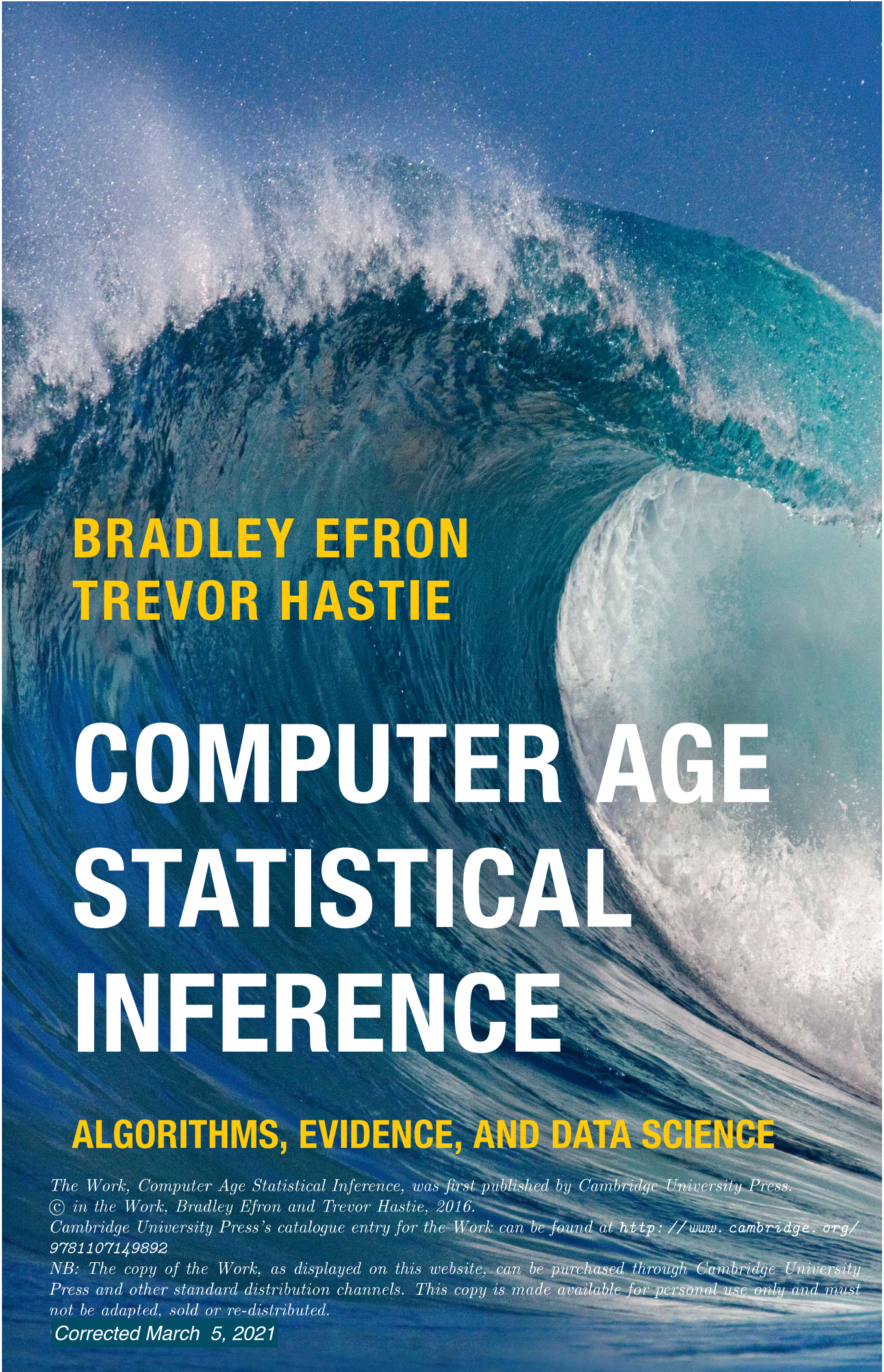

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE

Corrected March 5, 2021

Computer Age Statistical Inference

Algorithms, Evidence, and Data Science

Bradley Efron Trevor Hastie

Stanford University

To Donna and Lynda

Contents

Preface

Statistical inference is an unusually wide-ranging discipline, located as it is at the triple-point of mathematics, empirical science, and philosophy. The discipline can be said to date from 1763, with the publication of Bayes' rule (representing the philosophical side of the subject; the rule's early advocates considered it an argument for the existence of God). The most recent quarter of this 250-year history—from the 1950s to the present—is the "computer age" of our book's title, the time when computation, the traditional bottleneck of statistical applications, became faster and easier by a factor of a million.

The book is an examination of how statistics has evolved over the past sixty years—an aerial view of a vast subject, but seen from the height of a small plane, not a jetliner or satellite. The individual chapters take up a series of influential topics—generalized linear models, survival analysis, the jackknife and bootstrap, false-discovery rates, empirical Bayes, MCMC, neural nets, and a dozen more—describing for each the key methodological developments and their inferential justification.

Needless to say, the role of electronic computation is central to our story. This doesn't mean that every advance was computer-related. A land bridge had opened to a new continent but not all were eager to cross. Topics such as empirical Bayes and James–Stein estimation could have emerged just as well under the constraints of mechanical computation. Others, like the bootstrap and proportional hazards, were pureborn children of the computer age. Almost all topics in twenty-first-century statistics are now computer-dependent, but it will take our small plane a while to reach the new millennium.

Dictionary definitions of statistical inference tend to equate it with the entire discipline. This has become less satisfactory in the "big data" era of immense computer-based processing algorithms. Here we will attempt, not always consistently, to separate the two aspects of the statistical enterprise: algorithmic developments aimed at specific problem areas, for instance

random forests for prediction, as distinct from the inferential arguments offered in their support.

Very broadly speaking, algorithms are what statisticians do while inference says why they do them. A particularly energetic brand of the statistical enterprise has flourished in the new century, *data science*, emphasizing algorithmic thinking rather than its inferential justification. The later chapters of our book, where large-scale prediction algorithms such as boosting and deep learning are examined, illustrate the data-science point of view. (See the epilogue for a little more on the sometimes fraught statistics/data science marriage.)

There are no such subjects as Biological Inference or Astronomical Inference or Geological Inference. Why do we need "Statistical Inference"? The answer is simple: the natural sciences have nature to judge the accuracy of their ideas. Statistics operates one step back from Nature, most often interpreting the observations of natural scientists. Without Nature to serve as a disinterested referee, we need a system of mathematical logic for guidance and correction. Statistical inference is that system, distilled from two and a half centuries of data-analytic experience.

The book proceeds historically, in three parts. The great themes of classical inference, Bayesian, frequentist, and Fisherian, reviewed in Part I, were set in place before the age of electronic computation. Modern practice has vastly extended their reach without changing the basic outlines. (An analogy with classical and modern literature might be made.) Part II concerns early computer-age developments, from the 1950s through the 1990s. As a transitional period, this is the time when it is easiest to see the effects, or noneffects, of fast computation on the progress of statistical methodology, both in its theory and practice. Part III, "Twenty-First-Century topics," brings the story up to the present. Ours is a time of enormously ambitious algorithms ("machine learning" being the somewhat disquieting catchphrase). Their justification is the ongoing task of modern statistical inference.

Neither a catalog nor an encyclopedia, the book's topics were chosen as apt illustrations of the interplay between computational methodology and inferential theory. Some missing topics that might have served just as well include time series, general estimating equations, causal inference, graphical models, and experimental design. In any case, there is no implication that the topics presented here are the only ones worthy of discussion.

Also underrepresented are asymptotics and decision theory, the "math stat" side of the field. Our intention was to maintain a technical level of discussion appropriate to Masters'-level statisticians or first-year PhD stu-

dents. Inevitably, some of the presentation drifts into more difficult waters, more from the nature of the statistical ideas than the mathematics. Readers who find our aerial view circling too long over some topic shouldn't hesitate to move ahead in the book. For the most part, the chapters can be read independently of each other (though there is a connecting overall theme). This comment applies especially to nonstatisticians who have picked up the book because of interest in some particular topic, say survival analysis or boosting.

Useful disciplines that serve a wide variety of demanding clients run the risk of losing their center. Statistics has managed, for the most part, to maintain its philosophical cohesion despite a rising curve of outside demand. The center of the field has in fact moved in the past sixty years, from its traditional home in mathematics and logic toward a more computational focus. Our book traces that movement on a topic-by-topic basis. An answer to the intriguing question "What happens next?" won't be attempted here, except for a few words in the epilogue, where the rise of data science is discussed.

Acknowledgments

We are indebted to Cindy Kirby for her skillful work in the preparation of this book, and Galit Shmueli for her helpful comments on an earlier draft. At Cambridge University Press, a huge thank you to Steven Holt for his excellent copy editing, Clare Dennison for guiding us through the production phase, and to Diana Gillooly, our editor, for her unfailing support.

Bradley Efron
Trevor Hastie
Department of Statistics
Stanford University
May 2016

Notation

Throughout the book the numbered † sign indicates a technical note or reference element which is elaborated on at the end of the chapter. There, next to the number, the page number of the referenced location is given in parenthesis. For example, `lowess` in the notes on page 11 was referenced via a $\dagger_1$ on page 6. Matrices such as $\mathbf{\Sigma}$ are represented in bold font, as are certain vectors such as $\boldsymbol{y}$, a data vector with n elements. Most other vectors, such as coefficient vectors, are typically not bold. We use a dark green `typewriter` font to indicate data set names such as `prostate`, variable names such as `prog` from data sets, and `R` commands such as `glmnet` or `locfdr`. No bibliographic references are given in the body of the text; important references are given in the endnotes of each chapter.

Part I

Classic Statistical Inference

1

Algorithms and Inference

Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path toward Earth. It may seem surprising that any one theory can cover such an amorphous target as "learning from experience." In fact, there are *two* main statistical theories, Bayesianism and frequentism, whose connections and disagreements animate many of the succeeding chapters.

First, however, we want to discuss a less philosophical, more operational division of labor that applies to both theories: between the *algorithmic* and *inferential* aspects of statistical analysis. The distinction begins with the most basic, and most popular, statistical method, averaging. Suppose we have observed numbers $x_1, x_2, \ldots, x_n$ applying to some phenomenon of interest, perhaps the automobile accident rates in the $n = 50$ states. The *mean*

$$\bar{x} = \sum_{i=1}^{n} x_i / n \tag{1.1}$$

summarizes the results in a single number.

How accurate is that number? The textbook answer is given in terms of the *standard error*,

$$\widehat{\mathrm{se}} = \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n(n-1)) \right]^{1/2}. \tag{1.2}$$

Here *averaging* (1.1) is the algorithm, while the standard error provides an inference of the algorithm's accuracy. It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy.[1]

[1] "Inference" concerns more than accuracy: speaking broadly, algorithms say what the statistician does while inference says why he or she does it.

Of course, $\widehat{\text{se}}$ (1.2) is itself an algorithm, which could be (and is) subject to further inferential analysis concerning *its* accuracy. The point is that the algorithm comes first and the inference follows at a second level of statistical consideration. In practice this means that algorithmic invention is a more free-wheeling and adventurous enterprise, with inference playing catch-up as it strives to assess the accuracy, good or bad, of some hot new algorithmic methodology.

If the inference/algorithm race is a tortoise-and-hare affair, then modern electronic computation has bred a bionic hare. There are two effects at work here: computer-based technology allows scientists to collect enormous data sets, orders of magnitude larger than those that classic statistical theory was designed to deal with; huge data demands new methodology, and the demand is being met by a burst of innovative computer-based statistical algorithms. When one reads of "big data" in the news, it is usually these algorithms playing the starring roles.

Our book's title, *Computer Age Statistical Inference*, emphasizes the tortoise's side of the story. The past few decades have been a golden age of statistical methodology. It hasn't been, quite, a golden age for statistical inference, but it has not been a dark age either. The efflorescence of ambitious new algorithms has forced an evolution (though not a revolution) in inference, the theories by which statisticians choose among competing methods. The book traces the interplay between methodology and inference as it has developed since the 1950s, the beginning of our discipline's computer age. As a preview, we end this chapter with two examples illustrating the transition from classic to computer-age practice.

1.1 A Regression Example

Figure 1.1 concerns a study of kidney function. Data points (x_i, y_i) have been observed for $n = 157$ healthy volunteers, with x_i the ith volunteer's `age` in years, and y_i a composite measure "`tot`" of overall function. Kidney function generally declines with `age`, as evident in the downward scatter of the points. The rate of decline is an important question in kidney transplantation: in the past, potential donors past `age` 60 were prohibited, though, given a shortage of donors, this is no longer enforced.

The solid line in Figure 1.1 is a *linear regression*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.3}$$

fit to the data by *least squares*, that is by minimizing the sum of squared

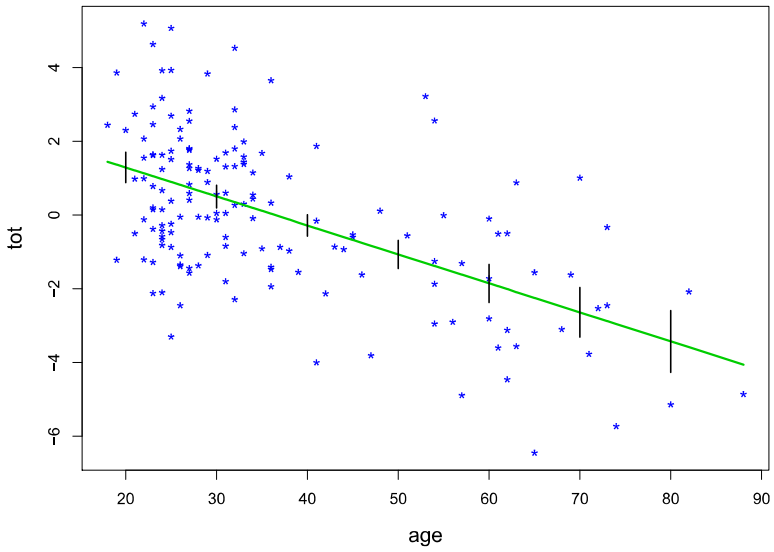

Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ±2 standard errors at selected values of `age`.

deviations

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.4}$$

over all choices of (β_0, β_1). The least squares algorithm, which dates back to Gauss and Legendre in the early 1800s, gives $\hat{\beta}_0 = 2.86$ and $\hat{\beta}_1 = -0.079$ as the least squares estimates. We can read off of the fitted line an estimated value of kidney fitness for any chosen `age`. The top line of Table 1.1 shows estimate 1.29 at `age` 20, down to −3.43 at `age` 80.

How accurate are these estimates? This is where inference comes in: an extended version of formula (1.2), also going back to the 1800s, provides the standard errors, shown in line 2 of the table. The vertical bars in Figure 1.1 are ± two standard errors, giving them about 95% chance of containing the true expected value of `tot` at each `age`.

That 95% coverage depends on the validity of the linear regression model (1.3). We might instead try a quadratic regression $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$, or a cubic, etc., all of this being well within the reach of pre-computer statistical theory.

Table 1.1 *Regression analysis of the kidney data; (1) linear regression estimates; (2) their standard errors; (3)* `lowess` *estimates; (4) their bootstrap standard errors.*

`age`	20	30	40	50	60	70	80
1. linear regression	1.29	.50	−.28	−1.07	−1.86	−2.64	−3.43
2. std error	.21	.15	.15	.19	.26	.34	.42
3. lowess	1.66	.65	−.59	−1.27	−1.91	−2.68	−3.50
4. bootstrap std error	.71	.23	.31	.32	.37	.47	.70

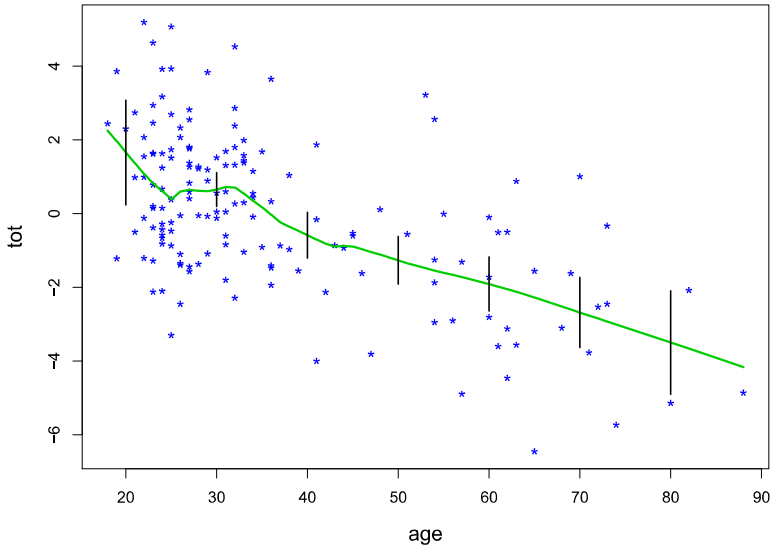

Figure 1.2 Local polynomial `lowess(x,y,1/3)` fit to the kidney-fitness data, with ±2 bootstrap standard deviations.

A modern computer-based algorithm `lowess` produced the somewhat bumpy regression curve in Figure 1.2. The `lowess`† [2] algorithm moves its attention along the x-axis, fitting local polynomial curves of differing degrees to nearby (x, y) points. (The 1/3 in the call[3] `lowess(x,y,1/3)`

†1

[2] Here and throughout the book, the numbered † sign indicates a technical note or reference element which is elaborated on at the end of the chapter.

[3] Here and in all our examples we are employing the language `R`, itself one of the key developments in computer-based statistical methodology.

determines the definition of local.) Repeated passes over the x-axis refine the fit, reducing the effects of occasional anomalous points. The fitted curve in Figure 1.2 is nearly linear at the right, but more complicated at the left where points are more densely packed. It is flat between ages 25 and 35, a potentially important difference from the uniform decline portrayed in Figure 1.1.

There is no formula such as (1.2) to infer the accuracy of the `lowess` curve. Instead, a computer-intensive inferential engine, the *bootstrap*, was used to calculate the error bars in Figure 1.2. A bootstrap data set is produced by resampling 157 pairs (x_i, y_i) from the original 157 *with replacement*, so perhaps (x_1, y_1) might show up twice in the bootstrap sample, (x_2, y_2) might be missing, (x_3, y_3) present once, etc. Applying `lowess` to the bootstrap sample generates a bootstrap replication of the original calculation.

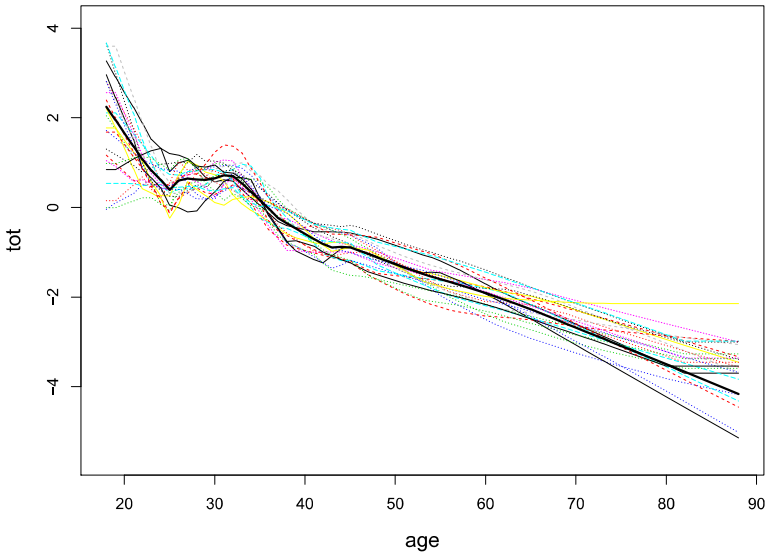

Figure 1.3 25 bootstrap replications of `lowess(x,y,1/3)`.

Figure 1.3 shows the first 25 (of 250) bootstrap `lowess` replications bouncing around the original curve from Figure 1.2. The variability of the replications at any one `age`, the *bootstrap standard deviation*, determined the original curve's accuracy. How and why the bootstrap works is discussed in Chapter 10. It has the great virtue of assessing estimation accu-

racy for *any* algorithm, no matter how complicated. The price is a hundred- or thousand-fold increase in computation, unthinkable in 1930, but routine now.

The bottom two lines of Table 1.1 show the `lowess` estimates and their standard errors. We have paid a price for the increased flexibility of `lowess`, its standard errors roughly doubling those for linear regression.

1.2 Hypothesis Testing

Our second example concerns the march of methodology and inference for *hypothesis testing* rather than estimation: 72 leukemia patients, 47 with `ALL` (acute lymphoblastic leukemia) and 25 with `AML` (acute myeloid leukemia, a worse prognosis) have each had genetic activity measured for a panel of 7,128 genes. The histograms in Figure 1.4 compare the genetic activities in the two groups for gene 136.

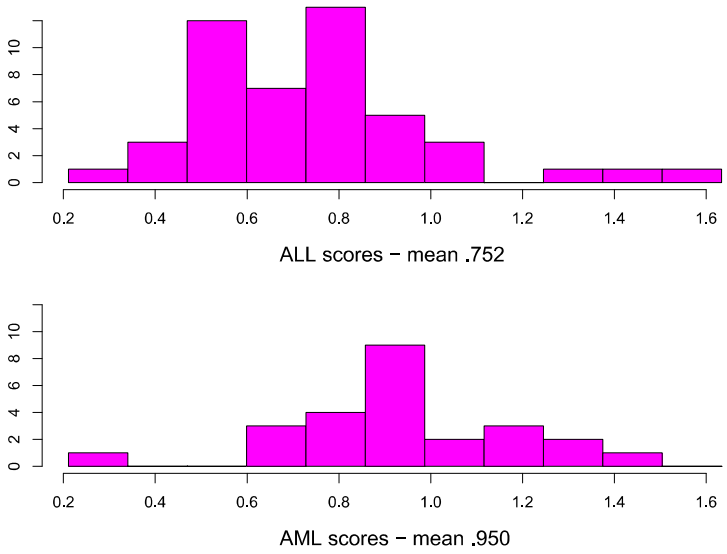

Figure 1.4 Scores for gene 136, leukemia data. Top `ALL` ($n = 47$), bottom `AML` ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

The `AML` group appears to show greater activity, the mean values being

$$\overline{\texttt{ALL}} = 0.752 \quad \text{and} \quad \overline{\texttt{AML}} = 0.950. \tag{1.5}$$

Is the perceived difference genuine, or perhaps, as people like to say, "a statistical fluke"? The classic answer to this question is via a *two-sample t-statistic*,

$$t = \frac{\overline{\texttt{AML}} - \overline{\texttt{ALL}}}{\widehat{\text{sd}}}, \tag{1.6}$$

where $\widehat{\text{sd}}$ is an estimate of the numerator's standard deviation.[4]

Dividing by $\widehat{\text{sd}}$ allows us (under Gaussian assumptions discussed in Chapter 5) to compare the observed value of t with a standard "null" distribution, in this case a Student's t distribution with 70 degrees of freedom. We obtain $t = 3.01$ from (1.6), which would classically be considered very strong evidence that the apparent difference (1.5) is genuine; in standard terminology, "with two-sided significance level 0.0036."

A small significance level (or "p-value") is a statement of statistical surprise: something very unusual has happened if in fact there is no difference in gene 136 expression levels between **ALL** and **AML** patients. We are less surprised by $t = 3.01$ if gene 136 is just one candidate out of thousands that might have produced "interesting" results.

That is the case here. Figure 1.5 shows the histogram of the two-sample t-statistics for the panel of 7128 genes. Now $t = 3.01$ looks less unusual; 400 other genes have t exceeding 3.01, about 5.6% of them.

This doesn't mean that gene 136 is "significant at the 0.056 level." There are two powerful complicating factors:

1. Large numbers of candidates, 7128 here, will produce some large t-values even if there is really no difference in genetic expression between **ALL** and **AML** patients.
2. The histogram implies that in this study there is something wrong with the theoretical null distribution ("Student's t with 70 degrees of freedom"), the smooth curve in Figure 1.5. It is much too narrow at the center, where presumably most of the genes are reporting non-significant results.

We will see in Chapter 15 that a low *false-discovery rate*, i.e., a low chance of crying wolf over an innocuous gene, requires t exceeding 6.16 in the **ALL**/**AML** study. Only 47 of the 7128 genes make the cut. False-discovery-rate theory is an impressive advance in statistical inference, incorporating Bayesian, frequentist, and empirical Bayesian (Chapter 6) el-

[4] Formally, a standard error is the standard deviation of a summary statistic, and $\widehat{\text{sd}}$ might better be called $\widehat{\text{se}}$, but we will follow the distinction less than punctiliously here.

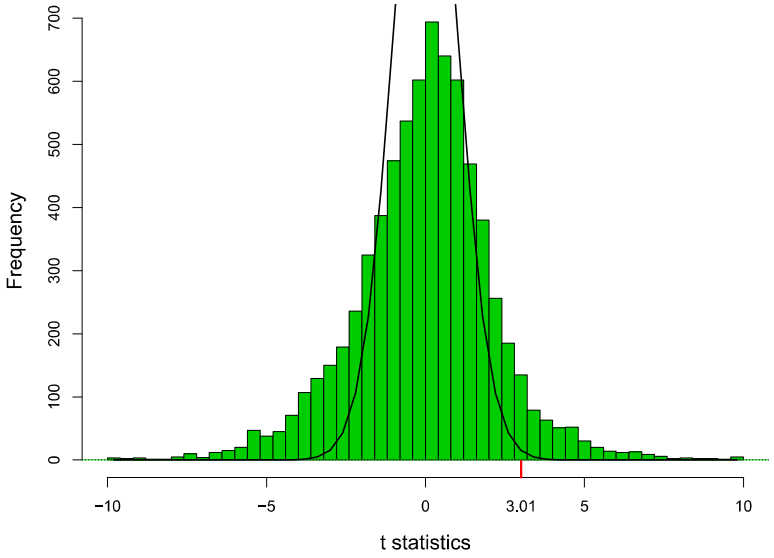

Figure 1.5 Two-sample t-statistics for 7128 genes, leukemia data. The smooth curve is the theoretical null density for the t-statistic.

ements. It was a *necessary* advance in a scientific world where computer-based technology routinely presents thousands of comparisons to be evaluated at once.

There is one more thing to say about the algorithm/inference statistical cycle. Important new algorithms often arise outside the world of professional statisticians: neural nets, support vector machines, and boosting are three famous examples. None of this is surprising. New sources of data, satellite imagery for example, or medical microarrays, inspire novel methodology from the observing scientists. The early literature tends toward the enthusiastic, with claims of enormous applicability and power.

In the second phase, statisticians try to locate the new metholodogy within the framework of statistical theory. In other words, they carry out the statistical inference part of the cycle, placing the new methodology within the known Bayesian and frequentist limits of performance. (Boosting offers a nice example, Chapter 17.) This is a healthy chain of events, good both for the hybrid vigor of the statistics profession and for the further progress of algorithmic technology.

1.3 Notes

Legendre published the least squares algorithm in 1805, causing Gauss to state that he had been using the method in astronomical orbit-fitting since 1795. Given Gauss' astonishing production of major mathematical advances, this says something about the importance attached to the least squares idea. Chapter 8 includes its usual algebraic formulation, as well as Gauss' formula for the standard errors, line 2 of Table 1.1.

Our division between algorithms and inference brings to mind Tukey's exploratory/confirmatory system. However the current algorithmic world is often bolder in its claims than the word "exploratory" implies, while to our minds "inference" conveys something richer than mere confirmation.

†1 [p. 6] `lowess` was devised by William Cleveland (Cleveland, 1981) and is available in the R statistical computing language. It is applied to the kidney data in Efron (2004). The kidney data originated in the nephrology laboratory of Dr. Brian Myers, Stanford University, and is available from this book's web site.

2

Frequentist Inference

Before the computer age there was the calculator age, and before "big data" there were small data sets, often a few hundred numbers or fewer, laboriously collected by individual scientists working under restrictive experimental constraints. Precious data calls for maximally efficient statistical analysis. A remarkably effective theory, feasible for execution on mechanical desk calculators, was developed beginning in 1900 by Pearson, Fisher, Neyman, Hotelling, and others, and grew to dominate twentieth-century statistical practice. The theory, now referred to as *classical*, relied almost entirely on frequentist inferential ideas. This chapter sketches a quick and simplified picture of frequentist inference, particularly as employed in classical applications.

We begin with another example from Dr. Myers' nephrology laboratory: 211 kidney patients have had their *glomerular filtration rates* measured, with the results shown in Figure 2.1; `gfr` is an important indicator of kidney function, with low values suggesting trouble. (It is a key component of `tot` in Figure 1.1.) The mean and standard error (1.1)–(1.2) are $\bar{x} = 54.25$ and $\widehat{\text{se}} = 0.95$, typically reported as

$$54.25 \pm 0.95; \tag{2.1}$$

± 0.95 denotes a frequentist inference for the accuracy of the estimate $\bar{x} = 54.25$, and suggests that we shouldn't take the ".25" very seriously, even the "4" being open to doubt. Where the inference comes from and what exactly it means remains to be said.

Statistical inference usually begins with the assumption that some probability model has produced the observed data $\boldsymbol{x}$, in our case the vector of $n = 211$ `gfr` measurements $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ indicate n independent draws from a probability distribution F, written

$$F \to \boldsymbol{X}, \tag{2.2}$$

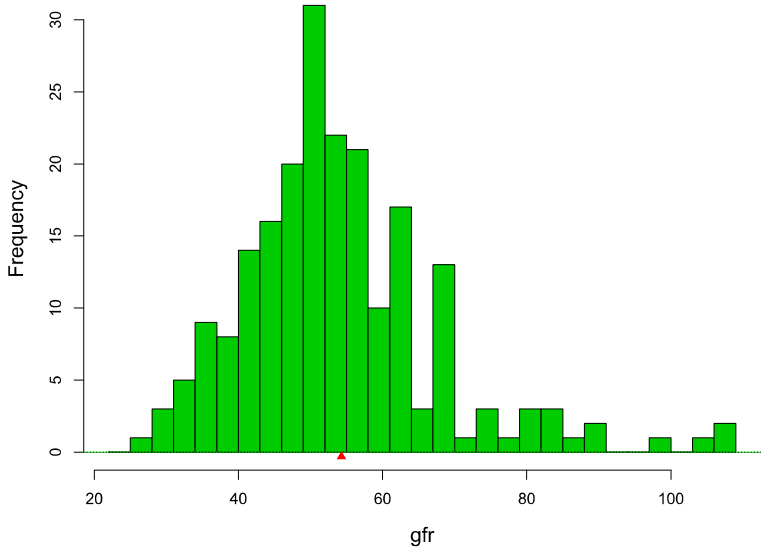

Figure 2.1 Glomerular filtration rates for 211 kidney patients; mean 54.25, standard error .95.

F being the underlying distribution of possible `gfr` scores here. A realization $\boldsymbol{X} = \boldsymbol{x}$ of (2.2) has been observed, and the statistician wishes to *infer* some property of the unknown distribution F.

Suppose the desired property is the *expectation* of a single random draw X from F, denoted

$$\theta = E_F\{X\} \tag{2.3}$$

(which also equals the expectation of the average $\bar{X} = \sum X_i/n$ of random vector (2.2)[1]). The obvious estimate of θ is $\hat{\theta} = \bar{x}$, the sample average. If n were enormous, say 10^{10}, we would expect $\hat{\theta}$ to nearly equal θ, but otherwise there is room for error. How much error is the inferential question.

The estimate $\hat{\theta}$ is calculated from $\boldsymbol{x}$ according to some known algorithm, say

$$\hat{\theta} = t(\boldsymbol{x}), \tag{2.4}$$

$t(\boldsymbol{x})$ in our example being the averaging function $\bar{x} = \sum x_i/n$; $\hat{\theta}$ is a

[1] The fact that $E_F\{\bar{X}\}$ equals $E_F\{X\}$ is a crucial, though easily proved, probabilistic result.

realization of

$$\hat{\Theta} = t(\boldsymbol{X}), \tag{2.5}$$

the output of $t(\cdot)$ applied to a theoretical sample $\boldsymbol{X}$ from F (2.2). We have chosen $t(\boldsymbol{X})$, we hope, to make $\hat{\Theta}$ a good estimator of θ, the desired property of F.

We can now give a first definition of frequentist inference: *the accuracy of an observed estimate* $\hat{\theta} = t(\boldsymbol{x})$ *is the probabilistic accuracy of* $\hat{\Theta} = t(\boldsymbol{X})$ *as an estimator of* θ. This may seem more a tautology than a definition, but it contains a powerful idea: $\hat{\theta}$ is just a single number but $\hat{\Theta}$ takes on a range of values whose spread can define measures of accuracy.

Bias and variance are familiar examples of frequentist inference. Define μ to be the expectation of $\hat{\Theta} = t(\boldsymbol{X})$ under model (2.2),

$$\mu = E_F\{\hat{\Theta}\}. \tag{2.6}$$

Then the bias and variance attributed to estimate $\hat{\theta}$ of parameter θ are

$$\text{bias} = \mu - \theta \quad \text{and} \quad \text{var} = E_F\left\{(\hat{\Theta} - \mu)^2\right\}. \tag{2.7}$$

Again, what keeps this from tautology is the attribution to the single number $\hat{\theta}$ of the probabilistic properties of $\hat{\Theta}$ following from model (2.2). If all of this seems too obvious to worry about, the Bayesian criticisms of Chapter 3 may come as a shock.

Frequentism is often defined with respect to "an infinite sequence of future trials." We imagine hypothetical data sets $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \ldots$ generated by the same mechanism as $\boldsymbol{x}$ providing corresponding values $\hat{\Theta}^{(1)}$, $\hat{\Theta}^{(2)}$, $\hat{\Theta}^{(3)}, \ldots$ as in (2.5). The frequentist principle is then to attribute for $\hat{\theta}$ the accuracy properties of the ensemble of $\hat{\Theta}$ values.[2] If the $\hat{\Theta}$s have empirical variance of, say, 0.04, then $\hat{\theta}$ is claimed to have standard error $0.2 = \sqrt{0.04}$, etc. This amounts to a more picturesque restatement of the previous definition.

2.1 Frequentism in Practice

Our working definition of frequentism is that *the probabilistic properties of a procedure of interest are derived and then applied verbatim to the procedure's output for the observed data*. This has an obvious defect: it requires calculating the properties of estimators $\hat{\Theta} = t(\boldsymbol{X})$ obtained from

[2] In essence, frequentists ask themselves "What would I see if I reran the same situation again (and again and again. . .)?"

the true distribution F, even though F is unknown. Practical frequentism uses a collection of more or less ingenious devices to circumvent the defect.

1. The plug-in principle. A simple formula relates the standard error of $\bar{X} = \sum X_i/n$ to $\mathrm{var}_F(X)$, the variance of a single X drawn from F,

$$\mathrm{se}\left(\bar{X}\right) = \left[\mathrm{var}_F(X)/n\right]^{1/2}. \tag{2.8}$$

But having observed $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ we can estimate $\mathrm{var}_F(X)$ without bias by

$$\widehat{\mathrm{var}}_F = \sum (x_i - \bar{x})^2 / (n-1). \tag{2.9}$$

Plugging formula (2.9) into (2.8) gives $\widehat{\mathrm{se}}$ (1.2), the usual estimate for the standard error of an average $\bar{x}$. In other words, the frequentist accuracy estimate for $\bar{x}$ is itself estimated from the observed data.[3]

2. Taylor-series approximations. Statistics $\hat{\theta} = t(\boldsymbol{x})$ more complicated than $\bar{x}$ can often be related back to the plug-in formula by local linear approximations, sometimes known as the "delta method."† For example, †1
$\hat{\theta} = \bar{x}^2$ has $d\hat{\theta}/d\bar{x} = 2\bar{x}$. Thinking of $2\bar{x}$ as a constant gives

$$\mathrm{se}\left(\bar{x}^2\right) \doteq 2\,|\bar{x}|\,\widehat{\mathrm{se}}, \tag{2.10}$$

with $\widehat{\mathrm{se}}$ as in (1.2). Large sample calculations, as sample size n goes to infinity, validate the delta method which, fortunately, often performs well in small samples.

3. Parametric families and maximum likelihood theory. Theoretical expressions for the standard error of a maximum likelihood estimate (MLE) are discussed in Chapters 4 and 5, in the context of parametric families of distributions. These combine Fisherian theory, Taylor-series approximations, and the plug-in principle in an easy-to-apply package.

4. Simulation and the bootstrap. Modern computation has opened up the possibility of numerically implementing the "infinite sequence of future trials" definition, except for the infinite part. An estimate $\hat{F}$ of F, perhaps the MLE, is found, and values $\hat{\Theta}^{(k)} = t(\boldsymbol{X}^{(k)})$ simulated from $\hat{F}$ for $k = 1, 2, \ldots, B$, say $B = 1000$. The empirical standard deviation of the $\hat{\Theta}$s is then the frequentist estimate of standard error for $\hat{\theta} = t(\boldsymbol{x})$, and similarly with other measures of accuracy.

This is a good description of the bootstrap, Chapter 10. (Notice that

[3] The most familiar example is the observed proportion p of heads in n flips of a coin having true probability π: the actual standard error is $[\pi(1-\pi)/n]^{1/2}$ but we can only report the plug-in estimate $[p(1-p)/n]^{1/2}$.

Table 2.1 *Three estimates of location for the* `gfr` *data, and their estimated standard errors; last two standard errors using the bootstrap,* $B = 1000$.

	Estimate	Standard error
mean	54.25	.95
25% Winsorized mean	52.61	.78
median	52.24	.87

here the plugging-in, of $\hat{F}$ for F, comes *first* rather than at the end of the process.) The classical methods 1–3 above are restricted to estimates $\hat{\theta} = t(\boldsymbol{x})$ that are smoothly defined functions of various sample means. Simulation calculations remove this restriction. Table 2.1 shows three "location" estimates for the `gfr` data, the mean, the 25% Winsorized mean,[4] and the median, along with their standard errors, the last two computed by the bootstrap. A happy feature of computer-age statistical inference is the tremendous expansion of useful and usable statistics $t(\boldsymbol{x})$ in the statistician's working toolbox, the `lowess` algorithm in Figures 1.2 and 1.3 providing a nice example.

5. Pivotal statistics. A pivotal statistic $\hat{\theta} = t(\boldsymbol{x})$ is one whose distribution does *not* depend upon the underlying probability distribution F. In such a case the theoretical distribution of $\hat{\Theta} = t(\boldsymbol{X})$ applies exactly to $\hat{\theta}$, removing the need for devices 1–4 above. The classic example concerns Student's two-sample t-test.

In a two-sample problem the statistician observes two sets of numbers,

$$\boldsymbol{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1}) \quad \boldsymbol{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2}), \tag{2.11}$$

and wishes to test the *null hypothesis* that they come from the same distribution (as opposed to, say, the second set tending toward larger values than the first). It is assumed that the distribution F_1 for $\boldsymbol{x}_1$ is *normal*, or *Gaussian*,

$$X_{1i} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_1, \sigma^2), \qquad i = 1, 2, \ldots, n_1, \tag{2.12}$$

the notation indicating n_1 independent draws from a normal distribution[5]

[4] All observations below the 25th percentile of the 211 observations are moved up to that point, similarly those above the 75th percentile are moved down, and finally the mean is taken.

[5] Each draw having probability density $(2\pi\sigma^2)^{-1/2} \exp\{-0.5 \cdot (x - \mu_1)^2/\sigma^2\}$.

with expectation μ_1 and variance σ^2. Likewise

$$X_{2i} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_2, \sigma^2) \qquad i = 1, 2, \ldots, n_2. \tag{2.13}$$

We wish to test the null hypothesis

$$H_0 : \mu_1 = \mu_2. \tag{2.14}$$

The obvious test statistic $\hat{\theta} = \bar{x}_2 - \bar{x}_1$, the difference of the means, has distribution

$$\hat{\theta} \sim \mathcal{N}\left(0, \sigma^2\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)\right) \tag{2.15}$$

under H_0. We could plug in the unbiased estimate of σ^2,

$$\hat{\sigma}^2 = \left[\sum_1^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_1^{n_2}(x_{2i} - \bar{x}_2)^2\right] \Big/ (n_1 + n_2 - 2), \tag{2.16}$$

but Student provided a more elegant solution: instead of $\hat{\theta}$, we test H_0 using the two-sample t-statistic

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\widehat{\text{sd}}}, \qquad \text{where } \widehat{\text{sd}} = \hat{\sigma}\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)^{1/2}. \tag{2.17}$$

Under H_0, t is pivotal, having the same distribution (Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom), no matter what the value of the "nuisance parameter" σ.

For $n_1 + n_2 - 2 = 70$, as in the leukemia example (1.5)–(1.6), Student's distribution gives

$$\Pr_{H_0}\{-1.99 \le t \le 1.99\} = 0.95. \tag{2.18}$$

The hypothesis test that rejects H_0 if $|t|$ exceeds 1.99 has probability exactly 0.05 of mistaken rejection. Similarly,

$$\bar{x}_2 - \bar{x}_1 \pm 1.99 \cdot \widehat{\text{sd}} \tag{2.19}$$

is an exact 0.95 *confidence interval* for the difference $\mu_2 - \mu_1$, covering the true value in 95% of repetitions of probability model (2.12)–(2.13).[6]

[6] Occasionally, one sees frequentism defined in careerist terms, e.g., "A statistician who always rejects null hypotheses at the 95% level will over time make only 5% errors of the first kind." This is not a comforting criterion for the statistician's clients, who are interested in their own situations, not everyone else's. Here we are only assuming hypothetical repetitions of the specific problem at hand.

What might be called the *strong definition of frequentism* insists on exact frequentist correctness under experimental repetitions. Pivotality, unfortunately, is unavailable in most statistical situations. Our looser definition of frequentism, supplemented by devices such as those above,[7] presents a more realistic picture of actual frequentist practice.

2.2 Frequentist Optimality

The popularity of frequentist methods reflects their relatively modest mathematical modeling assumptions: only a probability model F (more exactly a family of probabilities, Chapter 3) and an algorithm of choice $t(\boldsymbol{x})$. This flexibility is also a defect in that the principle of frequentist correctness doesn't help with the choice of algorithm. Should we use the sample mean to estimate the location of the `gfr` distribution? Maybe the 25% Winsorized mean would be better, as Table 2.1 suggests.

The years 1920–1935 saw the development of two key results on *frequentist optimality*, that is, finding the *best* choice of $t(\boldsymbol{x})$ given model F. The first of these was Fisher's theory of maximum likelihood estimation and the Fisher information bound: in parametric probability models of the type discussed in Chapter 4, the MLE is the optimum estimate in terms of minimum (asymptotic) standard error.

In the same spirit, the Neyman–Pearson lemma provides an optimum hypothesis-testing algorithm. This is perhaps the most elegant of frequentist constructions. In its simplest formulation, the NP lemma assumes we are trying to decide between two possible probability density functions for the observed data $\boldsymbol{x}$, a null hypothesis density $f_0(\boldsymbol{x})$ and an alternative density $f_1(\boldsymbol{x})$. A testing rule $t(\boldsymbol{x})$ says which choice, 0 or 1, we will make having observed data $\boldsymbol{x}$. Any such rule has two associated frequentist error probabilities: choosing f_1 when actually f_0 generated $\boldsymbol{x}$, and vice versa,

$$\begin{aligned} \alpha &= \Pr_{f_0}\{t(\boldsymbol{x}) = 1\}, \\ \beta &= \Pr_{f_1}\{t(\boldsymbol{x}) = 0\}. \end{aligned} \tag{2.20}$$

Let $L(\boldsymbol{x})$ be the *likelihood ratio*,

$$L(\boldsymbol{x}) = f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) \tag{2.21}$$

[7] The list of devices is not complete. Asymptotic calculations play a major role, as do more elaborate combinations of pivotality and the plug-in principle; see the discussion of approximate bootstrap confidence intervals in Chapter 11.

and define the testing rule $t_c(\boldsymbol{x})$ by

$$t_c(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \log L(\boldsymbol{x}) \geq c \\ 0 & \text{if } \log L(\boldsymbol{x}) < c. \end{cases} \tag{2.22}$$

There is one such rule for each choice of the cutoff c. The Neyman–Pearson lemma says that only rules of form (2.22) can be optimum; for any other rule $t(\boldsymbol{x})$ there will be a rule $t_c(\boldsymbol{x})$ having smaller errors of both kinds,[8]

$$\alpha_c < \alpha \quad \text{and} \quad \beta_c < \beta. \tag{2.23}$$

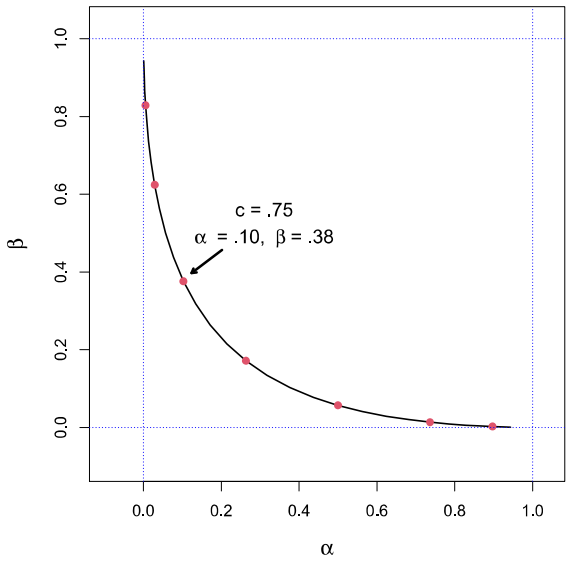

Figure 2.2 Neyman–Pearson alpha–beta curve for $f_0 \sim \mathcal{N}(0, 1)$, $f_1 \sim \mathcal{N}(.5, 1)$, and sample size $n = 10$. Red dots correspond to cutoffs $c = 2.75, 1.75, .75, \ldots, -3.25$.

Figure 2.2 graphs (α_c, β_c) as a function of the cutoff c, for the case where $\boldsymbol{x} = (x_1, x_2, \ldots, x_{10})$ is obtained by independent sampling from a normal distribution, $\mathcal{N}(0, 1)$ for f_0 versus $\mathcal{N}(0.5, 1)$ for f_1. The NP lemma says that any rule not of form (2.22) must have its (α, β) point lying above the curve.

[8] Here we are ignoring some minor definitional difficulties that can occur if f_0 and f_1 are discrete.

Frequentist optimality theory, both for estimation and for testing, anchored statistical practice in the twentieth century. The larger data sets and more complicated inferential questions of the current era have strained the capabilities of that theory. Computer-age statistical inference, as we will see, often displays an unsettling ad hoc character. Perhaps some contemporary Fishers and Neymans will provide us with a more capacious optimality theory equal to the challenges of current practice, but for now that is only a hope.

Frequentism cannot claim to be a seamless philosophy of statistical inference. Paradoxes and contradictions abound within its borders, as will be shown in the next chapter. That being said, frequentist methods have a natural appeal to working scientists, an impressive history of successful application, and, as our list of five "devices" suggests, the capacity to encourage clever methodology. The story that follows is not one of abandonment of frequentist thinking, but rather a broadening of connections with other methods.

2.3 Notes and Details

The name "frequentism" seems to have been suggested by Neyman as a statistical analogue of Richard von Mises' frequentist theory of probability, the connection being made explicit in his 1977 paper, "Frequentist probability and frequentist statistics." "Behaviorism" might have been a more descriptive name[9] since the theory revolves around the long-run behavior of statistics $t(\boldsymbol{x})$, but in any case "frequentism" has stuck, replacing the older (sometimes disparaging) term "objectivism." Neyman's attempt at a complete frequentist theory of statistical inference, "inductive behavior," is not much quoted today, but can claim to be an important influence on Wald's development of decision theory.

R. A. Fisher's work on maximum likelihood estimation is featured in Chapter 4. Fisher, arguably the founder of frequentist optimality theory, was not a pure frequentist himself, as discussed in Chapter 4 and Efron (1998), "R. A. Fisher in the 21st Century." (Now that we are well into the twenty-first century, the author's talents as a prognosticator can be frequentistically evaluated.)

†1 [p. 15] *Delta method.* The delta method uses a first-order Taylor series to approximate the variance of a function $s(\hat{\theta})$ of a statistic $\hat{\theta}$. Suppose $\hat{\theta}$ has mean/variance (θ, σ^2), and consider the approximation $s(\hat{\theta}) \approx s(\theta) +$

[9] That name is already spoken for in the psychology literature.

$s'(\theta)(\hat{\theta} - \theta)$. Hence $\text{var}\{s(\hat{\theta})\} \approx |s'(\theta)|^2\sigma^2$. We typically plug-in $\hat{\theta}$ for θ, and use an estimate for σ^2.

3

Bayesian Inference

The human mind is an inference machine: "It's getting windy, the sky is darkening, I'd better bring my umbrella with me." Unfortunately, it's not a very dependable machine, especially when weighing complicated choices against past experience. *Bayes' theorem* is a surprisingly simple mathematical guide to accurate inference. The theorem (or "rule"), now 250 years old, marked the beginning of statistical inference as a serious scientific subject. It has waxed and waned in influence over the centuries, now waxing again in the service of computer-age applications.

Bayesian inference, if not directly opposed to frequentism, is at least orthogonal. It reveals some worrisome flaws in the frequentist point of view, while at the same time exposing itself to the criticism of dangerous overuse. The struggle to combine the virtues of the two philosophies has become more acute in an era of massively complicated data sets. Much of what follows in succeeding chapters concerns this struggle. Here we will review some basic Bayesian ideas and the ways they impinge on frequentism.

The fundamental unit of statistical inference both for frequentists and for Bayesians is a *family* of probability densities

$$\mathcal{F} = \{f_\mu(x);\ x \in \mathcal{X},\ \mu \in \Omega\}; \tag{3.1}$$

x, the observed data, is a point[1] in the *sample space* $\mathcal{X}$, while the unobserved parameter μ is a point in the *parameter space* Ω. The statistician observes x from $f_\mu(x)$, and infers the value of μ.

Perhaps the most familiar case is the normal family

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} \tag{3.2}$$

[1] Both x and μ may be scalars, vectors, or more complicated objects. Other names for the generic "x" and "μ" occur in specific situations, for instance $\boldsymbol{x}$ for x in Chapter 2. We will also call $\mathcal{F}$ a "family of probability distributions."

(more exactly, the one-dimensional normal translation family[2] with variance 1), with both $\mathcal{X}$ and Ω equaling $\mathcal{R}^1$, the entire real line $(-\infty, \infty)$. Another central example is the Poisson family

$$f_\mu(x) = e^{-\mu}\mu^x/x!, \tag{3.3}$$

where $\mathcal{X}$ is the nonnegative integers $\{0, 1, 2, \ldots\}$ and Ω is the nonnegative real line $(0, \infty)$. (Here the "density" (3.3) specifies the atoms of probability on the discrete points of $\mathcal{X}$.)

Bayesian inference requires one crucial assumption in addition to the probability family $\mathcal{F}$, the knowledge of a *prior density*

$$g(\mu), \qquad \mu \in \Omega; \tag{3.4}$$

$g(\mu)$ represents prior information concerning the parameter μ, available to the statistician before the observation of x. For instance, in an application of the normal model (3.2), it could be known that μ is positive, while past experience shows it never exceeding 10, in which case we might take $g(\mu)$ to be the uniform density $g(\mu) = 1/10$ on the interval $[0, 10]$. Exactly what constitutes "prior knowledge" is a crucial question we will consider in ongoing discussions of Bayes' theorem.

Bayes' theorem is a rule for combining the prior knowledge in $g(\mu)$ with the current evidence in x. Let $g(\mu|x)$ denote the *posterior density* of μ, that is, our update of the prior density $g(\mu)$ after taking account of observation x. Bayes' rule provides a simple expression for $g(\mu|x)$ in terms of $g(\mu)$ and $\mathcal{F}$.

Bayes' Rule: $$g(\mu|x) = g(\mu)\, f_\mu(x)/f(x), \qquad \mu \in \Omega, \tag{3.5}$$

where $f(x)$ is the *marginal density* of x,

$$f(x) = \int_\Omega f_\mu(x) g(\mu)\, d\mu. \tag{3.6}$$

(The integral in (3.6) would be a sum if Ω were discrete.) The Rule is a straightforward exercise in conditional probability,[3] and yet has far-reaching and sometimes surprising consequences.

In Bayes' formula (3.5), x is fixed at its observed value while μ varies over Ω, just the opposite of frequentist calculations. We can emphasize this

[2] Standard notation is $x \sim \mathcal{N}(\mu, \sigma^2)$ for a normal distribution with expectation μ and variance σ^2, so (3.2) has $x \sim \mathcal{N}(\mu, 1)$.

[3] $g(\mu|x)$ is the ratio of $g(\mu)\, f_\mu(x)$, the joint probability of the pair (μ, x), and $f(x)$, the marginal probability of x.

by rewriting (3.5) as

$$g(\mu|x) = c_x L_x(\mu) g(\mu), \tag{3.7}$$

where $L_x(\mu)$ is the *likelihood function*, that is, $f_\mu(x)$ with x fixed and μ varying. Having computed $L_x(\mu)g(\mu)$, the constant c_x can be determined numerically from the requirement that $g(\mu|x)$ integrate to 1, obviating the calculation of $f(x)$ (3.6).

Note Multiplying the likelihood function by any fixed constant c_0 has no effect on (3.7) since c_0 can be absorbed into c_x. So for the Poisson family (3.3) we can take $L_x(\mu) = e^{-\mu}\mu^x$, ignoring the $x!$ factor, which acts as a constant in Bayes' rule. The luxury of ignoring factors depending only on x often simplifies Bayesian calculations.

For any two points μ_1 and μ_2 in Ω, the ratio of posterior densities is, by division in (3.5),

$$\frac{g(\mu_1|x)}{g(\mu_2|x)} = \frac{g(\mu_1)}{g(\mu_2)} \frac{f_{\mu_1}(x)}{f_{\mu_2}(x)} \tag{3.8}$$

(no longer involving the marginal density $f(x)$), that is, "the posterior odds ratio is the prior odds ratio times the likelihood ratio," a memorable restatement of Bayes' rule.

3.1 Two Examples

A simple but genuine example of Bayes' rule in action is provided by the story of the *Physicist's Twins*: thanks to sonograms, a physicist found out she was going to have twin boys. "What is the probability my twins will be *Identical*, rather than *Fraternal*?" she asked. The doctor answered that one-third of twin births were Identicals, and two-thirds Fraternals.

In this situation μ, the unknown parameter (or "state of nature") is either *Identical* or *Fraternal* with prior probability 1/3 or 2/3; X, the possible sonogram results for twin births, is either *Same Sex* or *Different Sexes*, and $x =$ *Same Sex* was observed. (We can ignore sex since that does not affect the calculation.) A crucial fact is that identical twins are always same-sex while fraternals have probability 0.5 of same or different, so *Same Sex* in the sonogram is twice as likely if the twins are Identical. Applying Bayes'

rule in ratio form (3.8) answers the physicist's question:

$$\begin{aligned}\frac{g(\text{Identical}\,|\,\text{Same})}{g(\text{Fraternal}\,|\,\text{Same})} &= \frac{g(\text{Identical})}{g(\text{Fraternal})}\cdot\frac{f_{\text{Identical}}(\text{Same})}{f_{\text{Fraternal}}(\text{Same})}\\ &= \frac{1/3}{2/3}\cdot\frac{1}{1/2} = 1.\end{aligned} \tag{3.9}$$

That is, the posterior odds are even, and the physicist's twins have equal probabilities 0.5 of being Identical or Fraternal.[4] Here the doctor's prior odds ratio, 2 to 1 in favor of Fraternal, is balanced out by the sonogram's likelihood ratio of 2 to 1 in favor of Identical.

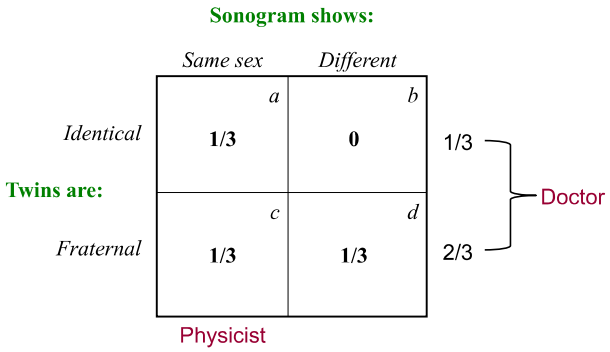

Figure 3.1 Analyzing the twins problem.

There are only four possible combinations of parameter μ and outcome x in the twins problem, labeled *a, b, c,* and *d* in Figure 3.1. Cell *b* has probability 0 since Identicals cannot be of Different Sexes. Cells *c* and *d* have equal probabilities because of the random sexes of Fraternals. Finally, $a + b$ must have total probability 1/3, and $c + d$ total probability 2/3, according to the doctor's prior distribution. Putting all this together, we can fill in the probabilities for all four cells, as shown. The physicist knows she is in the first column of the table, where the conditional probabilities of Identical or Fraternal are equal, just as provided by Bayes' rule in (3.9).

Presumably the doctor's prior distribution came from some enormous state or national database, say three million previous twin births, one million Identical pairs and two million Fraternals. We deduce that cells *a, c,* and *d* must have had one million entries each in the database, while cell *b* was empty. Bayes' rule can be thought of as a *big book* with one page

[4] They turned out to be Fraternal.

for each possible outcome x. (The book has only two pages in Figure 3.1.) The physicist turns to the page "Same Sex" and sees two million previous twin births, half Identical and half Fraternal, correctly concluding that the odds are equal in her situation.

Given any prior distribution $g(\mu)$ and any family of densities $f_\mu(x)$, Bayes' rule will always provide a version of the big book. That doesn't mean that the book's contents will always be equally convincing. The prior for the twins problems was based on a large amount of relevant previous experience. Such experience is most often unavailable. Modern Bayesian practice uses various strategies to construct an appropriate "prior" $g(\mu)$ in the absence of prior experience, leaving many statisticians unconvinced by the resulting Bayesian inferences. Our second example illustrates the difficulty.

Table 3.1 *Scores from two tests taken by 22 students,* `mechanics` *and* `vectors`.

	1	2	3	4	5	6	7	8	9	10	11
`mechanics`	7	44	49	59	34	46	0	32	49	52	44
`vectors`	51	69	41	70	42	40	40	45	57	64	61

	12	13	14	15	16	17	18	19	20	21	22
`mechanics`	36	42	5	22	18	41	48	31	42	46	63
`vectors`	59	60	30	58	51	63	38	42	69	49	63

Table 3.1 shows the scores on two tests, `mechanics` and `vectors`, achieved by $n = 22$ students. The sample correlation coefficient between the two scores is $\hat{\theta} = 0.498$,

$$\hat{\theta} = \sum_{i=1}^{22}(m_i - \bar{m})(v_i - \bar{v}) \Big/ \left[\sum_{i=1}^{22}(m_i - \bar{m})^2 \sum_{i=1}^{22}(v_i - \bar{v})^2\right]^{1/2}, \quad (3.10)$$

with m and v short for `mechanics` and `vectors`, $\bar{m}$ and $\bar{v}$ their averages. We wish to assign a Bayesian measure of posterior accuracy to the true correlation coefficient θ, "true" meaning the correlation for the hypothetical population of all students, of which we observed only 22.

If we assume that the joint (m, v) distribution is bivariate normal (as discussed in Chapter 5), then the density of $\hat{\theta}$ as a function of θ has a known form,[†] †1

$$f_\theta\left(\hat{\theta}\right) = \frac{(n-2)(1-\theta^2)^{(n-1)/2}\left(1-\hat{\theta}^2\right)^{(n-4)/2}}{\pi}\int_0^\infty \frac{dw}{\left(\cosh w - \theta\hat{\theta}\right)^{n-1}}. \tag{3.11}$$

In terms of our general Bayes notation, parameter μ is θ, observation x is $\hat{\theta}$, and family $\mathcal{F}$ is given by (3.11), with both Ω and $\mathcal{X}$ equaling the interval $[-1, 1]$. Formula (3.11) looks formidable to the human eye but not to the computer eye, which makes quick work of it.

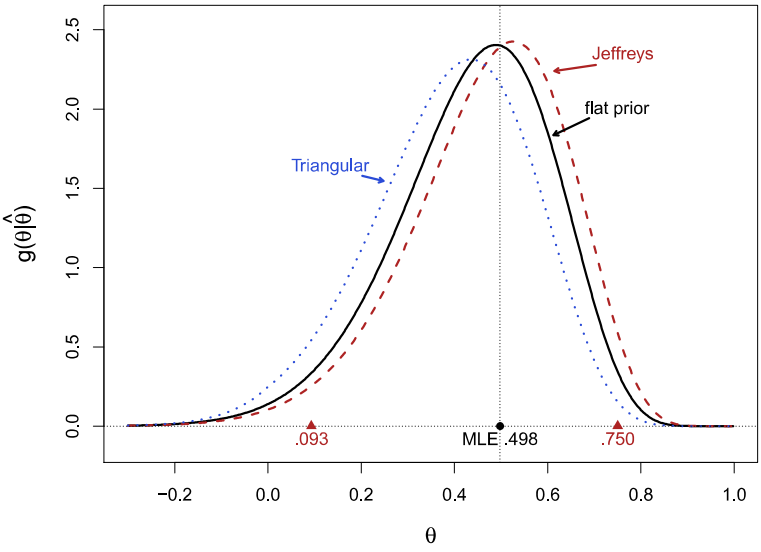

Figure 3.2 Student scores data; posterior density of correlation θ for three possible priors.

In this case, as in the majority of scientific situations, we don't have a trove of relevant past experience ready to provide a prior $g(\theta)$. One expedient, going back to Laplace, is the "principle of insufficient reason," that is, we take θ to be uniformly distributed over Ω,

$$g(\theta) = \tfrac{1}{2} \qquad \text{for } -1 \le \theta \le 1, \tag{3.12}$$

a "flat prior." The solid black curve in Figure 3.2 shows the resulting posterior density (3.5), which is just the likelihood $f_\theta(0.498)$ plotted as a function of θ (and scaled to have integral 1).

Jeffreys' prior,

$$g^{\text{Jeff}}(\theta) = 1/(1-\theta^2), \tag{3.13}$$

yields posterior density $g(\theta|\hat{\theta})$ shown by the dashed red curve. It suggests somewhat bigger values for the unknown parameter θ. Formula (3.13) arises from a theory of "uninformative priors" discussed in the next section, an improvement on the principle of insufficient reason; (3.13) is an *improper density* in that $\int_{-1}^{1} g(\theta)\, d\theta = \infty$, but it still provides proper posterior densities when deployed in Bayes' rule (3.5).

The dotted blue curve in Figure 3.2 is posterior density $g(\theta|\hat{\theta})$ obtained from the triangular-shaped prior

$$g(\theta) = 1 - |\theta|. \tag{3.14}$$

This is a primitive example of a *shrinkage* prior, one designed to favor smaller values of θ. Its effect is seen in the leftward shift of the posterior density. Shrinkage priors will play a major role in our discussion of large-scale estimation and testing problems, where we are hoping to find a few large effects hidden among thousands of negligible ones.

3.2 Uninformative Prior Distributions

Given a convincing prior distribution, Bayes' rule is easier to use and produces more satisfactory inferences than frequentist methods. The dominance of frequentist practice reflects the scarcity of useful prior information in day-to-day scientific applications. But the Bayesian impulse is strong, and almost from its inception 250 years ago there have been proposals for the construction of "priors" that permit the use of Bayes' rule in the absence of relevant experience.

One approach, perhaps the most influential in current practice, is the employment of *uninformative priors*. "Uninformative" has a positive connotation here, implying that the use of such a prior in Bayes' rule does not tacitly bias the resulting inference. Laplace's principle of insufficient reason, i.e., assigning uniform prior distributions to unknown parameters, is an obvious attempt at this goal. Its use went unchallenged for more than a century, perhaps because of Laplace's influence more than its own virtues.

Venn (of the Venn diagram) in the 1860s, and Fisher in the 1920s, attacking the routine use of Bayes' theorem, pointed out that Laplace's principle could not be applied consistently. In the student correlation example, for instance, a uniform prior distribution for θ would not be uniform if we

changed parameters to $\gamma = e^\theta$; posterior probabilities such as

$$\Pr\left\{\theta > 0|\hat{\theta}\right\} = \Pr\left\{\gamma > 1|\hat{\theta}\right\} \tag{3.15}$$

would depend on whether θ or γ was taken to be uniform a priori. Neither choice then could be considered uninformative.

A more sophisticated version of Laplace's principle was put forward by Jeffreys beginning in the 1930s. It depends, interestingly enough, on the frequentist notion of *Fisher information* (Chapter 4). For a *one-parameter family* $f_\mu(x)$, where the parameter space Ω is an interval of the real line $\mathcal{R}^1$, the Fisher information is defined to be

$$\mathcal{I}_\mu = E_\mu\left\{\left(\frac{\partial}{\partial\mu}\log f_\mu(x)\right)^2\right\}. \tag{3.16}$$

(For the Poisson family (3.3), $\partial/\partial\mu(\log f_\mu(x)) = x/\mu - 1$ and $\mathcal{I}_\mu = 1/\mu$.) The Jeffreys' prior $g^{\text{Jeff}}(\mu)$ is by definition

$$g^{\text{Jeff}}(\mu) = \mathcal{I}_\mu^{1/2}. \tag{3.17}$$

Because $1/\mathcal{I}_\mu$ equals, approximately, the variance σ_μ^2 of the MLE $\hat{\mu}$, an equivalent definition is

$$g^{\text{Jeff}}(\mu) = 1/\sigma_\mu. \tag{3.18}$$

Formula (3.17) does in fact transform correctly under parameter changes, avoiding the Venn–Fisher criticism.† It is known that $\hat{\theta}$ in family (3.11) has approximate standard deviation †2

$$\sigma_\theta = c(1-\theta^2), \tag{3.19}$$

yielding Jeffreys' prior (3.13) from (3.18), the constant factor c having no effect on Bayes' rule (3.5)–(3.6).

The red triangles in Figure 3.2 indicate the "95% credible interval" [0.093, 0.750] for θ, based on Jeffreys' prior. That is, the posterior probability $0.093 \le \theta \le 0.750$ equals 0.95,

$$\int_{0.093}^{0.750} g^{\text{Jeff}}\left(\theta|\hat{\theta}\right)\, d\theta = 0.95, \tag{3.20}$$

with probability 0.025 for $\theta < 0.093$ or $\theta > 0.750$. It is not an accident that this nearly equals the standard Neyman 95% confidence interval based on $f_\theta(\hat{\theta})$ (3.11). Jeffreys' prior tends to induce this nice connection between the Bayesian and frequentist worlds, at least in one-parameter families.

Multiparameter probability families, Chapter 4, make everything more

difficult. Suppose, for instance, the statistician observes 10 independent versions of the normal model (3.2), with possibly different values of μ,

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1) \qquad \text{for } i = 1, 2, \ldots, 10, \tag{3.21}$$

in standard notation. Jeffreys' prior is flat for any one of the 10 problems, which is reasonable for dealing with them separately, but the joint Jeffreys' prior

$$g(\mu_1, \mu_2, \ldots, \mu_{10}) = \text{constant}, \tag{3.22}$$

also flat, can produce disastrous overall results, as discussed in Chapter 13.

Computer-age applications are often more like (3.21) than (3.11), except with hundreds or thousands of cases rather than 10 to consider simultaneously. Uninformative priors of many sorts, including Jeffreys', are highly popular in current applications, as we will discuss. This leads to an interplay between Bayesian and frequentist methodology, the latter intended to control possible biases in the former, exemplifying our general theme of computer-age statistical inference.

3.3 Flaws in Frequentist Inference

Bayesian statistics provides an internally consistent ("coherent") program of inference. The same cannot be said of frequentism. The apocryphal story of the *meter reader* makes the point: an engineer measures the voltages on a batch of 12 tubes, using a voltmeter that is normally calibrated,

$$x \sim \mathcal{N}(\mu, 1), \tag{3.23}$$

x being any one measurement and μ the true batch voltage. The measurements range from 82 to 99, with an average of $\bar{x} = 92$, which he reports back as an unbiased estimate of μ.[†] †3

The next day he discovers a glitch in his voltmeter such that any voltage exceeding 100 would have been reported as $x = 100$. His frequentist statistician tells him that $\bar{x} = 92$ is no longer unbiased for the true expectation μ since (3.23) no longer completely describes the probability family. (The statistician says that 92 is a little too small.) The fact that the glitch didn't affect any of the actual measurements doesn't let him off the hook; $\bar{x}$ would not be unbiased for μ in future realizations of $\bar{X}$ from the actual probability model.

A Bayesian statistician comes to the meter reader's rescue. For any prior density $g(\mu)$, the posterior density $g(\mu|\boldsymbol{x}) = g(\mu) f_\mu(\boldsymbol{x})/f(\boldsymbol{x})$, where $\boldsymbol{x}$ is the vector of 12 measurements, depends only on the data $\boldsymbol{x}$ actually

observed, and *not on other potential data sets* $\boldsymbol{X}$ *that might have been seen*. The flat Jeffreys' prior $g(\mu) =$ constant yields posterior expectation $\bar{x} = 92$ for μ, irrespective of whether or not the glitch would have affected readings above 100.

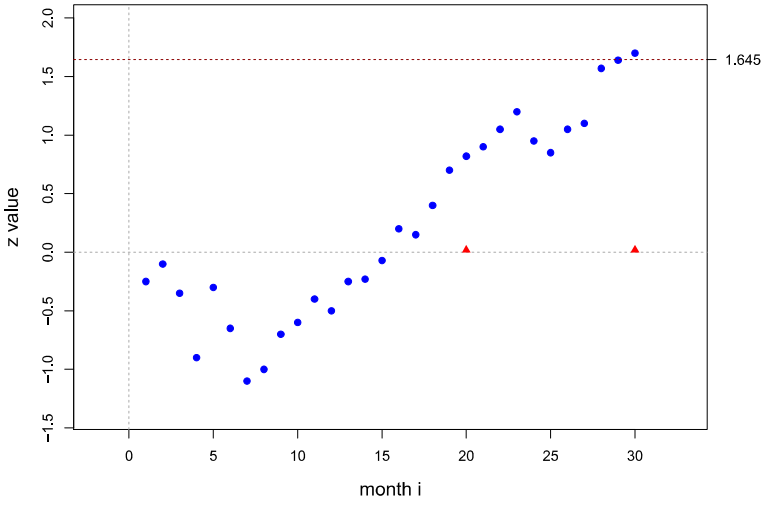

Figure 3.3 Z-values against null hypothesis $\mu = 0$ for months 1 through 30.

A less contrived version of the same phenomenon is illustrated in Figure 3.3. An ongoing experiment is being run. Each month i an independent normal variate is observed,

$$x_i \sim \mathcal{N}(\mu, 1), \tag{3.24}$$

with the intention of testing the null hypothesis $H_0 : \mu = 0$ versus the alternative $\mu > 0$. The plotted points are test statistics

$$Z_i = \sum_{j=1}^{i} x_j \Big/ \sqrt{i}\ , \tag{3.25}$$

a "z-value" based on all the data up to month i,

$$Z_i \sim \mathcal{N}\left(\sqrt{i}\,\mu, 1\right). \tag{3.26}$$

At month 30, the scheduled end of the experiment, $Z_{30} = 1.66$, just exceeding 1.645, the upper 95% point for a $\mathcal{N}(0, 1)$ distribution. Victory! The investigators get to claim "significant" rejection of H_0 at level 0.05.

Unfortunately, it turns out that the investigators broke protocol and peeked at the data at month 20, in the hope of being able to stop an expensive experiment early. This proved a vain hope, $Z_{20} = 0.79$ not being anywhere near significance, so they continued on to month 30 as originally planned. This means they effectively used the stopping rule "stop and declare significance if either Z_{20} or Z_{30} exceeds 1.645." Some computation shows that this rule had probability 0.074, not 0.05, of rejecting H_0 if it were true. Victory has turned into defeat according to the honored frequentist 0.05 criterion.

Once again, the Bayesian statistician is more lenient. The likelihood function for the full data set $\boldsymbol{x} = (x_1, x_2, \ldots, x_{30})$,

$$L_{\boldsymbol{x}}(\mu) = \prod_{i=1}^{30} e^{-\frac{1}{2}(x_i-\mu)^2}, \tag{3.27}$$

is the same irrespective of whether or not the experiment *might have* stopped early. The stopping rule doesn't affect the posterior distribution $g(\mu|\boldsymbol{x})$, which depends on $\boldsymbol{x}$ only through the likelihood (3.7).

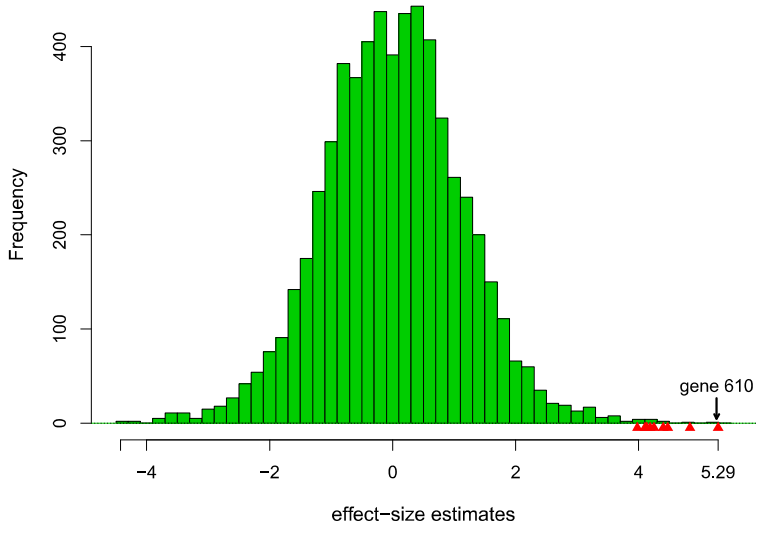

Figure 3.4 Unbiased effect-size estimates for 6033 genes, prostate cancer study. The estimate for gene 610 is $x_{610} = 5.29$. What is its effect size?

The lenient nature of Bayesian inference can look less benign in multi-

parameter settings. Figure 3.4 concerns a prostate cancer study comparing 52 patients with 50 healthy controls. Each man had his genetic activity measured for a panel of $N = 6033$ genes. A statistic x was computed for each gene,[5] comparing the patients with controls, say† †4

$$x_i \sim \mathcal{N}(\mu_i, 1) \qquad i = 1, 2, \ldots, N, \tag{3.28}$$

where μ_i represents the *true effect size* for gene i. Most of the genes, probably not being involved in prostate cancer, would be expected to have effect sizes near 0, but the investigators hoped to spot a few large μ_i values, either positive or negative.

The histogram of the 6033 x_i values does in fact reveal some large values, $x_{610} = 5.29$ being the winner. Question: what estimate should we give for μ_{610}? Even though x_{610} was individually unbiased for μ_{610}, a frequentist would (correctly) worry that focusing attention on the *largest* of 6033 values would produce an upward bias, and that our estimate should downwardly correct 5.29. "Selection bias," "regression to the mean," and "the winner's curse" are three names for this phenomenon.

Bayesian inference, surprisingly, is immune to selection bias.† Irrespective of whether gene 610 was prespecified for particular attention or only came to attention as the "winner," the Bayes' estimate for μ_{610} given all the data stays the same. This isn't obvious, but follows from the fact that any data-based selection process does not affect the likelihood function in (3.7). †5

What *does* affect Bayesian inference is the prior $g(\boldsymbol{\mu})$ for the full vector $\boldsymbol{\mu}$ of 6033 effect sizes. The flat prior, $g(\boldsymbol{\mu})$ constant, results in the dangerous overestimate $\hat{\mu}_{610} = x_{610} = 5.29$. A more appropriate uninformative prior appears as part of the empirical Bayes calculations of Chapter 15 (and gives $\hat{\mu}_{610} = 4.11$). The operative point here is that there is a price to be paid for the desirable properties of Bayesian inference. Attention shifts from choosing a good frequentist procedure to choosing an appropriate prior distribution. This can be a formidable task in high-dimensional problems, the very kinds featured in computer-age inference.

3.4 A Bayesian/Frequentist Comparison List

Bayesians and frequentists start out on the same playing field, a family of probability distributions $f_\mu(x)$ (3.1), but play the game in orthogonal

[5] The statistic was the two-sample t-statistic (2.17) transformed to normality (3.28); see the endnotes.

directions, as indicated schematically in Figure 3.5: Bayesian inference proceeds vertically, with x fixed, according to the posterior distribution $g(\mu|x)$, while frequentists reason horizontally, with μ fixed and x varying. Advantages and disadvantages accrue to both strategies, some of which are compared next.

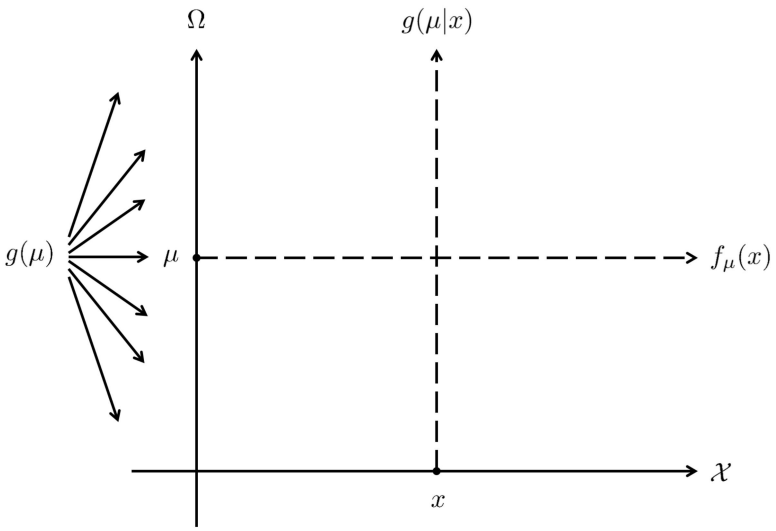

Figure 3.5 Bayesian inference proceeds vertically, given x; frequentist inference proceeds horizontally, given μ.

- Bayesian inference requires a prior distribution $g(\mu)$. When past experience provides $g(\mu)$, as in the twins example, there is every good reason to employ Bayes' theorem. If not, techniques such as those of Jeffreys still permit the use of Bayes' rule, but the results lack the full logical force of the theorem; the Bayesian's right to ignore selection bias, for instance, must then be treated with caution.
- Frequentism replaces the choice of a prior with the choice of a method, or algorithm, $t(x)$, designed to answer the specific question at hand. This adds an arbitrary element to the inferential process, and can lead to meter-reader kinds of contradictions. Optimal choice of $t(x)$ reduces arbitrary behavior, but computer-age applications typically move outside the safe waters of classical optimality theory, lending an ad-hoc character to frequentist analyses.
- Modern data-analysis problems are often approached via a favored meth-

odology, such as logistic regression or regression trees in the examples of Chapter 8. This plays into the methodological orientation of frequentism, which is more flexible than Bayes' rule in dealing with specific algorithms (though one always hopes for a reasonable Bayesian justification for the method at hand).

- Having chosen $g(\mu)$, only a single probability distribution $g(\mu|x)$ is in play for Bayesians. Frequentists, by contrast, must struggle to balance the behavior of $t(x)$ over a family of possible distributions, since μ in Figure 3.5 is unknown. The growing popularity of Bayesian applications (usually begun with uninformative priors) reflects their simplicity of application and interpretation.
- The simplicity argument cuts both ways. The Bayesian essentially bets it all on the choice of his or her prior being correct, or at least not harmful. Frequentism takes a more defensive posture, hoping to do well, or at least not poorly, whatever μ might be.
- A Bayesian analysis answers *all* possible questions at once, for example, estimating $E\{\text{gfr}\}$ or $\Pr\{\text{gfr} < 40\}$ or anything else relating to Figure 2.1. Frequentism focuses on the problem at hand, requiring different estimators for different questions. This is more work, but allows for more intense inspection of particular problems. In situation (2.9) for example, estimators of the form

$$\sum (x_i - \bar{x})^2/(n-c) \tag{3.29}$$

might be investigated for different choices of the constant c, hoping to reduce expected mean-squared error.

- The simplicity of the Bayesian approach is especially appealing in dynamic contexts, where data arrives sequentially and updating one's beliefs is a natural practice. Bayes' rule was used to devastating effect before the 2012 US presidential election, updating sequential polling results to correctly predict the outcome in all 50 states. Bayes' theorem is an excellent tool in general for combining statistical evidence from disparate sources, the closest frequentist analog being maximum likelihood estimation.
- In the absence of genuine prior information, a whiff of subjectivity[6] hangs over Bayesian results, even those based on uninformative priors. Classical frequentism claimed for itself the high ground of scientific objectivity, especially in contentious areas such as drug testing and approval, where skeptics as well as friends hang on the statistical details.

Figure 3.5 is soothingly misleading in its schematics: μ and x will

[6] Here we are not discussing the important subjectivist school of Bayesian inference, of Savage, de Finetti, and others, covered in Chapter 13.

typically be high-dimensional in the chapters that follow, sometimes *very* high-dimensional, straining to the breaking point both the frequentist and the Bayesian paradigms. Computer-age statistical inference at its most successful *combines* elements of the two philosophies, as for instance in the empirical Bayes methods of Chapter 6, and the lasso in Chapter 16. There are two potent arrows in the statistician's philosophical quiver, and faced, say, with 1000 parameters and 1,000,000 data points, there's no need to go hunting armed with just one of them.

3.5 Notes and Details

Thomas Bayes, if transferred to modern times, might well be employed as a successful professor of mathematics. Actually, he was a mid-eighteenth-century nonconformist English minister with substantial mathematical interests. Richard Price, a leading figure of letters, science, and politics, had Bayes' theorem published in the 1763 *Transactions of the Royal Society* (two years after Bayes' death), his interest being partly theological, with the rule somehow proving the existence of God. Bellhouse's (2004) biography includes some of Bayes' other mathematical accomplishments.

Harold Jeffreys was another part-time statistician, working from his day job as the world's premier geophysicist of the inter-war period (and fierce opponent of the theory of continental drift). What we called *uninformative* priors are also called *noninformative* or *objective*. Jeffreys' brand of Bayesianism had a dubious reputation among Bayesians in the period 1950–1990, with preference going to subjective analysis of the type advocated by Savage and de Finetti. The introduction of *Markov chain Monte Carlo* methodology was the kind of technological innovation that changes philosophies. MCMC (Chapter 13), being very well suited to Jeffreys-style analysis of Big Data problems, moved Bayesian statistics out of the textbooks and into the world of computer-age applications. Berger (2006) makes a spirited case for the objective Bayes approach.

†1 [p. 26] *Correlation coefficient density.* Formula (3.11) for the correlation coefficient density was R. A. Fisher's debut contribution to the statistics literature. Chapter 32 of Johnson and Kotz (1970b) gives several equivalent forms. The constant c in (3.19) is often taken to be $(n-3)^{-1/2}$, with n the sample size.

†2 [p. 29] *Jeffreys' prior and transformations.* Suppose we change parameters from μ to $\tilde{\mu}$ in a smoothly differentiable way. The new family $\tilde{f}_{\tilde{\mu}}(x)$

satisfies

$$\frac{\partial}{\partial\tilde{\mu}}\log\tilde{f}_{\tilde{\mu}}(x)=\frac{\partial\mu}{\partial\tilde{\mu}}\frac{\partial}{\partial\mu}\log f_{\mu}(x). \tag{3.30}$$

Then $\tilde{\mathcal{I}}_{\tilde{\mu}}=\left(\frac{\partial\mu}{\partial\tilde{\mu}}\right)^2\mathcal{I}_{\mu}$ (3.16) and $\tilde{g}^{\text{Jeff}}(\tilde{\mu})=\left|\frac{\partial\mu}{\partial\tilde{\mu}}\right|g^{\text{Jeff}}(\mu)$. But this just says that $g^{\text{Jeff}}(\mu)$ transforms correctly to $\tilde{g}^{\text{Jeff}}(\tilde{\mu})$.

†3 [p. 30] The *meter-reader* fable is taken from Edwards' (1992) book *Likelihood*, where he credits John Pratt. It nicely makes the point that frequentist inferences, which are calibrated in terms of possible observed data sets X, may be inappropriate for the actual observation x. This is the difference between working in the horizontal and vertical directions of Figure 3.5.

†4 [p. 33] *Two-sample t-statistic.* Applied to gene i's data in the prostate study, the two-sample t-statistic t_i (2.17) has theoretical null hypothesis distribution t_{100}, a Student's t distribution with 100 degrees of freedom; x_i in (3.28) is $\Phi^{-1}(F_{100}(t_i))$, where Φ and F_{100} are the cumulative distribution functions of standard normal and t_{100} variables. Section 7.4 of Efron (2010) motivates approximation (3.28).

†5 [p. 33] *Selection bias.* Senn (2008) discusses the immunity of Bayesian inferences to selection bias and other "paradoxes," crediting Phil Dawid for the original idea. The article catches the possible uneasiness of following Bayes' theorem too literally in applications.

The 22 students in Table 3.1 were randomly selected from a larger data set of 88 in Mardia *et al.* (1979) (which gave $\hat{\theta}=0.553$). Welch and Peers (1963) initiated the study of priors whose credible intervals, such as $[0.093, 0.750]$ in Figure 3.2, match frequentist confidence intervals. In one-parameter problems, Jeffreys' priors provide good matches, but not ususally in multiparameter situations. In fact, no single multiparameter prior can give good matches for all one-parameter subproblems, a source of tension between Bayesian and frequentist methods revisited in Chapter 11.

4

Fisherian Inference and Maximum Likelihood Estimation

Sir Ronald Fisher was arguably the most influential anti-Bayesian of all time, but that did not make him a conventional frequentist. His key data-analytic methods—analysis of variance, significance testing, and maximum likelihood estimation—were almost always applied frequentistically. Their Fisherian rationale, however, often drew on ideas neither Bayesian nor frequentist in nature, or sometimes the two in combination. Fisher's work held a central place in twentieth-century applied statistics, and some of it, particularly maximum likelihood estimation, has moved forcefully into computer-age practice. This chapter's brief review of Fisherian methodology sketches parts of its unique philosophical structure, while concentrating on those topics of greatest current importance.

4.1 Likelihood and Maximum Likelihood

Fisher's seminal work on estimation focused on the likelihood function, or more exactly its logarithm. For a family of probability densities $f_\mu(x)$ (3.1), the *log likelihood function* is

$$l_x(\mu) = \log\{f_\mu(x)\}, \tag{4.1}$$

the notation $l_x(\mu)$ emphasizing that the parameter vector μ is varying while the observed data vector x is fixed. The *maximum likelihood estimate* (MLE) is the value of μ in parameter space Ω that maximizes $l_x(\mu)$,

$$\text{MLE}: \quad \hat{\mu} = \arg\max_{\mu\in\Omega}\{l_x(\mu)\}. \tag{4.2}$$

It can happen that $\hat{\mu}$ doesn't exist or that there are multiple maximizers, but here we will assume the usual case where $\hat{\mu}$ exists uniquely. More careful references are provided in the endnotes.

Definition (4.2) is extended to provide maximum likelihood estimates

for a function $\theta = T(\mu)$ of μ according to the simple plug-in rule

$$\hat{\theta} = T(\hat{\mu}), \tag{4.3}$$

most often with θ being a scalar parameter of particular interest, such as the regression coefficient of an important covariate in a linear model.

Maximum likelihood estimation came to dominate classical applied estimation practice. Less dominant now, for reasons we will be investigating in subsequent chapters, the MLE algorithm still has iconic status, being often the method of first choice in any novel situation. There are several good reasons for its ubiquity.

1 The MLE algorithm is *automatic*: in theory, and almost in practice, a single numerical algorithm produces $\hat{\mu}$ without further statistical input. This contrasts with unbiased estimation, for instance, where each new situation requires clever theoretical calculations.
2 The MLE enjoys excellent frequentist properties. In large-sample situations, maximum likelihood estimates tend to be nearly unbiased, with the least possible variance. Even in small samples, MLEs are usually quite efficient, within say a few percent of the best possible performance.
3 The MLE also has reasonable Bayesian justification. Looking at Bayes' rule (3.7),

$$g(\mu|x) = c_x g(\mu) e^{l_x(\mu)}, \tag{4.4}$$

we see that $\hat{\mu}$ is the maximizer of the posterior density $g(\mu|x)$ if the prior $g(\mu)$ is flat, that is, constant. Because the MLE depends on the family $\mathcal{F}$ only through the likelihood function, anomalies of the meter-reader type are averted.

Figure 4.1 displays two maximum likelihood estimates for the `gfr` data of Figure 2.1. Here the data[1] is the vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, $n = 211$. We assume that $\boldsymbol{x}$ was obtained as a random sample of size n from a density $f_\mu(x)$,

$$x_i \overset{\text{iid}}{\sim} f_\mu(x) \qquad \text{for } i = 1, 2, \ldots, n, \tag{4.5}$$

"iid" abbreviating "independent and identically distributed." Two families are considered for the component density $f_\mu(x)$, the *normal*, with $\mu = (\theta, \sigma)$,

$$f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}, \tag{4.6}$$

[1] Now $\boldsymbol{x}$ is what we have been calling "x" before, while we will henceforth use x as a symbol for the individual components of $\boldsymbol{x}$.

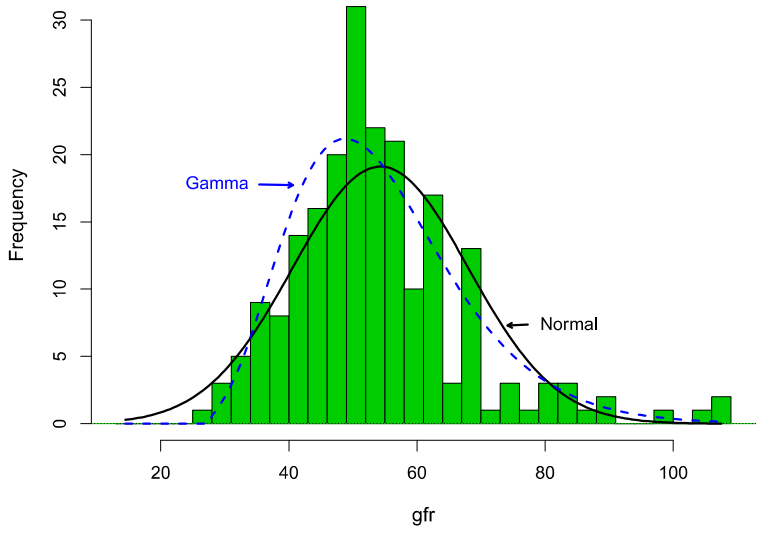

Figure 4.1 Glomerular filtration data of Figure 2.1 and two maximum-likelihood density estimates, normal (solid black), and gamma (dashed blue).

and the gamma,[2] with $\mu = (\lambda, \sigma, \nu)$,

$$f_\mu(x) = \frac{(x-\lambda)^{\nu-1}}{\sigma^\nu \Gamma(\nu)} e^{-\frac{x-\lambda}{\sigma}} \qquad \text{(for } x \geq \lambda, 0 \text{ otherwise).} \tag{4.7}$$

Since

$$f_\mu(\boldsymbol{x}) = \prod_{i=1}^{n} f_\mu(x_i) \tag{4.8}$$

under iid sampling, we have

$$l_{\boldsymbol{x}}(\mu) = \sum_{i=1}^{n} \log f_\mu(x_i) = \sum_{i=1}^{n} l_{x_i}(\mu). \tag{4.9}$$

Maximum likelihood estimates were found by maximizing $l_{\boldsymbol{x}}(\mu)$. For the normal model (4.6),

$$\left(\hat{\theta}, \hat{\sigma}\right) = (54.3, 13.7) = \left(\bar{x}, \left[\sum (x_i - \bar{x})^2 / n\right]^{1/2}\right). \tag{4.10}$$

[2] The gamma distribution is usually defined with $\lambda = 0$ as the lower limit of x. Here we are allowing the lower limit λ to vary as a free parameter.

There is no closed-form solution for gamma model (4.7), where numerical maximization gave

$$\left(\hat{\lambda}, \hat{\sigma}, \hat{\nu}\right) = (21.4, 5.47, 6.0). \tag{4.11}$$

The plotted curves in Figure 4.1 are the two MLE densities $f_{\hat{\mu}}(x)$. The gamma model gives a better fit than the normal, but neither is really satisfactory. (A more ambitious maximum likelihood fit appears in Figure 5.7.)

Most MLEs require numerical maximization, as for the gamma model. When introduced in the 1920s, maximum likelihood was criticized as computationally difficult, invidious comparisons being made with the older method of moments, which relied only on sample moments of various kinds.

There is a downside to maximum likelihood estimation that remained nearly invisible in classical applications: it is dangerous to rely upon in problems involving large numbers of parameters. If the parameter vector μ has 1000 components, each component individually may be well estimated by maximum likelihood, while the MLE $\hat{\theta} = T(\hat{\mu})$ for a quantity of particular interest can be grossly misleading.

For the prostate data of Figure 3.4, model (4.6) gives MLE $\hat{\mu}_i = x_i$ for each of the 6033 genes. This seems reasonable, but if we are interested in the maximum coordinate value

$$\theta = T(\mu) = \max_i\{\mu_i\}, \tag{4.12}$$

the MLE is $\hat{\theta} = 5.29$, almost certainly a flagrant overestimate. "Regularized" versions of maximum likelihood estimation more suitable for high-dimensional applications play an important role in succeeding chapters.

4.2 Fisher Information and the MLE

Fisher was not the first to suggest the maximum likelihood algorithm for parameter estimation. His paradigm-shifting work concerned the favorable inferential properties of the MLE, and in particular its achievement of the Fisher information bound. Only a brief heuristic review will be provided here, with more careful derivations referenced in the endnotes.

We begin[3] with a one-parameter family of densities

$$\mathcal{F} = \{f_\theta(x),\ \theta \in \Omega,\ x \in \mathcal{X}\}, \tag{4.13}$$

[3] The multiparameter case is considered in the next chapter.

where Ω is an interval of the real line, possibly infinite, while the sample space $\mathcal{X}$ may be multidimensional. (As in the Poisson example (3.3), $f_\theta(x)$ can represent a discrete density, but for convenience we assume here the continuous case, with the probability of set A equaling $\int_A f_\theta(x)\,dx$, etc.) The log likelihood function is $l_x(\theta) = \log f_\theta(x)$ and the MLE $\hat\theta = \arg\max\{l_x(\theta)\}$, with θ replacing μ in (4.1)–(4.2) in the one-dimensional case.

Dots will indicate differentiation with respect to θ, e.g., for the *score function*

$$\dot l_x(\theta) = \frac{\partial}{\partial\theta}\log f_\theta(x) = \dot f_\theta(x)/f_\theta(x). \tag{4.14}$$

The score function has expectation 0,

$$\begin{aligned}\int_{\mathcal{X}} \dot l_x(\theta)\, f_\theta(x)\,dx = \int_{\mathcal{X}} \dot f_\theta(x)\,dx &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f_\theta(x)\,dx \\ &= \frac{\partial}{\partial\theta}\,1 = 0,\end{aligned} \tag{4.15}$$

where we are assuming the regularity conditions necessary for differentiating under the integral sign at the third step.

The *Fisher information* $\mathcal{I}_\theta$ is defined to be the variance of the score function,

$$\mathcal{I}_\theta = \int_{\mathcal{X}} \dot l_x(\theta)^2\, f_\theta(x)\,dx, \tag{4.16}$$

the notation

$$\dot l_x(\theta) \sim (0, \mathcal{I}_\theta) \tag{4.17}$$

indicating that $\dot l_x(\theta)$ has mean 0 and variance $\mathcal{I}_\theta$. The term "information" is well chosen. The main result for maximum likelihood estimation, sketched next, is that the MLE $\hat\theta$ has an approximately normal distribution with mean θ and variance $1/\mathcal{I}_\theta$,

$$\hat\theta \mathrel{\dot\sim} \mathcal{N}(\theta, 1/\mathcal{I}_\theta), \tag{4.18}$$

and that no "nearly unbiased" estimator of θ can do better. In other words, bigger Fisher information implies smaller variance for the MLE.

The second derivative of the log likelihood function

$$\ddot l_x(\theta) = \frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{\ddot f_\theta(x)}{f_\theta(x)} - \left(\frac{\dot f_\theta(x)}{f_\theta(x)}\right)^2 \tag{4.19}$$

has expectation

$$E_\theta\left\{\ddot{l}_x(\theta)\right\} = -\mathcal{I}_\theta \tag{4.20}$$

(the $\ddot{f}_\theta(x)/f_\theta(x)$ term having expectation 0 as in (4.15)). We can write

$$-\ddot{l}_x(\theta) \sim (\mathcal{I}_\theta, \mathcal{J}_\theta), \tag{4.21}$$

where $\mathcal{J}_\theta$ is the variance of $\ddot{l}_x(\theta)$.

Now suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample from $f_\theta(x)$, as in (4.5), so that the total score function $\dot{l}_{\boldsymbol{x}}(\theta)$, as in (4.9), is

$$\dot{l}_{\boldsymbol{x}}(\theta) = \sum_{i=1}^{n} \dot{l}_{x_i}(\theta), \tag{4.22}$$

and similarly

$$-\ddot{l}_{\boldsymbol{x}}(\theta) = \sum_{i=1}^{n} -\ddot{l}_{x_i}(\theta). \tag{4.23}$$

The MLE $\hat{\theta}$ based on the full sample $\boldsymbol{x}$ satisfies the maximizing condition $\dot{l}_{\boldsymbol{x}}(\hat{\theta}) = 0$. A first-order Taylor series gives the approximation

$$0 = \dot{l}_{\boldsymbol{x}}\left(\hat{\theta}\right) \doteq \dot{l}_{\boldsymbol{x}}(\theta) + \ddot{l}_{\boldsymbol{x}}(\theta)\left(\hat{\theta} - \theta\right), \tag{4.24}$$

or

$$\hat{\theta} \doteq \theta + \frac{\dot{l}_{\boldsymbol{x}}(\theta)/n}{-\ddot{l}_{\boldsymbol{x}}(\theta)/n}. \tag{4.25}$$

Under reasonable regularity conditions, (4.17) and the central limit theorem imply that

$$\dot{l}_{\boldsymbol{x}}(\theta)/n \doteq \mathcal{N}(0, \mathcal{I}_\theta/n), \tag{4.26}$$

while the law of large numbers has $-\ddot{l}_{\boldsymbol{x}}(\theta)/n$ approaching the constant $\mathcal{I}_\theta$ (4.21).

Putting all of this together, (4.25) produces Fisher's fundamental theorem for the MLE, that in large samples

$$\hat{\theta} \doteq \mathcal{N}\left(\theta, 1/(n\mathcal{I}_\theta)\right). \tag{4.27}$$

This is the same as result (4.18) since the total Fisher information in an iid sample (4.5) is $n\mathcal{I}_\theta$, as can be seen by taking expectations in (4.23).

In the case of normal sampling,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n, \tag{4.28}$$

with σ^2 known, we compute the log likelihood

$$l_{\boldsymbol{x}}(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2). \tag{4.29}$$

This gives

$$\dot{l}_{\boldsymbol{x}}(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) \quad \text{and} \quad -\ddot{l}_{\boldsymbol{x}}(\theta) = \frac{n}{\sigma^2}, \tag{4.30}$$

yielding the familiar result $\hat{\theta} = \bar{x}$ and, since $\mathcal{I}_\theta = 1/\sigma^2$,

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2/n) \tag{4.31}$$

from (4.27).

This brings us to an aspect of Fisherian inference neither Bayesian nor frequentist. Fisher believed there was a "logic of inductive inference" that would produce the *correct* answer to any statistical question, in the same way ordinary logic solves deductive problems. His principal tactic was to logically reduce a complicated inferential question to a simple form where the solution should be obvious to all.

Fisher's favorite target for the obvious was (4.31), where a single scalar observation $\hat{\theta}$ is normally distributed around the unknown parameter of interest θ, with known variance σ^2/n. Then everyone should agree in the absence of prior information that $\hat{\theta}$ is the best estimate of θ, that θ has about 95% chance of lying in the interval $\hat{\theta} \pm 1.96\hat{\sigma}/\sqrt{n}$, etc.

Fisher was astoundingly resourceful at reducing statistical problems to the form (4.31). Sufficiency, efficiency, conditionality, and ancillarity were all brought to bear, with the maximum likelihood approximation (4.27) being the most influential example. Fisher's logical system is not in favor these days, but its conclusions remain as staples of conventional statistical practice.

Suppose that $\tilde{\theta} = t(\boldsymbol{x})$ is any *unbiased* estimate of θ based on an iid sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from $f_\theta(x)$. That is,

$$\theta = E_\theta\{t(\boldsymbol{x})\}. \tag{4.32}$$

Then the *Cramér–Rao lower bound*, described in the endnotes, says that
†1 the variance of $\tilde{\theta}$ exceeds the Fisher information bound (4.27),[†]

$$\mathrm{var}_\theta\left\{\tilde{\theta}\right\} \geq 1/(n\mathcal{I}_\theta). \tag{4.33}$$

A loose interpretation is that the MLE has variance at least as small as the best unbiased estimate of θ. The MLE is generally not unbiased, but

its bias is small (of order $1/n$, compared with standard deviation of order $1/\sqrt{n}$), making the comparison with unbiased estimates and the Cramér–Rao bound appropriate.

4.3 Conditional Inference

A simple example gets across the idea of conditional inference: an i.i.d. sample

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1), \qquad i = 1, 2, \ldots, n, \tag{4.34}$$

has produced estimate $\hat{\theta} = \bar{x}$. The investigators originally disagreed on an affordable sample size n and flipped a fair coin to decide,

$$n = \begin{cases} 25 & \text{probability } 1/2 \\ 100 & \text{probability } 1/2; \end{cases} \tag{4.35}$$

$n = 25$ won. Question: What is the standard deviation of $\bar{x}$?

If you answered $1/\sqrt{25} = 0.2$ then you, like Fisher, are an advocate of *conditional inference*. The *unconditional* frequentist answer says that $\bar{x}$ could have been $\mathcal{N}(\theta, 1/100)$ or $\mathcal{N}(\theta, 1/25)$ with equal probability, yielding standard deviation $[(0.01 + 0.04)/2]^{1/2} = 0.158$. Some less obvious (and less trivial) examples follow in this section, and in Chapter 9, where conditional inference plays a central role.

The data for a typical regression problem consists of pairs (x_i, y_i), $i = 1, 2, \ldots, n$, where x_i is a p-dimensional vector of covariates for the ith subject and y_i is a scalar response. In Figure 1.1, x_i is `age` and y_i the kidney fitness measure `tot`. Let $\boldsymbol{x}$ be the $n \times p$ matrix having x_i as its ith row, and $\boldsymbol{y}$ the vector of responses. A regression algorithm uses $\boldsymbol{x}$ and $\boldsymbol{y}$ to construct a function $r_{\boldsymbol{x},\boldsymbol{y}}(x)$ predicting y for any value of x, as in (1.3), where $\hat{\beta}_0$ and $\hat{\beta}_1$ were obtained using least squares.

How accurate is $r_{\boldsymbol{x},\boldsymbol{y}}(x)$? This question is usually answered under the assumption that $\boldsymbol{x}$ is fixed, not random: in other words, by *conditioning on the observed value of* $\boldsymbol{x}$. The standard errors in the second line of Table 1.1 are conditional in this sense; they are frequentist standard deviations of $\hat{\beta}_0 + \hat{\beta}_1 x$, assuming that the 157 values for `age` are fixed as observed. (A *correlation* analysis between `age` and `tot` would *not* make this assumption.)

Fisher argued for conditional inference on two grounds.

1 *More relevant inferences.* The conditional standard deviation in situation (4.35) seems obviously more relevant to the accuracy of the observed $\hat{\theta}$ for estimating θ. It is less obvious in the regression example, though arguably still the case.
2 *Simpler inferences.* Conditional inferences are often simpler to execute and interpret. This is the case with regression, where the statistician doesn't have to worry about correlation relationships among the covariates, and also with our next example, a Fisherian classic.

Table 4.1 shows the results of a randomized trial on 45 ulcer patients, comparing `new` and `old` surgical treatments. Was the `new` surgery significantly better? Fisher argued for carrying out the hypothesis test conditional on the marginals of the table $(16, 29, 21, 24)$. With the marginals fixed, the number y in the upper left cell determines the other three cells by subtraction. We need only test whether the number $y = 9$ is too big under the null hypothesis of no treatment difference, instead of trying to test the numbers in all four cells.[4]

Table 4.1 *Forty-five ulcer patients randomly assigned to either* `new` *or* `old` *surgery, with results evaluated as either* `success` *or* `failure`. *Was the* `new` *surgery significantly better?*

	`success`	`failure`	
`new`	9	12	21
`old`	7	17	24
	16	29	45

An ancillary statistic (again, Fisher's terminology) is one that contains no direct information by itself, but does determine the conditioning framework for frequentist calculations. Our three examples of ancillaries were the sample size n, the covariate matrix $\boldsymbol{x}$, and the table's marginals. "Contains no information" is a contentious claim. More realistically, the two advantages of conditioning, relevance and simplicity, are thought to outweigh the loss of information that comes from treating the ancillary statistic as nonrandom. Chapter 9 makes this case specifically for standard survival analysis methods.

[4] Section 9.3 gives the details of such tests; in the surgery example, the difference was not significant.

Our final example concerns the accuracy of a maximum likelihood estimate $\hat{\theta}$. Rather than

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta, 1\Big/\left(n\mathcal{I}_{\hat{\theta}}\right)\right), \tag{4.36}$$

the plug-in version of (4.27), Fisher suggested using

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta, 1/I(\boldsymbol{x})\right), \tag{4.37}$$

where $I(\boldsymbol{x})$ is the *observed Fisher information*

$$I(\boldsymbol{x}) = -\ddot{l}_{\boldsymbol{x}}\left(\hat{\theta}\right) = -\frac{\partial^2}{\partial\theta^2} l_{\boldsymbol{x}}(\theta)\Big|_{\hat{\theta}}. \tag{4.38}$$

The expectation of $I(\boldsymbol{x})$ is $n\mathcal{I}_\theta$, so in large samples the distribution (4.37) converges to (4.36). Before convergence, however, Fisher suggested that (4.37) gives a better idea of $\hat{\theta}$'s accuracy.

As a check, a simulation was run involving i.i.d. samples $\boldsymbol{x}$ of size $n = 20$ drawn from a Cauchy density

$$f_\theta(x) = \frac{1}{\pi}\,\frac{1}{1+(x-\theta)^2}. \tag{4.39}$$

10,000 samples $\boldsymbol{x}$ of size $n = 20$ were drawn (with $\theta = 0$) and the observed information bound $1/I(\boldsymbol{x})$ computed for each. The 10,000 $\hat{\theta}$ values were grouped according to deciles of $1/I(\boldsymbol{x})$, and the observed empirical variance of $\hat{\theta}$ within each group was then calculated.

This amounts to calculating a somewhat crude estimate of the conditional variance of the MLE $\hat{\theta}$, given the observed information bound $1/I(\boldsymbol{x})$. Figure 4.2 shows the results. We see that the conditional variance is close to $1/I(\boldsymbol{x})$, as Fisher predicted. The conditioning effect is quite substantial; the unconditional variance $1/n\mathcal{I}_\theta$ is 0.10 here, while the conditional variance ranges from 0.05 to 0.20.

The observed Fisher information $I(\boldsymbol{x})$ acts as an approximate ancillary, enjoying both of the virtues claimed by Fisher: it is more relevant than the unconditional information $n\mathcal{I}_{\hat{\theta}}$, and it is usually easier to calculate. Once $\hat{\theta}$ has been found, $I(\boldsymbol{x})$ is obtained by numerical second differentiation. Unlike $\mathcal{I}_\theta$, no probability calculations are required.

There is a strong Bayesian current flowing here. A narrow peak for the log likelihood function, i.e., a large value of $I(\boldsymbol{x})$, also implies a narrow posterior distribution for θ given $\boldsymbol{x}$. Conditional inference, of which Figure 4.2 is an evocative example, helps counter the central Bayesian criticism of frequentist inference: that the frequentist properties relate to data sets possibly much different than the one actually observed. The maximum

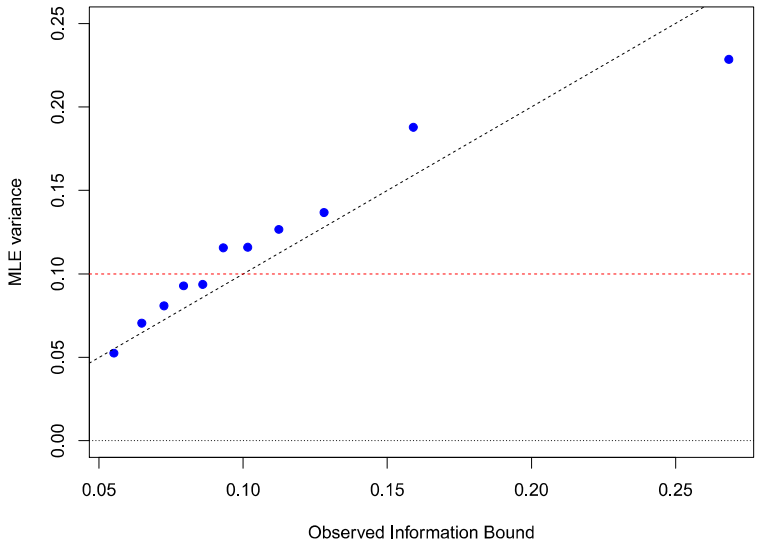

Figure 4.2 Conditional variance of MLE for Cauchy samples of size 20, plotted versus the observed information bound $1/I(\boldsymbol{x})$. Observed information bounds are grouped by quantile intervals for variance calculations (in percentages): (0–5), (5–15), . . . , (85–95), (95–100). The broken red horizontal line is the unconditional variance $1/n\mathcal{I}_\theta$.

likelihood algorithm can be interpreted both vertically and horizontally in Figure 3.5, acting as a connection between the Bayesian and frequentist worlds.

The equivalent of result (4.37) for multiparameter families, Section 5.3,

$$\hat{\mu} \doteq \mathcal{N}_p\left(\mu, I(\boldsymbol{x})^{-1}\right), \tag{4.40}$$

plays an important role in succeeding chapters, with $-I(\boldsymbol{x})$ the $p \times p$ matrix of second derivatives

$$I(\boldsymbol{x}) = -\ddot{l}_{\boldsymbol{x}}(\mu) = -\left[\frac{\partial^2}{\partial\mu_i\,\partial\mu_j}\log f_\mu(\boldsymbol{x})\right]_{\hat{\mu}}. \tag{4.41}$$

4.4 Permutation and Randomization

Fisherian methodology faced criticism for its overdependence on normal sampling assumptions. Consider the comparison between the 47 **ALL** and 25 **AML** patients in the gene 136 leukemia example of Figure 1.4. The two-sample t-statistic (1.6) had value 3.01, with two-sided significance level 0.0036 according to a Student-t null distribution with 70 degrees of freedom. All of this depended on the Gaussian, or normal, assumptions (2.12)–(2.13).

As an alternative significance-level calculation, Fisher suggested using permutations of the 72 data points. The 72 values are *randomly* divided into disjoint sets of size 47 and 25, and the two-sample t-statistic (2.17) is recomputed. This is done some large number B times, yielding permutation t-values $t_1^*, t_2^*, \ldots, t_B^*$. The two-sided permutation significance level for the original value t is then the proportion of the t_i^* values exceeding t in absolute value,

$$\#\{|t_i^*| \geq |t|\}/B. \tag{4.42}$$

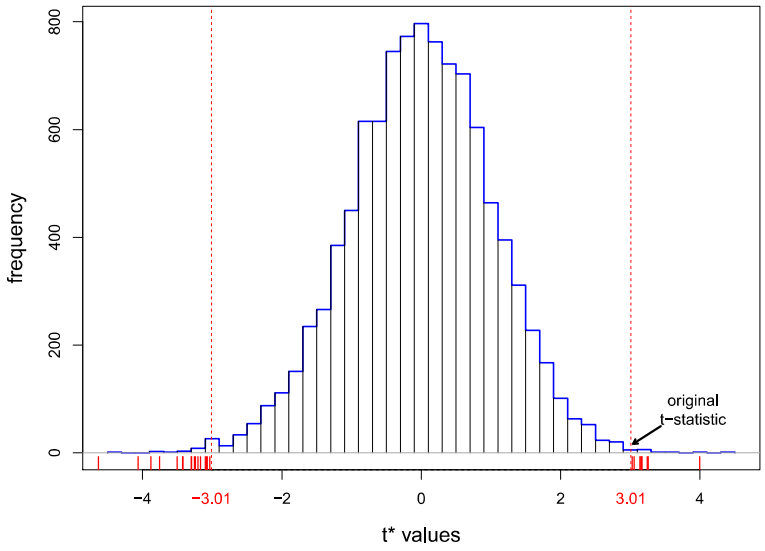

Figure 4.3 10,000 permutation t^*-values for testing **ALL** vs **AML**, for gene 136 in the **leukemia** data of Figure 1.3. Of these, 26 t^*-values (red ticks) exceeded in absolute value the observed t-statistic 3.01, giving permutation significance level 0.0026.

Figure 4.3 shows the histogram of $B = 10{,}000$ t_i^* values for the gene 136 data in Figure 1.3: 26 of these exceeded $t = 3.01$ in absolute value, yielding significance level 0.0026 against the null hypothesis of no **ALL**/**AML** difference, close to the normal-theory significance level 0.0036. (We were a little lucky here.)

Why should we believe the permutation significance level (4.42)? Fisher provided two arguments.

- Suppose we assume as a null hypothesis that the $n = 72$ observed measurements $\boldsymbol{x}$ are an iid sample obtained from the *same* distribution $f_\mu(x)$,

$$x_i \overset{\text{iid}}{\sim} f_\mu(x) \qquad \text{for } i = 1, 2, \ldots, n. \tag{4.43}$$

  (There is no normal assumption here, say that $f_\mu(x)$ is $\mathcal{N}(\theta, \sigma^2)$.)

  Let $\boldsymbol{o}$ indicate the *order statistic* of $\boldsymbol{x}$, i.e., the 72 numbers ordered from smallest to largest, with their **AML** or **ALL** labels removed. Then it can be shown that all $72!/(47!25!)$ ways of obtaining $\boldsymbol{x}$ by dividing $\boldsymbol{o}$ into disjoint subsets of sizes 47 and 25 are equally likely under null hypothesis (4.43). A small value of the permutation significance level (4.42) indicates that the actual division of **AML**/**ALL** measurements was *not* random, but rather resulted from negation of the null hypothesis (4.43). This might be considered an example of Fisher's logic of inductive inference, where the conclusion "should be obvious to all." It is certainly an example of conditional inference, now with conditioning used to avoid specific assumptions about the sampling density $f_\mu(x)$.
- In experimental situations, Fisher forcefully argued for *randomization*, that is for randomly assigning the experimental units to the possible treatment groups. Most famously, in a clinical trial comparing drug A with drug B, each patient should be randomly assigned to A or B.

  Randomization greatly strengthens the conclusions of a permutation test. In the **AML**/**ALL** gene-136 situation, where randomization wasn't feasible, we wind up almost certain that the **AML** group has systematically larger numbers, but cannot be certain that it is the different disease states causing the difference. Perhaps the **AML** patients are older, or heavier, or have more of some other characteristic affecting gene 136. Experimental randomization *almost* guarantees that age, weight, etc., will be well-balanced between the treatment groups. Fisher's RCT (randomized clinical trial) was and is the gold standard for statistical inference in medical trials.

Permutation testing is frequentistic: a statistician following the procedure has 5% chance of rejecting a valid null hypothesis at level 0.05, etc.

Randomization inference is somewhat different, amounting to a kind of forced frequentism, with the statistician imposing his or her preferred probability mechanism upon the data. Permutation methods are enjoying a healthy computer-age revival, in contexts far beyond Fisher's original justification for the t-test, as we will see in Chapter 15.

4.5 Notes and Details

On a linear scale that puts Bayesian on the left and frequentist on the right, Fisherian inference winds up somewhere in the middle. Fisher rejected Bayesianism early on, but later criticized as "wooden" the hard-line frequentism of the Neyman–Wald decision-theoretic school. Efron (1998) locates Fisher along the Bayes–frequentist scale for several different criteria; see in particular Figure 1 of that paper.

Bayesians, of course, believe there is only one true logic of inductive inference. Fisher disagreed. His most ambitious attempt to "enjoy the Bayesian omelette without breaking the Bayesian eggs"[5] was *fiducial inference*. The simplest example concerns the normal translation model $x \sim \mathcal{N}(\theta, 1)$, where $\theta - x$ has a standard $\mathcal{N}(0, 1)$ distribution, the fiducial distribution of θ given x then being $\mathcal{N}(x, 1)$. Among Fisher's many contributions, fiducial inference was the only outright popular bust. Nevertheless the idea has popped up again in the current literature under the name "confidence distribution;" see Efron (1993) and Xie and Singh (2013). A brief discussion appears in Chapter 11.

†1 [p. 44] For an unbiased estimator $\tilde{\theta} = t(\boldsymbol{x})$ (4.32), we have

$$\int_{\mathcal{X}} t(\boldsymbol{x})\dot{l}_{\boldsymbol{x}}(\theta) f_\theta(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathcal{X}} t(\boldsymbol{x})\dot{f}_\theta(\boldsymbol{x})\, d\boldsymbol{x} = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} t(\boldsymbol{x}) f_\theta(\boldsymbol{x})\, d\boldsymbol{x} = \frac{\partial}{\partial\theta}\,\theta = 1. \tag{4.44}$$

Here $\mathcal{X}$ is $\mathcal{X}^n$, the sample space of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, and we are assuming the conditions necessary for differentiating under the integral sign; (4.44) gives $\int (t(\boldsymbol{x}) - \theta)\dot{l}_{\boldsymbol{x}}(\theta) f_\theta(\boldsymbol{x})\, d\boldsymbol{x} = 1$ (since $\dot{l}_{\boldsymbol{x}}(\theta)$ has expectation

[5] Attributed to the important Bayesian theorist L. J. Savage.

0), and then, applying the *Cauchy–Schwarz inequality*,

$$\left[\int_{\mathcal{X}} (t(\boldsymbol{x}) - \theta)\, \dot{l}_{\boldsymbol{x}}(\theta)\, f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\right]^2 \leq \left[\int_{\mathcal{X}} (t(\boldsymbol{x}) - \theta)^2\, f_\theta(\boldsymbol{x}) d\boldsymbol{x}\right]\left[\int_{\mathcal{X}} \dot{l}_{\boldsymbol{x}}(\theta)^2 f_\theta(\boldsymbol{x})\, d\boldsymbol{x}\right], \quad (4.45)$$

or

$$1 \leq \text{var}_\theta\left\{\tilde{\theta}\right\} \mathcal{I}_\theta. \quad (4.46)$$

This verifies the Cramér–Rao lower bound (4.33): the optimal variance for an unbiased estimator is one over the Fisher information.

Optimality results are a sign of scientific maturity. Fisher information and its estimation bound mark the transition of statistics from a collection of ad-hoc techniques to a coherent discipline. (We have lost some ground recently, where, as discussed in Chapter 1, ad-hoc algorithmic coinages have outrun their inferential justification.) Fisher's information bound was a major mathematical innovation, closely related to and predating, Heisenberg's uncertainty principle and Shannon's information bound; see Dembo *et al.* (1991).

Unbiased estimation has strong appeal in statistical applications, where "biased," its opposite, carries a hint of self-interested data manipulation. In large-scale settings, such as the prostate study of Figure 3.4, one can, however, strongly argue for biased estimates. We saw this for gene 610, where the usual unbiased estimate $\hat{\mu}_{610} = 5.29$ is almost certainly too large. Biased estimation will play a major role in our subsequent chapters.

Maximum likelihood estimation is effectively unbiased in most situations. Under repeated sampling, the expected mean squared error

$$\text{MSE} = E\left\{\left(\hat{\theta} - \theta\right)^2\right\} = \text{variance} + \text{bias}^2 \quad (4.47)$$

has order-of-magnitude variance $= O(1/n)$ and $\text{bias}^2 = O(1/n^2)$, the latter usually becoming negligible as sample size n increases. (Important exceptions, where bias *is* substantial, can occur if $\hat{\theta} = T(\hat{\mu})$ when $\hat{\mu}$ is high-dimensional, as in the James–Stein situation of Chapter 7.) Section 10 of Efron (1975) provides a detailed analysis.

Section 9.2 of Cox and Hinkley (1974) gives a careful and wide-ranging account of the MLE and Fisher information. Lehmann (1983) covers the same ground, somewhat more technically, in his Chapter 6.

5

Parametric Models and Exponential Families

We have been reviewing classic approaches to statistical inference—frequentist, Bayesian, and Fisherian—with an eye toward examining their strengths and limitations in modern applications. Putting philosophical differences aside, there is a common methodological theme in classical statistics: a strong preference for low-dimensional parametric models; that is, for modeling data-analysis problems using parametric families of probability densities (3.1),

$$\mathcal{F} = \{f_\mu(x); x \in \mathcal{X}, \mu \in \Omega\}, \tag{5.1}$$

where the dimension of parameter μ is small, perhaps no greater than 5 or 10 or 20. The inverted nomenclature "nonparametric" suggests the predominance of classical parametric methods.

Two words explain the classic preference for parametric models: mathematical tractability. In a world of sliderules and slow mechanical arithmetic, mathematical formulation, by necessity, becomes the computational tool of choice. Our new computation-rich environment has unplugged the mathematical bottleneck, giving us a more realistic, flexible, and far-reaching body of statistical techniques. But the classic parametric families still play an important role in computer-age statistics, often assembled as small parts of larger methodologies (as with the generalized linear models of Chapter 8). This chapter[1] presents a brief review of the most widely used parametric models, ending with an overview of exponential families, the great connecting thread of classical theory and a player of continuing importance in computer-age applications.

[1] This chapter covers a large amount of technical material for use later, and may be reviewed lightly at first reading.

5.1 Univariate Families

Univariate parametric families, in which the sample space $\mathcal{X}$ of observation x is a subset of the real line $\mathcal{R}^1$, are the building blocks of most statistical analyses. Table 5.1 names and describes the five most familiar univariate families: normal, Poisson, binomial, gamma, and beta. (The chi-squared distribution with n degrees of freedom χ^2_n is also included since it is distributed as $2 \cdot \text{Gam}(n/2, 1)$.) The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is a shifted and scaled version of the $\mathcal{N}(0, 1)$ distribution[2] used in (3.27),

$$\mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma\mathcal{N}(0, 1). \tag{5.2}$$

Table 5.1 *Five familiar univariate densities, and their sample spaces $\mathcal{X}$, parameter spaces Ω, and expectations and variances; chi-squared distribution with n degrees of freedom is* $2\,\text{Gam}(n/2, 1)$.

Name, Notation	Density	$\mathcal{X}$	Ω	Expectation, Variance
Normal $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$\mathcal{R}^1$	$\mu \in \mathcal{R}^1$ $\sigma^2 > 0$	μ σ^2
Poisson $\text{Poi}(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}$	$\{0, 1, \dots\}$	$\mu > 0$	μ μ
Binomial $\text{Bi}(n, \pi)$	$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$	$\{0, 1, \dots, n\}$	$0 < \pi < 1$	$n\pi$ $n\pi(1-\pi)$
Gamma $\text{Gam}(\nu, \sigma)$	$\frac{x^{\nu-1}e^{-x/\sigma}}{\sigma^\nu\Gamma(\nu)}$	$x \geq 0$	$\nu > 0$ $\sigma > 0$	$\sigma\nu$ $\sigma^2\nu$
Beta $\text{Be}(\nu_1, \nu_2)$	$\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1}(1-x)^{\nu_2-1}$	$0 \leq x \leq 1$	$\nu_1 > 0$ $\nu_2 > 0$	$\nu_1/(\nu_1+\nu_2)$ $\frac{\nu_1\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$

Relationships abound among the table's families. For instance, independent gamma variables $\text{Gam}(\nu_1, \sigma)$ and $\text{Gam}(\nu_2, \sigma)$ yield a beta variate according to

$$\text{Be}(\nu_1, \nu_2) \sim \frac{\text{Gam}(\nu_1, \sigma)}{\text{Gam}(\nu_1, \sigma) + \text{Gam}(\nu_2, \sigma)}. \tag{5.3}$$

The binomial and Poisson are particularly close cousins. A $\text{Bi}(n, \pi)$ distribution (the number of heads in n independent flips of a coin with probabil-

[2] The notation in (5.2) indicates that if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(0, 1)$ then X and $\mu + \sigma Y$ have the same distribution.

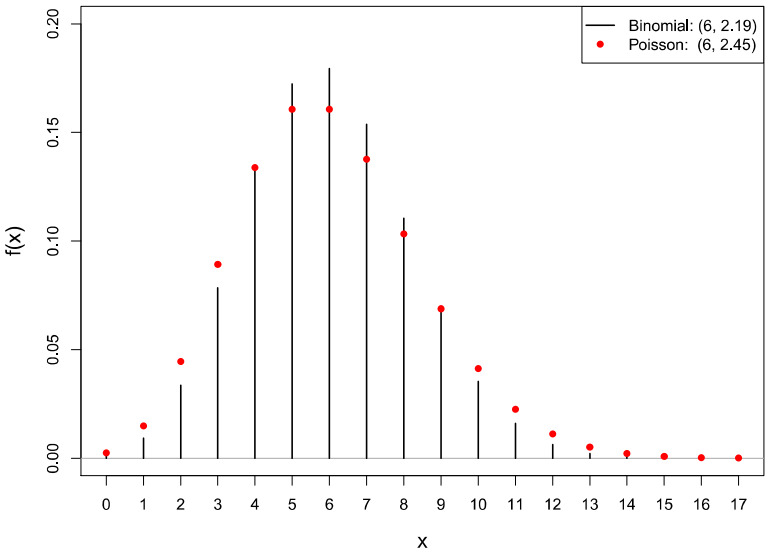

Figure 5.1 Comparison of the binomial distribution Bi(30, 0.2) (black lines) with the Poisson Poi(6) (red dots). In the legend we show the mean and standard deviation for each distribution.

ity of heads π) approaches a Poi($n\pi$) distribution,

$$\text{Bi}(n, \pi) \doteq \text{Poi}(n\pi) \tag{5.4}$$

as n grows large and π small, the notation $\doteq$ indicating approximate equality of the two distributions. Figure 5.1 shows the approximation already working quite effectively for $n = 30$ and $\pi = 0.2$.

The five families in Table 5.1 have five different sample spaces, making them appropriate in different situations. Beta distributions, for example, are natural candidates for modeling continuous data on the unit interval $[0, 1]$. Choices of the two parameters (ν_1, ν_2) provide a variety of possible shapes, as illustrated in Figure 5.2. Later we will discuss general exponential families, unavailable in classical theory, that greatly expand the catalog of possible shapes.

5.2 The Multivariate Normal Distribution

Classical statistics produced a less rich catalog of multivariate distributions, ones where the sample space $\mathcal{X}$ exists in $\mathcal{R}^p$, p-dimensional Eu-

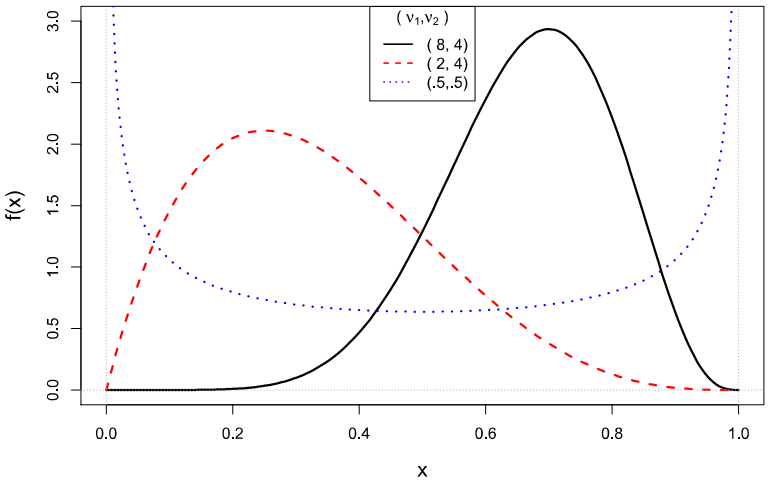

Figure 5.2 Three beta densities, with (ν_1, ν_2) indicated.

clidean space, $p > 1$. By far the greatest amount of attention focused on the multivariate normal distribution.

A random vector $x = (x_1, x_2, \ldots, x_p)'$, normally distributed or not, has *mean vector*

$$\mu = E\{x\} = \big(E\{x_1\}, E\{x_2\}, \ldots, E\{x_p\}\big)' \tag{5.5}$$

and $p \times p$ *covariance matrix*[3]

$$\boldsymbol{\Sigma} = E\left\{(x-\mu)(x-\mu)'\right\} = \left(E\left\{(x_i-\mu_i)(x_j-\mu_j)\right\}\right). \tag{5.6}$$

(The outer product uv' of vectors u and v is the matrix having elements $u_i v_j$.) We will use the convenient notation

$$x \sim (\mu, \boldsymbol{\Sigma}) \tag{5.7}$$

for (5.5) and (5.6), reducing to the familiar form $x \sim (\mu, \sigma^2)$ in the univariate case.

Denoting the entries of $\boldsymbol{\Sigma}$ by σ_{ij}, for i and j equaling $1, 2, \ldots, p$, the diagonal elements are variances,

$$\sigma_{ii} = \text{var}(x_i). \tag{5.8}$$

[3] The notation $\boldsymbol{\Sigma} = (\sigma_{ij})$ defines the ijth element of a matrix.

The off-diagonal elements relate to the correlations between the coordinates of x,

$$\text{cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{5.9}$$

The multivariate normal distribution extends the univariate definition $\mathcal{N}(\mu, \sigma^2)$ in Table 5.1. To begin with, let $z = (z_1, z_2, \ldots, z_p)'$ be a vector of p independent $\mathcal{N}(0, 1)$ variates, with probability density function

$$f(z) = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2}\sum_1^p z_i^2} = (2\pi)^{-\frac{p}{2}} e^{-\frac{1}{2}z'z} \tag{5.10}$$

according to line 1 of Table 5.1.

The multivariate normal family is obtained by linear transformations of z: let μ be a p-dimensional vector and $\boldsymbol{T}$ a $p \times p$ nonsingular matrix, and define the random vector

$$x = \mu + \boldsymbol{T}z. \tag{5.11}$$

Following the usual rules of probability transformations yields the density of x,

$$f_{\mu,\boldsymbol{\Sigma}}(x) = \frac{(2\pi)^{-p/2}}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\boldsymbol{\Sigma}^{-1}(x-\mu)}, \tag{5.12}$$

where $\boldsymbol{\Sigma}$ is the $p \times p$ symmetric positive definite matrix

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{T}' \tag{5.13}$$

and $|\boldsymbol{\Sigma}|$ its determinant;[†] $f_{\mu,\boldsymbol{\Sigma}}(x)$, the p-dimensional multivariate normal distribution with mean μ and covariance $\boldsymbol{\Sigma}$, is denoted [†1]

$$x \sim \mathcal{N}_p(\mu, \boldsymbol{\Sigma}). \tag{5.14}$$

Figure 5.3 illustrates the bivariate normal distribution with $\mu = (0, 0)'$ and $\boldsymbol{\Sigma}$ having $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = 0.5$ (so $\text{cor}(x_1, x_2) = 0.5$). The bell-shaped mountain on the left is a plot of density (5.12). The right panel shows a scatterplot of 2000 points drawn from this distribution. Concentric ellipses illustrate curves of constant density,

$$(x - \mu)'\boldsymbol{\Sigma}^{-1}(x - \mu) = \text{constant}. \tag{5.15}$$

Classical multivariate analysis was the study of the multivariate normal distribution, both of its probabilistic and statistical properties. The notes reference some important (and lengthy) multivariate texts. Here we will just recall a couple of results useful in the chapters to follow.

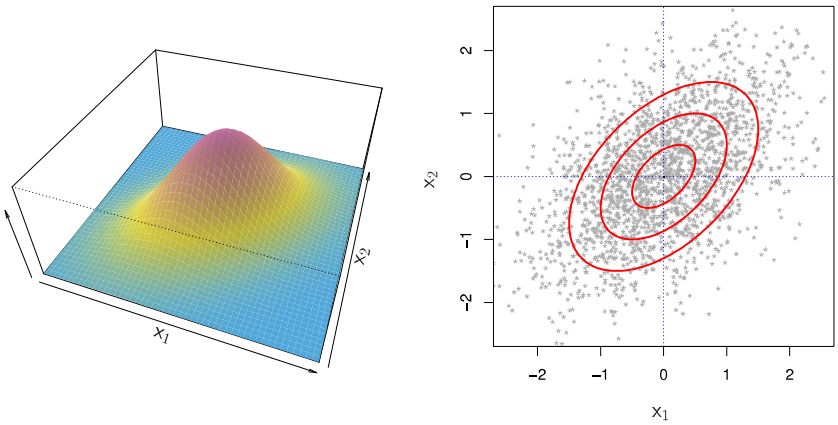

Figure 5.3 *Left:* bivariate normal density, with $\text{var}(x_1) = \text{var}(x_2) = 1$ and $\text{cor}(x_1, x_2) = 0.5$. *Right:* sample of 2000 (x_1, x_2) pairs from this bivariate normal density.

Suppose that $x = (x_1, x_2, \ldots, x_p)'$ is partitioned into

$$x_{(1)} = (x_1, x_2, \ldots, x_{p_1})' \quad \text{and} \quad x_{(2)} = (x_{p_1+1}, x_{p_1+2}, \ldots, x_{p_1+p_2})', \tag{5.16}$$

$p_1 + p_2 = p$, with μ and $\boldsymbol{\Sigma}$ similarly partitioned,

$$\begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{5.17}$$

(so $\boldsymbol{\Sigma}_{11}$ is $p_1 \times p_1$, $\boldsymbol{\Sigma}_{12}$ is $p_1 \times p_2$, etc.). Then the conditional distribution of $x_{(2)}$ given $x_{(1)}$ is itself normal,† †2

$$x_{(2)}|x_{(1)} \sim \mathcal{N}_{p_2} \left(\mu_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(x_{(1)} - \mu_{(1)}), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \right). \tag{5.18}$$

If $p_1 = p_2 = 1$, then (5.18) reduces to

$$x_2|x_1 \sim \mathcal{N}\left(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1 - \mu_1),\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right); \tag{5.19}$$

here σ_{12}/σ_{11} is familiar as the linear regression coefficient of x_2 as a function of x_1, while $\sigma_{12}^2/\sigma_{11}\sigma_{22}$ equals $\text{cor}(x_1, x_2)^2$, the squared proportion R^2 of the variance of x_2 explained by x_1. Hence we can write the (unexplained) variance term in (5.19) as $\sigma_{22}(1 - R^2)$.

Bayesian statistics also makes good use of the normal family. It helps to begin with the univariate case $x \sim \mathcal{N}(\mu, \sigma^2)$, where now we assume that

the expectation vector itself has a normal prior distribution $\mathcal{N}(M, A)$:

$$\mu \sim \mathcal{N}(M, A) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, \sigma^2). \tag{5.20}$$

Bayes' theorem and some algebra show that the posterior distribution of μ having observed x is normal,[†] [†3]

$$\mu|x \sim \mathcal{N}\left(M + \frac{A}{A+\sigma^2}(x-M), \frac{A\sigma^2}{A+\sigma^2}\right). \tag{5.21}$$

The posterior expectation $\hat{\mu}_{\text{Bayes}} = M + (A/(A+\sigma^2))(x-M)$ is a *shrinkage estimator* of μ: if, say, A equals σ^2, then $\hat{\mu}_{\text{Bayes}} = M + (x-M)/2$ is shrunk half the way back from the unbiased estimate $\hat{\mu} = x$ toward the prior mean M, while the posterior variance $\sigma^2/2$ of $\hat{\mu}_{\text{Bayes}}$ is only one-half that of $\hat{\mu}$.

The multivariate version of the Bayesian setup (5.20) is

$$\mu \sim \mathcal{N}_p(M, \boldsymbol{A}) \quad \text{and} \quad x|\mu \sim \mathcal{N}_p(\mu, \boldsymbol{\Sigma}), \tag{5.22}$$

now with M and μ p-vectors, and $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ positive definite $p \times p$ matrices. As indicated in the notes, the posterior distribution of μ given x is then

$$\mu|x \sim \mathcal{N}_p\left(M + \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{\Sigma})^{-1}(x-M),\ \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\right), \tag{5.23}$$

which reduces to (5.21) when $p = 1$.

5.3 Fisher's Information Bound for Multiparameter Families

The multivariate normal distribution plays its biggest role in applications as a large-sample approximation for maximum likelihood estimates. We suppose that the parametric family of densities $\{f_\mu(x)\}$, normal or not, is smoothly defined in terms of its p-dimensional parameter vector μ. (In terms of (5.1), Ω is a subset of $\mathcal{R}^p$.)

The MLE definitions and results are direct analogues of the single-parameter calculations beginning at (4.14) in Chapter 4. The *score function* $\dot{l}_x(\mu)$ is now defined as the gradient of $\log\{f_\mu(x)\}$,

$$\dot{l}_x(\mu) = \nabla_\mu \left\{\log f_\mu(x)\right\} = \left(\ldots, \frac{\partial \log f_\mu(x)}{\partial \mu_i}, \ldots\right)', \tag{5.24}$$

the p-vector of partial derivatives of $\log f_\mu(x)$ with respect to the coordinates of μ. It has mean zero,

$$E_\mu\left\{\dot{l}_x(\mu)\right\} = 0 = (0, 0, 0, \ldots, 0)'. \tag{5.25}$$

By definition, the Fisher information matrix $\mathcal{I}_\mu$ for μ is the $p \times p$ covariance matrix of $\dot{l}_x(\mu)$; using outer product notation,

$$\mathcal{I}_\mu = E_\mu\left\{\dot{l}_x(\mu)\dot{l}_x(\mu)'\right\} = \left(E_\mu\left\{\frac{\partial \log f_\mu(x)}{\partial \mu_i}\frac{\partial \log f_\mu(x)}{\partial \mu_j}\right\}\right). \quad (5.26)$$

The key result is that the MLE $\hat{\mu} = \arg\max_\mu\{f_\mu(x)\}$ has an approximately normal distribution with covariance matrix $\mathcal{I}_\mu^{-1}$,

$$\hat{\mu} \doteq \mathcal{N}_p(\mu, \mathcal{I}_\mu^{-1}). \quad (5.27)$$

Approximation (5.27) is justified by large-sample arguments, say with x an iid sample in $\mathcal{R}^p$, $(x_1, x_2, \ldots, x_n)$, n going to infinity.

Suppose the statistician is particularly interested in μ_1, the first coordinate of μ. Let $\mu_{(2)} = (\mu_2, \mu_3, \ldots, \mu_p)$ denote the other $p-1$ coordinates of μ, which are now "nuisance parameters" as far as the estimation of μ_1 goes. According to (5.27), the MLE $\hat{\mu}_1$, which is the first coordinate of $\hat{\mu}$, has

$$\hat{\mu}_1 \doteq \mathcal{N}\left(\mu_1, (\mathcal{I}_\mu^{-1})_{11}\right), \quad (5.28)$$

where the notation indicates the upper leftmost entry of $\mathcal{I}_\mu^{-1}$.

We can partition the information matrix $\mathcal{I}_\mu$ into the two parts corresponding to μ_1 and $\mu_{(2)}$,

$$\mathcal{I}_\mu = \begin{pmatrix} \mathcal{I}_{\mu 11} & \mathcal{I}_{\mu 1(2)} \\ \mathcal{I}_{\mu(2)1} & \mathcal{I}_{\mu(22)} \end{pmatrix} \quad (5.29)$$

(with $\mathcal{I}_{\mu 1(2)} = \mathcal{I}'_{\mu(2)1}$ of dimension $1\times(p-1)$ and $\mathcal{I}_{\mu(22)}$ $(p-1)\times(p-1)$).

†4 The endnotes show that†

$$(\mathcal{I}_\mu^{-1})_{11} = \left(\mathcal{I}_{\mu 11} - \mathcal{I}_{\mu 1(2)}\mathcal{I}_{\mu(22)}^{-1}\mathcal{I}_{\mu(2)1}\right)^{-1}. \quad (5.30)$$

The subtracted term on the right side of (5.30) is nonnegative, implying that

$$(\mathcal{I}_\mu^{-1})_{11} \geq \mathcal{I}_{\mu 11}^{-1}. \quad (5.31)$$

If $\mu_{(2)}$ were known to the statistician, rather than requiring estimation, then $f_{\mu_1\mu_{(2)}}(x)$ would be a one-parameter family, with Fisher information $\mathcal{I}_{\mu 11}$ for estimating μ_1, giving

$$\hat{\mu}_1 \doteq \mathcal{N}(\mu_1, \mathcal{I}_{\mu 11}^{-1}). \quad (5.32)$$

Comparing (5.28) with (5.32), (5.31) shows that the variance of the MLE $\hat{\mu}_1$ must always increase[4] in the presence of nuisance parameters.† †5

Maximum likelihood, and in fact any form of unbiased or nearly unbiased estimation, pays a nuisance tax for the presence of "other" parameters. Modern applications often involve thousands of *others*; think of regression fits with too many predictors. In some circumstances, biased estimation methods can reverse the situation, using the others to actually improve estimation of a target parameter; see Chapter 6 on empirical Bayes techniques, and Chapter 16 on ℓ_1 regularized regression models.

5.4 The Multinomial Distribution

Second in the small catalog of well-known classic multivariate distributions is the multinomial. The multinomial applies to situations in which the observations take on only a finite number of discrete values, say L of them. The 2×2 ulcer surgery of Table 4.1 is repeated in Table 5.2, now with the cells labeled $1, 2, 3,$ and 4. Here there are $L = 4$ possible outcomes for each patient: (`new`, `success`), (`new`, `failure`), (`old`, `success`), (`old`, `failure`).

Table 5.2 *The ulcer study of Table 4.1, now with the cells numbered 1 through 4 as shown.*

	`success`	`failure`
`new`	1 **9**	2 **12**
`old`	3 **7**	4 **17**

A number n of cases has been observed, $n = 45$ in Table 5.2. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_L)$ be the vector of counts for the L possible outcomes,

$$x_l = \#\{\text{cases having outcome } l\}, \tag{5.33}$$

$\boldsymbol{x} = (9, 12, 7, 17)'$ for the ulcer data. It is convenient to code the outcomes in terms of the coordinate vectors $\boldsymbol{e}_l$ of length L,

$$\boldsymbol{e}_l = (0, 0, \ldots, 0, 1, 0, \ldots, 0)', \tag{5.34}$$

with a 1 in the lth place.

[4] Unless $\mathcal{I}_{\mu 1(2)}$ is a vector of zeros, a condition that amounts to approximate independence of $\hat{\mu}_1$ and $\hat{\mu}_{(2)}$.

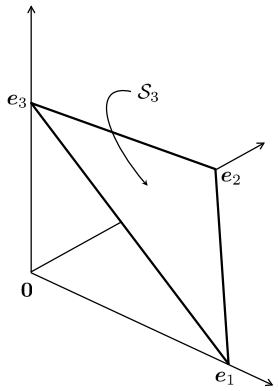

Figure 5.4 The simplex $\mathcal{S}_3$ is an equilateral triangle set at an angle to the coordinate axes in $\mathcal{R}^3$.

The multinomial probability model assumes that the n cases are independent of each other, with each case having probability π_l for outcome e_l,

$$\pi_l = \Pr\{\boldsymbol{e}_l\}, \qquad l = 1, 2, \ldots, L. \tag{5.35}$$

Let

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)' \tag{5.36}$$

indicate the vector of probabilities. The count vector $\boldsymbol{x}$ then follows the *multinomial distribution*,

$$f_{\boldsymbol{\pi}}(\boldsymbol{x}) = \frac{n!}{x_1! x_2! \ldots x_L!} \prod_{l=1}^{L} \pi_l^{x_l}, \tag{5.37}$$

denoted

$$\boldsymbol{x} \sim \text{Mult}_L(n, \boldsymbol{\pi}) \tag{5.38}$$

(for n observations, L outcomes, probability vector $\boldsymbol{\pi}$).

The parameter space Ω for $\boldsymbol{\pi}$ is the *simplex* $\mathcal{S}_L$,

$$\mathcal{S}_L = \left\{ \boldsymbol{\pi} : \pi_l \geq 0 \text{ and } \sum_{l=1}^{L} \pi_l = 1 \right\}. \tag{5.39}$$

Figure 5.4 shows $\mathcal{S}_3$, an equilateral triangle sitting at an angle to the coordinate axes $\boldsymbol{e}_1, \boldsymbol{e}_2,$ and $\boldsymbol{e}_3$. The midpoint of the triangle $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$

corresponds to a multinomial distribution putting equal probability on the three possible outcomes.

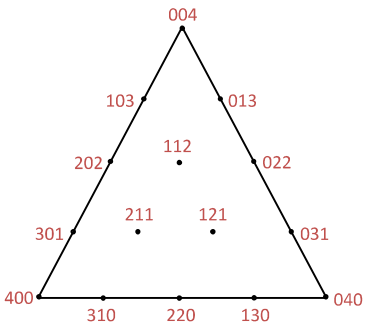

Figure 5.5 Sample space $\mathcal{X}$ for $\boldsymbol{x} \sim \text{Mult}_3(4, \boldsymbol{\pi})$; numbers indicate (x_1, x_2, x_3).

The sample space $\mathcal{X}$ for $\boldsymbol{x}$ is the subset of $n\mathcal{S}_L$ (the set of nonnegative vectors summing to n) having integer components. Figure 5.5 illustrates the case $n = 4$ and $L = 3$, now with the triangle of Figure 5.4 multiplied by 4 and set flat on the page. The point 121 indicates $\boldsymbol{x} = (1, 2, 1)$, with probability $12 \cdot \pi_1 \pi_2^2 \pi_3$ according to (5.37), etc.

In the *dichotomous* case, $L = 2$, the multinomial distribution reduces to the binomial, with (π_1, π_2) equaling $(\pi, 1 - \pi)$ in line 3 of Table 5.1, and (x_1, x_2) equaling $(x, n - x)$. The mean vector and covariance matrix of $\text{Mult}_L(n, \boldsymbol{\pi})$, for any value of L, are[†] †6

$$\boldsymbol{x} \sim \left(n\boldsymbol{\pi}, n\left[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]\right) \tag{5.40}$$

($\text{diag}(\boldsymbol{\pi})$ is the diagonal matrix with diagonal elements π_l), so $\text{var}(x_l) = n\pi_l(1 - \pi_l)$ and covariance $(x_l, x_j) = -n\pi_l\pi_j$; (5.40) generalizes the binomial mean and variance $(n\pi, n\pi(1 - \pi))$.

There is a useful relationship between the multinomial distribution and the Poisson. Suppose $S_1, S_2, \ldots, S_L$ are independent Poissons having possibly different parameters,

$$S_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_l), \qquad l = 1, 2, \ldots, L, \tag{5.41}$$

or, more concisely,

$$\boldsymbol{S} \sim \text{Poi}(\boldsymbol{\mu}) \tag{5.42}$$

with $\boldsymbol{S} = (S_1, S_2, \ldots, S_L)'$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_L)'$, the independence

being assumed in notation (5.42). Then the conditional distribution of $\boldsymbol{S}$ given the sum $S_+ = \sum S_l$ is multinomial,[†] †7

$$\boldsymbol{S}\,|\,S_+ \sim \text{Mult}_L(S_+, \boldsymbol{\mu}/\mu_+), \tag{5.43}$$

$\mu_+ = \sum \mu_l$.

Going in the other direction, suppose $N \sim \text{Poi}(n)$. Then the unconditional or marginal distribution of $\text{Mult}_L(N, \boldsymbol{\pi})$ is Poisson,

$$\text{Mult}_L(N, \boldsymbol{\pi}) \sim \text{Poi}(n\boldsymbol{\pi}) \qquad \text{if } N \sim \text{Poi}(n). \tag{5.44}$$

Calculations involving $\boldsymbol{x} \sim \text{Mult}_L(n, \boldsymbol{\pi})$ are sometimes complicated by the multinomial's correlations. The approximation $\boldsymbol{x} \doteq\sim \text{Poi}(n\boldsymbol{\pi})$ removes the correlations and is usually quite accurate if n is large.

There is one more important thing to say about the multinomial family: it contains *all* distributions on a sample space $\mathcal{X}$ composed of L discrete categories. In this sense it is a model for *nonparametric* inference on $\mathcal{X}$. The nonparametric bootstrap calculations of Chapter 10 use the multinomial in this way. Nonparametrics, and the multinomial, have played a larger role in the modern environment of large, difficult to model, data sets.

5.5 Exponential Families

Classic parametric families dominated statistical theory and practice for a century and more, with an enormous catalog of their individual properties—means, variances, tail areas, etc.—being compiled. A surprise, though a slowly emerging one beginning in the 1930s, was that all of them were examples of a powerful general construction: *exponential families*. What follows here is a brief introduction to the basic theory, with further development to come in subsequent chapters.

To begin with, consider the Poisson family, line 2 of Table 5.1. The ratio of Poisson densities at two parameter values μ and μ_0 is

$$\frac{f_\mu(x)}{f_{\mu_0}(x)} = e^{-(\mu-\mu_0)}\left(\frac{\mu}{\mu_0}\right)^x, \tag{5.45}$$

which can be re-expressed as

$$f_\mu(x) = e^{\alpha x - \psi(\alpha)} f_{\mu_0}(x), \tag{5.46}$$

where we have defined

$$\alpha = \log\{\mu/\mu_0\} \quad \text{and} \quad \psi(\alpha) = \mu_0(e^\alpha - 1). \tag{5.47}$$

Looking at (5.46), we can describe the Poisson family in three steps.

1 Start with any one Poisson distribution $f_{\mu_0}(x)$.
2 For any value of $\mu > 0$ let $\alpha = \log\{\mu/\mu_0\}$ and calculate

$$\tilde{f}_\mu(x) = e^{\alpha x} f_{\mu_0}(x) \qquad \text{for } x = 0, 1, 2, \ldots. \tag{5.48}$$

3 Finally, divide $\tilde{f}_\mu(x)$ by $\exp(\psi(\alpha))$ to get the Poisson density $f_\mu(x)$.

In other words, we "tilt" $f_{\mu_0}(x)$ with the exponential factor $e^{\alpha x}$ to get $\tilde{f}_\mu(x)$, and then renormalize $\tilde{f}_\mu(x)$ to sum to 1. Notice that (5.46) gives $\exp(-\psi(\alpha))$ as the renormalizing constant since

$$e^{\psi(\alpha)} = \sum_0^\infty e^{\alpha x} f_{\mu_0}(x). \tag{5.49}$$

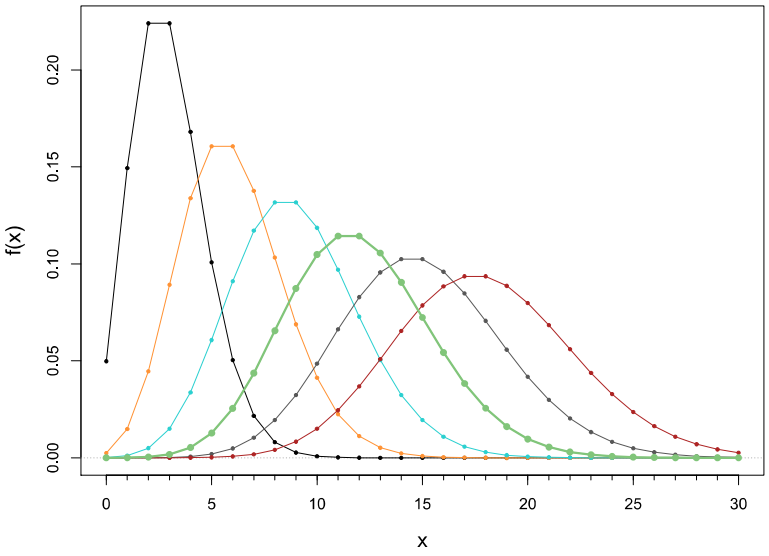

Figure 5.6 Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy green curve with dots for $\mu = 12$.

Figure 5.6 graphs the Poisson density $f_\mu(x)$ for $\mu = 3, 6, 9, 12, 15, 18$. Each Poisson density is a renormalized exponential tilt of any other Poisson density. So for instance $f_6(x)$ is obtained from $f_{12}(x)$ via the tilt $e^{\alpha x}$ with $\alpha = \log\{6/12\} = -0.693$.[5]

[5] Alternate expressions for $f_\mu(x)$ as an exponential family are available, for example $\exp(\alpha x - \psi(\alpha))\, f_0(x)$, where $\alpha = \log \mu$, $\psi(\alpha) = \exp(\alpha)$, and $f_0(x) = 1/x!$. (It isn't necessary for $f_0(x)$ to be a member of the family.)

The Poisson is a *one-parameter exponential family*, in that α and x in expression (5.46) are one-dimensional. A *p-parameter exponential family* has the form

$$f_\alpha(x) = e^{\alpha' y - \psi(\alpha)} f_0(x) \qquad \text{for } \alpha \in A, \tag{5.50}$$

where α and y are p-vectors and A is contained in $\mathcal{R}^p$. Here α is the "canonical" or "natural" parameter vector and $y = t(x)$ is the "sufficient statistic" vector. The normalizing function $\psi(\alpha)$, which makes $f_\alpha(x)$ integrate (or sum) to one, satisfies

$$e^{\psi(\alpha)} = \int_{\mathcal{X}} e^{\alpha' y} f_0(x)\, dx, \tag{5.51}$$

and it can be shown that the parameter space A for which the integral is finite is a convex set† in $\mathcal{R}^p$. As an example, the gamma family on line 4 of Table 5.1 is a two-parameter exponential family, with α and $y = t(x)$ given by

†8

$$(\alpha_1, \alpha_2) = \left(-\frac{1}{\sigma}, \nu\right), \quad (y_1, y_2) = (x, \log x), \tag{5.52}$$

and

$$\begin{aligned} \psi(\alpha) &= \nu \log\sigma + \log\Gamma(\nu) \\ &= -\alpha_2 \log\{-\alpha_1\} + \log\{\Gamma(\alpha_2)\}. \end{aligned} \tag{5.53}$$

The parameter space A is $\{\alpha_1 < 0 \text{ and } \alpha_2 > 0\}$.

Why are we interested in exponential tilting rather than some other transformational form? The answer has to do with repeated sampling. Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample from a p-parameter exponential family (5.50). Then, letting $y_i = t(x_i)$ denote the sufficient vector corresponding to x_i,

$$\begin{aligned} f_\alpha(\boldsymbol{x}) &= \prod_{i=1}^{n} e^{\alpha' y_i - \psi(\alpha)} f_0(x_i) \\ &= e^{n(\alpha'\bar{y} - \psi(\alpha))} f_0(\boldsymbol{x}), \end{aligned} \tag{5.54}$$

where $\bar{y} = \sum_1^n y_i/n$. This is still a p-parameter exponential family, now with natural parameter $n\alpha$, sufficient statistic $\bar{y}$, and normalizer $n\psi(\alpha)$. No matter how large n may be, the statistician can still compress all the inferential information into a p-dimensional statistic $\bar{y}$. Only exponential families enjoy this property.

Even though they were discovered and developed in quite different contexts, and at quite different times, all of the distributions discussed in this

chapter exist in exponential families. This isn't quite the coincidence it seems. Mathematical tractability was the prized property of classic parametric distributions, and tractability was greatly facilitated by exponential structure, even if that structure went unrecognized.

In one-parameter exponential families, the normalizer $\psi(\alpha)$ is also known as the *cumulant generating function*. Derivatives of $\psi(\alpha)$ yield the cumulants of y,[6] the first two giving the mean and variance† †9

$$\dot{\psi}(\alpha) = E_\alpha\{y\} \quad \text{and} \quad \ddot{\psi}(\alpha) = \text{var}_\alpha\{y\}. \tag{5.55}$$

Similarly, in p-parametric families

$$\dot{\psi}(\alpha) = (\ldots \partial\psi/\partial\alpha_j \ldots)' = E_\alpha\{y\} \tag{5.56}$$

and

$$\ddot{\psi}(\alpha) = \left(\frac{\partial^2\psi(\alpha)}{\partial\alpha_j\,\partial\alpha_k}\right) = \text{cov}_\alpha\{y\}. \tag{5.57}$$

The p-dimensional *expectation parameter*, denoted

$$\beta = E_\alpha\{y\}, \tag{5.58}$$

is a one-to-one function of the natural parameter α. Let V_α indicate the $p \times p$ covariance matrix,

$$V_\alpha = \text{cov}_\alpha\{y\}. \tag{5.59}$$

Then the $p \times p$ derivative matrix of β with respect to α is

$$\frac{d\beta}{d\alpha} = (\partial\beta_j/\partial\alpha_k) = V_\alpha, \tag{5.60}$$

this following from (5.56)–(5.57), the inverse mapping being $d\alpha/d\beta = V_\alpha^{-1}$. As a one-parameter example, the Poisson in Table 5.1 has $\alpha = \log\mu$, $\beta = \mu$, $y = x$, and $d\beta/d\alpha = 1/(d\alpha/d\beta) = \mu = V_\alpha$.

The maximum likelihood estimate for the expectation parameter β is simply y (or $\bar{y}$ under repeated sampling (5.54)), which makes it immediate to calculate in most situations.† Less immediate is the MLE for the natural †10 parameter α: the one-to-one mapping $\beta = \dot{\psi}(\alpha)$ (5.56) has inverse $\alpha = \dot{\psi}^{-1}(\beta)$, so

$$\hat{\alpha} = \dot{\psi}^{-1}(y), \tag{5.61}$$

[6] The simplified dot notation leads to more compact expressions: $\dot{\psi}(\alpha) = d\psi(\alpha)/d\alpha$ and $\ddot{\psi}(\alpha) = d^2\psi(\alpha)/d\alpha^2$.

e.g., $\hat{\alpha} = \log y$ for the Poisson. The trouble is that $\dot{\psi}^{-1}(\cdot)$ is usually unavailable in closed form. Numerical approximation algorithms are necessary to calculate $\hat{\alpha}$ in most cases.

All of the classic exponential families have closed-form expressions for $\psi(\alpha)$ (and $f_\alpha(x)$), yielding pleasant formulas for the mean β and covariance V_α, (5.56)–(5.57). Modern computational technology allows us to work with general exponential families, designed for specific tasks, without concern for mathematical tractability.

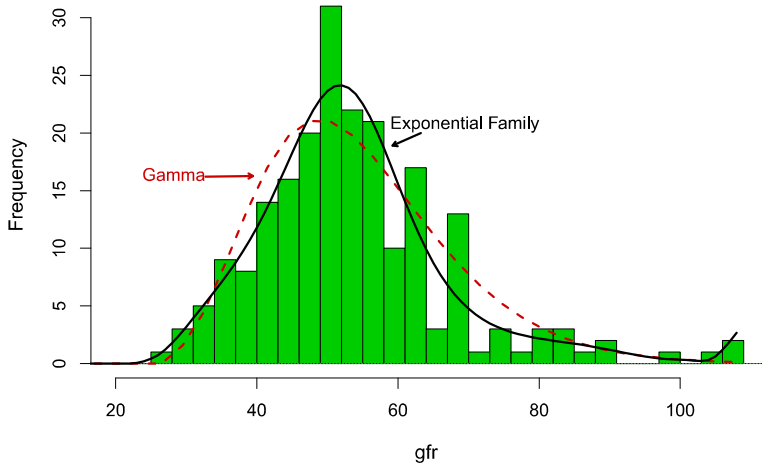

Figure 5.7 A seven-parameter exponential family fit to the **gfr** data of Figure 2.1 (solid) compared with gamma fit of Figure 4.1 (dashed).

As an example we again consider fitting the **gfr** data of Figure 2.1. For our exponential family of possible densities we take $f_0(x) \equiv 1$, and sufficient statistic vector

$$y(x) = (x, x^2, \ldots, x^7), \tag{5.62}$$

so $\alpha' y$ in (5.50) can represent all 7th-order polynomials in x, the **gfr** measurement.[7] (Stopping at power 2 gives the $\mathcal{N}(\mu, \sigma^2)$ family, which we already know fits poorly from Figure 4.1.) The heavy curve in Figure 5.7 shows the MLE fit $f_{\hat{\alpha}}(x)$ now following the **gfr** histogram quite closely. Chapter 10 discusses "Lindsey's method," a simplified algorithm for calculating the MLE $\hat{\alpha}$.

[7] Any intercept in the polynomial is absorbed into the $\psi(\alpha)$ term in (5.57).

A more exotic example concerns the generation of random graphs on a fixed set of N nodes. Each possible graph has a certain total number E of edges, and T of triangles. A popular choice for generating such graphs is the two-parameter exponential family having $y = (E, T)$, so that larger values of α_1 and α_2 yield more connections.

5.6 Notes and Details

The notion of *sufficient statistics*, ones that contain all available inferential information, was perhaps Fisher's happiest contribution to the classic corpus. He noticed that in the exponential family form (5.50), the fact that the parameter α interacts with the data x only through the factor $\exp(\alpha' y)$ makes $y(x)$ sufficient for estimating α. In 1935–36, a trio of authors, working independently in different countries, Pitman, Darmois, and Koopmans, showed that exponential families are the only ones that enjoy fixed-dimensional sufficient statistics under repeated independent sampling. Until the late 1950s such distributions were called Pitman–Darmois–Koopmans families, the long name suggesting infrequent usage.

Generalized linear models, Chapter 8, show the continuing impact of sufficiency on statistical practice. Peter Bickel has pointed out that *data compression*, a lively topic in areas such as image transmission, is a modern, less stringent, version of sufficiency.

Our only nonexponential family so far was (4.39), the Cauchy translational model. Efron and Hinkley (1978) analyze the Cauchy family in terms of *curved exponential families*, a generalization of model (5.50).

Properties of classical distributions (lots of properties and lots of distributions) are covered in Johnson and Kotz's invaluable series of reference books, 1969–1972. Two classic multivariate analysis texts are Anderson (2003) and Mardia *et al.* (1979).

†1 [p. 57] *Formula* (5.12). From $z = \boldsymbol{T}^{-1}(x - \mu)$ we have $dz/dx = \boldsymbol{T}^{-1}$ and

$$f_{\mu,\boldsymbol{\Sigma}}(x) = f(z)|\boldsymbol{T}^{-1}| = (2\pi)^{-\frac{p}{2}}|\boldsymbol{T}^{-1}|e^{-\frac{1}{2}(x-\mu)'\boldsymbol{T}^{-1'}\boldsymbol{T}^{-1}(x-\mu)}, \quad (5.63)$$

so (5.12) follows from $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{\Sigma}$ and $|\boldsymbol{T}| = |\boldsymbol{\Sigma}|^{1/2}$.

†2 [p. 58] *Formula* (5.18). Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ be partitioned as in (5.17). Then

$$\begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Lambda}_{22} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Lambda}_{11} & \left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)^{-1} \end{pmatrix}, \quad (5.64)$$

direct multiplication showing that $\boldsymbol{\Lambda}\boldsymbol{\Sigma} = \boldsymbol{I}$, the identity matrix. If $\boldsymbol{\Sigma}$ is

symmetric then $\mathbf{\Lambda}_{21} = \mathbf{\Lambda}_{12}'$. By redefining x to be $x - \mu$ we can set $\mu_{(1)}$ and $\mu_{(2)}$ equal to zero in (5.18). The quadratic form in the exponent of (5.12) is

$$(x_{(1)}', x_{(2)}')\mathbf{\Lambda}\left(x_{(1)}, x_{(2)}\right) = x_{(2)}'\mathbf{\Lambda}_{22}x_{(2)} + 2x_{(1)}'\mathbf{\Lambda}_{12}x_{(2)} + x_{(1)}'\mathbf{\Lambda}_{11}x_{(1)}. \tag{5.65}$$

But, using (5.64), this matches the quadratic form from (5.18),

$$\left(x_{(2)} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}x_{(1)}\right)'\mathbf{\Lambda}_{22}\left(x_{(2)} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}x_{(1)}\right) \tag{5.66}$$

except for an added term that does *not* involve $x_{(2)}$. For a multivariate normal distribution, this is sufficient to show that the conditional distribution of $x_{(2)}$ given $x_{(1)}$ is indeed (5.18) (see †3).

†3 [p. 59] *Formulas* (5.21) *and* (5.23). Suppose that the continuous univariate random variable z has density of the form

$$f(z) = c_0 e^{-\frac{1}{2}Q(z)}, \qquad \text{where } Q(z) = az^2 + 2bz + c_1, \tag{5.67}$$

a, b, c_0 and c_1 constants, $a > 0$. Then, by "completing the square,"

$$f(z) = c_2 e^{-\frac{1}{2}a\left(z - \frac{b}{a}\right)^2}, \tag{5.68}$$

and we see that $z \sim \mathcal{N}(b/a, 1/a)$. The key point is that form (5.67) specifies z as normal, with mean and variance uniquely determined by a and b. The multivariate version of this fact was used in the derivation of formula (5.18).

By redefining μ and x as $\mu - M$ and $x - M$, we can take $M = 0$ in (5.21). Setting $B = A/(A + \sigma^2)$, density (5.21) for $\mu|x$ is of form (5.67), with

$$Q(\mu) = \frac{\mu^2}{B\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{Bx^2}{\sigma^2}. \tag{5.69}$$

But Bayes' rule says that the density of $\mu|x$ is proportional to $g(\mu)\, f_\mu(x)$, also of form (5.67), now with

$$Q(\mu) = \left(\frac{1}{A} + \frac{1}{\sigma^2}\right)\mu^2 - \frac{2x\mu}{\sigma^2} + \frac{x^2}{\sigma^2}. \tag{5.70}$$

A little algebra shows that the quadratic and linear coefficients of μ match in (5.69)–(5.70), verifying (5.21).

We verify the multivariate result (5.23) using a different argument. The $2p$ vector $(\mu, x)'$ has joint distribution

$$\mathcal{N}\left(\begin{pmatrix} M \\ M \end{pmatrix}, \begin{pmatrix} \boldsymbol{A} & \boldsymbol{A} \\ \boldsymbol{A} & \boldsymbol{A} + \mathbf{\Sigma} \end{pmatrix}\right). \tag{5.71}$$

Now we employ (5.18) and a little manipulation to get (5.23).

†4 [p. 60] *Formula* (5.30). This is the matrix identity (5.64), now with $\boldsymbol{\Sigma}$ equaling $\mathcal{I}_\mu$.

†5 [p. 61] *Multivariate Gaussian and nuisance parameters.* The cautionary message here—that increasing the number of unknown nuisance parameters decreases the accuracy of the estimate of interest—can be stated more positively: if some nuisance parameters are actually known, then the MLE of the parameter of interest becomes more accurate. Suppose, for example, we wish to estimate μ_1 from a sample of size n in a bivariate normal model $x \sim \mathcal{N}_2(\mu, \boldsymbol{\Sigma})$ (5.14). The MLE $\bar{x}_1$ has variance σ_{11}/n in notation (5.19). But if μ_2 is known then the MLE of μ_1 becomes $\bar{x}_1 - (\sigma_{12}/\sigma_{22})(\bar{x}_2 - \mu_2)$ with variance $(\sigma_{11}/n)\cdot(1-\rho^2)$, ρ being the correlation $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

†6 [p. 63] *Formula* (5.40). $\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{x}_i$, where the $\boldsymbol{x}_i$ are iid observations having $\Pr\{\boldsymbol{x}_i = \boldsymbol{e}_i\} = \pi_l$, as in (5.35). The mean and covariance of each $\boldsymbol{x}_i$ are

$$E\{\boldsymbol{x}_i\} = \sum_{1}^{L} \pi_l \boldsymbol{e}_l = \boldsymbol{\pi} \tag{5.72}$$

and

$$\begin{aligned} \operatorname{cov}\{\boldsymbol{x}_i\} &= E\{\boldsymbol{x}_i \boldsymbol{x}_i'\} - E\{\boldsymbol{x}_i\}E\{\boldsymbol{x}_i'\} = \sum \pi_l \boldsymbol{e}_l \boldsymbol{e}_l' - \boldsymbol{\pi}\boldsymbol{\pi}' \\ &= \operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'. \end{aligned} \tag{5.73}$$

Formula (5.40) follows from $E\{\boldsymbol{x}\} = \sum E\{\boldsymbol{x}_i\}$ and $\operatorname{cov}\{\boldsymbol{x}\} = \sum \operatorname{cov}\{\boldsymbol{x}_i\}$.

†7 [p. 64] *Formula* (5.43). The densities of $\boldsymbol{S}$ (5.42) and $S_+ = \sum S_l$ are

$$f_{\boldsymbol{\mu}}(\boldsymbol{S}) = \prod_{l=1}^{L} e^{-\mu_l} \mu_l^{S_l}/S_l! \quad \text{and} \quad f_{\mu_+}(S_+) = e^{-\mu_+}\mu_+^{S_+}/S_+!. \tag{5.74}$$

The conditional density of $\boldsymbol{S}$ given S_+ is the ratio

$$f_{\boldsymbol{\mu}}(\boldsymbol{S}\,|S_+) = \left(\frac{S_+!}{\prod_1^L S_l!}\right) \prod_{l=1}^{L} \left(\frac{\mu_l}{\mu_+}\right)^{S_l}, \tag{5.75}$$

which is (5.43).

†8 [p. 66] *Formula* (5.51) *and the convexity of A.* Suppose α_1 and α_2 are any two points in A, i.e., values of α having the integral in (5.51) finite. For any value of c in the interval $[0, 1]$, and any value of y, we have

$$c e^{\alpha_1' y} + (1-c) e^{\alpha_2' y} \geq e^{[c\alpha_1 + (1-c)\alpha_2]' y} \tag{5.76}$$

because of the convexity in c of the function on the right (verified by showing that its second derivative is positive). Integrating both sides of (5.76)

over $\mathcal{X}$ with respect to $f_0(x)$ shows that the integral on the right must be finite: that is, $c\alpha_1 + (1-c)\alpha_2$ is in A, verifying A's convexity.

†9 [p. 67] *Formula* (5.55). In the univariate case, differentiating both sides of (5.51) with respect to α gives

$$\dot{\psi}(\alpha)e^{\psi(\alpha)} = \int_{\mathcal{X}} y e^{\alpha y} f_0(x)\, dx; \tag{5.77}$$

dividing by $e^{\psi(\alpha)}$ shows that $\dot{\psi}(\alpha) = E_\alpha\{y\}$. Differentiating (5.77) again gives

$$\left(\ddot{\psi}(\alpha) + \dot{\psi}(\alpha)^2\right) e^{\psi(\alpha)} = \int_{\mathcal{X}} y^2 e^{\alpha y} f_0(x)\, dx, \tag{5.78}$$

or

$$\ddot{\psi}(\alpha) = E_\alpha\{y^2\} - E_\alpha\{y\}^2 = \mathrm{var}_\alpha\{y\}. \tag{5.79}$$

Successive derivatives of $\psi(\alpha)$ yield the higher cumulants of y, its skewness, kurtosis, etc.

†10 [p. 67] *MLE for* β. The gradient with respect to α of $\log f_\alpha(y)$ (5.50) is

$$\nabla_\alpha\left(\alpha' y - \psi(\alpha)\right) = y - \dot{\psi}(\alpha) = y - E_\alpha\{y^*\}, \tag{5.80}$$

(5.56), where y^* represents a hypothetical realization $y(x^*)$ drawn from $f_\alpha(\cdot)$. We achieve the MLE $\hat{\alpha}$ at $\nabla_{\hat{\alpha}} = 0$, or

$$E_{\hat{\alpha}}\{y^*\} = y. \tag{5.81}$$

In other words the MLE $\hat{\alpha}$ is the value of α that makes the expectation $E_\alpha\{y^*\}$ match the observed y. Thus (5.58) implies that the MLE of parameter β is y.

Part II

Early Computer-Age Methods

6

Empirical Bayes

The constraints of slow mechanical computation molded classical statistics into a mathematically ingenious theory of sharply delimited scope. Emerging after the Second World War, electronic computation loosened the computational stranglehold, allowing a more expansive and useful statistical methodology.

Some revolutions start slowly. The journals of the 1950s continued to emphasize classical themes: pure mathematical development typically centered around the normal distribution. Change came gradually, but by the 1990s a new statistical technology, computer enabled, was firmly in place. Key developments from this period are described in the next several chapters. The ideas, for the most part, would not startle a pre-war statistician, but their computational demands, factors of 100 or 1000 times those of classical methods, would. More factors of a thousand lay ahead, as will be told in Part III, the story of statistics in the twenty-first century.

Empirical Bayes methodology, this chapter's topic, has been a particularly slow developer despite an early start in the 1940s. The roadblock here was not so much the computational demands of the theory as a lack of appropriate data sets. Modern scientific equipment now provides ample grist for the empirical Bayes mill, as will be illustrated later in the chapter, and more dramatically in Chapters 15–21.

6.1 Robbins' Formula

Table 6.1 shows one year of claims data for a European automobile insurance company; 7840 of the 9461 policy holders made no claims during the year, 1317 made a single claim, 239 made two claims each, etc., with Table 6.1 continuing to the one person who made seven claims. Of course the insurance company is concerned about the claims each policy holder will make in the *next* year.

Bayes' formula seems promising here. We suppose that x_k, the number

Table 6.1 *Counts y_x of number of claims x made in a single year by 9461 automobile insurance policy holders. Robbins' formula* (6.7) *estimates the number of claims expected in a succeeding year, for instance* 0.168 *for a customer in the $x = 0$ category. Parametric maximum likelihood analysis based on a gamma prior gives less noisy estimates.*

Claims x	0	1	2	3	4	5	6	7
Counts y_x	7840	1317	239	42	14	4	4	1
Formula (6.7)	.168	.363	.527	1.33	1.43	6.00	1.75	
Gamma MLE	.164	.398	.633	.87	1.10	1.34	1.57	

of claims to be made in a single year by policy holder k, follows a Poisson distribution with parameter θ_k,

$$\Pr\{x_k = x\} = p_{\theta_k}(x) = e^{-\theta_k}\theta_k^x/x!, \tag{6.1}$$

for $x = 0, 1, 2, 3, \ldots$; θ_k is the expected value of x_k. A good customer, from the company's point of view, has a small value of θ_k, though in any one year his or her actual number of accidents x_k will vary randomly according to probability density (6.1).

Suppose we knew the prior density $g(\theta)$ for the customers' θ values. Then Bayes' rule (3.5) would yield

$$E\{\theta|x\} = \frac{\int_0^\infty \theta p_\theta(x) g(\theta)\, d\theta}{\int_0^\infty p_\theta(x) g(\theta)\, d\theta} \tag{6.2}$$

for the expected value of θ of a customer observed to make x claims in a single year. This would answer the insurance company's question of what number of claims X to expect the next year from the same customer, since $E\{\theta|x\}$ is also $E\{X|x\}$ (θ being the expectation of X).

Formula (6.2) is just the ticket if the prior $g(\theta)$ is known to the company, but what if it is not? A clever rewriting of (6.2) provides a way forward. Using (6.1), (6.2) becomes

$$\begin{aligned} E\{\theta|x\} &= \frac{\int_0^\infty \left[e^{-\theta}\theta^{x+1}/x!\right] g(\theta)\, d\theta}{\int_0^\infty \left[e^{-\theta}\theta^x/x!\right] g(\theta)\, d\theta} \\ &= \frac{(x+1)\int_0^\infty \left[e^{-\theta}\theta^{x+1}/(x+1)!\right] g(\theta)\, d\theta}{\int_0^\infty \left[e^{-\theta}\theta^x/x!\right] g(\theta)\, d\theta}. \end{aligned} \tag{6.3}$$

The *marginal density* of x, integrating $p_\theta(x)$ over the prior $g(\theta)$, is

$$f(x) = \int_0^\infty p_\theta(x)g(\theta)\,d\theta = \int_0^\infty \left[e^{-\theta}\theta^x/x!\right] g(\theta)\,d\theta. \tag{6.4}$$

Comparing (6.3) with (6.4) gives *Robbins' formula*,

$$E\{\theta|x\} = (x+1)f(x+1)/f(x). \tag{6.5}$$

The surprising and gratifying fact is that, even with no knowledge of the prior density $g(\theta)$, the insurance company can estimate $E\{\theta|x\}$ (6.2) from formula (6.5). The obvious estimate of the marginal density $f(x)$ is the proportion of total counts in category x,

$$\hat{f}(x) = y_x/N, \quad \text{with } N = \textstyle\sum_x y_x, \text{ the total count,} \tag{6.6}$$

$\hat{f}(0) = 7840/9461$, $\hat{f}(1) = 1317/9461$, etc. This yields an empirical version of Robbins' formula,

$$\hat{E}\{\theta|x\} = (x+1)\hat{f}(x+1)\big/\hat{f}(x) = (x+1)y_{x+1}/y_x, \tag{6.7}$$

the final expression not requiring N. Table 6.1 gives $\hat{E}\{\theta|0\} = 0.168$: customers who made zero claims in one year had expectation 0.168 of a claim the next year; those with one claim had expectation 0.363, and so on.

Robbins' formula came as a surprise[1] to the statistical world of the 1950s: the expectation $E\{\theta_k|x_k\}$ for a single customer, unavailable without the prior $g(\theta)$, somehow becomes available in the context of a large study. The terminology *empirical Bayes* is apt here: Bayesian formula (6.5) for a single subject is estimated empirically (i.e., frequentistically) from a collection of similar cases. The crucial point, and the surprise, is that *large data sets of parallel situations carry within them their own Bayesian information*. Large parallel data sets are a hallmark of twenty-first-century scientific investigation, promoting the popularity of empirical Bayes methods.

Formula (6.7) goes awry at the right end of Table 6.1, where it is destabilized by small count numbers. A parametric approach gives more dependable results: now we assume that the prior density $g(\theta)$ for the customers' θ_k values has a gamma form (Table 5.1)

$$g(\theta) = \frac{\theta^{\nu-1}e^{-\theta/\sigma}}{\sigma^\nu\Gamma(\nu)}, \qquad \text{for } \theta \geq 0, \tag{6.8}$$

but with parameters ν and σ unknown. Estimates $(\hat{\nu}, \hat{\sigma})$ are obtained by

[1] Perhaps it shouldn't have; estimation methods similar to (6.7) were familiar in the actuarial literature.

maximum likelihood fitting to the counts y_x, yielding a parametrically estimated marginal density[†] †1

$$\hat{f}(x) = f_{\hat{\nu},\hat{\sigma}}(x), \tag{6.9}$$

or equivalently $\hat{y}_x = N f_{\hat{\nu},\hat{\sigma}}(x)$.

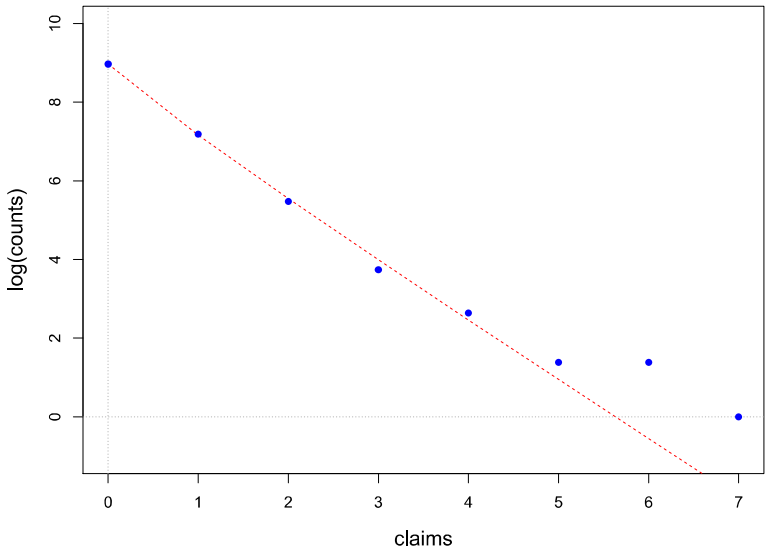

Figure 6.1 Auto accident data; log(counts) vs claims for 9461 auto insurance policies. The dashed line is a gamma MLE fit.

The bottom row of Table 6.1 gives parametric estimates $E_{\hat{\nu},\hat{\sigma}}\{\theta|x\} = (x+1)\hat{y}_{x+1}/\hat{y}_x$, which are seen to be less eccentric for large x. Figure 6.1 compares (on the log scale) the raw counts y_x with their parametric cousins $\hat{y}_x$.

6.2 The Missing-Species Problem

The very first empirical Bayes success story related to the butterfly data of Table 6.2. Even in the midst of World War II Alexander Corbet, a leading naturalist, had been trapping butterflies for two years in Malaysia (then Malaya): 118 species were so rare that he had trapped only one specimen each, 74 species had been trapped twice each, Table 6.2 going on to show that 44 species were trapped three times each, and so on. Some of the more

common species had appeared hundreds of times each, but of course Corbet was interested in the rarer specimens.

Table 6.2 *Butterfly data; number y of species seen x times each in two years of trapping; 118 species trapped just once, 74 trapped twice each, etc.*

x	1	2	3	4	5	6	7	8	9	10	11	12
y	118	74	44	24	29	22	20	19	20	15	12	14
x	13	14	15	16	17	18	19	20	21	22	23	24
y	6	12	6	9	9	6	10	10	11	5	3	3

Corbet then asked a seemingly impossible question: if he trapped for one additional year, how many new species would he expect to capture? The question relates to the *absent* entry in Table 6.2, $x = 0$, the species that haven't been seen yet. Do we really have any evidence at all for answering Corbet? Fortunately he asked the right man: R. A. Fisher, who produced a surprisingly satisfying solution for the "missing-species problem."

Suppose there are S species in all, seen or unseen, and that x_k, the number of times species k is trapped in one time unit,[2] follows a Poisson distribution with parameter θ_k as in (6.1),

$$x_k \sim \text{Poi}(\theta_k), \qquad \text{for } k = 1, 2, \ldots, S. \tag{6.10}$$

The entries in Table 6.2 are

$$y_x = \#\{x_k = x\}, \qquad \text{for } x = 1, 2, \ldots, 24, \tag{6.11}$$

the number of species trapped exactly x times each.

Now consider a further trapping period of t time units, $t = 1/2$ in Corbet's question, and let $x_k(t)$ be the number of times species k is trapped in the new period. Fisher's key assumption is that

$$x_k(t) \sim \text{Poi}(\theta_k t) \tag{6.12}$$

independently of x_k. That is, any one species is trapped independently over time[3] at a rate proportional to its parameter θ_k.

The probability that species k is *not* seen in the initial trapping period

[2] One time unit equals two years in Corbet's situation.

[3] This is the definition of a *Poisson process*.

but *is* seen in the new period, that is $x_k = 0$ and $x_k(t) > 0$, is

$$e^{-\theta_k}\left(1 - e^{-\theta_k t}\right), \tag{6.13}$$

so that $E(t)$, the expected number of new species seen in the new trapping period, is

$$E(t) = \sum_{k=1}^{S} e^{-\theta_k}\left(1 - e^{-\theta_k t}\right). \tag{6.14}$$

It is convenient to write (6.14) as an integral,

$$E(t) = S\int_0^\infty e^{-\theta}\left(1 - e^{-\theta t}\right) g(\theta)\, d\theta, \tag{6.15}$$

where $g(\theta)$ is the "empirical density" putting probability $1/S$ on each of the θ_k values. (Later we will think of $g(\theta)$ as a continuous prior density on the possible θ_k values.)

Expanding $1 - e^{-\theta t}$ gives

$$E(t) = S\int_0^\infty e^{-\theta}\left[\theta t - (\theta t)^2/2! + (\theta t)^3/3! - \cdots\right] g(\theta)\, d\theta. \tag{6.16}$$

Notice that the expected value e_x of y_x is the sum of the probabilities of being seen exactly x times in the initial period,

$$\begin{aligned} e_x = E\{y_x\} &= \sum_{k=1}^{S} e^{-\theta_k}\theta_k^x/x! \\ &= S\int_0^\infty \left[e^{-\theta}\theta^x/x!\right] g(\theta)\, d\theta. \end{aligned} \tag{6.17}$$

Comparing (6.16) with (6.17) provides a surprising result,

$$E(t) = e_1 t - e_2 t^2 + e_3 t^3 - \cdots. \tag{6.18}$$

We don't know the e_x values but, as in Robbins' formula, we can estimate them by the y_x values, yielding an answer to Corbet's question,

$$\hat{E}(t) = y_1 t - y_2 t^2 + y_3 t^3 - \cdots. \tag{6.19}$$

Corbet specified $t = 1/2$, so[4]

$$\begin{aligned} \hat{E}(1/2) &= 118(1/2) - 74(1/2)^2 + 44(1/2)^3 - \cdots \\ &= 45.2. \end{aligned} \tag{6.20}$$

[4] This may have been discouraging; there were no new trapping results reported.

Table 6.3 *Expectation* (6.19) *and its standard error* (6.21) *for the number of new species captured in t additional fractional units of trapping time.*

t	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$E(t)$	0	11.10	20.96	29.79	37.79	45.2	52.1	58.9	65.6	71.6	75.0
$\widehat{\mathrm{sd}}(t)$	0	2.24	4.48	6.71	8.95	11.2	13.4	15.7	17.9	20.1	22.4

Formulas (6.18) and (6.19) do not require the butterflies to arrive independently. If we are willing to add the assumption that the x_k's are mutually independent, we can calculate† †2

$$\widehat{\mathrm{sd}}(t) = \left(\sum_{x=1}^{24} y_x t^{2x}\right)^{1/2} \tag{6.21}$$

as an approximate standard error for $\hat{E}(t)$. Table 6.3 shows $\hat{E}(t)$ and $\widehat{\mathrm{sd}}(t)$ for $t = 0, 0.1, 0.2, \ldots, 1$; in particular,

$$\hat{E}(0.5) = 45.2 \pm 11.2. \tag{6.22}$$

Formula (6.19) becomes unstable for $t > 1$. This is our price for substituting the nonparametric estimates y_x for e_x in (6.18). Fisher actually answered Corbet using a parametric empirical Bayes model in which the prior $g(\theta)$ for the Poisson parameters θ_k (6.12) was assumed to be of the gamma form (6.8). It can be shown† that then $E(t)$ (6.15) is given by †3

$$E(t) = e_1 \{1 - (1 + \gamma t)^{-\nu}\} \big/ (\gamma \nu), \tag{6.23}$$

where $\gamma = \sigma/(1 + \sigma)$. Taking $\hat{e}_1 = y_1$, maximum likelihood estimation gave

$$\hat{\nu} = 0.104 \quad \text{and} \quad \hat{\sigma} = 89.79. \tag{6.24}$$

Figure 6.2 shows that the parametric estimate of $E(t)$ (6.23) using $\hat{e}_1$, $\hat{\nu}$, and $\hat{\sigma}$ is just slightly greater than the nonparametric estimate (6.19) over the range $0 \le t \le 1$. Fisher's parametric estimate, however, gives reasonable results for $t > 1$, $\hat{E}(2) = 123$ for instance, for a future trapping period of 2 units (4 years). "Reasonable" does not necessarily mean dependable. The gamma prior is a mathematical convenience, not a fact of nature; projections into the far future fall into the category of educated guessing.

The missing-species problem encompasses more than butterflies. There are 884,647 words in total in the recognized Shakespearean canon, of which 14,376 are so rare they appear just once each, 4343 appear twice each, etc.,

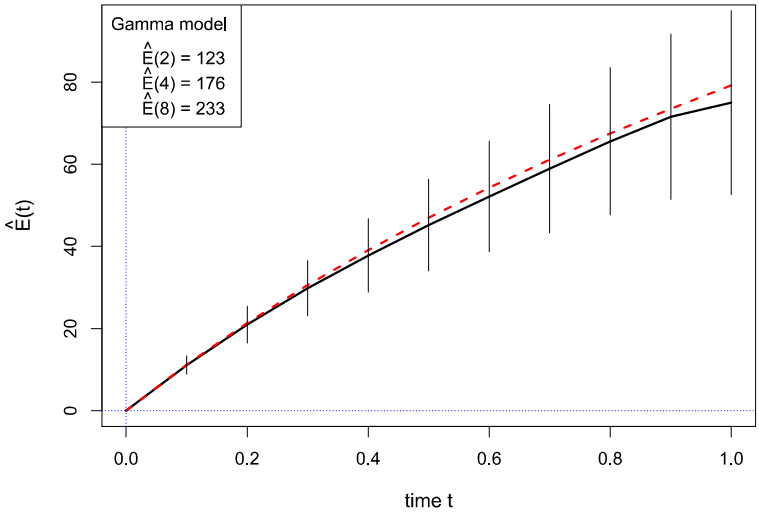

Figure 6.2 Butterfly data; expected number of new species in t units of additional trapping time. Nonparametric fit (solid) $\pm$ 1 standard deviation; gamma model (dashed).

Table 6.4 *Shakespeare's word counts; 14,376 distinct words appeared once each in the canon, 4343 distinct words twice each, etc. The canon has 884,647 words in total, counting repeats.*

	1	2	3	4	5	6	7	8	9	10
0+	14376	4343	2292	1463	1043	837	638	519	430	364
10+	305	259	242	223	187	181	179	130	127	128
20+	104	105	99	112	93	74	83	76	72	63
30+	73	47	56	59	53	45	34	49	45	52
40+	49	41	30	35	37	21	41	30	28	19
50+	25	19	28	27	31	19	19	22	23	14
60+	30	19	21	18	15	10	15	14	11	16
70+	13	12	10	16	18	11	8	15	12	7
80+	13	12	11	8	10	11	7	12	9	8
90+	4	7	6	7	10	10	15	7	7	5

as in Table 6.4, which goes on to the five words appearing 100 times each. All told, 31,534 distinct words appear (including those that appear more than 100 times each), this being the observed size of Shakespeare's vocabulary. But what of the words Shakespeare knew but didn't use? These are the "missing species" in Table 6.4.

Suppose another quantity of previously unknown Shakespeare manuscripts was discovered, comprising $884647 \cdot t$ words (so $t = 1$ would represent a new canon just as large as the old one). How many previously unseen distinct words would we expect to discover?

Employing formulas (6.19) and (6.21) gives

$$11430 \pm 178 \tag{6.25}$$

for the expected number of distinct new words if $t = 1$. This is a very conservative lower bound on how many words Shakespeare knew but didn't use. We can imagine t rising toward infinity, revealing ever more unseen vocabulary. Formula (6.19) fails for $t > 1$, and Fisher's gamma assumption is just that, but more elaborate empirical Bayes calculations give a firm lower bound of $35,000+$ on Shakespeare's unseen vocabulary, exceeding the visible portion!

Missing mass is an easier version of the missing-species problem, in which we only ask for the proportion of the total sum of θ_k values corresponding to the species that went unseen in the original trapping period,

$$M = \sum_{\text{unseen}} \theta_k \Big/ \sum_{\text{all}} \theta_k. \tag{6.26}$$

The numerator has expectation

$$\sum_{\text{all}} \theta_k e^{-\theta_k} = S \int_0^\infty \theta e^{-\theta} g(\theta) d\theta = e_1 \tag{6.27}$$

as in (6.17), while the expectation of the denominator is

$$\sum_{\text{all}} \theta_k = \sum_{\text{all}} E\{x_s\} = E\left\{\sum_{\text{all}} x_s\right\} = E\{N\}, \tag{6.28}$$

where N is the total number of butterflies trapped. The obvious missing-mass estimate is then

$$\hat{M} = y_1/N. \tag{6.29}$$

For the Shakespeare data,

$$\hat{M} = 14376/884647 = 0.016. \tag{6.30}$$

We have seen most of Shakespeare's vocabulary, as weighted by his usage, though not by his vocabulary count.

All of this seems to live in the rarefied world of mathematical abstraction, but in fact some previously unknown Shakespearean work *might* have

been discovered in 1985. A short poem, "Shall I die?," was found in the archives of the Bodleian Library and, controversially, attributed to Shakespeare by some but not all experts.

The poem of 429 words provided a new "trapping period" of length only

$$t = 429/884647 = 4.85 \cdot 10^{-4}, \tag{6.31}$$

and a prediction from (6.19) of

$$E\{t\} = 6.97 \tag{6.32}$$

new "species," i.e., distinct words not appearing in the canon. In fact there were nine such words in the poem. Similar empirical Bayes predictions for the number of words appearing once each in the canon, twice each, etc., showed reasonable agreement with the poem's counts, but not enough to stifle doubters. "Shall I die?" is currently grouped with other canonical apocrypha by a majority of experts.

6.3 A Medical Example

The reader may have noticed that our examples so far have not been particularly computer intensive; all of the calculations could have been (and originally were) done by hand.[5] This section discusses a medical study where the empirical Bayes analysis is more elaborate.

Cancer surgery sometimes involves the removal of surrounding lymph nodes as well as the primary target at the site. Figure 6.3 concerns $N = 844$ surgeries, each reporting

$$n = \#\text{ nodes removed} \quad \text{and} \quad x = \#\text{ nodes found positive}, \tag{6.33}$$

"positive" meaning malignant. The ratios

$$p_k = x_k/n_k, \qquad k = 1, 2, \ldots, N, \tag{6.34}$$

are described in the histogram. A large proportion of them, 340/844 or 40%, were zero, the remainder spreading unevenly between zero and one. The denominators n_k ranged from 1 to 69, with a mean of 19 and standard deviation of 11.

We suppose that each patient has some true probability of a node being

[5] Not so collecting the data. Corbet's work was pre-computer but Shakespeare's word counts were done electronically. Twenty-first-century scientific technology excels at the production of the large parallel-structured data sets conducive to empirical Bayes analysis.

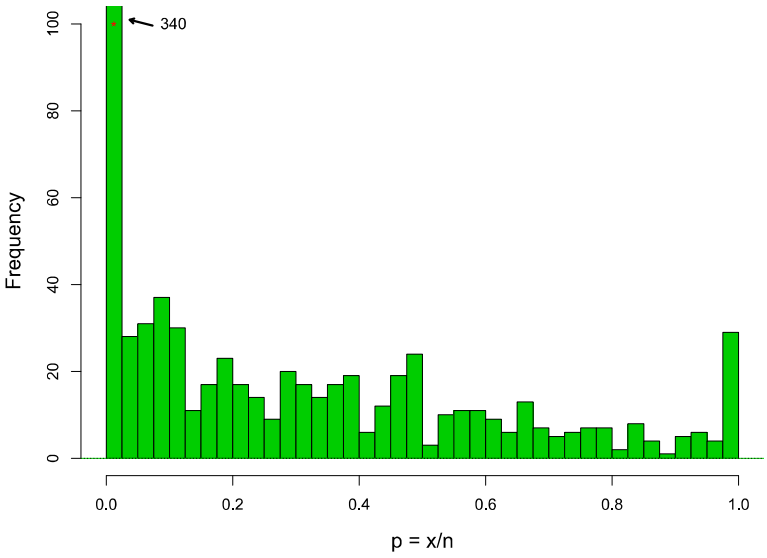

Figure 6.3 Nodes study; ratio $p = x/n$ for 844 patients; $n =$ number of nodes removed, $x =$ number positive.

positive, say probability θ_k for patient k, and that his or her nodal results occur independently of each other, making x_k binomial,

$$x_k \sim \text{Bi}(n_k, \theta_k). \tag{6.35}$$

This gives $p_k = x_k/n_k$ with mean and variance

$$p_k \sim (\theta_k, \theta_k(1-\theta_k)/n_k)\,, \tag{6.36}$$

so that θ_k is estimated more accurately when n_k is large.

A Bayesian analysis would begin with the assumption of a prior density $g(\theta)$ for the θ_k values,

$$\theta_k \sim g(\theta), \qquad \text{for } k = 1, 2, \ldots, N = 844. \tag{6.37}$$

We don't know $g(\theta)$, but the parallel nature of the nodes data set—844 similar cases—suggests an empirical Bayes approach. As a first try for the nodes study, we assume that $\log\{g(\theta)\}$ is a fourth-degree polynomial in θ,

$$\log\{g_\alpha(\theta)\} = a_0 + \sum_{j=1}^{4} \alpha_j \theta^j; \tag{6.38}$$

$g_\alpha(\theta)$ is determined by the parameter vector $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ since, given α, a_0 can be calculated from the requirement that

$$\int_0^1 g_\alpha(\theta)\, d\theta = 1 = \int_0^1 \exp\left\{a_0 + \sum_1^4 \alpha_j \theta^j\right\}\, d\theta. \tag{6.39}$$

For a given choice of α, let $f_\alpha(x_k)$ be the marginal probability of the observed value x_k for patient k,

$$f_\alpha(x_k) = \int_0^1 \binom{n_k}{x_k} \theta^{x_k} (1-\theta)^{n_k - x_k} g_\alpha(\theta)\, d\theta. \tag{6.40}$$

The maximum likelihood estimate of α is the maximizer

$$\hat{\alpha} = \arg\max_\alpha \left\{\sum_{k=1}^N \log f_\alpha(x_k)\right\}. \tag{6.41}$$

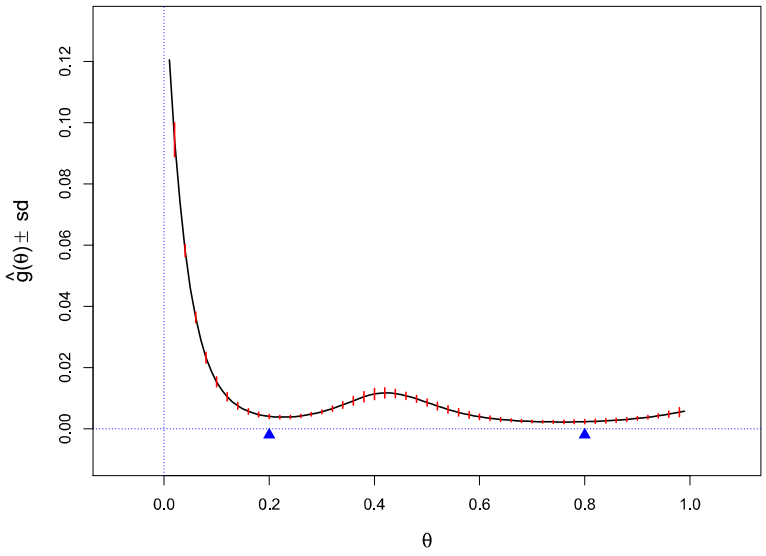

Figure 6.4 Estimated prior density $g(\theta)$ for the nodes study; 59% of patients have $\theta \leq 0.2$, 7% have $\theta \geq 0.8$.

Figure 6.4 graphs $g_{\hat{\alpha}}(\theta)$, the empirical Bayes estimate for the prior distribution of the θ_k values. The huge spike at zero in Figure 6.3 is now reduced: $\Pr\{\theta_k \leq 0.01\} = 0.12$ compared with the 38% of the p_k values

less than 0.01. Small θ values are still the rule though, for instance

$$\int_0^{0.20} g_{\hat{\alpha}}(\theta)\, d\theta = 0.59 \text{ compared with } \int_{0.80}^{1.00} g_{\hat{\alpha}}(\theta)\, d\theta = 0.07. \quad (6.42)$$

The vertical bars in Figure 6.4 indicate $\pm$ one standard error for the estimation of $g(\theta)$. The curve seems to have been estimated very accurately, at least if we assume the adequacy of model (6.37). Chapter 21 describes the computations involved in Figure 6.4.

The posterior distribution of θ_k given x_k and n_k is estimated according to Bayes' rule (3.5) to be

$$\hat{g}(\theta|x_k, n_k) = g_{\hat{\alpha}}(\theta)\binom{n_k}{x_k}\theta^{x_k}(1-\theta)^{n_k-x_k} \Big/ f_{\hat{\alpha}}(x_k), \quad (6.43)$$

with $f_{\hat{\alpha}}(x_k)$ from (6.40).

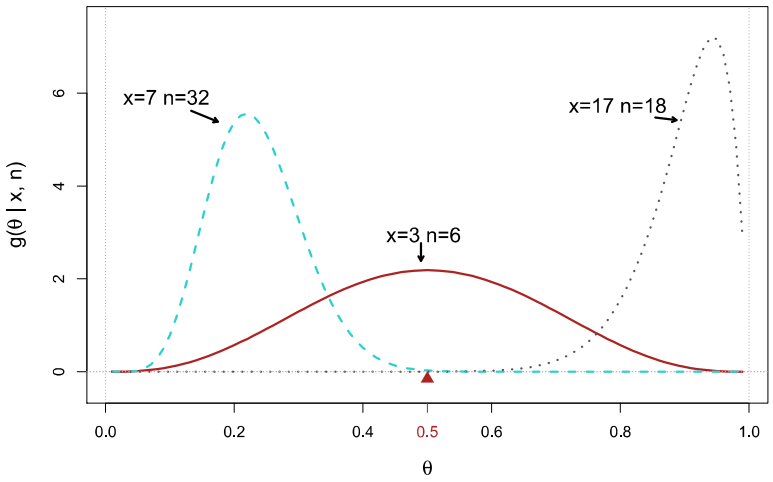

Figure 6.5 Empirical Bayes posterior densities of θ for three patients, given x = number of positive nodes, n = number of nodes.

Figure 6.5 graphs $\hat{g}(\theta|x_k, n_k)$ for three choices of (x_k, n_k): $(7, 32)$, $(3, 6)$, and $(17, 18)$. If we take $\theta \geq 0.50$ as indicating poor prognosis (and suggesting more aggressive follow-up therapy), then the first patient is almost surely on safe ground, the third patient almost surely needs more follow-up therapy and the situation of the second is uncertain.

6.4 Indirect Evidence 1

A good definition of a statistical argument is one in which many small pieces of evidence, often contradictory, are combined to produce an overall conclusion. In the clinical trial of a new drug, for instance, we don't expect the drug to cure every patient, or the placebo to always fail, but eventually perhaps we will obtain convincing evidence of the new drug's efficacy.

The clinical trial is collecting *direct* statistical evidence, in which each subject's success or failure bears directly upon the question of interest. Direct evidence, interpreted by frequentist methods, was the dominant mode of statistical application in the twentieth century, being strongly connected to the idea of scientific objectivity.

Bayesian inference provides a theoretical basis for incorporating *indirect* evidence, for example the doctor's prior experience with twin sexes in Section 3.1. The assertion of a prior density $g(\theta)$ amounts to a claim for the relevance of past data to the case at hand.

Empirical Bayes removes the Bayes scaffolding. In place of a reassuring prior $g(\theta)$, the statistician must put his or her faith in the relevance of the "other" cases in a large data set to the case of direct interest. For the second patient in Figure 6.5, the direct estimate of his θ value is $\hat{\theta} = 3/6 = 0.50$. The empirical Bayes estimate is a little less,

$$\hat{\theta}^{\text{EB}} = \int_0^1 \theta \hat{g}(\theta | x_k = 3, n_k = 6) = 0.446. \tag{6.44}$$

A small difference, but we will see bigger ones in succeeding chapters.

The changes in twenty-first-century statistics have largely been demand driven, responding to the massive data sets enabled by modern scientific equipment. Philosophically, as opposed to methodologically, the biggest change has been the increased acceptance of indirect evidence, especially as seen in empirical Bayes and objective ("uninformative") Bayes applications. *False-discovery rates*, Chapter 15, provide a particularly striking shift from direct to indirect evidence in hypothesis testing. Indirect evidence in estimation is the subject of our next chapter.

6.5 Notes and Details

Robbins (1956) introduced the term "empirical Bayes" as well as rule (6.7) as part of a general theory of empirical Bayes estimation. 1956 was also the publication year for Good and Toulmin's solution (6.19) to the missing-species problem. Good went out of his way to credit his famous Bletchley

colleague Alan Turing for some of the ideas. The auto accident data is taken from Table 3.1 of Carlin and Louis (1996), who provide a more complete discussion. Empirical Bayes estimates such as 11430 in (6.25) do not depend on independence among the "species," but accuracies such as ± 178 do; and similarly for the error bars in Figures 6.2 and 6.4.

Corbet's enormous efforts illustrate the difficulties of amassing large data sets in pre-computer times. *Dependable* data is still hard to come by, but these days it is often the statistician's job to pry it out of enormous databases. Efron and Thisted (1976) apply formula (6.19) to the Shakespeare word counts, and then use linear programming methods to bound Shakespeare's unseen vocabulary from below at 35,000 words. (Shakespeare was actually less "wordy" than his contemporaries, Marlow and Donne.) "Shall I die," the possibly Shakespearean poem recovered in 1985, is analyzed by a variety of empirical Bayes techniques in Thisted and Efron (1987). Comparisons are made with other Elizabethan authors, none of whom seem likely candidates for authorship.

The Shakespeare word counts are from Spevack's (1968) concordance. (The first concordance was compiled by hand in the mid 1800s, listing every word Shakespeare wrote and where it appeared, a full life's labor.)

The nodes example, Figure 6.3, is taken from Gholami *et al.* (2015).

†1 [p. 78] *Formula* (6.9). For any positive numbers c and d we have

$$\int_0^\infty \theta^{c-1} e^{-\theta/d}\, d\theta = d^c \Gamma(c), \tag{6.45}$$

so combining gamma prior (6.8) with Poisson density (6.1) gives marginal density

$$\begin{aligned} f_{\nu,\sigma}(x) &= \frac{\int_0^\infty \theta^{\nu+x-1} e^{-\theta/\gamma}\, d\theta}{\sigma^\nu \Gamma(\nu) x!} \\ &= \frac{\gamma^{\nu+x}\Gamma(\nu+x)}{\sigma^\nu \Gamma(\nu) x!}, \end{aligned} \tag{6.46}$$

where $\gamma = \sigma/(1+\sigma)$. Assuming independence among the counts y_x (which is exactly true if the customers act independently of each other and N, the total number of them, is itself Poisson), the log likelihood function for the accident data is

$$\sum_{x=0}^{x_{\max}} y_x \log\{f_{\nu,\sigma}(x)\}. \tag{6.47}$$

Here $x_{\max}$ is some notional upper bound on the maximum possible number

of accidents for a single customer; since $y_x = 0$ for $x > 7$ the choice of $x_{\max}$ is irrelevant. The values $(\hat{\nu}, \hat{\sigma})$ in (6.8) maximize (6.47).

†2 [p. 81] *Formula* (6.21). If $N = \sum y_x$, the total number trapped, is assumed to be Poisson, and if the N observed values x_k are mutually independent, then a useful property of the Poisson distribution implies that the counts y_x are themselves approximately independent Poisson variates

$$y_x \overset{\text{ind}}{\sim} \text{Poi}(e_x), \qquad \text{for } x = 0, 1, 2, \ldots, \tag{6.48}$$

in notation (6.17). Formula (6.19) and $\text{var}\{y_x\} = e_x$ then give

$$\text{var}\left\{\hat{E}(t)\right\} = \sum_{x \geq 1} e_x t^{2x}. \tag{6.49}$$

Substituting y_x for e_x produces (6.21). Section 11.5 of Efron (2010) shows that (6.49) is an upper bound on $\text{var}\{\hat{E}(t)\}$ if N is considered fixed rather than Poisson.

†3 [p. 81] *Formula* (6.23). Combining the case $x = 1$ in (6.17) with (6.15) yields

$$E(t) = \frac{e_1\left[\int_0^\infty e^{-\theta} g(\theta)\, d\theta - \int_0^\infty e^{-\theta(1+t)} g(\theta)\, d\theta\right]}{\int_0^\infty \theta e^{-\theta} g(\theta)\, d\theta}. \tag{6.50}$$

Substituting the gamma prior (6.8) for $g(\theta)$, and using (6.45) three times, gives formula (6.23).

7

James–Stein Estimation and Ridge Regression

If Fisher had lived in the era of "apps," maximum likelihood estimation might have made him a billionaire. Arguably the twentieth century's most influential piece of applied mathematics, maximum likelihood continues to be a prime method of choice in the statistician's toolkit. Roughly speaking, maximum likelihood provides nearly unbiased estimates of nearly minimum variance, and does so in an automatic way.

That being said, maximum likelihood estimation has shown itself to be an inadequate and dangerous tool in many twenty-first-century applications. Again speaking roughly, unbiasedness can be an unaffordable luxury when there are hundreds or thousands of parameters to estimate at the same time.

The James–Stein estimator made this point dramatically in 1961, and made it in the context of just a few unknown parameters, not hundreds or thousands. It begins the story of *shrinkage estimation*, in which deliberate biases are introduced to improve overall performance, at a possible danger to individual estimates. Chapters 7 and 21 will carry on the story in its modern implementations.

7.1 The James–Stein Estimator

Suppose we wish to estimate a single parameter μ from observation x in the Bayesian situation

$$\mu \sim \mathcal{N}(M, A) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, 1), \tag{7.1}$$

in which case μ has posterior distribution

$$\mu|x \sim \mathcal{N}\left(M + B(x - M), B\right) \qquad [B = A/(A + 1)] \tag{7.2}$$

as given in (5.21) (where we take $\sigma^2 = 1$ for convenience). The Bayes estimator of μ,

$$\hat{\mu}^{\text{Bayes}} = M + B(x - M), \tag{7.3}$$

has expected squared error

$$E\left\{\left(\hat{\mu}^{\text{Bayes}}-\mu\right)^2\right\}=B, \tag{7.4}$$

compared with 1 for the MLE $\hat{\mu}^{\text{MLE}}=x$,

$$E\left\{\left(\hat{\mu}^{\text{MLE}}-\mu\right)^2\right\}=1. \tag{7.5}$$

If, say, $A=1$ in (7.1) then $B=1/2$ and $\hat{\mu}^{\text{Bayes}}$ has only half the risk of the MLE.

The same calculation applies to a situation where we have N independent versions of (7.1), say

$$\boldsymbol{\mu}=(\mu_1,\mu_2,\ldots,\mu_N)' \quad \text{and} \quad \boldsymbol{x}=(x_1,x_2,\ldots,x_N)', \tag{7.6}$$

with

$$\mu_i\sim\mathcal{N}(M,A) \quad \text{and} \quad x_i|\mu_i\sim\mathcal{N}(\mu_i,1), \tag{7.7}$$

independently for $i=1,2,\ldots,N$. (Notice that the μ_i differ from each other, and that this situation is not the same as (5.22)–(5.23).) Let $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ indicate the vector of individual Bayes estimates $\hat{\mu}_i^{\text{Bayes}}=M+B(x_i-M)$,

$$\hat{\boldsymbol{\mu}}^{\text{Bayes}}=\boldsymbol{M}+B(\boldsymbol{x}-\boldsymbol{M}), \qquad \left[\boldsymbol{M}=(M,M,\ldots,M)'\right], \tag{7.8}$$

and similarly

$$\hat{\boldsymbol{\mu}}^{\text{MLE}}=\boldsymbol{x}.$$

Using (7.4) the total squared error risk of $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ is

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{Bayes}}-\boldsymbol{\mu}\right\|^2\right\}=E\left\{\sum_{i=1}^{N}\left(\hat{\mu}_i^{\text{Bayes}}-\mu_i\right)^2\right\}=N\cdot B \tag{7.9}$$

compared with

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{MLE}}-\boldsymbol{\mu}\right\|^2\right\}=N. \tag{7.10}$$

Again, $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ has only B times the risk of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$.

This is fine if we know M and A (or equivalently M and B) in (7.1). If not, we might try to estimate them from $\boldsymbol{x}=(x_1,x_2,\ldots,x_N)$. Marginally, (7.7) gives

$$x_i\overset{\text{ind}}{\sim}\mathcal{N}(M,A+1). \tag{7.11}$$

Then $\hat{M}=\bar{x}$ is an unbiased estimate of M. Moreover,

$$\hat{B}=1-(N-3)/S \qquad \left[S=\sum_{i=1}^{N}(x_i-\bar{x})^2\right] \tag{7.12}$$

unbiasedly estimates B, as long as $N > 3$.† The James–Stein estimator is †1
the plug-in version of (7.3),

$$\hat{\mu}_i^{\text{JS}} = \hat{M} + \hat{B}\left(x_i - \hat{M}\right) \qquad \text{for } i = 1, 2, \ldots, N, \tag{7.13}$$

or equivalently $\hat{\boldsymbol{\mu}}^{\text{JS}} = \hat{\boldsymbol{M}} + \hat{B}(\boldsymbol{x} - \hat{\boldsymbol{M}})$, with $\hat{\boldsymbol{M}} = (\hat{M}, \hat{M}, \ldots, \hat{M})'$.

At this point the terminology "empirical Bayes" seems especially apt: Bayesian model (7.7) leads to the Bayes estimator (7.8), which itself is estimated empirically (i.e., frequentistically) from all the data $\boldsymbol{x}$, and then applied to the individual cases. Of course $\hat{\boldsymbol{\mu}}^{\text{JS}}$ cannot perform as well as the actual Bayes' rule $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$, but the increased risk is surprisingly modest. The expected squared risk of $\hat{\boldsymbol{\mu}}^{\text{JS}}$ under model (7.7) is† †2

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{JS}} - \boldsymbol{\mu}\right\|^2\right\} = NB + 3(1 - B). \tag{7.14}$$

If, say, $N = 20$ and $A = 1$, then (7.14) equals 11.5, compared with true Bayes risk 10 from (7.9), much less than risk 20 for $\hat{\boldsymbol{\mu}}^{\text{MLE}}$.

A defender of maximum likelihood might respond that none of this is surprising: Bayesian model (7.7) specifies the parameters μ_i to be clustered more or less closely around a central point M, while $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ makes no such assumption, and cannot be expected to perform as well. Wrong! Removing the Bayesian assumptions does not rescue $\hat{\boldsymbol{\mu}}^{\text{MLE}}$, as James and Stein proved in 1961:

James–Stein Theorem *Suppose that*

$$x_i | \mu_i \sim \mathcal{N}(\mu_i, 1) \tag{7.15}$$

independently for $i = 1, 2, \ldots, N$, *with* $N \geq 4$. *Then*

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{JS}} - \boldsymbol{\mu}\right\|^2\right\} < N = E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{MLE}} - \boldsymbol{\mu}\right\|^2\right\} \tag{7.16}$$

for all choices of $\boldsymbol{\mu} \in \mathcal{R}^N$. (The expectations in (7.16) are with $\boldsymbol{\mu}$ fixed and $\boldsymbol{x}$ varying according to (7.15).)

In the language of decision theory, equation (7.16) says that $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ is *inadmissible*:† its total squared error risk exceeds that of $\hat{\boldsymbol{\mu}}^{\text{JS}}$ no matter †3
what $\boldsymbol{\mu}$ may be. This is a strong frequentist form of defeat for $\hat{\boldsymbol{\mu}}^{\text{MLE}}$, not depending on Bayesian assumptions.

The James–Stein theorem came as a rude shock to the statistical world of 1961. First of all, the defeat came on MLE's home field: normal observations with squared error loss. Fisher's "logic of inductive inference," Chapter 4, claimed that $\hat{\mu}^{\text{MLE}} = x$ was the obviously correct estimator in the univariate case, an assumption tacitly carried forward to multiparameter linear

regression problems, where versions of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ were predominant. There are still some good reasons for sticking with $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ in low-dimensional problems, as discussed in Section 7.4. But shrinkage estimation, as exemplified by the James–Stein rule, has become a necessity in the high-dimensional situations of modern practice.

7.2 The Baseball Players

The James–Stein theorem doesn't say by how much $\hat{\boldsymbol{\mu}}^{\text{JS}}$ beats $\hat{\boldsymbol{\mu}}^{\text{MLE}}$. If the improvement were infinitesimal nobody except theorists would be interested. In favorable situations the gains can in fact be substantial, as suggested by (7.14). One such situation appears in Table 7.1. The batting averages[1] of 18 Major League players have been observed over the 1970 season. The column labeled **MLE** reports the player's observed average over his first 90 at bats; **TRUTH** is the average over the remainder of the 1970 season (370 further at bats on average). We would like to predict **TRUTH** from the early-season observations.

The column labeled **JS** in Table 7.1 is from a version of the James–Stein estimator applied to the 18 MLE numbers. We suppose that each player's **MLE** value p_i (his batting average in the first 90 tries) is a binomial proportion,

$$p_i \sim \text{Bi}(90, P_i)/90. \tag{7.17}$$

Here P_i is his *true average*, how he would perform over an infinite number of tries; $\texttt{TRUTH}_i$ is itself a binomial proportion, taken over an average of 370 more tries per player.

At this point there are two ways to proceed. The simplest uses a normal approximation to (7.17),

$$p_i \doteq \mathcal{N}(P_i, \sigma_0^2), \tag{7.18}$$

where σ_0^2 is the binomial variance

$$\sigma_0^2 = \bar{p}(1-\bar{p})/90, \tag{7.19}$$

with $\bar{p} = 0.254$ the average of the p_i values. Letting $x_i = p_i/\sigma_0$, applying (7.13), and transforming back to $\hat{p}_i^{\text{JS}} = \sigma_0 \hat{\mu}_i^{\text{JS}}$, gives James–Stein estimates

$$\hat{p}_i^{\text{JS}} = \bar{p} + \left[1 - \frac{(N-3)\sigma_0^2}{\sum(p_i - \bar{p})^2}\right](p_i - \bar{p}). \tag{7.20}$$

[1] Batting average = # hits /# at bats, that is, the success rate. For example, Player 1 hits successfully 31 times in his first 90 tries, for batting average $31/90 = 0.345$. This data is based on 1970 Major League performances, but is partly artificial; see the endnotes.

Table 7.1 *Eighteen baseball players;* `MLE` *is batting average in first 90 at bats;* `TRUTH` *is average in remainder of 1970 season; James–Stein estimator* `JS` *is based on arcsin transformation of MLEs. Sum of squared errors for predicting* `TRUTH`*:* `MLE` *.0425,* `JS` *.0218.*

Player	MLE	JS	TRUTH	x
1	.345	.283	.298	11.96
2	.333	.279	.346	11.74
3	.322	.276	.222	11.51
4	.311	.272	.276	11.29
5	.289	.265	.263	10.83
6	.289	.264	.273	10.83
7	.278	.261	.303	10.60
8	.255	.253	.270	10.13
9	.244	.249	.230	9.88
10	.233	.245	.264	9.64
11	.233	.245	.264	9.64
12	.222	.242	.210	9.40
13	.222	.241	.256	9.39
14	.222	.241	.269	9.39
15	.211	.238	.316	9.14
16	.211	.238	.226	9.14
17	.200	.234	.285	8.88
18	.145	.212	.200	7.50

A second approach begins with the *arcsin transformation*

$$x_i = 2(n+0.5)^{1/2} \sin^{-1}\left[\left(\frac{np_i + 0.375}{n+0.75}\right)^{1/2}\right], \tag{7.21}$$

$n = 90$ (column labeled `x` in Table 7.1), a classical device that produces approximate normal deviates of variance 1,

$$x_i \doteq \mathcal{N}(\mu_i, 1), \tag{7.22}$$

where μ_i is transformation (7.21) applied to $\texttt{TRUTH}_i$. Using (7.13) gives $\hat{\mu}_i^{\text{JS}}$, which is finally inverted back to the binomial scale,

$$\hat{p}_i^{\text{JS}} = \frac{1}{n}\left[(n+0.75)\left(\sin\left(\frac{\hat{\mu}_i^{\text{JS}}}{2\sqrt{n+0.5}}\right)\right)^2 - 0.375\right] \tag{7.23}$$

Formulas (7.20) and (7.23) yielded nearly the same estimates for the baseball players; the `JS` column in Table 7.1 is from (7.23). James and Stein's theorem requires normality, but the James–Stein estimator often

works perfectly well in less ideal situations. That is the case in Table 7.1:

$$\sum_{i=1}^{18}(\texttt{MLE}_i-\texttt{TRUTH}_i)^2 = 0.0425 \quad \text{while} \quad \sum_{i=1}^{18}(\texttt{JS}_i-\texttt{TRUTH}_i)^2 = 0.0218. \tag{7.24}$$

In other words, the James–Stein estimator reduced total predictive squared error by about 50%.

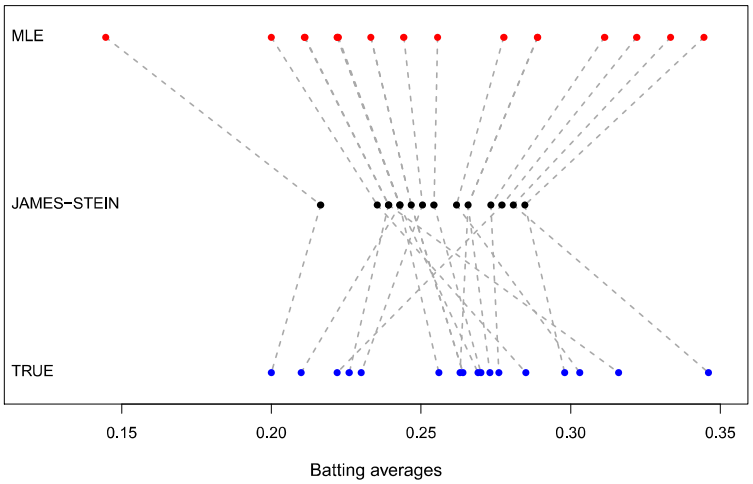

Figure 7.1 Eighteen baseball players; top line MLE, middle James–Stein, bottom true values. Only 13 points are visible, since there are ties.

The James–Stein rule describes a *shrinkage estimator*, each MLE value x_i being shrunk by factor $\hat{B}$ toward the grand mean $\hat{M} = \bar{x}$ (7.13). ($\hat{B} = 0.34$ in (7.20).) Figure 7.1 illustrates the shrinking process for the baseball players.

To see why shrinking might make sense, let us return to the original Bayes model (7.8) and take $M = 0$ for simplicity, so that the x_i are marginally $\mathcal{N}(0, A+1)$ (7.11). Even though each x_i is unbiased for its parameter μ_i, as a group they are "overdispersed,"

$$E\left\{\sum_{i=1}^{N} x_i^2\right\} = N(A+1) \quad \text{compared with} \quad E\left\{\sum_{i=1}^{N} \mu_i^2\right\} = NA. \tag{7.25}$$

The sum of squares of the MLEs exceeds that of the true values by expected amount N; shrinkage improves group estimation by removing the excess.

In fact the James–Stein rule *overshrinks* the data, as seen in the bottom two lines of Figure 7.1, a property it inherits from the underlying Bayes model: the Bayes estimates $\hat{\mu}_i^{\text{Bayes}} = Bx_i$ have

$$E\left\{\sum_{i=1}^{N}\left(\hat{\mu}_i^{\text{Bayes}}\right)^2\right\} = NB^2(A+1) = NA\frac{A}{A+1}, \tag{7.26}$$

overshrinking $E(\sum \mu_i^2) = NA$ by factor $A/(A+1)$. We could use the less extreme shrinking rule $\tilde{\mu}_i = \sqrt{B}x_i$, which gives the correct expected sum of squares NA, but a larger expected sum of squared estimation errors $E\{\sum(\tilde{\mu}_i - \mu_i)^2|\boldsymbol{x}\}$.

The most extreme shrinkage rule would be "all the way," that is, to

$$\hat{\mu}_i^{\text{NULL}} = \bar{x} \qquad \text{for } i = 1, 2, \ldots, N, \tag{7.27}$$

NULL indicating that in a classical sense we have accepted the null hypothesis of no differences among the μ_i values. (This gave $\sum(P_i - \bar{p})^2 = 0.0266$ for the baseball data (7.24).) The James–Stein estimator is a data-based rule for compromising between the null hypothesis of no differences and the MLE's tacit assumption of no relationship at all among the μ_i values. In this sense it blurs the classical distinction between hypothesis testing and estimation.

7.3 Ridge Regression

Linear regression, perhaps the most widely used estimation technique, is based on a version of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$. In the usual notation, we observe an n-dimensional vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ from the linear model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{7.28}$$

Here $\boldsymbol{X}$ is a known $n \times p$ *structure matrix*, β is an unknown p-dimensional parameter vector, while the *noise vector* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ has its components uncorrelated and with constant variance σ^2,

$$\boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{7.29}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Often $\boldsymbol{\epsilon}$ is assumed to be multivariate normal,

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{7.30}$$

but that is not required for most of what follows.

The *least squares estimate* $\hat{\beta}$, going back to Gauss and Legendre in the early 1800s, is the minimizer of the total sum of squared errors,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 \right\}. \tag{7.31}$$

It is given by

$$\hat{\beta} = S^{-1}X'y, \tag{7.32}$$

where S is the $p \times p$ inner product matrix

$$S = X'X; \tag{7.33}$$

$\hat{\beta}$ is unbiased for β and has covariance matrix $\sigma^2 S^{-1}$,

$$\hat{\beta} \sim \left(\beta, \sigma^2 S^{-1}\right). \tag{7.34}$$

In the normal case (7.30) $\hat{\beta}$ is the MLE of β. Before 1950 a great deal of effort went into designing matrices X such that S^{-1} could be feasibly calculated, which is now no longer a concern.

A great advantage of the linear model is that it reduces the number of unknown parameters to p (or $p + 1$ including σ^2), no matter how large n may be. In the kidney data example of Section 1.1, $n = 157$ while $p = 2$. In modern applications, however, p has grown larger and larger, sometimes into the thousands or more, as we will see in Part III, causing statisticians again to confront the limitations of high-dimensional unbiased estimation.

Ridge regression is a shrinkage method designed to improve the estimation of β in linear models. By transformations[†] we can *standardize* (7.28) so that the columns of X each have mean 0 and sum of squares 1, that is,

†4

$$S_{ii} = 1 \qquad \text{for } i = 1, 2, \ldots, p. \tag{7.35}$$

(This puts the regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$ on comparable scales.) For convenience, we also assume $\bar{y} = 0$. A ridge regression estimate $\hat{\beta}(\lambda)$ is defined, for $\lambda \geq 0$, to be

$$\hat{\beta}(\lambda) = (S + \lambda I)^{-1}X'y = (S + \lambda I)^{-1}S\hat{\beta} \tag{7.36}$$

(using (7.32)); $\hat{\beta}(\lambda)$ is a shrunken version of $\hat{\beta}$, the bigger λ the more extreme the shrinkage: $\hat{\beta}(0) = \hat{\beta}$ while $\hat{\beta}(\infty)$ equals the vector of zeros.

Ridge regression effects can be quite dramatic. As an example, consider the diabetes data, partially shown in Table 7.2, in which 10 prediction variables measured at baseline—`age`, `sex`, `bmi` (body mass index), `map` (mean arterial blood pressure), and six blood serum measurements—have

Table 7.2 *First 7 of $n = 442$ patients in the diabetes study; we wish to predict disease progression at one year "prog" from the 10 baseline measurements* `age`, `sex`, ..., `glu`.

age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	prog
59	1	32.1	101	157	93.2	38	4	2.11	87	151
48	0	21.6	87	183	103.2	70	3	1.69	69	75
72	1	30.5	93	156	93.6	41	4	2.03	85	141
24	0	25.3	84	198	131.4	40	5	2.12	89	206
50	0	23.0	101	192	125.4	52	4	1.86	80	135
23	0	22.6	89	139	64.8	61	2	1.82	68	97
36	1	22.0	90	160	99.6	50	3	1.72	82	138
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

been obtained for $n = 442$ patients. We wish to use the 10 variables to predict `prog`, a quantitative assessment of disease progression one year after baseline. In this case $\boldsymbol{X}$ is the 442×10 matrix of standardized predictor variables, and $\boldsymbol{y}$ is `prog` with its mean subtracted off.

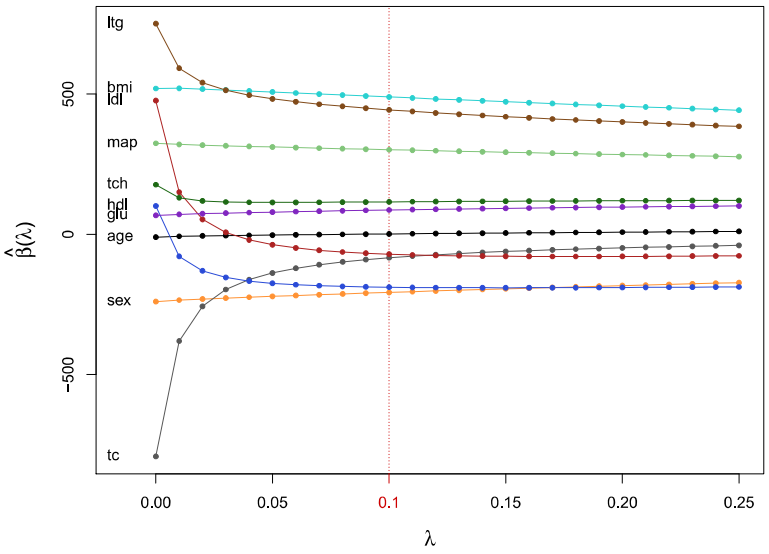

Figure 7.2 Ridge coefficient trace for the standardized diabetes data.

Table 7.3 *Ordinary least squares estimate $\hat{\beta}(0)$ compared with ridge regression estimate $\hat{\beta}(0.1)$ with $\lambda = 0.1$. The columns sd(0) and sd(0.1) are their estimated standard errors. (Here σ was taken to be 54.1, the usual OLS estimate based on model (7.28).)*

	$\hat{\beta}(0)$	$\hat{\beta}(0.1)$	sd(0)	sd(0.1)
`age`	−10.0	1.3	59.7	52.7
`sex`	−239.8	−207.2	61.2	53.2
`bmi`	519.8	489.7	66.5	56.3
`map`	324.4	301.8	65.3	55.7
`tc`	−792.2	−83.5	416.2	43.6
`ldl`	476.7	−70.8	338.6	52.4
`hdl`	101.0	−188.7	212.3	58.4
`tch`	177.1	115.7	161.3	70.8
`ltg`	751.3	443.8	171.7	58.4
`glu`	67.6	86.7	65.9	56.6

Figure 7.2 vertically plots the 10 coordinates of $\hat{\beta}(\lambda)$ as the ridge parameter λ increases from 0 to 0.25. Four of the coefficients change rapidly at first. Table 7.3 compares $\hat{\beta}(0)$, that is the usual estimate $\hat{\beta}$, with $\hat{\beta}(0.1)$. Positive coefficients predict increased disease progression. Notice that `ldl`, the "bad cholesterol" measurement, goes from being a strongly positive predictor in $\hat{\beta}$ to a mildly negative one in $\hat{\beta}(0.1)$.

There is a Bayesian rationale for ridge regression. Assume that the noise vector $\boldsymbol{\epsilon}$ is normal as in (7.30), so that

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2 \boldsymbol{S}^{-1}\right) \tag{7.37}$$

rather than just (7.34). Then the Bayesian prior

$$\beta \sim \mathcal{N}_p\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\boldsymbol{I}\right) \tag{7.38}$$

makes

$$E\left\{\beta|\hat{\beta}\right\} = (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}\hat{\beta}, \tag{7.39}$$

the same as the ridge regression estimate $\hat{\beta}(\lambda)$ (using (5.23) with $M = 0$, $\boldsymbol{A} = (\sigma^2/\lambda)\boldsymbol{I}$, and $\boldsymbol{\Sigma} = (\boldsymbol{S}/\sigma^2)^{-1}$). Ridge regression amounts to an increased prior belief that β lies near 0.

†5 The last two columns of Table 7.3 compare the standard deviations† of $\hat{\beta}$ and $\hat{\beta}(0.1)$. Ridging has greatly reduced the variability of the estimated

regression coefficients. This does *not* guarantee that the corresponding estimate of $\boldsymbol{\mu} = \boldsymbol{X}\beta$,

$$\hat{\boldsymbol{\mu}}(\lambda) = \boldsymbol{X}\hat{\beta}(\lambda), \tag{7.40}$$

will be more accurate than the ordinary least squares estimate $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\beta}$. We have (deliberately) introduced bias, and the squared bias term counteracts some of the advantage of reduced variability. The C_p calculations of Chapter 12 suggest that the two effects nearly offset each other for the diabetes data. However, if interest centers on the coefficients of β, then ridging can be crucial, as Table 7.3 emphasizes.

By current standards, $p = 10$ is a small number of predictors. Data sets with p in the thousands, and more, will show up in Part III. In such situations the scientist is often looking for a few interesting predictor variables hidden in a sea of uninteresting ones: the prior belief is that most of the β_i values lie near zero. Biasing the maximum likelihood estimates $\hat{\beta}_i$ toward zero then becomes a necessity.

There is still another way to motivate the ridge regression estimator $\hat{\beta}(\lambda)$:

$$\hat{\beta}(\lambda) = \arg\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2\}. \tag{7.41}$$

Differentiating the term in brackets with respect to β shows that $\hat{\beta}(\lambda) = (\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ as in (7.36). If $\lambda = 0$ then (7.41) describes the ordinary least squares algorithm; $\lambda > 0$ *penalizes* choices of β having $\|\beta\|$ large, biasing $\hat{\beta}(\lambda)$ toward the origin.

Various terminologies are used to describe algorithms such as (7.41): *penalized least squares*; *penalized likelihood*; *maximized a-posteriori probability* (MAP);[†]and, generically, *regularization* describes almost any method [†6] that tamps down statistical variability in high-dimensional estimation or prediction problems.

A wide variety of penalty terms are in current use, the most influential one involving the "ℓ_1 norm" $\|\beta\|_1 = \sum_1^p |\beta_j|$,

$$\tilde{\beta}(\lambda) = \arg\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|_1\}, \tag{7.42}$$

the so-called *lasso* estimator, Chapter 16. Despite the Bayesian provenance, most regularization research is carried out frequentistically, with various penalty terms investigated for their probabilistic behavior regarding estimation, prediction, and variable selection.

If we apply the James–Stein rule to the normal model (7.37), we get a different shrinkage rule[†] for $\hat{\beta}$, say $\tilde{\beta}^{\mathrm{JS}}$, [†7]

$$\tilde{\beta}^{\mathrm{JS}} = \left[1 - \frac{(p-2)\sigma^2}{\hat{\beta}' \boldsymbol{S} \hat{\beta}}\right] \hat{\beta}. \tag{7.43}$$

Letting $\tilde{\boldsymbol{\mu}}^{\mathrm{JS}} = \boldsymbol{X}\tilde{\beta}^{\mathrm{JS}}$ be the corresponding estimator of $\boldsymbol{\mu} = E\{\boldsymbol{y}\}$ in (7.28), the James–Stein Theorem guarantees that

$$E\left\{\left\|\tilde{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} < p\sigma^2 \tag{7.44}$$

no matter what β is, as long as $p \geq 3$.[2] There is no such guarantee for ridge regression, and no foolproof way to choose the ridge parameter λ. On the other hand, $\tilde{\beta}^{\mathrm{JS}}$ does not stabilize the coordinate standard deviations, as in the sd(0.1) column of Table 7.3. The main point here is that at present there is no optimality theory for shrinkage estimation. Fisher provided an elegant theory for optimal unbiased estimation. It remains to be seen whether biased estimation can be neatly codified.

7.4 Indirect Evidence 2

There is a downside to shrinkage estimation, which we can examine by returning to the baseball data of Table 7.1. One thousand simulations were run, each one generating simulated batting averages

$$p_i^* \sim \mathrm{Bi}(90, \mathtt{TRUTH}_i)/90 \qquad i = 1, 2, \ldots, 18. \tag{7.45}$$

These gave corresponding James–Stein (JS) estimates (7.20), with $\sigma_0^2 = \bar{p}^*(1-\bar{p}^*)/90$.

Table 7.4 shows the root mean square error for the MLE and JS estimates over 1000 simulations for each of the 18 players,

$$\left[\frac{1}{1000}\sum_{j=1}^{1000}(p_{ij}^* - \mathtt{TRUTH}_i)^2\right]^{1/2} \text{ and } \left[\frac{1}{1000}\sum_{j=1}^{1000}(\hat{p}_{ij}^{*\,\mathrm{JS}} - \mathtt{TRUTH}_i)^2\right]^{1/2} \tag{7.46}$$

As foretold by the James–Stein Theorem, the JS estimates are easy victors in terms of total squared error (summing over all 18 players). However, $\hat{p}_i^{*\,\mathrm{JS}}$ loses to $\hat{p}_i^{*\,\mathrm{MLE}} = p_i^*$ for 4 of the 18 players, losing badly in the case of player 2.

Histograms comparing the 1000 simulations of p_i^* with those of $\hat{p}_i^{*\,\mathrm{JS}}$ for player 2 appear in Figure 7.3. Strikingly, all 1000 of the $\hat{p}_{2j}^{*\,\mathrm{JS}}$ values lie

[2] Of course we are assuming σ^2 is known in (7.43); if it is estimated, some of the improvement erodes away.

Table 7.4 *Simulation study comparing root mean square errors for MLE and JS estimators* (7.20) *as estimates of* `TRUTH`*. Total mean square errors .0384 (*`MLE`*) and .0235 (*`JS`*). Asterisks indicate four players for whom* `rmsJS` *exceeded* `rmsMLE`*; these have two largest and two smallest* `TRUTH` *values (player 2 is Clemente). Column* `rmsJS1` *is for the limited translation version of* `JS` *that bounds shrinkage to within one standard deviation of the* `MLE`*.*

Player	`TRUTH`	`rmsMLE`	`rmsJS`	`rmsJS1`
1	.298	.046	.033	.032
2	.346*	.049	.077	.056
3	.222	.044	.042	.038
4	.276	.048	.015	.023
5	.263	.047	.011	.020
6	.273	.046	.014	.021
7	.303	.047	.037	.035
8	.270	.049	.012	.022
9	.230	.044	.034	.033
10	.264	.047	.011	.021
11	.264	.047	.012	.020
12	.210*	.043	.053	.044
13	.256	.045	.014	.020
14	.269	.048	.012	.021
15	.316*	.048	.049	.043
16	.226	.045	.038	.036
17	.285	.046	.022	.026
18	.200*	.043	.062	.048

below $\texttt{TRUTH}_2 = 0.346$. Player 2 could have had a legitimate complaint if the James–Stein estimate were used to set his next year's salary.

The four losing cases for $\hat{p}_i^{*\,\mathrm{JS}}$ are the players with the two largest and two smallest values of the `TRUTH`. Shrinkage estimators work against cases that are genuinely outstanding (in a positive or negative sense). Player 2 was Roberto Clemente. A better informed Bayesian, that is, a baseball fan, would know that Clemente had led the league in batting over the previous several years, and shouldn't be thrown into a shrinkage pool with 17 ordinary hitters.

Of course the James–Stein estimates *were* more accurate for 14 of the 18 players. Shrinkage estimation tends to produce better results *in general*, at the possible expense of extreme cases. Nobody cares much about Cold War batting averages, but if the context were the efficacies of 18 new anticancer drugs the stakes would be higher.

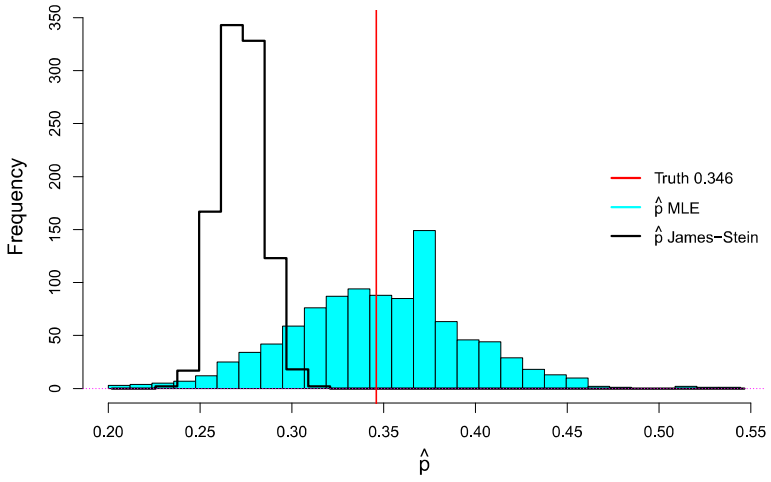

Figure 7.3 Comparing MLE estimates (solid) with JS estimates (line) for Clemente; 1000 simulations, 90 at bats each.

Compromise methods are available. The `rmsJS1` column of Table 7.4 refers to a *limited translation* version of $\hat{p}_i^{\,\mathrm{JS}}$ in which shrinkage is not allowed to diverge more than one σ_0 unit from $\hat{p}_i$; in formulaic terms,

$$\hat{p}_i^{\,\mathrm{JS}\,1} = \min\left\{\max\left(\hat{p}_i^{\,\mathrm{JS}}, \hat{p}_i - \sigma_0\right), \hat{p}_i + \sigma_0\right\}. \tag{7.47}$$

This mitigates the Clemente problem while still gaining most of the shrinkage advantages.

The use of indirect evidence amounts to *learning from the experience of others*, each batter learning from the 17 others in the baseball examples. "Which others?" is a key question in applying computer-age methods. Chapter 15 returns to the question in the context of false-discovery rates.

7.5 Notes and Details

The Bayesian motivation emphasized in Chapters 6 and 7 is anachronistic: originally the work emerged mainly from frequentist considerations and was justified frequentistically, as in Robbins (1956). Stein (1956) proved the inadmissibility of $\hat{\boldsymbol{\mu}}^{\mathrm{MLE}}$, the neat version of $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ appearing in James and Stein (1961) (Willard James was Stein's graduate student); $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ is itself inadmissible, being everywhere improvable by changing $\hat{B}$ in (7.13)

to $\max(\hat{B}, 0)$. This in turn is inadmissible, but further gains tend to the minuscule.

In a series of papers in the early 1970s, Efron and Morris emphasized the empirical Bayes motivation of the James–Stein rule, Efron and Morris (1972) giving the limited translation version (7.47). The baseball data in its original form appears in Table 1.1 of Efron (2010). Here the original 45 at bats recorded for each player have been artificially augmented by adding 45 binomial draws, $\text{Bi}(45, \texttt{TRUTH}_i)$ for player i. This gives a somewhat less optimistic view of the James–Stein rule's performance.

"Stein's paradox in statistics," Efron and Morris' title for their 1977 *Scientific American* article, catches the statistics world's sense of discomfort with the James–Stein theorem. Why should our estimate for Player A go up or down depending on the other players' performances? This is the question of direct versus indirect evidence, raised again in the context of hypothesis testing in Chapter 15. Unbiased estimation has great scientific appeal, so the argument is by no means settled.

Ridge regression was introduced into the statistics literature by Hoerl and Kennard (1970). It appeared previously in the numerical analysis literature as Tikhonov regularization.

†1 [p. 93] *Formula* (7.12). If Z has a chi-squared distribution with ν degrees of freedom, $Z \sim \chi^2_\nu$ (that is, $Z \sim \text{Gam}(\nu/2, 2)$ in Table 5.1), it has density

$$f(z) = \frac{z^{\nu/2-1}e^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)} \qquad \text{for } z \geq 0, \tag{7.48}$$

yielding

$$E\left\{\frac{1}{z}\right\} = \int_0^\infty \frac{z^{\nu/2-2}e^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)}\,dz = \frac{2^{\nu/2-1}}{2^{\nu/2}}\frac{\Gamma(\nu/2-1)}{\Gamma(\nu/2)} = \frac{1}{\nu-2}. \tag{7.49}$$

But standard results, starting from (7.11), show that $S \sim (A+1)\chi^2_{N-1}$. With $\nu = N-1$ in (7.49),

$$E\left\{\frac{N-3}{S}\right\} = \frac{1}{A+1}, \tag{7.50}$$

verifying (7.12).

†2 [p. 93] *Formula* (7.14). First consider the simpler situation where M in (7.11) is known to equal zero, in which case the James–Stein estimator is

$$\hat{\mu}_i^{\text{JS}} = \hat{B}x_i \qquad \text{with } \hat{B} = 1-(N-2)/S, \tag{7.51}$$

where $S = \sum_1^N x_i^2$. For convenient notation let

$$\hat{C} = 1-\hat{B} = (N-2)/S \quad \text{and} \quad C = 1-B = 1/(A+1). \tag{7.52}$$

The conditional distribution $\mu_i|\boldsymbol{x} \sim \mathcal{N}(Bx_i, B)$ gives

$$E\left\{\left(\hat{\mu}_i^{\mathrm{JS}} - \mu_i\right)^2 \middle| \boldsymbol{x}\right\} = B + (\hat{C} - C)^2 x_i^2, \tag{7.53}$$

and, adding over the N coordinates,

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2 \middle| \boldsymbol{x}\right\} = NB + (\hat{C} - C)^2 S. \tag{7.54}$$

The marginal distribution $S \sim (A+1)\chi_N^2$ and (7.49) yields, after a little calculation,

$$E\left\{(\hat{C} - C)^2 S\right\} = 2(1 - B), \tag{7.55}$$

and so

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} = NB + 2(1 - B). \tag{7.56}$$

By orthogonal transformations, in situation (7.7), where M is not assumed to be zero, $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ can be represented as the sum of two parts: a JS estimate in $N - 1$ dimensions but with $M = 0$ as in (7.51), and a MLE estimate of the remaining one coordinate. Using (7.56) this gives

$$\begin{aligned} E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} &= (N-1)B + 2(1-B) + 1 \\ &= NB + 3(1-B), \end{aligned} \tag{7.57}$$

which is (7.14).

†3 [p. 93] *The James–Stein Theorem.* Stein (1981) derived a simpler proof of the JS Theorem that appears in Section 1.2 of Efron (2010).

†4 [p. 98] *Transformations to form* (7.35). The linear regression model (7.28) is *equivariant* under scale changes of the variables $\boldsymbol{x}_j$. What this means is that the space of fits using linear combinations of the $\boldsymbol{x}_j$ is the same as the space of linear combinations using scaled versions $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j / s_j$, with $s_j > 0$. Furthermore, the least squares fits are the same, and the coefficient estimates map in the obvious way: $\hat{\tilde{\beta}}_j = s_j \hat{\beta}_j$.

Not so for ridge regression. Changing the scales of the columns of $\boldsymbol{X}$ will generally lead to different fits. Using the penalty version (7.41) of ridge regression, we see that the penalty term $\|\beta\|^2 = \sum_j \beta_j^2$ treats all the coefficients as equals. This penalty is most natural if all the variables are measured on the same scale. Hence we typically use for s_j the standard deviation of variable $\boldsymbol{x}_j$, which leads to (7.35). Furthermore, with ridge regression we typically do not penalize the intercept. This can be achieved

by *centering* and scaling each of the variables, $\tilde{\boldsymbol{x}}_j = (\boldsymbol{x}_j - \mathbf{1}\bar{x}_j)/s_j$, where

$$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n \quad \text{and} \quad s_j = \left[\frac{1}{n}\sum (x_{ij} - \bar{x}_j)^2\right]^{1/2}, \tag{7.58}$$

with $\mathbf{1}$ the n-vector of 1s. We now work with $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_p)$ rather than $\boldsymbol{X}$, and the intercept is estimated separately as $\bar{y}$.

†5 [p. 100] *Standard deviations in Table 7.3.* From the first equality in (7.36) we calculate the covariance matrix of $\hat{\beta}(\lambda)$ to be

$$\text{Cov}_\lambda = \sigma^2(\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}(\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}. \tag{7.59}$$

The entries sd(0.1) in Table 7.3 are square roots of the diagonal elements of Cov_λ, substituting the ordinary least squares estimate $\hat{\sigma} = 54.1$ for σ^2.

†6 [p. 101] *Penalized likelihood and MAP.* With σ^2 fixed and known in the normal linear model $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I})$, minimizing $\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$ is the same as maximizing the log density function

$$\log f_\beta(y) = -\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \text{ constant}. \tag{7.60}$$

In this sense, the term $\lambda\|\beta\|^2$ in (7.41) *penalizes* the likelihood $\log f_\beta(y)$ connected with β in proportion to the magnitude $\|\beta\|^2$. Under the prior distribution (7.38), the log posterior density of β given $\boldsymbol{y}$ (the log of (3.5)) is

$$-\frac{1}{2\sigma^2}\left\{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2\right\}, \tag{7.61}$$

plus a term that doesn't depend on β. That makes the maximizer of (7.41) also the maximizer of the posterior density of β given $\boldsymbol{y}$, or the MAP.

†7 [p. 101] *Formula* (7.43). Let $\gamma = (\boldsymbol{S}^{1/2}/\sigma)\beta$ and $\hat{\gamma} = (\boldsymbol{S}^{1/2}/\sigma)\hat{\beta}$ in (7.37), where $\boldsymbol{S}^{1/2}$ is a matrix square root of $\boldsymbol{S}$, $(\boldsymbol{S}^{1/2})^2 = \boldsymbol{S}$. Then

$$\hat{\gamma} \sim \mathcal{N}_p(\gamma, \boldsymbol{I}), \tag{7.62}$$

and the $M = 0$ form of the James–Stein rule (7.51) is

$$\hat{\gamma}^{\text{JS}} = \left[1 - \frac{p-2}{\|\hat{\gamma}\|^2}\right]\hat{\gamma}. \tag{7.63}$$

Transforming back to the β scale gives (7.43).

8

Generalized Linear Models and Regression Trees

Indirect evidence is not the sole property of Bayesians. Regression models are the frequentist method of choice for incorporating the experience of "others." As an example, Figure 8.1 returns to the kidney fitness data of Section 1.1. A potential new donor, aged 55, has appeared, and we wish to assess his kidney fitness without subjecting him to an arduous series of medical tests. Only one of the 157 previously tested volunteers was age 55, his `tot` score being -0.01 (the upper large dot in Figure 8.1). Most applied statisticians, though, would prefer to read off the height of the least squares regression line at age $= 55$ (the green dot on the regression line), $\widehat{\texttt{tot}} = -1.46$. The former is the only direct evidence we have, while the

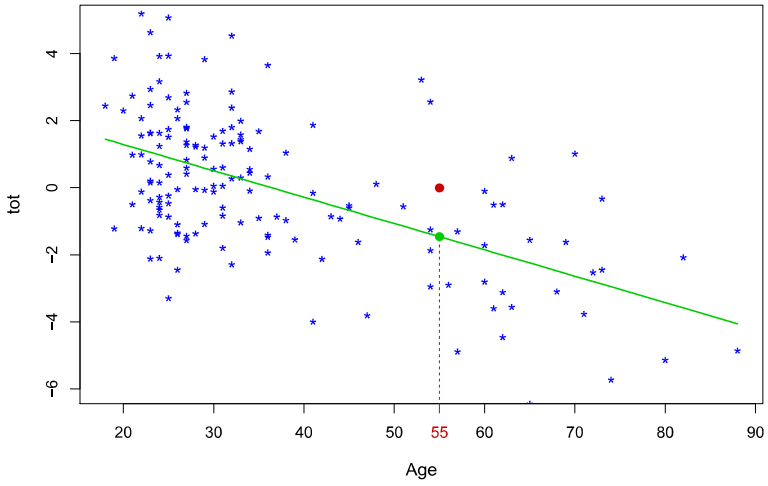

Figure 8.1 Kidney data; a new volunteer donor is aged 55. Which prediction is preferred for his kidney function?

regression line lets us incorporate indirect evidence for age 55 from all 157 previous cases.

Increasingly aggressive use of regression techniques is a hallmark of modern statistical practice, "aggressive" applying to the number and type of predictor variables, the coinage of new methodology, and the sheer size of the target data sets. Generalized linear models, this chapter's main topic, have been the most pervasively influential of the new methods. The chapter ends with a brief review of regression trees, a completely different regression methodology that will play an important role in the prediction algorithms of Chapter 17.

8.1 Logistic Regression

An experimental new anti-cancer drug called **Xilathon** is under development. Before human testing can begin, animal studies are needed to determine safe dosages. To this end, a *bioassay* or dose–response experiment was carried out: 11 groups of $n = 10$ mice each were injected with increasing amounts of **Xilathon**, dosages coded[1] $1, 2, \ldots, 11$.

Let

$$y_i = \#\text{ mice dying in } i\text{th group.} \tag{8.1}$$

The points in Figure 8.2 show the proportion of deaths

$$p_i = y_i/10, \tag{8.2}$$

lethality generally increasing with dose. The counts y_i are modeled as independent binomials,

$$y_i \overset{\text{ind}}{\sim} \text{Bi}(n_i, \pi_i) \qquad \text{for } i = 1, 2, \ldots, N, \tag{8.3}$$

$N = 11$ and all n_i equaling 10 here; π_i is the true death rate in group i, estimated unbiasedly by p_i, the direct evidence for π_i. The regression curve in Figure 8.2 uses *all* the doses to give a better picture of the true dose–response relation.

Logistic regression is a specialized technique for regression analysis of count or proportion data. The *logit* parameter λ is defined as

$$\lambda = \log\left\{\frac{\pi}{1-\pi}\right\}, \tag{8.4}$$

[1] Dose would usually be labeled on a log scale, each one, say, 50% larger than its predecessor.

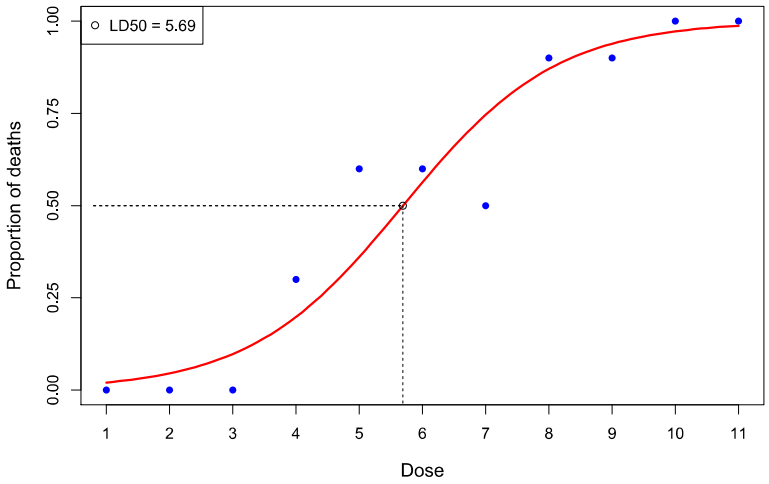

Figure 8.2 Dose–response study; groups of 10 mice exposed to increasing doses of experimental drug. The points are the observed proportions that died in each group. The fitted curve is the maximum-likelihoood estimate of the linear logistic regression model. The open circle on the curve is the LD50, the estimated dose for 50% mortality.

with λ increasing from $-\infty$ to ∞ as π increases from 0 to 1. A linear logistic regression dose–response analysis begins with binomial model (8.3), and assumes that the logit is a linear function of dose,

$$\lambda_i = \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \alpha_0 + \alpha_1 x_i. \tag{8.5}$$

Maximum likelihood gives estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$, and fitted curve

$$\hat{\lambda}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x. \tag{8.6}$$

Since the inverse transformation of (8.4) is

$$\pi = \left(1 + e^{-\lambda}\right)^{-1} \tag{8.7}$$

we obtain from (8.6) the linear logistic regression curve

$$\hat{\pi}(x) = \left(1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 x)}\right)^{-1} \tag{8.8}$$

pictured in Figure 8.2.

Table 8.1 compares the standard deviation of the estimated regression

Table 8.1 *Standard deviation estimates for $\hat{\pi}(x)$ in Figure 8.1. The first row is for the linear logistic regression fit* (8.8)*; the second row is based on the individual binomial estimates p_i.*

x	1	2	3	4	5	6	7	8	9	10	11
sd $\hat{\pi}(x)$	.015	.027	.043	.061	.071	.072	.065	.050	.032	.019	.010
sd p_i	.045	.066	.094	.126	.152	.157	.138	.106	.076	.052	.035

curve (8.8) at $x = 1, 2, \ldots, 11$ (as discussed in the next section) with the usual binomial standard deviation estimate $[p_i(1-p_i)/10]^{1/2}$ obtained by considering the 11 doses separately.[2] Regression has reduced error by better than 50%, the price being possible bias if model (8.5) goes seriously wrong.

One advantage of the logit transformation is that λ isn't restricted to the range $[0, 1]$, so model (8.5) never verges on forbidden territory. A better reason has to do with the exploitation of exponential family properties. We can rewrite the density function for $\text{Bi}(n, y)$ as

$$\binom{n}{y}\pi^y(1-\pi)^{n-y} = e^{\lambda y - n\psi(\lambda)}\binom{n}{y} \tag{8.9}$$

with λ the logit parameter (8.4) and

$$\psi(\lambda) = \log\{1 + e^{\lambda}\}; \tag{8.10}$$

(8.9) is a one-parameter exponential family[3] as described in Section 5.5, with λ the natural parameter, called α there.

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ denote the full data set, $N = 11$ in Figure 8.2. Using (8.5), (8.9), and the independence of the y_i gives the probability density of $\boldsymbol{y}$ as a function of (α_0, α_1),

$$\begin{aligned} f_{\alpha_0,\alpha_1}(\boldsymbol{y}) &= \prod_{i=1}^{N} e^{\lambda_i y_i - n_i\psi(\lambda_i)}\binom{n_i}{y_i} \\ &= e^{\alpha_0 S_0 + \alpha_1 S_1} \cdot e^{-\sum_1^N n_i\psi(\alpha_0+\alpha_1 x_i)} \cdot \prod_{i=1}^{N}\binom{n_i}{y_i}, \end{aligned} \tag{8.11}$$

[2] For the separate-dose standard error, p_i was taken equal to the fitted value from the curve in Figure 8.2.

[3] It is not necessary for $f_{\mu_0}(x)$ in (5.46) on page 64 to be a probability density function, only that it not depend on the parameter μ.

where

$$S_0 = \sum_{i=1}^{N} y_i \quad \text{and} \quad S_1 = \sum_{i=1}^{N} x_i y_i. \tag{8.12}$$

Formula (8.11) expresses $f_{\alpha_0,\alpha_1}(\boldsymbol{y})$ as the product of three factors,

$$f_{\alpha_0,\alpha_1}(\boldsymbol{y}) = g_{\alpha_0,\alpha_1}(S_0, S_1)h(\alpha_0, \alpha_1)\, j(\boldsymbol{y}), \tag{8.13}$$

only the first of which involves both the parameters and the data. This implies that (S_0, S_1) is a *sufficient statistic*:† no matter how large N might be (later we will have N in the thousands), just the two numbers (S_0, S_1) contain all of the experiment's information. Only the logistic parameterization (8.4) makes this happen.[4]

†1

A more intuitive picture of logistic regression depends on $D(p_i, \hat{\pi}_i)$, the *deviance* between an observed proportion p_i (8.2) and an estimate $\hat{\pi}_i$,

$$D\,(p_i, \hat{\pi}_i) = 2n_i \left[p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (1 - p_i) \log\left(\frac{1 - p_i}{1 - \hat{\pi}_i}\right)\right]. \tag{8.14}$$

The deviance[5] is zero if $\hat{\pi}_i = p_i$, otherwise it increases as $\hat{\pi}_i$ departs further from p_i.

The logistic regression MLE value $(\hat{\alpha}_0, \hat{\alpha}_1)$ also turns out to be the choice of (α_0, α_1) minimizing the total deviance between the N points p_i and their corresponding estimates $\hat{\pi}_i = \pi_{\hat{\alpha}_0,\hat{\alpha}_1}(x_i)$ (8.8):

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \operatorname*{arg\,min}_{(\alpha_0,\alpha_1)} \sum_{i=1}^{N} D\,(p_i, \pi_{\alpha_0,\alpha_1}(x_i))\,. \tag{8.15}$$

The solid line in Figure 8.2 is the linear logistic curve coming closest to the 11 points, when distance is measured by total deviance. In this way the 200-year-old notion of least squares is generalized to binomial regression, as discussed in the next section. A more sophisticated notion of distance between data and models is one of the accomplishments of modern statistics.

Table 8.2 reports on the data for a more structured logistic regression analysis. Human muscle cell colonies were infused with mouse nuclei in five different ratios, cultured over time periods ranging from one to five

[4] Where the name "logistic regression" comes from is explained in the endnotes, along with a description of its nonexponential family predecessor *probit analysis*.

[5] Deviance is analogous to squared error in ordinary regression theory, as discussed in what follows. It is twice the "Kullback–Leibler distance," the preferred name in the information-theory literature.

Table 8.2 *Cell infusion data; human cell colonies infused with mouse nuclei in five ratios over 1 to 5 days and observed to see whether they did or did not thrive. Green numbers are estimates $\hat{\pi}_{ij}$ from the logistic regression model. For example, 5 of 31 colonies in the lowest ratio/days category thrived, with observed proportion $5/31 = 0.16$, and logistic regression estimate $\hat{\pi}_{11} = 0.11$.*

Ratio	Time 1	2	3	4	5
1	5/31 **.11**	3/28 **.25**	20/45 **.42**	24/47 **.54**	29/35 **.75**
2	15/77 **.24**	36/78 **.45**	43/71 **.64**	56/71 **.74**	66/74 **.88**
3	48/126 **.38**	68/116 **.62**	145/171 **.77**	98/119 **.85**	114/129 **.93**
4	29/92 **.32**	35/52 **.56**	57/85 **.73**	38/50 **.81**	72/77 **.92**
5	11/53 **.18**	20/52 **.37**	20/48 **.55**	40/55 **.67**	52/61 **.84**

days, and observed to see whether they thrived. For example, of the 126 colonies having the third ratio and shortest time period, 48 thrived.

Let π_{ij} denote the true probability of thriving for ratio i during time period j, and λ_{ij} its logit $\log\{\pi_{ij}/(1-\pi_{ij})\}$. A two-way additive logistic regression was fit to the data,[6]

$$\lambda_{ij} = \mu + \alpha_i + \beta_j, \qquad i = 1, 2, \ldots, 5,\ j = 1, 2, \ldots, 5. \tag{8.16}$$

The green numbers in Table 8.2 show the maximum likelihood estimates

$$\hat{\pi}_{ij} = 1 \Big/ \left[1 + e^{-\left(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j\right)}\right]. \tag{8.17}$$

Model (8.16) has nine free parameters (taking into account the constraints $\sum \alpha_i = \sum \beta_j = 0$ necessary to avoid definitional difficulties) compared with just two in the dose–response experiment. The count can easily go much higher these days.

Table 8.3 reports on a 57-variable logistic regression applied to the **spam** data. A researcher (named George) labeled $N = 4601$ of his email mes-

[6] Using the statistical computing language R; see the endnotes.

Table 8.3 *Logistic regression analysis of the* `spam` *data, model* (8.17)*; estimated regression coefficients, standard errors, and* z = *estimate*/*se, for 57 keyword predictors. The notation* `char$` *means the relative number of times* `$` *appears, etc. The last three entries measure characteristics such as length of capital-letter strings. The word* `george` *is special, since the recipient of the email is named George, and the goal here is to build a customized spam filter.*

	Estimate	se	z-value		Estimate	se	z-value
`intercept`	−12.27	1.99	−6.16	`lab`	−1.48	.89	−1.66
`make`	−.12	.07	−1.68	`labs`	−.15	.14	−1.05
`address`	−.19	.09	−2.10	`telnet`	−.07	.19	− .35
`all`	.06	.06	1.03	`857`	.84	1.08	.78
`3d`	3.14	2.10	1.49	`data`	−.41	.17	−2.37
`our`	.38	.07	5.52	`415`	.22	.53	.42
`over`	.24	.07	3.53	`85`	−1.09	.42	−2.61
`remove`	.89	.13	6.85	`technology`	.37	.12	2.99
`internet`	.23	.07	3.39	`1999`	.02	.07	.26
`order`	.20	.08	2.58	`parts`	−.13	.09	−1.41
`mail`	.08	.05	1.75	`pm`	−.38	.17	−2.26
`receive`	−.05	.06	− .86	`direct`	−.11	.13	− .84
`will`	−.12	.06	−1.87	`cs`	−16.27	9.61	−1.69
`people`	−.02	.07	− .35	`meeting`	−2.06	.64	−3.21
`report`	.05	.05	1.06	`original`	−.28	.18	−1.55
`addresses`	.32	.19	1.70	`project`	−.98	.33	−2.97
`free`	.86	.12	7.13	`re`	−.80	.16	−5.09
`business`	.43	.10	4.26	`edu`	−1.33	.24	−5.43
`email`	.06	.06	1.03	`table`	−.18	.13	−1.40
`you`	.14	.06	2.32	`conference`	−1.15	.46	−2.49
`credit`	.53	.27	1.95	`char;`	−.31	.11	−2.92
`your`	.29	.06	4.62	`char(`	−.05	.07	− .75
`font`	.21	.17	1.24	`char_`	−.07	.09	− .78
`000`	.79	.16	4.76	`char!`	.28	.07	3.89
`money`	.19	.07	2.63	`char$`	1.31	.17	7.55
`hp`	−3.21	.52	−6.14	`char#`	1.03	.48	2.16
`hpl`	−.92	.39	−2.37	`cap.ave`	.38	.60	.64
`george`	−39.62	7.12	−5.57	`cap.long`	1.78	.49	3.62
`650`	.24	.11	2.24	`cap.tot`	.51	.14	3.75

sages as either `spam` or `ham` (nonspam[7]), say

$$y_i = \begin{cases} 1 & \text{if email } i \text{ is } \texttt{spam} \\ 0 & \text{if email } i \text{ is } \texttt{ham} \end{cases} \tag{8.18}$$

[7] "Ham" refers to "nonspam" or good email; this is a playful connection to the processed

(40% of the messages were **spam**). The $p = 57$ predictor variables represent the most frequently used words and tokens in George's corpus of email (excluding trivial words such as articles), and are in fact the relative frequencies of these chosen words in each email (standardized by the length of the email). The goal of the study was to predict whether future emails are **spam** or **ham** using these keywords; that is, to build a customized *spam filter*.

Let x_{ij} denote the relative frequency of keyword j in email i, and π_i represent the probability that email i is **spam**. Letting λ_i be the logit transform $\log\{\pi_i/(1-\pi_i)\}$, we fit the additive logistic model

$$\lambda_i = \alpha_0 + \sum_{j=1}^{57} \alpha_j x_{ij}. \tag{8.19}$$

Table 8.3 shows $\hat{\alpha}_i$ for each word—for example, −0.12 for **make**—as well as the estimated standard error and the *z-value*: estimate/se.

It looks like certain words, such as **free** and **your**, are good **spam** predictors. However, the table as a whole has an unstable appearance, with occasional very large estimates $\hat{\alpha}_i$ accompanied by very large standard deviations.[8] The dangers of high-dimensional maximum likelihood estimation are apparent here. Some sort of shrinkage estimation is called for, as discussed in Chapter 16.

Regression analysis, either in its classical form or in modern formulations, requires covariate information x to put the various cases into some sort of geometrical relationship. Given such information, regression is the statistician's most powerful tool for bringing "other" results to bear on a case of primary interest: for instance, the age-55 volunteer in Figure 8.1.

Empirical Bayes methods do not require covariate information but may be improvable if it exists. If, for example, the player's age were an important covariate in the baseball example of Table 7.1, we might first regress the MLE values on age, and then shrink them toward the regression line rather than toward the grand mean $\bar{p}$ as in (7.20). In this way, two different sorts of indirect evidence would be brought to bear on the estimation of each player's ability.

spam that was fake ham during WWII, and has been adopted by the machine-learning community.

[8] The 4601×57 $\boldsymbol{X}$ matrix (x_{ij}) was standardized, so disparate scalings are not the cause of these discrepancies. Some of the features have mostly "zero" observations, which may account for their unstable estimation.

8.2 Generalized Linear Models[9]

Logistic regression is a special case of *generalized linear models* (GLMs), a key 1970s methodology having both algorithmic and inferential influence. GLMs extend ordinary linear regression, that is least squares curve-fitting, to situations where the response variables are binomial, Poisson, gamma, beta, or in fact any exponential family form.

We begin with a one-parameter exponential family,

$$\left\{ f_\lambda(y) = e^{\lambda y - \gamma(\lambda)} f_0(y),\ \lambda \in \Lambda \right\}, \tag{8.20}$$

as in (5.46) (now with α and x replaced by λ and y, and $\psi(\alpha)$ replaced by $\gamma(\lambda)$, for clearer notation in what follows). Here λ is the *natural parameter* and y the *sufficient statistic*, both being one-dimensional in usual applications; λ takes its values in an interval of the real line. Each coordinate y_i of an observed data set $\boldsymbol{y} = (y_1, y_2, \ldots, y_i, \ldots, y_N)'$ is assumed to come from a member of family (8.20),

$$y_i \sim f_{\lambda_i}(\cdot) \text{ independently for } i = 1, 2, \ldots, N. \tag{8.21}$$

Table 8.4 lists λ and y for the first four families in Table 5.1, as well as their deviance and normalizing functions.

By itself, model (8.21) requires N parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$, usually too many for effective individual estimation. A key GLM tactic is to specify the λs in terms of a linear regression equation. Let $\boldsymbol{X}$ be an $N \times p$ "structure matrix," with ith row say x_i', and α an unknown vector of p parameters; the N-vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$ is then specified by

$$\boldsymbol{\lambda} = \boldsymbol{X}\alpha. \tag{8.22}$$

In the dose–response experiment of Figure 8.2 and model (8.5), $\boldsymbol{X}$ is $N \times 2$ with ith row $(1, x_i)$ and parameter vector $\alpha = (\alpha_0, \alpha_1)$.

The probability density function $f_\alpha(\boldsymbol{y})$ of the data vector $\boldsymbol{y}$ is

$$f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} f_{\lambda_i}(y_i) = e^{\sum_1^N (\lambda_i y_i - \gamma(\lambda_i))} \prod_{i=1}^{N} f_0(y_i), \tag{8.23}$$

which can be written as

$$f_\alpha(\boldsymbol{y}) = e^{\alpha' z - \psi(\alpha)} f_0(\boldsymbol{y}), \tag{8.24}$$

[9] Some of the more technical points raised in this section are referred to in later chapters, and can be scanned or omitted at first reading.

Table 8.4 *Exponential family form for first four cases in Table 5.1; natural parameter λ, sufficient statistic y, deviance* (8.31) *between family members f_1 and f_2, $D(f_1, f_2)$, and normalizing function $\gamma(\lambda)$.*

		λ	y	$D(f_1, f_2)$	$\gamma(\lambda)$
1.	*Normal* $\mathcal{N}(\mu, \sigma^2)$, σ^2 known	μ/σ^2	x	$\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2$	$\sigma^2\lambda^2/2$
2.	*Poisson* Poi(μ)	$\log\mu$	x	$2\mu_1\left[\left(\frac{\mu_2}{\mu_1}-1\right)-\log\frac{\mu_2}{\mu_1}\right]$	e^λ
3.	*Binomial* Bi(n, π)	$\log\frac{\pi}{1-\pi}$	x	$2n\left[\pi_1\log\frac{\pi_1}{\pi_2}+(1-\pi_1)\log\frac{1-\pi_1}{1-\pi_2}\right]$	$n\log(1+e^\lambda)$
4.	*Gamma* Gam(ν, σ), ν known	$-1/\sigma$	x	$2\nu\left[\left(\frac{\sigma_1}{\sigma_2}-1\right)-\log\frac{\sigma_1}{\sigma_2}\right]$	$-\nu\log(-\lambda)$

where

$$z = \boldsymbol{X}'\boldsymbol{y} \quad \text{and} \quad \psi(\alpha) = \sum_{i=1}^{N} \gamma(\boldsymbol{x}_i'\alpha), \tag{8.25}$$

a p-parameter exponential family (5.50), with natural parameter vector α and sufficient statistic vector z. The main point is that all the information from a p-parameter GLM is summarized in the p-dimensional vector z, no matter how large N may be, making it easier both to understand and to analyze.

We have now reduced the N-parameter model (8.20)–(8.21) to the p-parameter exponential family (8.24), with p usually much smaller than N, in this way avoiding the difficulties of high-dimensional estimation. The moments of the one-parameter constituents (8.20) determine the estimation properties in model (8.22)–(8.24). Let $(\mu_\lambda, \sigma_\lambda^2)$ denote the expectation and variance of univariate density $f_\lambda(y)$ (8.20),

$$y \sim (\mu_\lambda, \sigma_\lambda^2), \tag{8.26}$$

for instance $(\mu_\lambda, \sigma_\lambda^2) = (e^\lambda, e^\lambda)$ for the Poisson. The N-vector $\boldsymbol{y}$ obtained from GLM (8.22) then has mean vector and covariance matrix

$$\boldsymbol{y} \sim (\boldsymbol{\mu}(\alpha), \boldsymbol{\Sigma}(\alpha)), \tag{8.27}$$

where $\boldsymbol{\mu}(\alpha)$ is the vector with ith component μ_{λ_i} with $\lambda_i = x_i'\alpha$, and $\boldsymbol{\Sigma}(\alpha)$ is the $N \times N$ diagonal matrix having diagonal elements $\sigma^2_{\lambda_i}$.

The maximum likelihood estimate $\hat{\alpha}$ of the parameter vector α can be †2 shown to satisfy the simple equation†

$$\boldsymbol{X}'[\boldsymbol{y} - \boldsymbol{\mu}(\hat{\alpha})] = 0. \tag{8.28}$$

For the normal case where $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ in (8.21), that is, for ordinary linear regression, $\boldsymbol{\mu}(\hat{\alpha}) = \boldsymbol{X}\hat{\alpha}$ and (8.28) becomes $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\alpha}) = 0$, with the familiar solution

$$\hat{\alpha} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}; \tag{8.29}$$

otherwise, $\boldsymbol{\mu}(\alpha)$ is a nonlinear function of α, and (8.28) must be solved by numerical iteration. This is made easier by the fact that, for GLMs, $\log f_\alpha(\boldsymbol{y})$, the likelihood function we wish to maximize, is a *concave function of* α. The MLE $\hat{\alpha}$ has approximate expectation and covariance† †3

$$\hat{\alpha} \doteq (\alpha, \left(\boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}\right)^{-1}), \tag{8.30}$$

†4 similar to the exact OLS result $\hat{\alpha} \sim (\alpha, \sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})^{-1})$.†

Generalizing the binomial definition (8.14), the *deviance* between densities $f_1(y)$ and $f_2(y)$ is defined to be

$$D(f_1, f_2) = 2\int_{\mathcal{Y}} f_1(y) \log\left\{\frac{f_1(y)}{f_2(y)}\right\} dy, \tag{8.31}$$

the integral (or sum for discrete distributions) being over their common sample space $\mathcal{Y}$. $D(f_1, f_2)$ is always nonnegative, equaling zero only if f_1 and f_2 are the same; in general $D(f_1, f_2)$ does not equal $D(f_2, f_1)$. Deviance does not depend on how the two densities are named, for example (8.14) having the same expression as the *Binomial* entry in Table 8.4.

In what follows it will sometimes be useful to label the family (8.20) by its *expectation parameter* $\mu = E_\lambda\{y\}$ rather than by the natural parameter λ:

$$f_\mu(y) = e^{\lambda y - \gamma(\lambda)} f_0(y), \tag{8.32}$$

meaning the same thing as (8.20), only the names attached to the individual family members being changed. In this notation it is easy to show a fundamental result sometimes known as

†5 **Hoeffding's Lemma**† *The maximum likelihood estimate of μ given y is y itself, and the log likelihood* $\log f_\mu(y)$ *decreases from its maximum* $\log f_y(y)$ *by an amount that depends on the deviance* $D(y, \mu)$*,*

$$f_\mu(y) = f_y(y)e^{-D(y,\mu)/2}. \tag{8.33}$$

Returning to the GLM framework (8.21)–(8.22), parameter vector α gives $\boldsymbol{\lambda}(\alpha) = \boldsymbol{X}\alpha$, which in turn gives the vector of expectation parameters

$$\boldsymbol{\mu}(\alpha) = (\ldots \mu_i(\alpha) \ldots)', \tag{8.34}$$

for instance $\mu_i(\alpha) = \exp\{\lambda_i(\alpha)\}$ for the Poisson family. Multiplying Hoeffding's lemma (8.33) over the N cases $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ yields

$$f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} f_{\mu_i(\alpha)}(y_i) = \left[\prod_{i=1}^{N} f_{y_i}(y_i)\right] e^{-\sum_1^N D(y_i, \mu_i(\alpha))}. \tag{8.35}$$

This has an important consequence: *the MLE $\hat{\alpha}$ is the choice of α that minimizes the total deviance* $\sum_1^N D(y_i, \mu_i(\alpha))$. As in Figure 8.2, GLM maximum likelihood fitting is "least total deviance" in the same way that ordinary linear regression is least sum of squares.

The inner circle of Figure 8.3 represents normal theory, the preferred venue of classical applied statistics. Exact inferences—t-tests, F distributions, most of multivariate analysis—were feasible within the circle. Outside the circle was a general theory based mainly on asymptotic (large-sample) approximations involving Taylor expansions and the central limit theorem.

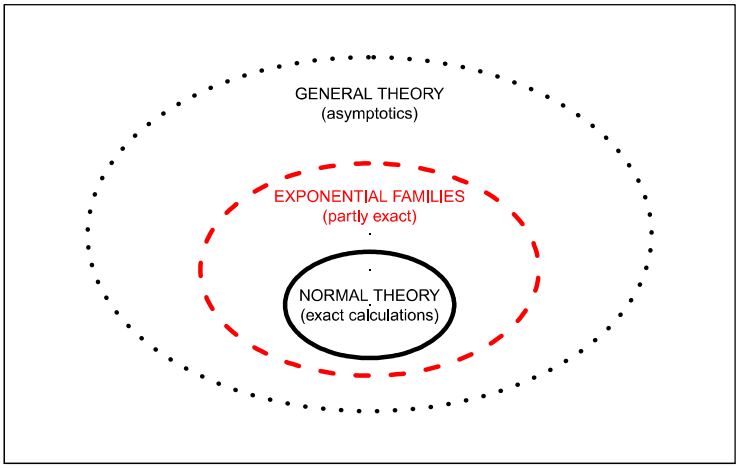

Figure 8.3 Three levels of statistical modeling.

A few useful exact results lay outside the normal theory circle, relating

to a few special families: the binomial, Poisson, gamma, beta, and others less well known. Exponential family theory, the second circle in Figure 8.3, unified the special cases into a coherent whole. It has a "partly exact" flavor, with some ideal counterparts to normal theory—convex likelihood surfaces, least deviance regression—but with some approximations necessary, as in (8.30). Even the approximations, though, are often more convincing than those of general theory, exponential families' fixed-dimension sufficient statistics making the asymptotics more transparent.

Logistic regression has banished its predecessors (such as probit analysis) almost entirely from the field, and not only because of estimating efficiencies and computational advantages (which are actually rather modest), but also because it is seen as a clearer analogue to ordinary least squares, our 200-year-old dependable standby. GLM research development has been mostly frequentist, but with a substantial admixture of likelihood-based reasoning, and a hint of Fisher's "logic of inductive inference."

Helping the statistician choose between competing methodologies is the job of statistical inference. In the case of generalized linear models the choice has been made, at least partly, in terms of aesthetics as well as philosophy.

8.3 Poisson Regression

The third most-used member of the GLM family, after normal theory least squares and logistic regression, is Poisson regression. N independent Poisson variates are observed,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i), \qquad i = 1, 2, \ldots, N, \tag{8.36}$$

where $\lambda_i = \log \mu_i$ is assumed to follow a linear model,

$$\boldsymbol{\lambda}(\alpha) = \boldsymbol{X}\alpha, \tag{8.37}$$

where $\boldsymbol{X}$ is a known $N \times p$ structure matrix and α an unknown p-vector of regression coefficients. That is, $\lambda_i = x_i'\alpha$ for $i = 1, 2, \ldots, N$, where x_i' is the ith row of $\boldsymbol{X}$.

In the chapters that follow we will see Poisson regression come to the rescue in what at first appear to be awkward data-analytic situations. Here we will settle for an example involving density estimation from a spatially truncated sample.

†6 Table 8.5 shows galaxy counts† from a small portion of the sky: 487 galaxies have had their redshifts r and apparent magnitudes m measured.

Table 8.5 *Counts for a truncated sample of 487 galaxies, binned by redshift and magnitude.*

magnitude (dimmer) ↑ \ redshift (farther) →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
18	1	6	6	3	1	4	6	8	8	20	10	7	16	9	4
17	3	2	3	4	0	5	7	6	6	7	5	7	6	8	5
16	3	2	3	3	3	2	9	9	6	3	5	4	5	2	1
15	1	1	4	3	4	3	2	3	8	9	4	3	4	1	1
14	1	3	2	3	3	4	5	7	6	7	3	4	0	0	1
13	3	2	4	5	3	6	4	3	2	2	5	1	0	0	0
12	2	0	2	4	5	4	2	3	3	0	1	2	0	0	1
11	4	1	1	4	7	3	3	1	2	0	1	1	0	0	0
10	1	0	0	2	2	2	1	2	0	0	0	1	2	0	0
9	1	1	0	2	2	2	0	0	0	0	1	0	0	0	0
8	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0
7	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
6	0	0	3	1	1	0	0	0	0	0	0	0	0	0	0
5	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Distance from earth is an increasing function of r, while apparent brightness is a decreasing function[10] of m. In this survey, counts were limited to galaxies having

$$1.22 \le r \le 3.32 \quad \text{and} \quad 17.2 \le m \le 21.5, \tag{8.38}$$

the upper limit reflecting the difficulty of measuring very dim galaxies.

The range of $\log r$ has been divided into 15 equal intervals and likewise 18 equal intervals for m. Table 8.5 gives the counts of the 487 galaxies in the $18 \times 15 = 270$ bins. (The lower right corner of the table is empty because distant galaxies always appear dim.) The multinomial/Poisson connection (5.44) helps motivate model (8.36), picturing the table as a multinomial observation on 270 categories, in which the sample size N was itself Poisson.

We can imagine Table 8.5 as a small portion of a much more extensive table, hypothetically available if the data were *not* truncated. Experience suggests that we might then fit an appropriate bivariate normal density to the data, as in Figure 5.3. It seems like it might be awkward to fit part of a bivariate normal density to truncated data, but Poisson regression offers an easy solution.

[10] An object of the second magnitude is less bright than one of the first, and so on, a classification system owing to the Greeks.

Let $\boldsymbol{r}$ be the 270-vector listing the values of r in each bin of the table (in column order), and likewise $\boldsymbol{m}$ for the 270 m values—for instance $\boldsymbol{m} = (18, 17, \ldots, 1)$ repeated 15 times—and define the 270×5 matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = [\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{r}^2, \boldsymbol{r}\boldsymbol{m}, \boldsymbol{m}^2], \tag{8.39}$$

where $\boldsymbol{r}^2$ is the vector whose components are the square of $\boldsymbol{r}$'s, etc. The log density of a bivariate normal distribution in (r, m) is of the form $\alpha_1 r + \alpha_2 m + \alpha_3 r^2 + \alpha_4 rm + \alpha_5 m^2$, agreeing with $\log \mu_i = x_i'\alpha$ as specified by (8.39). We can use a Poisson GLM, with y_i the ith bin's count, to estimate the portion of our hypothesized bivariate normal distribution in the truncation region (8.38).

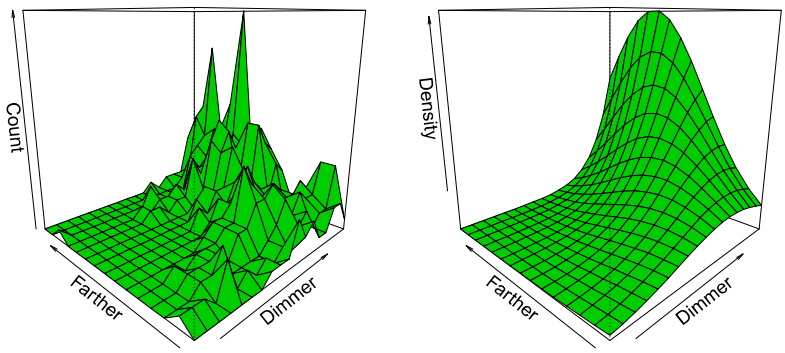

Figure 8.4 *Left* galaxy data; binned counts. *Right* Poisson GLM density estimate.

The left panel of Figure 8.4 is a perspective picture of the raw counts in Table 8.5. On the right is the fitted density from the Poisson regression. Irrespective of density estimation, Poisson regression has done a useful job of smoothing the raw bin counts.

Contours of equal value of the fitted log density

$$\hat{\alpha}_0 + \hat{\alpha}_1 r + \hat{\alpha}_2 m + \hat{\alpha}_3 r^2 + \hat{\alpha}_4 rm + \hat{\alpha}_5 m^2 \tag{8.40}$$

are shown in Figure 8.5. One can imagine the contours as truncated portions of ellipsoids, of the type shown in Figure 5.3. The right panel of Figure 8.4 makes it clear that we are nowhere near the center of the hypothetical bivariate normal density, which must lie well beyond our dimness limit.

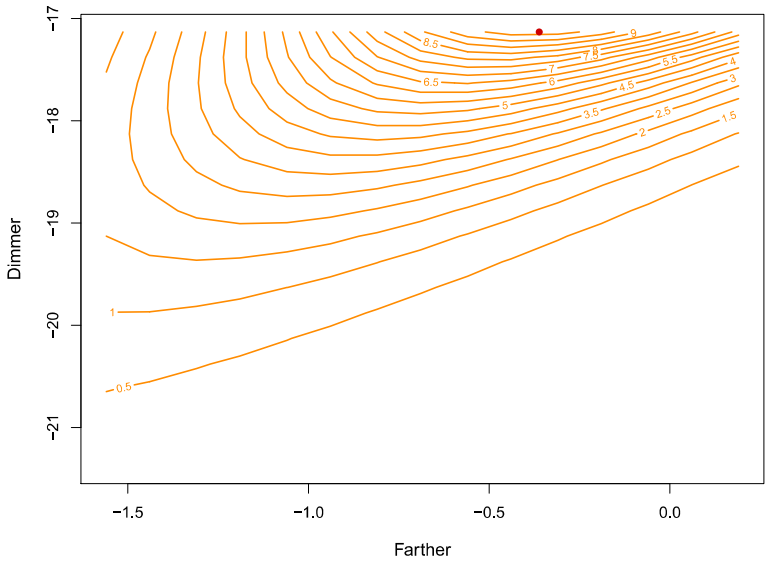

Figure 8.5 Contour curves for Poisson GLM density estimate for the galaxy data. The red dot shows the point of maximum density.

The Poisson *deviance residual* Z between an observed count y and a fitted value $\hat{\mu}$ is

$$Z = \text{sign}(y - \hat{\mu})D(y, \hat{\mu})^{1/2}, \tag{8.41}$$

with D the Poisson deviance from Table 8.4. Z_{jk}, the deviance residual between the count y_{ij} in the ijth bin of Table 8.5 and the fitted value $\hat{\mu}_{jk}$ from the Poisson GLM, was calculated for all 270 bins. Standard frequentist GLM theory says that $S = \sum_{jk} Z_{jk}^2$ should be about 270 if the bivariate normal model (8.39) is correct.[11] Actually the fit was poor: $S = 610$.

In practice we might try adding columns to $\boldsymbol{X}$ in (8.39), e.g., $\boldsymbol{rm}^2$ or $\boldsymbol{r}^2\boldsymbol{m}^2$, improving the fit where it was worst, near the boundaries of the table. Chapter 12 demonstrates some other examples of Poisson density estimation. In general, Poisson GLMs reduce density estimation to regression model fitting, a familiar and flexible inferential technology.

[11] This is a modern version of the classic chi-squared goodness-of-fit test.

8.4 Regression Trees

The data set $\boldsymbol{d}$ for a regression problem typically consists of N pairs (x_i, y_i),

$$\boldsymbol{d} = \{(x_i, y_i),\ i = 1, 2, \ldots, N\}, \tag{8.42}$$

where x_i is a vector of *predictors*, or "covariates," taking its value in some space $\mathcal{X}$, and y_i is the *response*, assumed to be univariate in what follows. The regression algorithm, perhaps a Poisson GLM, inputs $\boldsymbol{d}$ and outputs a *rule* $r_{\boldsymbol{d}}(x)$: for any value of x in $\mathcal{X}$, $r_{\boldsymbol{d}}(x)$ produces an estimate $\hat{y}$ for a possible future value of y,

$$\hat{y} = r_{\boldsymbol{d}}(x). \tag{8.43}$$

In the logistic regression example (8.8), $r_{\boldsymbol{d}}(x)$ is $\hat{\pi}(x)$.

There are three principal uses for the rule $r_{\boldsymbol{d}}(x)$.

1 For *prediction*: Given a new observation of x, but not of its corresponding y, we use $\hat{y} = r_{\boldsymbol{d}}(x)$ to predict y. In the `spam` example, the 57 keywords of an incoming message could be used to predict whether or not it is spam.[12] (See Chapter 12.)

2 For *estimation*: The rule $r_{\boldsymbol{d}}(x)$ describes a "regression surface" $\hat{S}$ over $\mathcal{X}$,

$$\hat{S} = \{r_{\boldsymbol{d}}(x),\ x \in \mathcal{X}\}. \tag{8.44}$$

The right panel of Figure 8.4 shows $\hat{S}$ for the galaxy example. $\hat{S}$ can be thought of as estimating S, the *true* regression surface, often defined in the form of conditional expectation,

$$S = \{E\{y|x\},\ x \in \mathcal{X}\}. \tag{8.45}$$

(In a dichotomous situation where y is coded as 0 or 1, $S = \{\Pr\{y = 1|x\},\ x \in \mathcal{X}\}$.)

For estimation, but not necessarily for prediction, we want $\hat{S}$ to accurately portray S. The right panel of Figure 8.4 shows the estimated galaxy density still increasing monotonically in `dimmer` at the top end of the truncation region, but not so in `farther`, perhaps an important clue for directing future search counts.[13] The flat region in the kidney function regression curve of Figure 1.2 makes almost no difference to prediction, but is of scientific interest if accurate.

[12] Prediction of dichotomous outcomes is often called "classification."

[13] Physicists call a regression-based search for new objects "bump hunting."

3 For *explanation*: The 10 predictors for the diabetes data of Section 7.3, **age**, **sex**, **bmi**,..., were selected by the researcher in the hope of explaining the etiology of diabetes progression. The relative contribution of the different predictors to $r_{\boldsymbol{d}}(x)$ is then of interest. *How* the regression surface is composed is of prime concern in this use, but not in use 1 or 2 above.

The three different uses of $r_{\boldsymbol{d}}(x)$ raise different inferential questions. Use 1 calls for estimates of prediction error. In a dichotomous situation such as the **spam** study, we would want to know both error probabilities

$$\Pr\{\hat{y} = \texttt{spam} | y = \texttt{ham}\} \quad \text{and} \quad \Pr\{\hat{y} = \texttt{ham} | y = \texttt{spam}\}. \tag{8.46}$$

For estimation, the accuracy of $r_{\boldsymbol{d}}(x)$ as a function of x, perhaps in standard deviation terms,

$$\text{sd}(x) = \text{sd}(\hat{y}|x), \tag{8.47}$$

would tell how closely $\hat{S}$ approximates S. Use 3, explanation, requires more elaborate inferential tools, saying for example which of the regression coefficients α_i in (8.19) can safely be set to zero.

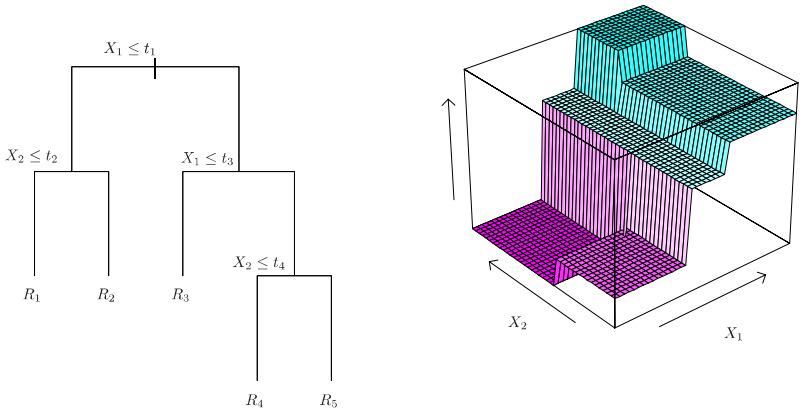

Figure 8.6 *Left* a hypothetical regression tree based on two predictors X_1 and X_2. *Right* corresponding regression surface.

Regression trees use a simple but intuitively appealing technique to form a regression surface: recursive partitioning. The left panel of Figure 8.6 illustrates the method for a hypothetical situation involving two predictor variables, X_1 and X_2 (e.g., r and m in the galaxy example). At the top of

the tree, the sample population of N cases has been split into two groups: those with X_1 equal to or less than value t_1 go to the left, those with $X_1 > t_1$ to the right. The leftward group is itself then divided into two groups depending on whether or not $X_2 \le t_2$. The division stops there, leaving two *terminal nodes* R_1 and R_2. On the tree's right side, two other splits give terminal nodes R_3, R_4, and R_5.

A prediction value $\hat{y}_{R_j}$ is attached to each terminal node R_j. The prediction $\hat{y}$ applying to a new observation $x = (x_1, x_2)$ is calculated by starting x at the top of the tree and following the splits downward until a terminal node, and its attached prediction $\hat{y}_{R_j}$, is reached. The corresponding regression surface $\hat{S}$ is shown in the right panel of Figure 8.6 (here the $\hat{y}_{R_j}$ happen to be in ascending order).

Various algorithmic rules are used to decide which variable to split and which splitting value t to take at each step of the tree's construction. Here is the most common method: suppose at step k of the algorithm, group_k of N_k cases remains to be split, those cases having mean and sum of squares

$$m_k = \sum_{i \in \text{group}_k} y_i/N_k \quad \text{and} \quad s_k^2 = \sum_{i \in \text{group}_k} (y_i - m_k)^2. \tag{8.48}$$

Dividing group_k into $\text{group}_{k,\text{left}}$ and $\text{group}_{k,\text{right}}$ produces means $m_{k,\text{left}}$ and $m_{k,\text{right}}$, and corresponding sums of squares $s^2_{k,\text{left}}$ and $s^2_{k,\text{right}}$. The algorithm proceeds by choosing the splitting variable X_k and the threshold t_k to minimize

$$s^2_{k,\text{left}} + s^2_{k,\text{right}}. \tag{8.49}$$

In other words, it splits group_k into two groups that are as different from each other as possible.† †7

Cross-validation estimates of prediction error, Chapter 12, are used to decide when the splitting process should stop. If group_k is not to be further divided, it becomes terminal node R_k, with prediction value $\hat{y}_{R_k} = m_k$. None of this would be feasible without electronic computation, but even quite large prediction problems can be short work for modern computers.

Figure 8.7 shows a regression tree analysis[14] of the **spam** data, Table 8.3. There are seven terminal nodes, labeled 0 or 1 for decision **ham** or **spam**. The leftmost node, say R_1, is a 0, and contains 2462 **ham** cases and 275 **spam** (compared with 2788 and 1813 in the full data set). Starting at the top of the tree, R_1 is reached if it has a low proportion of \$ symbols

[14] Using the R program **rpart**, in classification mode, employing a different splitting rule than the version based on (8.49).

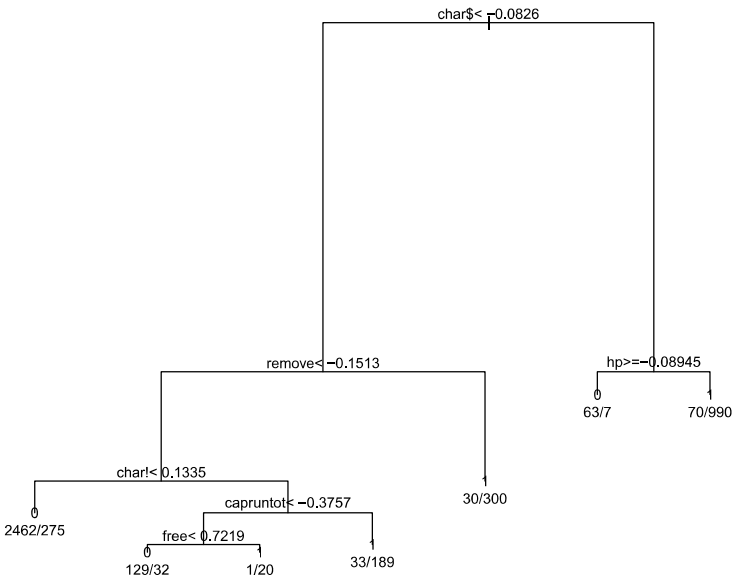

Figure 8.7 Regression tree on the `spam` data; 0 = `ham`, 1 = `spam`. Error rates: `ham` 5.2%, `spam` 17.4%. Captions indicate leftward (ham) moves.

`char$`, a low proportion of the word `remove`, and a low proportion of exclamation marks `char!`.

Regression trees are easy to interpret ("Too many dollar signs means spam!") seemingly suiting them for use 3, explanation. Unfortunately, they are also easy to overinterpret, with a reputation for being unstable in practice. Discontinuous regression surfaces $\hat{S}$, as in Figure 8.6, disqualify them for use 2, estimation. Their principal use in what follows will be as key parts of prediction algorithms, use 1. The tree in Figure 8.6 has apparent error rates (8.46) of 5.2% and 17.4%. This can be much improved upon by "bagging" (bootstrap aggregation), Chapters 17 and 20, and by other computer-intensive techniques.

Compared with generalized linear models, regression trees represent a break from classical methodology that is more stark. First of all, they are totally nonparametric; bigger but less structured data sets have promoted nonparametrics in twenty-first-century statistics. Regression trees are more computer-intensive and less efficient than GLMs but, as will be seen in Part III, the availability of massive data sets and modern computational equip-

ment has diminished the appeal of efficiency in favor of easy assumption-free application.

8.5 Notes and Details

Computer-age algorithms depend for their utility on statistical computing languages. After a period of evolution, the language **S** (Becker *et al.*, 1988) and its open-source successor **R** (R Core Team, 2015), have come to dominate applied practice.[15] Generalized linear models are available from a single **R** command, e.g.,

```
glm(y~X,family=binomial)
```

for logistic regression (Chambers and Hastie, 1993), and similarly for regression trees and hundreds of other applications.

The classic version of bioassay, *probit analysis*, assumes that each test animal has its own lethal dose level X, and that the population distribution of X is normal,

$$\Pr\{X \le x\} = \Phi(\alpha_0 + \alpha_1 x) \tag{8.50}$$

for unknown parameters (α_0, α_1) and standard normal cdf Φ. Then the number of animals dying at dose x is binomial $\mathrm{Bi}(n_x, \pi_x)$ as in (8.3), with $\pi_x = \Phi(\alpha_0 + \alpha_1 x)$, or

$$\Phi^{-1}(\pi_x) = \alpha_0 + \alpha_1 x. \tag{8.51}$$

Replacing the standard normal cdf $\Phi(z)$ with the logistic cdf $1/(1 + e^{-z})$ (which resembles Φ), changes (8.51) into logistic regression (8.5). The usual goal of bioassay was to estimate "LD50," the dose lethal to 50% of the test population; it is indicated by the open circle in Figure 8.2.

Cox (1970), the classic text on logistic regression, lists Berkson (1944) as an early practitioner. Wedderburn (1974) is credited with generalized linear models in McCullagh and Nelder's influential text of that name, first edition 1983; Birch (1964) developed an important and suggestive special case of GLM theory.

The twenty-first century has seen an efflorescence of computer-based regression techniques, as described extensively in Hastie *et al.* (2009). The discussion of regression trees here is taken from their Section 9.2, including our Figure 8.6. They use the **spam** data as a central example; it is publicly

[15] Previous computer packages such as SAS and SPSS continue to play a major role in application areas such as the social sciences, biomedical statistics, and the pharmaceutical industry.

available at `ftp.ics.uci.edu`. Breiman *et al.* (1984) propelled regression trees into wide use with their CART algorithm.

†1 [p. 112] *Sufficiency as in* (8.13). The Fisher–Neyman criterion says that if $f_\alpha(\boldsymbol{x}) = h_\alpha(S(\boldsymbol{x}))g(\boldsymbol{x})$, when $g(\cdot)$ does not depend on α, then $S(\boldsymbol{x})$ is sufficient for α.

†2 [p. 118] *Equation* (8.28). From (8.24)–(8.25) we have the log likelihood function

$$l_\alpha(\boldsymbol{y}) = \alpha' z - \psi(\alpha) \tag{8.52}$$

with sufficient statistic $z = \boldsymbol{X}'\boldsymbol{y}$ and $\psi(\alpha) = \sum_{i=1}^{N} \gamma(x_i'\alpha)$. Differentiating with respect to α,

$$\dot{l}_\alpha(\boldsymbol{y}) = z - \dot{\psi}(\alpha) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\mu}(\alpha), \tag{8.53}$$

where we have used $d\gamma/d\lambda = \mu_\lambda$ (5.55), so $\dot{\gamma}(x_i'\alpha) = x_i'\mu_i(\alpha)$. But (8.53) says $\dot{l}_\alpha(\boldsymbol{y}) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}(\alpha))$, verifying the MLE equation (8.28).

†3 [p. 118] *Concavity of the log likelihood.* From (8.53), the second derivative matrix $\ddot{l}_\alpha(\boldsymbol{y})$ with respect to α is

$$-\ddot{\psi}(\alpha) = -\operatorname{cov}_\alpha(z), \tag{8.54}$$

(5.57)–(5.59). But $z = \boldsymbol{X}'\boldsymbol{y}$ has

$$\operatorname{cov}_\alpha(z) = \boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}, \tag{8.55}$$

a positive definite $p \times p$ matrix, verifying the concavity of $l_\alpha(\boldsymbol{y})$ (which in fact applies to any exponential family, not only GLMs).

†4 [p. 118] *Formula* (8.30). The sufficient statistic z has mean vector and covariance matrix

$$z \sim (\beta, V_\alpha), \tag{8.56}$$

with $\beta = E_\alpha\{z\}$ (5.58) and $V_\alpha = \boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}$ (8.55). Using (5.60), the first-order Taylor series for $\hat{\alpha}$ as a function of z is

$$\hat{\alpha} \doteq \alpha + V_\alpha^{-1}(z - \beta). \tag{8.57}$$

Taken literally, (8.57) gives (8.30). In the OLS formula, we have σ^{-2} rather than σ^2 since the natural parameter α for the Normal entry in Table 8.4 is μ/σ^2.

†5 [p. 118] *Formula* (8.33). This formula, attributed to Hoeffding (1965), is a key result in the interpretation of GLM fitting. Applying definition (8.31)

to family (8.32) gives

$$\begin{aligned}\frac{1}{2}D(\lambda_1,\lambda_2) &= E_{\lambda_1}\{(\lambda_1-\lambda_2)y-[\gamma(\lambda_1)-\gamma(\lambda_2)]\}\\ &= (\lambda_1-\lambda_2)\mu_1-[\gamma(\lambda_1)-\gamma(\lambda_2)]\,.\end{aligned}\tag{8.58}$$

If λ_1 is the MLE $\hat{\lambda}$ then $\mu_1 = y$ (from the maximum likelihood equation $0 = d[\log f_\lambda(y)]/d\lambda = y - \dot{\gamma}(\lambda) = y - \mu_\lambda$), giving[16]

$$\frac{1}{2}D\left(\hat{\lambda},\lambda\right) = \left(\hat{\lambda}-\lambda\right)y - \left[\gamma\left(\hat{\lambda}\right)-\gamma(\lambda)\right] \tag{8.59}$$

for any choice of λ. But the right-hand side of (8.59) is $-\log[f_\lambda(y)/f_y(y)]$, verifying (8.33).

†6 [p. 120] *Table 8.5.* The galaxy counts are from Loh and Spillar's 1988 redshift survey, as discussed in Efron and Petrosian (1992).

†7 [p. 126] *Criteria* (8.49). Abbreviating "left" and "right" by l and r, we have

$$s_k^2 = s_{kl}^2 + s_{kr}^2 + \frac{N_{kl}N_{kr}}{N_k}(m_{kl}-m_{kr})^2, \tag{8.60}$$

with N_{kl} and N_{kr} the subgroup sizes, showing that minimizing (8.49) is the same as maximizing the last term in (8.60). Intuitively, a *good* split is one that makes the left and right groups as different as possible, the ideal being all 0s on the left and all 1s on the right, making the terminal nodes "pure."

[16] In some cases $\hat{\lambda}$ is undefined; for example, when $y = 0$ for a Poisson response, $\hat{\lambda} = \log(y)$ which is undefined. But, in (8.59), we assume that $\hat{\lambda}y = 0$. Similarly for binary y and the binomial family.

9

Survival Analysis and the EM Algorithm

Survival analysis had its roots in governmental and actuarial statistics, spanning centuries of use in assessing life expectancies, insurance rates, and annuities. In the 20 years between 1955 and 1975, survival analysis was adapted by statisticians for application to biomedical studies. Three of the most popular post-war statistical methodologies emerged during this period: the Kaplan–Meier estimate, the log-rank test,[1] and Cox's proportional hazards model, the succession showing increased computational demands along with increasingly sophisticated inferential justification. A connection with one of Fisher's ideas on maximum likelihood estimation leads in the last section of this chapter to another statistical method that has "gone platinum," the EM algorithm.

9.1 Life Tables and Hazard Rates

An insurance company's *life table* appears in Table 9.1, showing its number of clients (that is, life insurance policy holders) by age, and the number of deaths during the past year in each age group,[2] for example five deaths among the 312 clients aged 59. The column labeled $\hat{S}$ is of great interest to the company's actuaries, who have to set rates for new policy holders. It is an estimate of survival probability: probability 0.893 of a person aged 30 (the beginning of the table) surviving past age 59, etc. $\hat{S}$ is calculated according to an ancient but ingenious algorithm.

Let X represent a typical lifetime, so

$$f_i = \Pr\{X = i\} \tag{9.1}$$

[1] Also known as the Mantel–Haenszel or Cochran–Mantel–Haenszel test.

[2] The insurance company is fictitious but the deaths y are based on the true 2010 rates for US men, per Social Security Administration data.

Table 9.1 *Insurance company life table; at each age, n = number of policy holders, y = number of deaths, $\hat{h}$ = hazard rate y/n, $\hat{S}$ = survival probability estimate* (9.6).

Age	n	y	$\hat{h}$	$\hat{S}$	Age	n	y	$\hat{h}$	$\hat{S}$
30	116	0	.000	1.000	60	231	1	.004	.889
31	44	0	.000	1.000	61	245	5	.020	.871
32	95	0	.000	1.000	62	196	5	.026	.849
33	97	0	.000	1.000	63	180	4	.022	.830
34	120	0	.000	1.000	64	170	2	.012	.820
35	71	1	.014	.986	65	114	0	.000	.820
36	125	0	.000	.986	66	185	5	.027	.798
37	122	0	.000	.986	67	127	2	.016	.785
38	82	0	.000	.986	68	127	5	.039	.755
39	113	0	.000	.986	69	158	2	.013	.745
40	79	0	.000	.986	70	100	3	.030	.723
41	90	0	.000	.986	71	155	4	.026	.704
42	154	0	.000	.986	72	92	1	.011	.696
43	103	0	.000	.986	73	90	1	.011	.689
44	144	0	.000	.986	74	110	2	.018	.676
45	192	2	.010	.976	75	122	5	.041	.648
46	153	1	.007	.969	76	138	8	.058	.611
47	179	1	.006	.964	77	46	0	.000	.611
48	210	0	.000	.964	78	75	4	.053	.578
49	259	2	.008	.956	79	69	6	.087	.528
50	225	2	.009	.948	80	95	4	.042	.506
51	346	1	.003	.945	81	124	6	.048	.481
52	370	2	.005	.940	82	67	7	.104	.431
53	568	4	.007	.933	83	112	12	.107	.385
54	1081	8	.007	.927	84	113	8	.071	.358
55	1042	2	.002	.925	85	116	12	.103	.321
56	1094	10	.009	.916	86	124	17	.137	.277
57	597	4	.007	.910	87	110	21	.191	.224
58	359	1	.003	.908	88	63	9	.143	.192
59	312	5	.016	.893	89	79	10	.127	.168

is the probability of dying at age i, and

$$S_i = \sum_{j \geq i} f_j = \Pr\{X \geq i\} \tag{9.2}$$

is the probability of surviving past age $i - 1$. The *hazard rate* at age i is by

definition

$$h_i = f_i/S_i = \Pr\{X = i | X \geq i\}, \tag{9.3}$$

the probability of dying at age i given survival past age $i-1$.

A crucial observation is that the probability S_{ij} of surviving past age j given survival past age $i-1$ is the product of surviving each intermediate year,

$$S_{ij} = \prod_{k=i}^{j}(1-h_k) = \Pr\{X > j | X \geq i\}; \tag{9.4}$$

first you have to survive year i, probability $1-h_i$; then year $i+1$, probability $1-h_{i+1}$, etc., up to year j, probability $1-h_j$. Notice that S_i (9.2) equals $S_{1,i-1}$.

$\hat{S}$ in Table 9.1 is an estimate of S_{ij} for $i = 30$. First, each h_i was estimated as the binomial proportion of the number of deaths y_i among the n_i clients,

$$\hat{h}_i = y_i/n_i, \tag{9.5}$$

and then we set

$$\hat{S}_{30,j} = \prod_{k=30}^{j}\left(1-\hat{h}_k\right). \tag{9.6}$$

The insurance company doesn't have to wait 50 years to learn the probability of a 30-year-old living past 80 (estimated to be 0.506 in the table). One year's data suffices.[3]

Hazard rates are more often described in terms of a *continuous* positive random variable T (often called "time"), having density function $f(t)$ and "reverse cdf," or survival function,

$$S(t) = \int_t^\infty f(x)\,dx = \Pr\{T \geq t\}. \tag{9.7}$$

The hazard rate

$$h(t) = f(t)/S(t) \tag{9.8}$$

satisfies

$$h(t)dt \doteq \Pr\{T \in (t, t+dt) | T \geq t\} \tag{9.9}$$

for $dt \to 0$, in analogy with (9.3). The analog of (9.4) is† †1

[3] Of course the estimates can go badly wrong if the hazard rates change over time.

$$\Pr\{T \geq t_1 | T \geq t_0\} = \exp\left\{-\int_{t_0}^{t_1} h(x)\, dx\right\} \tag{9.10}$$

so in particular the reverse cdf (9.7) is given by

$$S(t) = \exp\left\{-\int_0^t h(x)\, dx\right\}. \tag{9.11}$$

A one-sided exponential density

$$f(t) = (1/c)e^{-t/c} \qquad \text{for } t \geq 0 \tag{9.12}$$

has $S(t) = \exp\{-t/c\}$ and constant hazard rate

$$h(t) = 1/c. \tag{9.13}$$

The name "memoryless" is quite appropriate for density (9.12): having survived to any time t, the probability of surviving dt units more is always the same, about $1 - dt/c$, no matter what t is. If human lifetimes were exponential there wouldn't be old or young people, only lucky or unlucky ones.

9.2 Censored Data and the Kaplan–Meier Estimate

Table 9.2 reports the survival data from a randomized clinical trial run by NCOG (the Northern California Oncology Group) comparing two treatments for head and neck cancer: Arm_A, chemotherapy, versus Arm_B, chemotherapy plus radiation. The response for each patient is survival time in days. The + sign following some entries indicates *censored data*, that is, survival times known only to exceed the reported value. These are patients "lost to followup," mostly because the NCOG experiment ended with some of the patients still alive.

This is what the experimenters hoped to see of course, but it complicates the comparison. Notice that there is more censoring in Arm_B. In the absence of censoring we could run a simple two-sample test, maybe Wilcoxon's test, to see whether the more aggressive treatment of Arm_B was increasing the survival times. *Kaplan–Meier* curves provide a graphical comparison that takes proper account of censoring. (The next section describes an appropriate censored data two-sample test.) Kaplan–Meier curves have become familiar friends to medical researchers, a *lingua franca* for reporting clinical trial results.

Life table methods are appropriate for censored data. Table 9.3 puts the Arm_A results into the same form as the insurance study of Table 9.1, now

Table 9.2 *Censored survival times in days, from two arms of the* NCOG *study of head/neck cancer.*

Arm_A: Chemotherapy

7	34	42	63	64	74+	83	84	91
108	112	129	133	133	139	140	140	146
149	154	157	160	160	165	173	176	185+
218	225	241	248	273	277	279+	297	319+
405	417	420	440	523	523+	583	594	1101
1116+	1146	1226+	1349+	1412+	1417			

Arm_B: Chemotherapy+Radiation

37	84	92	94	110	112	119	127	130
133	140	146	155	159	169+	173	179	194
195	209	249	281	319	339	432	469	519
528+	547+	613+	633	725	759+	817	1092+	1245+
1331+	1557	1642+	1771+	1776	1897+	2023+	2146+	2297+

with the time unit being months. Of the 51 patients enrolled[4] in Arm_A, $y_1 = 1$ was observed to die in the first month after treatment; this left 50 at risk, $y_2 = 2$ of whom died in the second month; $y_3 = 5$ of the remaining 48 died in their third month after treatment, and one was lost to followup, this being noted in the l column of the table, leaving $n_4 = 40$ patients "at risk" at the beginning of month 5, etc.

$\hat{S}$ here is calculated as in (9.6) except starting at time 1 instead of 30. There is nothing wrong with this estimate, but binning the NCOG survival data by months is arbitrary. Why not go down to days, as the data was originally presented in Table 9.2? A Kaplan–Meier survival curve is the limit of life table survival estimates as the time unit goes to zero.

Observations z_i for censored data problems are of the form

$$z_i = (t_i, d_i), \tag{9.14}$$

where t_i equals the observed survival time while d_i indicates whether or not there was censoring,

$$d_i = \begin{cases} 1 & \text{if death observed} \\ 0 & \text{if death not observed} \end{cases} \tag{9.15}$$

[4] The patients were enrolled at different calendar times, as they entered the study, but for each patient "time zero" in the table is set at the beginning of his or her treatment.

Table 9.3 `Arm_A` *of the* `NCOG` *head/neck cancer study, binned by month; n = number at risk, y = number of deaths, l = lost to followup, h = hazard rate y/n; $\hat{S}$ = life table survival estimate.*

Month	n	y	l	h	$\hat{S}$	Month	n	y	l	h	$\hat{S}$
1	51	1	0	.020	.980	25	7	0	0	.000	.184
2	50	2	0	.040	.941	26	7	0	0	.000	.184
3	48	5	1	.104	.843	27	7	0	0	.000	.184
4	42	2	0	.048	.803	28	7	0	0	.000	.184
5	40	8	0	.200	.642	29	7	0	0	.000	.184
6	32	7	0	.219	.502	30	7	0	0	.000	.184
7	25	0	1	.000	.502	31	7	0	0	.000	.184
8	24	3	0	.125	.439	32	7	0	0	.000	.184
9	21	2	0	.095	.397	33	7	0	0	.000	.184
10	19	2	1	.105	.355	34	7	0	0	.000	.184
11	16	0	1	.000	.355	35	7	0	0	.000	.184
12	15	0	0	.000	.355	36	7	0	0	.000	.184
13	15	0	0	.000	.355	37	7	1	1	.143	.158
14	15	3	0	.200	.284	38	5	1	0	.200	.126
15	12	1	0	.083	.261	39	4	0	0	.000	.126
16	11	0	0	.000	.261	40	4	0	0	.000	.126
17	11	0	0	.000	.261	41	4	0	1	.000	.126
18	11	1	1	.091	.237	42	3	0	0	.000	.126
19	9	0	0	.000	.237	43	3	0	0	.000	.126
20	9	2	0	.222	.184	44	3	0	0	.000	.126
21	7	0	0	.000	.184	45	3	0	1	.000	.126
22	7	0	0	.000	.184	46	2	0	0	.000	.126
23	7	0	0	.000	.184	47	2	1	1	.500	.063
24	7	0	0	.000	.184						

(so $d_i = 0$ corresponds to a + in Table 9.2). Let

$$t_{(1)} < t_{(2)} < t_{(3)} < \ldots < t_{(n)} \tag{9.16}$$

denote the *ordered* survival times,[5] censored or not, with corresponding indicator $d_{(k)}$ for $t_{(k)}$. The *Kaplan–Meier estimate* for survival probability †2 $S_{(j)} = \Pr\{X > t_{(j)}\}$ is then† the life table estimate

$$\hat{S}_{(j)} = \prod_{k \le j} \left(\frac{n-k}{n-k+1} \right)^{d_{(k)}}. \tag{9.17}$$

[5] Assuming no ties among the survival times, which is convenient but not crucial for what follows.

$\hat{S}$ jumps downward at death times t_j, and is constant between observed deaths.

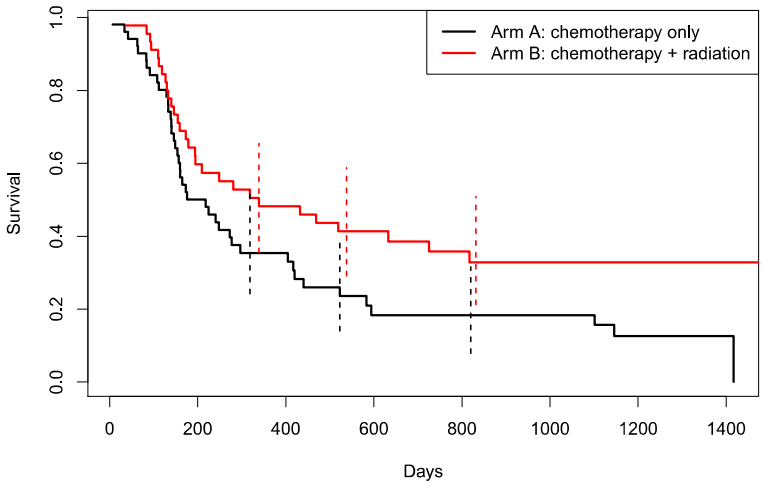

Figure 9.1 `NCOG` Kaplan–Meier survival curves; lower `Arm_A` (chemotherapy only); upper `Arm_B` (chemotherapy+radiation). Vertical lines indicate approximate 95% confidence intervals.

The Kaplan–Meier curves for both arms of the `NCOG` study are shown in Figure 9.1. `Arm_B`, the more aggressive treatment, looks better: its 50% survival estimate occurs at 324 days, compared with 182 days for `Arm_A`. The answer to the inferential question—is `B` really better than `A` or is this just random variability?—is less clear-cut.

The accuracy of $\hat{S}_{(j)}$ can be estimated from Greenwood's formula† for its standard deviation (now back in life table notation), †3

$$\text{sd}\left(\hat{S}_{(j)}\right) = \hat{S}_{(j)}\left[\sum_{k\le j}\frac{y_k}{n_k(n_k-y_k)}\right]^{1/2}. \tag{9.18}$$

The vertical bars in Figure 9.1 are approximate 95% confidence limits for the two curves based on Greenwood's formula. They overlap enough to cast doubt on the superiority of `Arm_B` at any one choice of "days," but the two-sample test of the next section, which compares survival at all timepoints, will provide more definitive evidence.

Life tables and the Kaplan–Meier estimate seem like a textbook example of frequentist inference as described in Chapter 2: a useful probabilistic

result is derived (9.4), and then implemented by the plug-in principle (9.6). There is more to the story though, as discussed below.

Life table curves are nonparametric, in the sense that no particular relationship is assumed between the hazard rates h_i. A parametric approach can greatly improve the curves' accuracy.† Reverting to the life table form of Table 9.3, we assume that the death counts y_k are independent binomials,

†4

$$y_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k), \tag{9.19}$$

and that the logits $\lambda_k = \log\{h_k/(1-h_k)\}$ satisfy some sort of regression equation

$$\boldsymbol{\lambda} = \boldsymbol{X}\alpha, \tag{9.20}$$

as in (8.22). A cubic regression for instance would set $x_k = (1, k, k^2, k^3)'$ for the kth row of $\boldsymbol{X}$, with $\boldsymbol{X}$ 47×4 for Table 9.3.

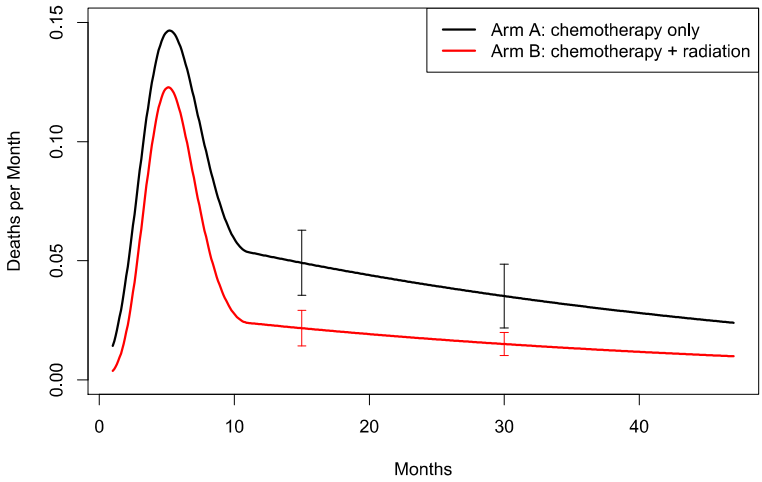

Figure 9.2 Parametric hazard rate estimates for the `NCOG` study. `Arm_A`, black curve, has about 2.5 times higher hazard than `Arm_B` for all times more than a year after treatment. Standard errors shown at 15 and 30 months.

The parametric hazard-rate estimates in Figure 9.2 were instead based on a "cubic-linear spline,"

$$x_k = \left(1, k, (k-11)_+^2, (k-11)_+^3\right)', \tag{9.21}$$

where $(k-11)_+$ equals $k-11$ for $k \le 11$, and 0 for $k \ge 11$. The vector

$\boldsymbol{\lambda} = \boldsymbol{X}\alpha$ describes a curve that is cubic for $k \leq 11$, linear for $k \geq 11$, and joined smoothly at 11. The logistic regression maximum likelihood estimate $\hat{\alpha}$ produced hazard rate curves

$$\hat{h}_k = 1 \Big/ \left(1 + e^{-x_k'\hat{\alpha}}\right) \tag{9.22}$$

as in (8.8). The black curve in Figure 9.2 traces $\hat{h}_k$ for **Arm_A**, while the red curve is that for **Arm_B**, fit separately.

Comparison in terms of hazard rates is more informative than the survival curves of Figure 9.1. Both arms show high initial hazards, peaking at five months, and then a long slow decline.[6] **Arm_B** hazard is always below **Arm_A**, in a ratio of about 2.5 to 1 after the first year. Approximate 95% confidence limits, obtained as in (8.30), don't overlap, indicating superiority of **Arm_B** at 15 and 30 months after treatment.

In addition to its frequentist justification, survival analysis takes us into the Fisherian realm of conditional inference, Section 4.3. The y_k's in model (9.19) are considered *conditionally* on the n_k's, effectively treating the n_k values in Table 9.3 as *ancillaries*, that is as fixed constants, by themselves containing no statistical information about the unknown hazard rates. We will examine this tactic more carefully in the next two sections.

9.3 The Log-Rank Test

A randomized clinical trial, interpreted by a two-sample test, remains the gold standard of medical experimentation. Interpretation usually involves Student's two-sample t-test or its nonparametric cousin Wilcoxon's test, but neither of these is suitable for censored data. The *log-rank test* † †5 employs an ingenious extension of life tables for the nonparametric two-sample comparison of censored survival data.

Table 9.4 compares the results of the **NCOG** study for the first six months[7] after treatment. At the beginning[8] of month 1 there were 45 patients "at risk" in **Arm_B**, none of whom died, compared with 51 at risk and 1 death in **Arm_A**. This left 45 at risk in **Arm_B** at the beginning of month 2, and 50 in **Arm_A**, with 1 and 2 deaths during the month respectively. (Losses

[6] The cubic–linear spline (9.21) is designed to show more detail in the early months, where there is more available patient data and where hazard rates usually change more quickly.

[7] A month is defined here as 365/12=30.4 days.

[8] The "beginning of month 1" is each patient's initial treatment time, at which all 45 patients ever enrolled in **Arm_B** were at risk, that is, available for observation.

Table 9.4 *Life table comparison for the first six months of the* NCOG *study. For example, at the beginning of the sixth month after treatment, there were 33 remaining* Arm_B *patients, of whom 4 died during the month, compared with 32 at risk and 7 dying in* Arm_A*. The conditional expected number of deaths in* Arm_A*, assuming the null hypothesis of equal hazard rates in both arms, was 5.42, using expression* (9.24).

Month	Arm_B		Arm_A		Expected number Arm_A deaths
	At risk	Died	At risk	Died	
1	45	0	51	1	.53
2	45	1	50	2	1.56
3	44	1	48	5	3.13
4	43	5	42	2	3.46
5	38	5	40	8	6.67
6	33	4	32	7	5.42

to followup were assumed to occur at the *end* of each month; there was 1 such at the end of month 3, reducing the number at risk in Arm_A to 42 for month 4.)

The month 6 data is displayed in two-by-two tabular form in Table 9.5, showing the notation used in what follows: n_A for the number at risk in Arm_A, n_d for the number of deaths, etc.; y indicates the number of Arm_A deaths. If the marginal totals n_A, n_B, n_d, and n_s are given, then y determines the other three table entries by subtraction, so we are not losing any information by focusing on y.

Table 9.5 *Two-by-two display of month-6 data for the* NCOG *study. E is the expected number of* Arm_A *deaths assuming the null hypothesis of equal hazard rates (last column of Table 9.4).*

	Died	Survived	
Arm_A	$y = 7$ $E = 5.42$	25	$n_A = 32$
Arm_B	4	29	$n_B = 33$
	$n_d = 11$	$n_s = 54$	$n = 65$

Consider the null hypothesis that the hazard rates (9.3) for month 6 are

the same in **Arm_A** and **Arm_B**,

$$H_0(6) : h_{A6} = h_{B6}. \tag{9.23}$$

Under $H_0(6)$, y has mean E and variance V,

$$\begin{aligned} E &= n_A n_d / n \\ V &= n_A n_B n_d n_s \big/ \left[n^2(n-1)\right], \end{aligned} \tag{9.24}$$

as calculated according to the *hypergeometric distribution*.† $E = 5.42$ and †6
$V = 2.28$ in Table 9.5.

We can form a two-by-two table for each of the $N = 47$ months of the **NCOG** study, calculating y_i, E_i, and V_i for month i. The log-rank statistic Z is then defined to be† †7

$$Z = \sum_{i=1}^{N} (y_i - E_i) \Bigg/ \left(\sum_{i=1}^{N} V_i\right)^{1/2}. \tag{9.25}$$

The idea here is simple but clever. Each month we test the null hypothesis of equal hazard rates

$$H_0(i) : h_{Ai} = h_{Bi}. \tag{9.26}$$

The numerator $y_i - E_i$ has expectation 0 under $H_0(i)$, but, if h_{Ai} is greater than h_{Bi}, that is, if treatment B is superior, then the numerator has a positive expectation. Adding up the numerators gives us power to detect a general superiority of treatment B over A, against the null hypothesis of equal hazard rates, $h_{Ai} = h_{Bi}$ for all i.

For the **NCOG** study, binned by months,

$$\sum_{i=1}^{N} y_i = 42, \quad \sum_{i=1}^{N} E_i = 32.9, \quad \sum_{i=1}^{N} V_i = 16.0, \tag{9.27}$$

giving log-rank test statistic

$$Z = 2.27. \tag{9.28}$$

Asymptotic calculations based on the central limit theorem suggest

$$Z \doteq \mathcal{N}(0, 1) \tag{9.29}$$

under the null hypothesis that the two treatments are equally effective, i.e., that $h_{Ai} = h_{Bi}$ for $i = 1, 2, \ldots, N$. In the usual interpretation, $Z = 2.27$ is significant at the one-sided 0.012 level, providing moderately strong evidence in favor of treatment B.

An impressive amount of inferential guile goes into the log-rank test.

1 Working with hazard rates instead of densities or cdfs is essential for survival data.
2 Conditioning at each period on the numbers at risk, n_A and n_B in Table 9.5, finesses the difficulties of censored data; censoring only changes the at-risk numbers in future periods.
3 Also conditioning on the number of deaths and survivals, n_d and n_s in Table 9.5, leaves only the *univariate* statistic y to interpret at each period, which is easily done through the null hypothesis of equal hazard rates (9.26).
4 Adding the discrepancies $y_i - E_i$ in the numerator of (9.25) (rather than say, adding the individual Z values $Z_i = (y_i - E_i)/V_i^{1/2}$, or adding the Z_i^2 values) accrues power for the natural alternative hypothesis "$h_{Ai} > h_{Bi}$ for all i," while avoiding destabilization from small values of V_i.

Each of the four tactics had been used separately in classical applications. Putting them together into the log-rank test was a major inferential accomplishment, foreshadowing a still bigger step forward, the *proportional hazards model*, our subject in the next section.

Conditional inference takes on an aggressive form in the log-rank test. Let $\boldsymbol{D}_i$ indicate all the data except y_i available at the end of the ith period. For month 6 in the NCOG study, $\boldsymbol{D}_6$ includes all data for months 1–5 in Table 9.4, and the marginals n_A, n_B, n_d, and n_s in Table 9.5, but not the y value for month 6. The key assumption is that, under the null hypothesis of equal hazard rates (9.26),

$$y_i | \boldsymbol{D}_i \overset{\text{ind}}{\sim} (E_i, V_i), \tag{9.30}$$

"ind" here meaning that the y_i's can be treated as independent quantities with means and variances (9.24). In particular, we can add the variances V_i to get the denominator of (9.25). (A "partial likelihood" argument, described in the endnotes, justifies adding the variances.)

The purpose of all this Fisherian conditioning is to simplify the inference: the conditional distribution $y_i | \boldsymbol{D}_i$ depends only on the hazard rates h_{Ai} and h_{Bi}; "nuisance parameters," relating to the survival times and censoring mechanism of the data in Table 9.2, are hidden away. There is a price to pay in testing power, though usually a small one. The lost-to-followup values l in Table 9.3 have been ignored, even though they might contain useful information, say if all the early losses occurred in one arm.

9.4 The Proportional Hazards Model

The Kaplan–Meier estimator is a one-sample device, dealing with data coming from a single distribution. The log-rank test makes two-sample comparisons. *Proportional hazards* ups the ante to allow for a full regression analysis of censored data. Now the individual data points z_i are of the form

$$z_i = (c_i, t_i, d_i), \tag{9.31}$$

where t_i and d_i are observed survival time and censoring indicator, as in (9.14)–(9.15), and c_i is a known $1 \times p$ vector of covariates whose effect on survival we wish to assess. Both of the previous methods are included here: for the log-rank test, c_i indicates treatment, say c_i equals 0 or 1 for `Arm_A` or `Arm_B`, while c_i is absent for Kaplan–Meier.

Table 9.6 *Pediatric cancer data, first 20 of 1620 children.* `Sex` *1 = male, 2 = female;* `race` *1 = white, 2 = nonwhite;* `age` *in years;* `entry` *= calendar date of entry in days since July 1, 2001;* `far` *= home distance from treatment center in miles;* `t` *= survival time in days;* `d` *= 1 if death observed, 0 if not.*

sex	race	age	entry	far	t	d
1	1	2.50	710	108	325	0
2	1	10.00	1866	38	1451	0
2	2	18.17	2531	100	221	0
2	1	3.92	2210	100	2158	0
1	1	11.83	875	78	760	0
2	1	11.17	1419	0	168	0
2	1	5.17	1264	28	2976	0
2	1	10.58	670	120	1833	0
1	1	1.17	1518	73	131	0
2	1	6.83	2101	104	2405	0
1	1	13.92	1239	0	969	0
1	1	5.17	518	117	1894	0
1	1	2.50	1849	99	193	1
1	1	.83	2758	38	1756	0
2	1	15.50	2004	12	682	0
1	1	17.83	986	65	1835	0
2	1	3.25	1443	58	2993	0
1	1	10.75	2807	42	1616	0
1	2	18.08	1229	23	1302	0
2	2	5.83	2727	23	174	1

Medical studies regularly produce data of form (9.31). An example, the *pediatric cancer* data, is partially listed in Table 9.6. The first 20 of $n = 1620$ cases are shown. There are five explanatory covariates (defined in the table's caption): **sex**, **race**, **age** at entry, calendar date of **entry** into the study, and **far**, the distance of the child's home from the treatment center. The response variable t is survival in days from time of treatment until death. Happily, only 160 of the children were observed to die ($d = 1$). Some left the study for various reasons, but most of the $d = 0$ cases were those children still alive at the end of the study period. Of particular interest was the effect of **far** on survival. We wish to carry out a regression analysis of this heavily censored data set.

The proportional hazards model assumes that the hazard rate $h_i(t)$ for the ith individual (9.8) is

$$h_i(t) = h_0(t)e^{c_i'\beta}. \tag{9.32}$$

Here $h_0(t)$ is a baseline hazard (which we need not specify) and β is an unknown p-parameter vector we want to estimate. For concise notation, let

$$\theta_i = e^{c_i'\beta}; \tag{9.33}$$

model (9.32) says that individual i's hazard is a constant nonnegative factor θ_i times the baseline hazard. Equivalently, from (9.11), the ith survival function $S_i(t)$ is a power of the baseline survival function $S_0(t)$,

$$S_i(t) = S_0(t)^{\theta_i}. \tag{9.34}$$

Larger values of θ_i lead to more quickly declining survival curves, i.e., to worse survival (as in (9.11)).

Let J be the number of observed deaths, $J = 160$ here, occurring at times

$$T_{(1)} < T_{(2)} < \ldots < T_{(J)}, \tag{9.35}$$

again for convenience assuming no ties.[9] Just before time $T_{(j)}$ there is a *risk set* of individuals still under observation, whose indices we denote by $\mathcal{R}_j$,

$$\mathcal{R}_j = \{i : t_i \geq T_{(j)}\}. \tag{9.36}$$

Let i_j be the index of the individual observed to die at time $T_{(j)}$. The key to proportional hazards regression is the following result.

[9] More precisely, assuming only one event, a death, occurred at $T_{(j)}$, with none of the other individuals being lost to followup at exact time $T_{(j)}$.

Lemma † *Under the proportional hazards model* (9.32)*, the conditional probability, given the risk set* $\mathcal{R}_j$*, that individual* i *in* $\mathcal{R}_j$ *is the one observed to die at time* $T_{(j)}$ *is* †8

$$\Pr\{i_j = i | \mathcal{R}_j\} = e^{c_i'\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k'\beta}. \tag{9.37}$$

To put it in words, given that one person dies at time $T_{(j)}$*, the probability it is individual* i *is proportional to* $\exp(c_i'\beta)$*, among the set of individuals at risk.*

For the purpose of estimating the parameter vector β in model (9.32), we multiply factors (9.37) to form the *partial likelihood*

$$L(\beta) = \prod_{j=1}^{J} \left(e^{c_{i_j}'\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k'\beta} \right). \tag{9.38}$$

$L(\beta)$ is then treated as an ordinary likelihood function, yielding an approximately unbiased MLE-like estimate

$$\hat{\beta} = \arg\max_{\beta} \{L(\beta)\}, \tag{9.39}$$

with an approximate covariance obtained from the second-derivative matrix of $l(\beta) = \log L(\beta)$,† as in Section 4.3, †9

$$\hat{\beta} \doteq \left(\beta, \left[-\ddot{l}\left(\hat{\beta}\right)\right]^{-1}\right). \tag{9.40}$$

Table 9.7 shows the proportional hazards analysis of the pediatric cancer data, with the covariates **age**, **entry**, and **far** standardized to have mean 0 and standard deviation 1 for the 1620 cases.[10] Neither **sex** nor **race** seems to make much difference. We see that **age** is a mildly significant factor, with older children doing better (i.e., the estimated regression coefficient is negative). However, the dramatic effects are date of **entry** and **far**. Individuals who entered the study later survived longer—perhaps the treatment protocol was being improved—while children living farther away from the treatment center did worse.

Justification of the partial likelihood calculations is similar to that for the log-rank test, but there are some important differences, too: the proportional hazards model is semiparametric ("semi" because we don't have to specify $h_0(t)$ in (9.32)), rather than nonparametric as before; and the

[10] Table 9.7 was obtained using the **R** program **coxph**.

Table 9.7 *Proportional hazards analysis of pediatric cancer data (*`age`*,* `entry` *and* `far` *standardized).* `Age` *significantly negative, older children doing better;* `entry` *very significantly negative, showing hazard rate declining with calendar date of entry;* `far` *very significantly positive, indicating worse results for children living farther away from the treatment center. Last two columns show limits of approximate 95% confidence intervals for* $\exp(\beta)$.

	β	sd	z-value	p-value	$\exp(\beta)$	Lower	Upper
`sex`	−.023	.160	−.142	.887	.98	.71	1.34
`race`	.282	.169	1.669	.095	1.33	.95	1.85
`age`	−.235	.088	−2.664	.008	.79	.67	.94
`entry`	−.460	.079	−5.855	.000	.63	.54	.74
`far`	.296	.072	4.117	.000	1.34	1.17	1.55

emphasis on likelihood has increased the Fisherian nature of the inference, moving it further away from pure frequentism. Still more Fisherian is the emphasis on likelihood inference in (9.38)–(9.40), rather than the direct frequentist calculations of (9.24)–(9.25).

The conditioning argument here is less obvious than that for the Kaplan–Meier estimate or the log-rank test. Has its convenience possibly come at too high a price? In fact it can be shown that inference based on the partial likelihood is highly efficient, assuming of course the correctness of the proportional hazards model (9.32).

9.5 Missing Data and the EM Algorithm

Censored data, the motivating factor for survival analysis, can be thought of as a special case of a more general statistical topic, *missing data*. What's missing, in Table 9.2 for example, are the actual survival times for the + cases, which are known only to exceed the tabled values. If the data were *not* missing, we could use standard statistical methods, for instance Wilcoxon's test, to compare the two arms of the `NCOG` study. The EM algorithm is an iterative technique for solving missing-data inferential problems using only standard methods.

A missing-data situation is shown in Figure 9.3: $n = 40$ points have been independently sampled from a bivariate normal distribution (5.12),

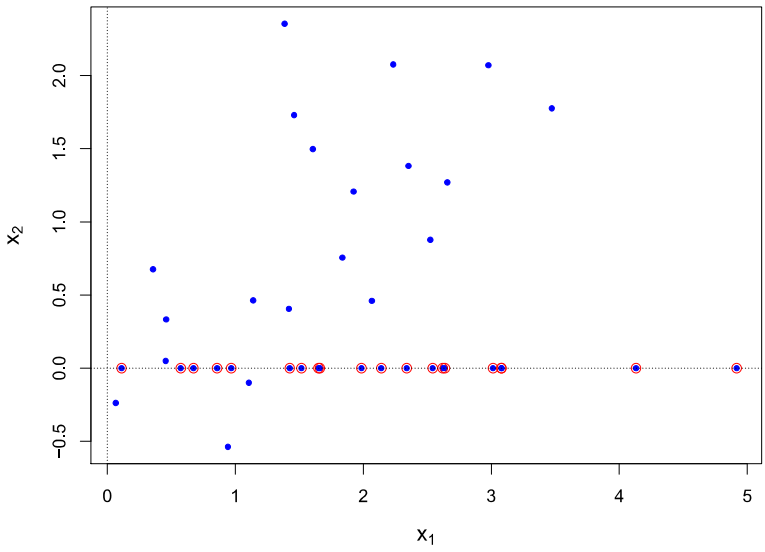

Figure 9.3 Forty points from a bivariate normal distribution, the last 20 with x_2 missing (circled).

means (μ_1, μ_2), variances (σ_1^2, σ_2^2), and correlation ρ,

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \overset{\text{ind}}{\sim} \mathcal{N}_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \right). \tag{9.41}$$

However, the second coordinates of the last 20 points have been lost. These are represented by the circled points in Figure 9.3, with their x_2 values arbitrarily set to 0.

We wish to find the maximum likelihood estimate of the parameter vector $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. The standard maximum likelihood estimates

$$\begin{aligned} \hat{\mu}_1 &= \sum_{i=1}^{40} x_{1i}/40, \quad \hat{\mu}_2 = \sum_{i=1}^{40} x_{2i}/40, \\ \hat{\sigma}_1 &= \left[\sum_{i=1}^{40} (x_{1i} - \hat{\mu}_1)^2 /40 \right]^{1/2}, \quad \hat{\sigma}_2 = \left[\sum_{i=1}^{40} (x_{2i} - \hat{\mu}_2)^2 /40 \right]^{1/2}, \\ \hat{\rho} &= \left[\sum_{i=1}^{40} (x_{1i} - \hat{\mu}_1)(x_{2i} - \hat{\mu}_2) /40 \right] \Big/ (\hat{\sigma}_1 \hat{\sigma}_2), \end{aligned} \tag{9.42}$$

are unavailable for μ_2, σ_2, and ρ because of the missing data.

The EM algorithm begins by filling in the missing data in some way, say by setting $x_{2i} = 0$ for the 20 missing values, giving an artificially complete data set $data^{(0)}$. Then it proceeds as follows.

- The standard method (9.42) is applied to the filled-in $data^{(0)}$ to produce $\hat{\theta}^{(0)} = (\hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_1^{(0)}, \hat{\sigma}_2^{(0)}, \hat{\rho}^{(0)})$; this is the M ("maximizing") step.[11]
- Each of the missing values is replaced by its conditional expectation (assuming $\theta = \hat{\theta}^{(0)}$) given the nonmissing data; this is the E ("expectation") step. In our case the missing values x_{2i} are replaced by

$$\hat{\mu}_2^{(0)} + \hat{\rho}^{(0)} \frac{\hat{\sigma}_2^{(0)}}{\hat{\sigma}_1^{(0)}} \left(x_{1i} - \hat{\mu}_1^{(0)} \right). \tag{9.43}$$

- The E and M steps are repeated, at the jth stage giving a new artificially complete data set $data^{(j)}$ and an updated estimate $\hat{\theta}^{(j)}$. The iteration stops when $\|\hat{\theta}^{(j+1)} - \hat{\theta}^{(j)}\|$ is suitably small.

Table 9.8 shows the EM algorithm at work on the bivariate normal example of Figure 9.3. In exponential families the algorithm is guaranteed to converge to the MLE $\hat{\theta}$ based on just the observed data $\boldsymbol{o}$; moreover, the likelihood $f_{\hat{\theta}^{(j)}}(\boldsymbol{o})$ increases with every step j. (The convergence can be sluggish, as it is here for $\hat{\sigma}_2$ and $\hat{\rho}$.)

The EM algorithm ultimately derives from the *fake-data principle*, a property of maximum likelihood estimation going back to Fisher that can
†10 only briefly be summarized here.† Let $\boldsymbol{x} = (\boldsymbol{o}, \boldsymbol{u})$ represent the "complete data," of which $\boldsymbol{o}$ is observed while $\boldsymbol{u}$ is unobserved or missing. Write the density for $\boldsymbol{x}$ as

$$f_\theta(\boldsymbol{x}) = f_\theta(\boldsymbol{o})\, f_\theta(\boldsymbol{u}|\boldsymbol{o}), \tag{9.44}$$

and let $\hat{\theta}(\boldsymbol{o})$ be the MLE of θ based just on $\boldsymbol{o}$.

Suppose we now generate simulations of $\boldsymbol{u}$ by sampling from the conditional distribution $f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o})$,

$$\boldsymbol{u}^{*k} \sim f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o}) \qquad \text{for } k = 1, 2, \ldots, K \tag{9.45}$$

(the stars indicating creation by the statistician and not by observation), giving fake complete-data values $\boldsymbol{x}^{*k} = (\boldsymbol{o}, \boldsymbol{u}^{*k})$. Let

$$data^* = \{\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*K}\}, \tag{9.46}$$

[11] In this example, $\hat{\mu}_1^{(0)}$ and $\hat{\sigma}_1^{(0)}$ are available as the complete-data estimates in (9.42), and, as in Table 9.8, stay the same in subsequent steps of the algorithm.

Table 9.8 *EM algorithm for estimating means, standard deviations, and the correlation of the bivariate normal distribution that gave the data in Figure 9.3.*

Step	μ_1	μ_2	σ_1	σ_2	ρ
1	1.86	.463	1.08	.738	.162
2	1.86	.707	1.08	.622	.394
3	1.86	.843	1.08	.611	.574
4	1.86	.923	1.08	.636	.679
5	1.86	.971	1.08	.667	.736
6	1.86	1.002	1.08	.694	.769
7	1.86	1.023	1.08	.716	.789
8	1.86	1.036	1.08	.731	.801
9	1.86	1.045	1.08	.743	.808
10	1.86	1.051	1.08	.751	.813
11	1.86	1.055	1.08	.756	.816
12	1.86	1.058	1.08	.760	.819
13	1.86	1.060	1.08	.763	.820
14	1.86	1.061	1.08	.765	.821
15	1.86	1.062	1.08	.766	.822
16	1.86	1.063	1.08	.767	.822
17	1.86	1.064	1.08	.768	.823
18	1.86	1.064	1.08	.768	.823
19	1.86	1.064	1.08	.769	.823
20	1.86	1.064	1.08	.769	.823

whose notional likelihood $\prod_1^K f_\theta(\boldsymbol{x}^{*k})$ yields MLE $\hat{\theta}^*$. It then turns out that $\hat{\theta}^*$ goes to $\hat{\theta}(\boldsymbol{o})$ as K goes to infinity. In other words, maximum likelihood estimation is *self-consistent*: generating artificial data from the MLE density $f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o})$ doesn't change the MLE. Moreover, any value $\hat{\theta}^{(0)}$ not equal to the MLE $\hat{\theta}(\boldsymbol{o})$ cannot be self-consistent: carrying through (9.45)–(9.46) using $f_{\hat{\theta}^{(0)}}(\boldsymbol{u}|\boldsymbol{o})$ leads to hypothetical MLE $\hat{\theta}^{(1)}$ having $f_{\hat{\theta}^{(1)}}(\boldsymbol{o}) > f_{\hat{\theta}^{(0)}}(\boldsymbol{o})$, etc., a more general version of the EM algorithm.[12]

Modern technology allows social scientists to collect huge data sets, perhaps hundreds of responses for each of thousands or even millions of individuals. Inevitably, some entries of the individual responses will be missing. *Imputation* amounts to employing some version of the fake-data principle to fill in the missing values. Imputation's goal goes beyond find-

[12] Simulation (9.45) is unnecessary in exponential families, where at each stage *data** can be replaced by $(\boldsymbol{o}, E^{(j)}(\boldsymbol{u}|\boldsymbol{o}))$, with $E^{(j)}$ indicating expectation with respect to $\hat{\theta}^{(j)}$, as in (9.43).

ing the MLE, to the creation of graphs, confidence intervals, histograms, and more, using only convenient, standard complete-data methods.

Finally, returning to survival analysis, the Kaplan–Meier estimate (9.17) †11 is itself self-consistent.† Consider the **Arm_A** censored observation 74+ in Table 9.2. We know that that patient's survival time exceeded 74. Suppose we distribute his probability mass (1/51 of the **Arm_A** sample) to the right, in accordance with the conditional distribution for $x > 74$ defined by the **Arm_A** Kaplan–Meier survival curve. It turns out that redistributing all the censored cases does not change the original Kaplan–Meier survival curve; Kaplan–Meier is self-consistent, leading to its identification as the "nonparametric MLE" of a survival function.

9.6 Notes and Details

The progression from life tables, Kaplan–Meier curves, and the log-rank test to proportional hazards regression was modest in its computational demands, until the final step. Kaplan–Meier curves lie within the capabilities of mechanical calculators. Not so for proportional hazards, which is emphatically a child of the computer age. As the algorithms grew more intricate, their inferential justification deepened in scope and sophistication. This is a pattern we also saw in Chapter 8, in the progression from bioassay to logistic regression to generalized linear models, and will reappear as we move from the jackknife to the bootstrap in Chapter 10.

Censoring is not the same as truncation. For the truncated galaxy data of Section 8.3, we learn of the existence of a galaxy only if it falls into the observation region (8.38). The censored individuals in Table 9.2 are known to exist, but with imperfect knowledge of their lifetimes. There is a version of the Kaplan–Meier curve applying to truncated data, which was developed in the astronomy literature by Lynden-Bell (1971).

The methods of this chapter apply to data that is left-truncated as well as right-censored. In a survival time study of a new HIV drug, for instance, subject i might not enter the study until some time τ_i after his or her initial diagnosis, in which case t_i would be left-truncated at τ_i, as well as possibly later right-censored. This only modifies the composition of the various risk sets. However, other missing-data situations, e.g., left- *and* right-censoring, require more elaborate, less elegant, treatments.

†1 [p. 133] *Formula* (9.10). Let the interval $[t_0, t_1]$ be partitioned into a large number of subintervals of length dt, with t_k the midpoint of subinterval k.

As in (9.4), using (9.9),

$$\begin{aligned}\Pr\{T \geq t_1 | T \geq t_0\} &\doteq \prod (1 - h(t_i)\, dt) \\ &= \exp\left\{\sum \log(1 - h(t_i)\, dt)\right\} \\ &\doteq \exp\left\{-\sum h(t_i)\, dt\right\},\end{aligned} \tag{9.47}$$

which, as $dt \to 0$, goes to (9.10).

†2 [p. 136] *Kaplan–Meier estimate.* In the life table formula (9.6) (with $k = 1$), let the time unit be small enough to make each bin contain at most one value $t_{(k)}$ (9.16). Then at $t_{(k)}$,

$$\hat{h}_{(k)} = \frac{d_{(k)}}{n - k + 1}, \tag{9.48}$$

giving expression (9.17).

†3 [p. 137] *Greenwood's formula* (9.18). In the life table formulation of Section 9.1, (9.6) gives

$$\log \hat{S}_j = \sum_1^j \log\left(1 - \hat{h}_k\right). \tag{9.49}$$

From $n_k \hat{h}_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k)$ we get

$$\begin{aligned}\text{var}\left\{\log \hat{S}_j\right\} &= \sum_1^j \text{var}\left\{\log\left(1 - \hat{h}_k\right)\right\} \doteq \sum_1^j \frac{\text{var}\, \hat{h}_k}{(1 - h_k)^2} \\ &= \sum_1^j \frac{h_k}{1 - h_k} \frac{1}{n_k},\end{aligned} \tag{9.50}$$

where we have used the delta-method approximation $\text{var}\{\log X\} \doteq \text{var}\{X\}/E\{X\}^2$. Plugging in $h_k = y_k/n_k$ yields

$$\text{var}\left\{\log \hat{S}_j\right\} \doteq \sum_1^j \frac{y_k}{n_k(n_k - y_k)}. \tag{9.51}$$

Then the inverse approximation $\text{var}\{X\} = E\{X\}^2 \text{var}\{\log X\}$ gives Greenwood's formula (9.18).

The censored data situation of Section 9.2 does not enjoy independence between the $\hat{h}_k$ values. However, successive conditional independence, given the n_k values, is enough to verify the result, as in the partial likelihood calculations below. *Note*: the confidence intervals in Figure 9.1 were obtained

by exponentiating the intervals,

$$\log \hat{S}_j \pm 1.96 \left[\operatorname{var} \left\{ \log \hat{S}_j \right\} \right]^{1/2}. \tag{9.52}$$

†4 [p. 138] *Parametric life tables analysis.* Figure 9.2 and the analysis behind it is developed in Efron (1988), where it is called "partial logistic regression" in analogy with partial likelihood.

†5 [p. 139] *The log-rank test.* This chapter featured an all-star cast, including four of the most referenced papers of the post-war era: Kaplan and Meier (1958), Cox (1972) on proportional hazards, Dempster *et al.* (1977) codifying and naming the EM algorithm, and Mantel and Haenszel (1959) on the log-rank test. (Cox (1958) gives a careful, and early, analysis of the Mantel–Haenszel idea.) The not very helpful name "log-rank" does at least remind us that the test depends only on the ranks of the survival times, and will give the same result if all the observed survival times t_i are monotonically transformed, say to $\exp(t_i)$ or $t_i^{1/2}$. It is often referred to as the Mantel–Haenszel or Cochran–Mantel–Haenszel test in older literature. Kaplan–Meier and proportional hazards are also rank-based procedures.

†6 [p. 141] *Hypergeometric distribution.* Hypergeometric calculations, as for Table 9.5, are often stated as follows: n marbles are placed in an urn, n_A labeled A and n_B labeled B; n_d marbles are drawn out at random; y is the number of these labeled A. Elementary (but not simple) calculations then produce the conditional distribution of y given the table's marginals n_A, n_B, n, n_d, and n_s,

$$\Pr\{y|\text{marginals}\} = \binom{n_A}{y}\binom{n_B}{n_d - y} \Big/ \binom{n}{n_d} \tag{9.53}$$

for

$$\max(n_A - n_s, 0) \le y \le \min(n_d, n_A),$$

and expressions (9.24) for the mean and variance. If n_A and n_B go to infinity such that $n_A/n \to p_A$ and $n_B/n \to 1 - p_A$, then $V \to n_d\, p_A(1 - p_A)$, the variance of $y \sim \text{Bi}(n_d, p_A)$.

†7 [p. 141] *Log-rank statistic* Z (9.25). Why is $(\sum_1^N V_i)^{1/2}$ the correct denominator for Z? Let $u_i = y_i - E_i$ in (9.30), so Z's numerator is $\sum_1^N u_i$, with

$$u_i | \boldsymbol{D}_i \sim (0, V_i) \tag{9.54}$$

under the null hypothesis of equal hazard rates. This implies that, unconditionally, $E\{u_i\} = 0$. For $j < i$, u_j is a function of $\boldsymbol{D}_i$ (since y_j and

E_j are), so $E\{u_j u_i | \boldsymbol{D}_i\} = 0$, and, again unconditionally, $E\{u_j u_i\} = 0$. Therefore, assuming equal hazard rates,

$$E\left(\sum_1^N u_i\right)^2 = E\left(\sum_1^N u_i^2\right) = \sum_1^N \text{var}\{u_i\} \doteq \sum_1^N V_i. \tag{9.55}$$

The last approximation, replacing unconditional variances $\text{var}\{u_i\}$ with conditional variances V_i, is justified in Crowley (1974), as is the asymptotic normality (9.29).

†8 [p. 145] *Lemma* (9.37). For $i \in \mathcal{R}_j$, the probability p_i that death occurs in the infinitesimal interval $(T_{(j)}, T_{(j)} + dT)$ is $h_i(T_{(j)})\, dT$, so

$$p_i = h_0(T_{(j)}) e^{c_i'\beta}\, dT, \tag{9.56}$$

and the probability of event A_i that individual i dies while the others don't is

$$P_i = p_i \prod_{k \in \mathcal{R}_j - i} (1 - p_k). \tag{9.57}$$

But the A_i are disjoint events, so, given that $\cup A_i$ has occurred, the probability that it is individual i who died is

$$P_i \Big/ \sum_{\mathcal{R}_j} P_j \doteq e^{c_i\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k\beta}, \tag{9.58}$$

this becoming exactly (9.37) as $dT \to 0$.

†9 [p. 145] *Partial likelihood* (9.40). Cox (1975) introduced partial likelihood as inferential justification for the proportional hazards model, which had been questioned in the literature. Let $\boldsymbol{D}_j$ indicate all the observable information available just before time $T_{(j)}$(9.35), including all the death or loss times for individuals having $t_i < T_{(j)}$. (Notice that $\boldsymbol{D}_j$ determines the risk set $\mathcal{R}_j$.) By successive conditioning we write the full likelihood $f_\theta(\text{data})$ as

$$\begin{aligned} f_\theta(\text{data}) &= f_\theta(\boldsymbol{D}_1) f_\theta(i_1|\mathcal{R}_1) f_\theta(\boldsymbol{D}_2|\boldsymbol{D}_1) f_\theta(i_2|\mathcal{R}_2) \ldots \\ &= \prod_{j=1}^J f_\theta(\boldsymbol{D}_j|\boldsymbol{D}_{j-1}) \prod_{j=1}^J f_\theta(i_j|\mathcal{R}_j). \end{aligned} \tag{9.59}$$

Letting $\theta = (\alpha, \beta)$, where α is a nuisance parameter vector having to do

with the occurrence and timing of events between observed deaths,

$$f_{\alpha,\beta}(\text{data}) = \left[\prod_{j=1}^{J} f_{\alpha,\beta}(\boldsymbol{D}_j|\boldsymbol{D}_{j-1})\right] L(\beta), \tag{9.60}$$

where $L(\beta)$ is the partial likelihood (9.38).

The proportional hazards model simply ignores the bracketed factor in (9.60); $l(\beta) = \log L(\beta)$ is treated as a genuine likelihood, maximized to give $\hat{\beta}$, and assigned covariance matrix $(-\ddot{l}(\hat{\beta}))^{-1}$ as in Section 4.3. Efron (1977) shows this tactic is highly efficient for the estimation of β.

†10 [p. 148] *Fake-data principle.* For any two values of the parameters θ_1 and θ_2 define

$$l_{\theta_1}(\theta_2) = \int [\log f_{\theta_2}(\boldsymbol{o}, \boldsymbol{u})]\, f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o})\, d\boldsymbol{u}, \tag{9.61}$$

this being the limit as $K \to \infty$ of

$$l_{\theta_1}(\theta_2) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \log f_{\theta_2}(\boldsymbol{o}, \boldsymbol{u}^{*k}), \tag{9.62}$$

the fake-data log likelihood (9.46) under θ_2, if θ_1 were the true value of θ. Using $f_\theta(\boldsymbol{o}, \boldsymbol{u}) = f_\theta(\boldsymbol{o})\, f_\theta(\boldsymbol{u}|\boldsymbol{o})$, definition (9.61) gives

$$\begin{aligned} l_{\theta_1}(\theta_2) - l_{\theta_1}(\theta_1) &= \log\left(\frac{f_{\theta_2}(\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{o})}\right) + \int \log\left(\frac{f_{\theta_2}(\boldsymbol{u}|\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o})}\right) f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o}) \\ &= \log\left(\frac{f_{\theta_2}(\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{o})}\right) - \frac{1}{2} D\left(f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o}),\, f_{\theta_2}(\boldsymbol{u}|\boldsymbol{o})\right), \end{aligned} \tag{9.63}$$

with D the deviance (8.31), which is always positive unless $\boldsymbol{u}|\boldsymbol{o}$ has the same distribution under θ_1 and θ_2, which we will assume doesn't happen.

Suppose we begin the EM algorithm at $\theta = \theta_1$ and find the value θ_2 maximizing $l_{\theta_1}(\theta)$. Then $l_{\theta_1}(\theta_2) > l_{\theta_1}(\theta_1)$ and $D > 0$ implies $f_{\theta_2}(\boldsymbol{o}) > f_{\theta_1}(\boldsymbol{o})$ in (9.63); that is, we have increased the likelihood of the observed data. Now take $\theta_1 = \hat{\theta} = \arg\max_\theta f_\theta(\boldsymbol{o})$. Then the right side of (9.63) is negative, implying $l_{\hat{\theta}}(\hat{\theta}) > l_{\hat{\theta}}(\theta_2)$ for any θ_2 not equaling $\theta_1 = \hat{\theta}$. Putting this together,[13] successively computing $\theta_1, \theta_2, \theta_3, \ldots$ by fake-data MLE calculations increases $f_\theta(\boldsymbol{o})$ at every step, and the only stable point of the algorithm is at $\theta = \hat{\theta}(\boldsymbol{o})$.

†11 [p. 150] *Kaplan–Meier self-consistency.* This property was verified in Efron (1967), where the name was coined.

[13] Generating the fake data is equivalent to the E step of the algorithm, the M step being the maximization of $l_{\theta_j}(\theta)$.

10

The Jackknife and the Bootstrap

A central element of frequentist inference is the *standard error*. An algorithm has produced an estimate of a parameter of interest, for instance the mean $\bar{x} = 0.752$ for the 47 **ALL** scores in the top panel of Figure 1.4. How accurate is the estimate? In this case, formula (1.2) for the standard deviation[1] of a sample mean gives estimated standard error

$$\widehat{\text{se}} = 0.040, \tag{10.1}$$

so one can't take the third digit of $\bar{x} = 0.752$ very seriously, and even the 5 is dubious.

Direct standard error formulas like (1.2) exist for various forms of averaging, such as linear regression (7.34), and for hardly anything else. Taylor series approximations ("device 2" of Section 2.1) extend the formulas to smooth functions of averages, as in (8.30). Before computers, applied statisticians needed to be Taylor series experts in laboriously pursuing the accuracy of even moderately complicated statistics.

The jackknife (1957) was a first step toward a computation-based, nonformulaic approach to standard errors. The bootstrap (1979) went further toward automating a wide variety of inferential calculations, including standard errors. Besides sparing statisticians the exhaustion of tedious routine calculations the jackknife and bootstrap opened the door for more complicated estimation algorithms, which could be pursued with the assurance that their accuracy would be easily assessed. This chapter focuses on standard errors, with more adventurous bootstrap ideas deferred to Chapter 11. We end with a brief discussion of accuracy estimation for robust statistics.

[1] We will use the terms "standard error" and "standard deviation" interchangeably.

10.1 The Jackknife Estimate of Standard Error

The basic applications of the jackknife apply to *one-sample problems*, where the statistician has observed an independent and identically distributed (iid) sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ from an unknown probability distribution F on some space $\mathcal{X}$,

$$x_i \overset{\text{iid}}{\sim} F \qquad \text{for } i = 1, 2, \ldots, n. \tag{10.2}$$

$\mathcal{X}$ can be anything: the real line, the plane, a function space.[2] A *real-valued* statistic $\hat{\theta}$ has been computed by applying some algorithm $s(\cdot)$ to $\boldsymbol{x}$,

$$\hat{\theta} = s(\boldsymbol{x}), \tag{10.3}$$

and we wish to assign a standard error to $\hat{\theta}$. That is, we wish to estimate the standard deviation of $\hat{\theta} = s(\boldsymbol{x})$ under sampling model (10.2).

Let $\boldsymbol{x}_{(i)}$ be the sample with x_i removed,

$$\boldsymbol{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)', \tag{10.4}$$

and denote the corresponding value of the statistic of interest as

$$\hat{\theta}_{(i)} = s(\boldsymbol{x}_{(i)}). \tag{10.5}$$

Then the *jackknife estimate of standard error* for $\hat{\theta}$ is

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{n-1}{n}\sum_1^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2}, \quad \text{with } \hat{\theta}_{(\cdot)} = \sum_1^n \hat{\theta}_{(i)}/n. \tag{10.6}$$

In the case where $\hat{\theta}$ is the mean $\bar{x}$ of real values $x_1, x_2, \ldots, x_n$ (i.e., $\mathcal{X}$ is an interval of the real line), $\hat{\theta}_{(i)}$ is their average excluding x_i, which can be expressed as

$$\hat{\theta}_{(i)} = (n\bar{x} - x_i)/(n-1). \tag{10.7}$$

Equation (10.7) gives $\hat{\theta}_{(\cdot)} = \bar{x}$, $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = (\bar{x} - x_i)/(n-1)$, and

$$\widehat{\text{se}}_{\text{jack}} = \left[\sum_{i=1}^n (x_i - \bar{x})^2/\,(n(n-1))\right]^{1/2}, \tag{10.8}$$

exactly the same as the classic formula (1.2). This is no coincidence. The fudge factor $(n-1)/n$ in definition (10.6) was inserted to make $\widehat{\text{se}}_{\text{jack}}$ agree with (1.2) when $\hat{\theta}$ is $\bar{x}$.

[2] If $\mathcal{X}$ is an interval of the real line we might take F to be the usual cumulative distribution function, but here we will just think of F as any full description of the probability distribution for an x_i on $\mathcal{X}$.

The advantage of $\widehat{\text{se}}_{\text{jack}}$ is that definition (10.6) can be applied in an automatic way to *any* statistic $\hat{\theta} = s(\boldsymbol{x})$. All that is needed is an algorithm that computes $s(\cdot)$ for the deleted data sets $\boldsymbol{x}_{(i)}$. Computer power is being substituted for theoretical Taylor series calculations. Later we will see that the underlying inferential ideas—plug-in estimation of frequentist standard errors—haven't changed, only their implementation.

As an example, consider the kidney function data set of Section 1.1. Here the data consists of $n = 157$ points (x_i, y_i), with x = age and y = tot in Figure 1.1. (So the generic x_i in (10.2) now represents the pair (x_i, y_i), and F describes a distribution in the plane.) Suppose we are interested in the correlation between age and tot, estimated by the usual sample correlation $\hat{\theta} = s(\boldsymbol{x})$,

$$s(\boldsymbol{x}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Bigg/ \left[\sum_{1}^{n}(x_i - \bar{x})^2 \sum_{1}^{n}(y_i - \bar{y})^2\right]^{1/2}, \quad (10.9)$$

computed to be $\hat{\theta} = -0.572$ for the kidney data.

Applying (10.6) gave $\widehat{\text{se}}_{\text{jack}} = 0.058$ for the accuracy of $\hat{\theta}$. Nonparametric bootstrap computations, Section 10.2, also gave estimated standard error 0.058. The classic Taylor series formula looks quite formidable in this case,

$$\widehat{\text{se}}_{\text{taylor}} = \left\{\frac{\hat{\theta}^2}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}}\right]\right\}^{1/2} \quad (10.10)$$

where

$$\hat{\mu}_{hk} = \sum_{i=1}^{n}(x_i - \bar{x})^h(y_i - \bar{y})^k/n. \quad (10.11)$$

It gave $\widehat{\text{se}} = 0.057$.

It is worth emphasizing some features of the jackknife formula (10.6).

- It is nonparametric; no special form of the underlying distribution F need be assumed.
- It is completely automatic: a single master algorithm can be written that inputs the data set $\boldsymbol{x}$ and the function $s(\boldsymbol{x})$, and outputs $\widehat{\text{se}}_{\text{jack}}$.
- The algorithm works with data sets of size $n-1$, not n. There is a hidden assumption of smooth behavior across sample sizes. This can be worrisome for statistics like the sample median that have a different definition for odd and even sample size.

- The jackknife standard error is upwardly biased as an estimate of the true standard error.† †1
- The connection of the jackknife formula (10.6) with Taylor series methods is closer than it appears. We can write

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{\sum_1^n D_i^2}{n^2}\right]^{1/2}, \qquad \text{where } D_i = \frac{\hat\theta_{(i)} - \hat\theta_{(\cdot)}}{1/\sqrt{n(n-1)}}. \tag{10.12}$$

As discussed in Section 10.3, the D_i are approximate *directional derivatives*, measures of how fast the statistic $s(\boldsymbol{x})$ is changing as we decrease the weight on data point x_i. So $\text{se}^2_{\text{jack}}$ is proportional to the sum of squared derivatives of $s(\boldsymbol{x})$ in the n component directions. Taylor series expressions such as (10.10) amount to doing the derivatives by formula rather than numerically.

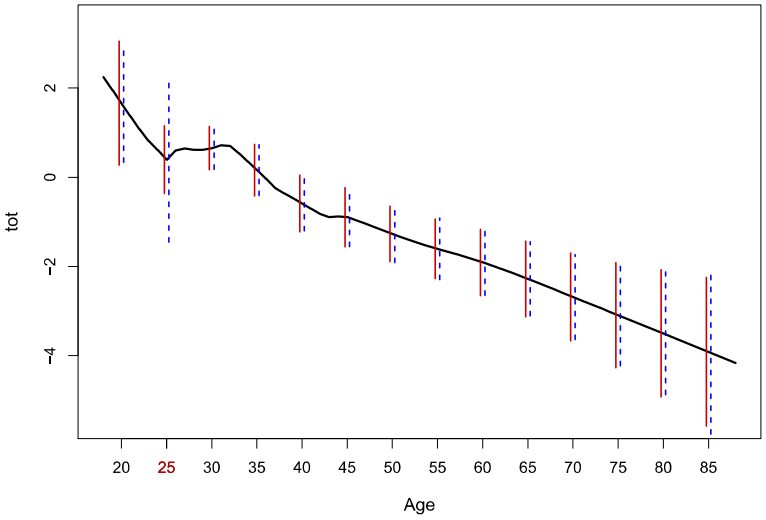

Figure 10.1 The `lowess` curve for the kidney data of Figure 1.2. Vertical bars indicate ±2 standard errors: *jackknife* (10.6) blue dashed; *bootstrap* (10.16) red solid. The jackknife greatly overestimates variability at age 25.

The principal weakness of the jackknife is its dependence on local derivatives. Unsmooth statistics $s(\boldsymbol{x})$, such as the kidney data `lowess` curve in Figure 1.2, can result in erratic behavior for $\widehat{\text{se}}_{\text{jack}}$. Figure 10.1 illustrates the point. The dashed blue vertical bars indicate ±2 jackknife standard er-

rors for the `lowess` curve evaluated at ages $20, 25, \ldots, 85$. For the most part these agree with the dependable bootstrap standard errors, solid red bars, described in Section 10.2. But things go awry at age 25, where the local derivatives greatly overstate the sensitivity of the `lowess` curve to global changes in the sample $\boldsymbol{x}$.

10.2 The Nonparametric Bootstrap

From the point of view of the bootstrap, the jackknife was a halfway house between classical methodology and a full-throated use of electronic computation. (The term "computer-intensive statistics" was coined to describe the bootstrap.) The frequentist standard error of an estimate $\hat{\theta} = s(\boldsymbol{x})$ is, ideally, the standard deviation we would observe by repeatedly sampling new versions of $\boldsymbol{x}$ from F. This is impossible since F is unknown. Instead, the bootstrap ("ingenious device" number 4 in Section 2.1) substitutes an estimate $\hat{F}$ for F and then estimates the frequentist standard error by direct simulation, a feasible tactic only since the advent of electronic computation.

The bootstrap estimate of standard error for a statistic $\hat{\theta} = s(\boldsymbol{x})$ computed from a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ (10.2) begins with the notion of a *bootstrap sample*

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*), \tag{10.13}$$

where each x_i^* is drawn randomly with equal probability and with replacement from $\{x_1, x_2, \ldots, x_n\}$. Each bootstrap sample provides a *bootstrap replication* of the statistic of interest,[3]

$$\hat{\theta}^* = s(\boldsymbol{x}^*). \tag{10.14}$$

Some large number B of bootstrap samples are independently drawn ($B = 500$ in Figure 10.1). The corresponding bootstrap replications are calculated, say

$$\hat{\theta}^{*b} = s(\boldsymbol{x}^{*b}) \qquad \text{for } b = 1, 2, \ldots, B. \tag{10.15}$$

The resulting bootstrap estimate of standard error for $\hat{\theta}$ is the empirical

[3] The star notation $\boldsymbol{x}^*$ is intended to avoid confusion with the original data $\boldsymbol{x}$, which stays fixed in bootstrap computations, and likewise $\hat{\theta}^*$ vis-a-vis $\hat{\theta}$.

standard deviation of the $\hat{\theta}^{*b}$ values,

$$\widehat{\text{se}}_{\text{boot}} = \left[\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot}\right)^2 \Big/ (B-1)\right]^{1/2}, \quad \text{with } \hat{\theta}^{*\cdot} = \sum_{b=1}^{B} \hat{\theta}^{*b}/B. \tag{10.16}$$

Motivation for $\widehat{\text{se}}_{\text{boot}}$ begins by noting that $\hat{\theta}$ is obtained in two steps: first $\boldsymbol{x}$ is generated by iid sampling from probability distribution F, and then $\hat{\theta}$ is calculated from $\boldsymbol{x}$ according to algorithm $s(\cdot)$,

$$F \xrightarrow{\text{iid}} \boldsymbol{x} \xrightarrow{s} \hat{\theta}. \tag{10.17}$$

We don't know F, but we can estimate it by the *empirical probability distribution* $\hat{F}$ that puts probability $1/n$ on each point x_i (e.g., weight $1/157$ on each point (x_i, y_i) in Figure 1.2). Notice that a bootstrap sample $\boldsymbol{x}^*$ (10.13) *is an iid sample drawn from* $\hat{F}$, since then each $\boldsymbol{x}^*$ independently has equal probability of being any member of $\{x_1, x_2, \ldots, x_n\}$. It can be shown that $\hat{F}$ maximizes the probability of obtaining the observed sample $\boldsymbol{x}$ under all possible choices of F in (10.2), i.e., it is the *nonparametric MLE* of F.

Bootstrap replications $\hat{\theta}^*$ are obtained by a process analogous to (10.17),

$$\hat{F} \xrightarrow{\text{iid}} \boldsymbol{x}^* \xrightarrow{s} \hat{\theta}^*. \tag{10.18}$$

In the real world (10.17) we only get to see the single value $\hat{\theta}$, but the bootstrap world (10.18) is more generous: we can generate as many bootstrap replications $\hat{\theta}^{*b}$ as we want, or have time for, and directly estimate their variability as in (10.16). The fact that $\hat{F}$ approaches F as n grows large suggests, correctly in most cases, that $\widehat{\text{se}}_{\text{boot}}$ approaches the true standard error of $\hat{\theta}$.

The true standard deviation of $\hat{\theta}$, i.e., its standard error, can be thought of as a function of the probability distribution F that generates the data, say $\text{Sd}(F)$. Hypothetically, $\text{Sd}(F)$ inputs F and outputs the standard deviation of $\hat{\theta}$, which we can imagine being evaluated by independently running (10.17) some enormous number of times N, and then computing the empirical standard deviation of the resulting $\hat{\theta}$ values,

$$\text{Sd}(F) = \left[\sum_{j=1}^{N}\left(\hat{\theta}^{(j)} - \hat{\theta}^{(\cdot)}\right)^2 \Big/ (N-1)\right]^{1/2}, \quad \text{with } \hat{\theta}^{(\cdot)} = \sum_{1}^{N} \hat{\theta}^{(j)}/N. \tag{10.19}$$

The bootstrap standard error of $\hat{\theta}$ is the plug-in estimate

$$\widehat{\text{se}}_{\text{boot}} = \text{Sd}(\hat{F}). \tag{10.20}$$

More exactly, $\text{Sd}(\hat{F})$ is the *ideal bootstrap estimate* of standard error, what we would get by letting the number of bootstrap replications B go to infinity. In practice we have to stop at some finite value of B, as discussed in what follows.

As with the jackknife, there are several important points worth emphasizing about $\widehat{\text{se}}_{\text{boot}}$.

- It is completely automatic. Once again, a master algorithm can be written that inputs the data $\boldsymbol{x}$ and the function $s(\cdot)$, and outputs $\widehat{\text{se}}_{\text{boot}}$.
- We have described the *one-sample nonparametric bootstrap*. Parametric and multisample versions will be taken up later.
- Bootstrapping "shakes" the original data more violently than jackknifing, producing nonlocal deviations of $\boldsymbol{x}^*$ from $\boldsymbol{x}$. The bootstrap is more dependable than the jackknife for unsmooth statistics since it doesn't depend on local derivatives.
- $B = 200$ is usually sufficient† for evaluating $\widehat{\text{se}}_{\text{boot}}$. Larger values, 1000 or 2000, will be required for the bootstrap confidence intervals of Chapter 11. †2
- There is nothing special about standard errors. We could just as well use the bootstrap replications to estimate the expected absolute error $E\{|\hat{\theta} - \theta|\}$, or any other accuracy measure.
- Fisher's MLE formula (4.27) is applied in practice via

$$\widehat{\text{se}}_{\text{fisher}} = (n\mathcal{I}_{\hat{\theta}})^{-1/2}, \tag{10.21}$$

that is, by plugging in $\hat{\theta}$ for θ after a theoretical calculation of se. The bootstrap operates in the same way at (10.20), though the plugging in is done before rather than after the calculation. The connection with Fisherian theory is more obvious for the parametric bootstrap of Section 10.4.

The jackknife is a completely frequentist device, both in its assumptions and in its applications (standard errors and biases). The bootstrap is also basically frequentist, but with a touch of the Fisherian as in the relation with (10.21). Its versatility has led to applications in a variety of estimation and prediction problems, with even some Bayesian connections.† †3
Unusual applications can also pop up for the jackknife; see the jackknife-after-bootstrap comment in the chapter endnotes.† †4

From a classical point of view, the bootstrap is an incredible computational spendthrift. Classical statistics was fashioned to minimize the hard

labor of mechanical computation. The bootstrap seems to go out of its way to multiply it, by factors of $B = 200$ or 2000 or more. It is nice to report that all this computational largesse can have surprising data analytic payoffs.

Table 10.1 *Correlation matrix for the student score data. The eigenvalues are 3.463, 0.660, 0.447, 0.234, and 0.197. The eigenratio statistic* $\hat{\theta} = 0.693$*, and its bootstrap standard error estimate is 0.075* ($B = 2000$).

	mechanics	vectors	algebra	analytics	statistics
mechanics	1.00	.50	.76	.65	.54
vectors	.50	1.00	.59	.51	.38
algebra	.76	.59	1.00	.76	.67
analysis	.65	.51	.76	1.00	.74
statistics	.54	.38	.67	.74	1.00

The 22 students of Table 3.1 actually each took five tests, **mechanics**, **vectors**, **algebra**, **analytics**, and **statistics**. Table 10.1 shows the sample correlation matrix and also its eigenvalues. The "eigenratio" statistic,

$$\hat{\theta} = \text{ largest eigenvalue/sum eigenvalues,} \tag{10.22}$$

measures how closely the five scores can be predicted by a single linear combination, essentially an IQ score for each student: $\hat{\theta} = 0.693$ here, indicating strong predictive power for the IQ score. How accurate is 0.693?

$B = 2000$ bootstrap replications (10.15) yielded bootstrap standard error estimate (10.16) $\widehat{\text{se}}_{\text{boot}} = 0.075$. (This was 10 times more bootstraps than necessary for $\widehat{\text{se}}_{\text{boot}}$, but will be needed for Chapter 11's bootstrap confidence interval calculations.) The jackknife (10.6) gave a bigger estimate, $\widehat{\text{se}}_{\text{jack}} = 0.083$.

Standard errors are usually used to suggest approximate confidence intervals, often $\hat{\theta} \pm 1.96\widehat{\text{se}}$ for 95% coverage. These are based on an assumption of normality for $\hat{\theta}$. The histogram of the 2000 bootstrap replications of $\hat{\theta}$, as seen in Figure 10.2, disabuses belief in even approximate normality. Compared with classical methods, a massive amount of computation has gone into the histogram, but this will pay off in Chapter 11 with more accurate confidence limits. We can claim a double reward here for bootstrap methods: much wider applicability and improved inferences. The bootstrap

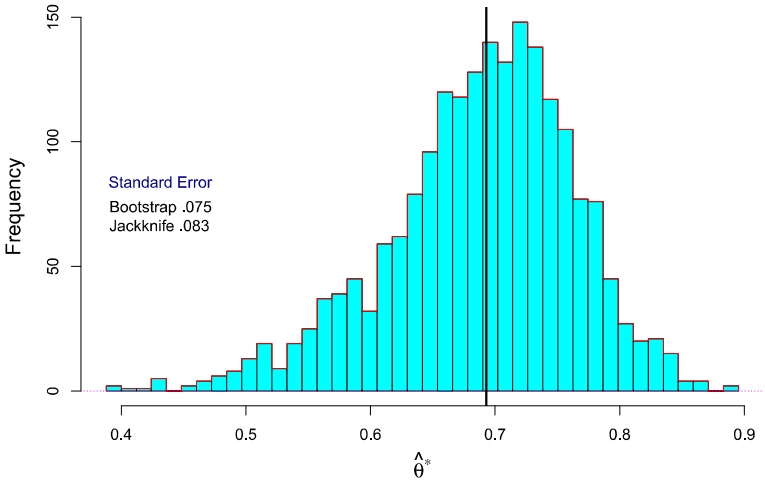

Figure 10.2 Histogram of $B = 2000$ bootstrap replications $\hat{\theta}^*$ for the eigenratio statistic (10.22) for the student score data. The vertical black line is at $\hat{\theta} = .693$. The long left tail shows that normality is a dangerous assumption in this case.

histogram—invisible to classical statisticians—nicely illustrates the advantages of computer-age statistical inference.

10.3 Resampling Plans

There is a second way to think about the jackknife and the bootstrap: as algorithms that reweight, or *resample*, the original data vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$. At the price of a little more abstraction, resampling connects the two algorithms and suggests a class of other possibilities.

A *resampling vector* $\boldsymbol{P} = (P_1, P_2, \ldots, P_n)'$ is by definition a vector of nonnegative weights summing to 1,

$$\boldsymbol{P} = (P_1, P_2, \ldots, P_n)' \qquad \text{with } P_i \geq 0 \text{ and } \sum_{i=1}^{n} P_i = 1. \tag{10.23}$$

That is, $\boldsymbol{P}$ is a member of the simplex $\mathcal{S}_n$ (5.39). Resampling plans operate by holding the original data set $\boldsymbol{x}$ fixed, and seeing how the statistic of interest $\hat{\theta}$ changes as the weight vector $\boldsymbol{P}$ varies across $\mathcal{S}_n$.

We denote the value of $\hat{\theta}$ for a vector putting weight P_i on x_i as

$$\hat{\theta}^* = S(\boldsymbol{P}), \tag{10.24}$$

the star notation now indicating any reweighting, not necessarily from bootstrapping; $\hat{\theta} = s(\boldsymbol{x})$ describes the behavior of $\hat{\theta}$ in the real world (10.17), while $\hat{\theta}^* = S(\boldsymbol{P})$ describes it in the resampling world. For the sample mean $s(\boldsymbol{x}) = \bar{x}$, we have $S(\boldsymbol{P}) = \sum_1^n P_i x_i$. The unbiased estimate of variance $s(\boldsymbol{x}) = \sum_i^n (x_i - \bar{x})^2/(n-1)$ can be seen to have

$$S(\boldsymbol{P}) = \frac{n}{n-1}\left[\sum_{i=1}^{n} P_i x_i^2 - \left(\sum_{i=1}^{n} P_i x_i\right)^2\right]. \tag{10.25}$$

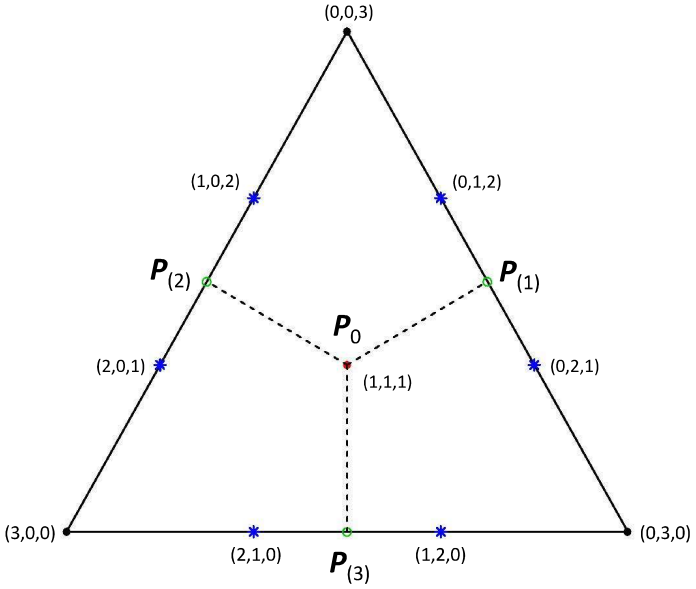

Figure 10.3 Resampling simplex for sample size $n = 3$. The center point is $\boldsymbol{P}_0$ (10.26); the green circles are the jackknife points $\boldsymbol{P}_{(i)}$ (10.28); triples indicate bootstrap resampling numbers (N_1, N_2, N_3) (10.29). The *bootstrap probabilities* are $6/27$ for $\boldsymbol{P}_0$, $1/27$ for each corner point, and $3/27$ for each of the six starred points.

Letting

$$\boldsymbol{P}_0 = (1, 1, \ldots, 1)'/n, \tag{10.26}$$

the resampling vector putting equal weight on each value x_i, we require in the definition of $S(\cdot)$ that

$$S(\boldsymbol{P}_0) = s(\boldsymbol{x}) = \hat{\theta}, \tag{10.27}$$

the original estimate. The ith jackknife value $\hat{\theta}_{(i)}$ (10.5) corresponds to resampling vector

$$\boldsymbol{P}_{(i)} = (1, 1, \ldots, 1, 0, 1, \ldots, 1)'/(n-1), \tag{10.28}$$

with 0 in the ith place. Figure 10.3 illustrates the resampling simplex $\mathcal{S}_3$ applying to sample size $n = 3$, with the center point being $\boldsymbol{P}_0$ and the open circles the three possible jackknife vectors $\boldsymbol{P}_{(i)}$.

With $n = 3$ sample points $\{x_1, x_2, x_3\}$ there are only 10 distinct bootstrap vectors (10.13), also shown in Figure 10.3. Let

$$N_i = \#\{x_j^* = x_i\}, \tag{10.29}$$

the number of bootstrap draws in $\boldsymbol{x}^*$ equaling x_i. The triples in the figure are (N_1, N_2, N_3), for example $(1, 0, 2)$ for $\boldsymbol{x}^*$ having x_1 once and x_3 twice.[4] The bootstrap resampling vectors are of the form

$$\boldsymbol{P}^* = (N_1, N_2, \ldots, N_n)'/n, \tag{10.30}$$

where the N_i are nonnegative integers summing to n. According to definition (10.13) of bootstrap sampling, the vector $\boldsymbol{N} = (N_1, N_2, \ldots, N_n)'$ follows a multinomial distribution (5.38) with n draws on n equally likely categories,

$$\boldsymbol{N} \sim \text{Mult}_n(n, \boldsymbol{P}_0). \tag{10.31}$$

This gives bootstrap probability (5.37)

$$\frac{n!}{N_1!N_2!\ldots N_n!}\,\frac{1}{n^n} \tag{10.32}$$

on $\boldsymbol{P}^*$ (10.30).

Figure 10.3 is misleading in that the jackknife vectors $\boldsymbol{P}_{(i)}$ appear only slightly closer to $\boldsymbol{P}_0$ than are the bootstrap vectors $\boldsymbol{P}^*$. As n grows large they are, in fact, an order of magnitude closer. Subtracting (10.26) from (10.28) gives Euclidean distance

$$\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\| = 1\Big/\sqrt{n(n-1)}. \tag{10.33}$$

[4] A hidden assumption of definition (10.24) is that $\hat{\theta} = s(\boldsymbol{x})$ has the same value for any permutation of $\boldsymbol{x}$, so for instance $s(x_1, x_3, x_3) = s(x_3, x_1, x_3) = S(1/3, 0, 2/3)$.

For the bootstrap, notice that N_i in (10.29) has a binomial distribution,

$$N_i \sim \text{Bi}\left(n, \frac{1}{n}\right), \tag{10.34}$$

with mean 1 and variance $(n-1)/n$. Then $P_i^* = N_i/n$ has mean and variance $(1/n, (n-1)/n^3)$. Adding over the n coordinates gives the expected root mean square distance for bootstrap vector $\boldsymbol{P}^*$,

$$\left(E\|\boldsymbol{P}^* - \boldsymbol{P}_0\|^2\right)^{1/2} = \sqrt{(n-1)/n^2}, \tag{10.35}$$

an order of magnitude $\sqrt{n}$ times further than (10.33).

The function $S(\boldsymbol{P})$ has approximate directional derivative

$$D_i = \frac{S(\boldsymbol{P}_{(i)}) - S(\boldsymbol{P}_0)}{\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\|} \tag{10.36}$$

in the direction from $\boldsymbol{P}_0$ toward $\boldsymbol{P}_{(i)}$ (measured along the dashed lines in Figure 10.3). D_i measures the slope of function $S(\boldsymbol{P})$ at $\boldsymbol{P}_0$, in the direction of $\boldsymbol{P}_{(i)}$. Formula (10.12) shows $\widehat{\text{se}}_{\text{jack}}$ as proportional to the root mean square of the slopes.

If $S(\boldsymbol{P})$ is a *linear* function of $\boldsymbol{P}$, as it is for the sample mean, it turns out that $\widehat{\text{se}}_{\text{jack}}$ equals $\widehat{\text{se}}_{\text{boot}}$ (except for the fudge factor $(n-1)/n$ in (10.6)). Most statistics are not linear, and then the local jackknife resamples may provide a poor approximation to the full resampling behavior of $S(\boldsymbol{P})$. This was the case at one point in Figure 10.1.

With only 10 possible resampling points $\boldsymbol{P}^*$, we can easily evaluate the *ideal* bootstrap standard error estimate

$$\widehat{\text{se}}_{\text{boot}} = \left[\sum_{k=1}^{10} p_k \left(\hat{\theta}^{*k} - \hat{\theta}^{*\cdot}\right)^2\right]^{1/2}, \quad \hat{\theta}^{*\cdot} = \sum_{k=1}^{10} p_k \hat{\theta}^{*k}, \tag{10.37}$$

with $\hat{\theta}^{*k} = S(\boldsymbol{P}^k)$ and p_k the probability from (10.32) (listed in Figure 10.3). This rapidly becomes impractical. The number of distinct bootstrap samples for n points turns out to be

$$\binom{2n-1}{n}. \tag{10.38}$$

For $n = 10$ this is already 92,378, while $n = 20$ gives 6.9×10^{10} distinct possible resamples. Choosing B vectors $\boldsymbol{P}^*$ at random, which is what algorithm (10.13)–(10.15) effectively is doing, makes the un-ideal bootstrap standard error estimate (10.16) almost as accurate as (10.37) for B as small as 200 or even less.

The luxury of examining the resampling surface provides a major advantage to modern statisticians, both in inference and methodology. A variety of other resampling schemes have been proposed, a few of which follow.

The Infinitesimal Jackknife

Looking at Figure 10.3 again, the vector

$$\boldsymbol{P}_i(\epsilon) = (1-\epsilon)\boldsymbol{P}_0 + \epsilon\boldsymbol{P}_{(i)} = \boldsymbol{P}_0 + \epsilon(\boldsymbol{P}_{(i)} - \boldsymbol{P}_0) \tag{10.39}$$

lies proportion ϵ of the way from $\boldsymbol{P}_0$ to $\boldsymbol{P}_{(i)}$. Then

$$\tilde{D}_i = \lim_{\epsilon\to 0} \frac{S\left(\boldsymbol{P}_i(\epsilon)\right) - S(\boldsymbol{P}_0)}{\epsilon\|\boldsymbol{P}_{(i)} - \boldsymbol{P}_0\|} \tag{10.40}$$

exactly defines the direction derivative at $\boldsymbol{P}_0$ in the direction of $\boldsymbol{P}_{(i)}$. The infinitesimal jackknife estimate of standard error is

$$\widehat{\text{se}}_{\text{IJ}} = \left(\sum_{i=1}^{n} \tilde{D}_i^2/n^2\right)^{1/2}, \tag{10.41}$$

usually evaluated numerically by setting ϵ to some small value in (10.40)–(10.41) (rather than $\epsilon = 1$ in (10.12)). We will meet the infinitesimal jackknife again in Chapters 17 and 20.

Multisample Bootstrap

The median difference between the `AML` and the `ALL` scores in Figure 1.4 is

$$\texttt{mediff} = 0.968 - 0.733 = 0.235. \tag{10.42}$$

How accurate is 0.235? An appropriate form of bootstrapping draws 25 times with replacement from the 25 `AML` patients, 47 times with replacement from the 47 `ALL` patients, and computes $\texttt{mediff}^*$ as the difference between the medians of the two bootstrap samples. (Drawing one bootstrap sample of size 72 from all the patients would result in random sample sizes for the $\texttt{AML}^*/\texttt{ALL}^*$ groups, adding inappropriate variability to the frequentist standard error estimate.)

A histogram of $B = 500$ $\texttt{mediff}^*$ values appears in Figure 10.4. They give $\widehat{\text{se}}_{\text{boot}} = 0.074$. The estimate (10.42) is 3.18 $\widehat{\text{se}}$ units above zero, agreeing surprisingly well with the usual two-sample t-statistic 3.01 (based on *mean* differences), and its permutation histogram Figure 4.3. Permutation testing can be considered another form of resampling.

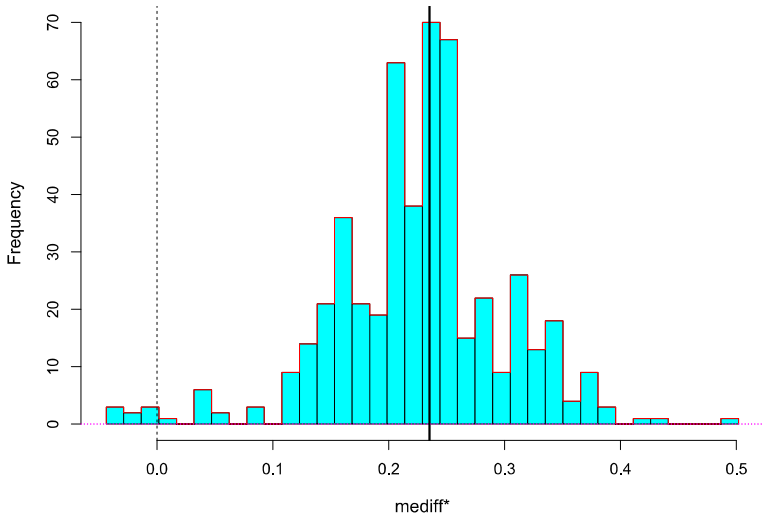

Figure 10.4 $B = 500$ bootstrap replications for the median difference between the `AML` and `ALL` scores in Figure 1.4, giving $\widehat{\text{se}}_{\text{boot}} = 0.074$. The observed value `mediff` $= 0.235$ (vertical black line) is more than 3 standard errors above zero.

Moving Blocks Bootstrap

Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, instead of being an iid sample (10.2), is a time series. That is, the x values occur in a meaningful order, perhaps with nearby observations highly correlated with each other. Let $\mathcal{B}_m$ be the set of contiguous blocks of length m, for example

$$\mathcal{B}_3 = \{(x_1, x_2, x_3), (x_2, x_3, x_4), \ldots, (x_{n-2}, x_{n-1}, x_n)\}. \tag{10.43}$$

Presumably, m is chosen large enough that correlations between x_i and x_j, $|j - i| > m$, are neglible. The moving block bootstrap first selects n/m blocks from $\mathcal{B}_m$, and assembles them in random order to construct a bootstrap sample $\boldsymbol{x}^*$. Having constructed B such samples, $\widehat{\text{se}}_{\text{boot}}$ is calculated as in (10.15)–(10.16).

The Bayesian Bootstrap

Let $G_1, G_2, \ldots, G_n$ be independent one-sided exponential variates (denoted Gam(1,1) in Table 5.1), each having density $\exp(-x)$ for $x > 0$.

The Bayesian bootstrap uses resampling vectors

$$\boldsymbol{P}^* = (G_1, G_2, \ldots, G_n) \Big/ \sum_1^n G_i. \tag{10.44}$$

It can be shown that $\boldsymbol{P}^*$ is then uniformly distributed over the resampling simplex $\mathcal{S}_n$; for $n = 3$, uniformly distributed over the triangle in Figure 10.3. Prescription (10.44) is motivated by assuming a Jeffreys-style uninformative prior distribution (Section 3.2) on the unknown distribution F (10.2).

Distribution (10.44) for $\boldsymbol{P}^*$ has mean vector and covariance matrix

$$\boldsymbol{P}^* \sim \left[\boldsymbol{P}_0, \frac{1}{n+1}\left(\text{diag}(\boldsymbol{P}_0) - \boldsymbol{P}_0\boldsymbol{P}_0'\right)\right]. \tag{10.45}$$

This is almost identical to the mean and covariance of bootstrap resamples $\boldsymbol{P}^* \sim \text{Mult}_n(n, \boldsymbol{P}_0)/n$,

$$\boldsymbol{P}^* \sim \left[\boldsymbol{P}_0, \frac{1}{n}\left(\text{diag}(\boldsymbol{P}_0) - \boldsymbol{P}_0\boldsymbol{P}_0'\right)\right], \tag{10.46}$$

(5.40). The Bayesian bootstrap and the ordinary bootstrap tend to agree, at least for smoothly defined statistics $\hat\theta^* = S(\boldsymbol{P}^*)$.

There was some Bayesian disparagement of the bootstrap when it first appeared because of its blatantly frequentist take on estimation accuracy. And yet connections like (10.45)–(10.46) have continued to pop up, as we will see in Chapter 13.

10.4 The Parametric Bootstrap

In our description (10.18) of bootstrap resampling,

$$\hat F \xrightarrow{\text{iid}} \boldsymbol{x}^* \longrightarrow \hat\theta^*, \tag{10.47}$$

there is no need to insist that $\hat F$ be the nonparametric MLE of F. Suppose we are willing to assume that the observed data vector $\boldsymbol{x}$ comes from a *parametric family* $\mathcal{F}$ as in (5.1),

$$\mathcal{F} = \left\{f_\mu(\boldsymbol{x}),\ \mu \in \Omega\right\}. \tag{10.48}$$

Let $\hat\mu$ be the MLE of μ. The *bootstrap parametric* resamples from $f_{\hat\mu}(\cdot)$,

$$f_{\hat\mu} \longrightarrow \boldsymbol{x}^* \longrightarrow \hat\theta^*, \tag{10.49}$$

and proceeds as in (10.14)–(10.16) to calculate $\widehat{\text{se}}_{\text{boot}}$.

As an example, suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample of size n from a normal distribution,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1), \qquad i = 1, 2, \ldots, n. \tag{10.50}$$

Then $\hat{\mu} = \bar{x}$, and a parametric bootstrap sample is $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$, where

$$x_i^* \overset{\text{iid}}{\sim} \mathcal{N}(\bar{x}, 1), \qquad i = 1, 2, \ldots, n. \tag{10.51}$$

More adventurously, if $\mathcal{F}$ were a family of time series models for $\boldsymbol{x}$, algorithm (10.49) would still apply (now without any iid structure): $\boldsymbol{x}^*$ would be a time series sampled from model $f_{\hat{\mu}}(\cdot)$, and $\hat{\theta}^* = s(\boldsymbol{x}^*)$ the resampled statistic of interest. B independent realizations $\boldsymbol{x}^{*b}$ would give $\hat{\theta}^{*b}$, $b = 1, 2, \ldots, B$, and $\widehat{\text{se}}_{\text{boot}}$ from (10.16).

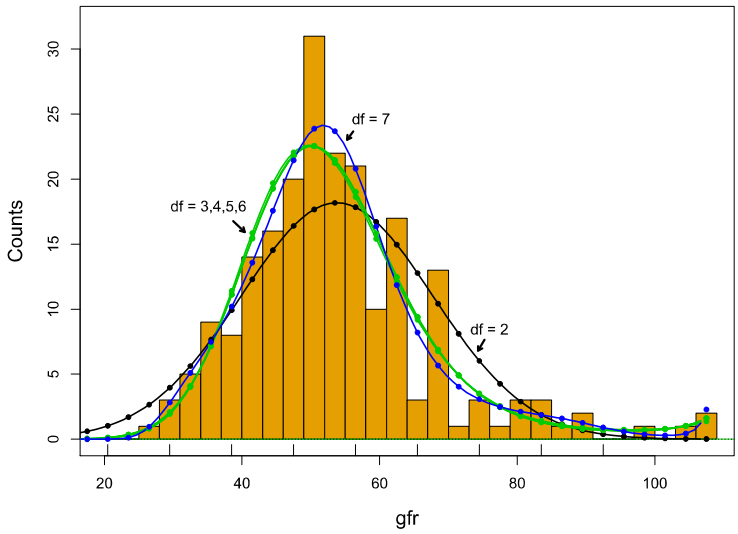

Figure 10.5 The `gfr` data of Figure 5.7 (histogram). Curves show the MLE fits from polynomial Poisson models, for degrees of freedom df $= 2, 3, \ldots, 7$. The points on the curves show the fits computed at the centers $x_{(j)}$ of the bins, with the responses being the counts in the bins. The dashes at the base of the plot show the nine `gfr` values appearing in Table 10.2.

As an example of parametric bootstrapping, Figure 10.5 expands the `gfr` investigation of Figure 5.7. In addition to the seventh-degree polynomial fit (5.62), we now show lower-degree polynomial fits for 2, 3, 4, 5,

and 6 degrees of freedom; df = 2 obviously gives a poor fit; df = 3, 4, 5, 6 give nearly identical curves; df = 7 gives only a slightly better fit to the raw data.

The plotted curves were obtained from the Poisson regression method used in Section 8.3, which we refer to as "Lindsey's method".

- The x-axis was partitioned into $K = 32$ bins, with endpoints 13, 16, 19, ..., 109, and centerpoints, say,

$$\boldsymbol{x}_{(\,)} = (x_{(1)}, x_{(2)}, \ldots, x_{(K)}), \tag{10.52}$$

  $x_{(1)} = 14.5$, $x_{(2)} = 17.5$, etc.
- Count vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$ was computed

$$y_k = \#\{x_i \text{ in bin}_k\} \tag{10.53}$$

  (so $\boldsymbol{y}$ gives the heights of the bars in Figure 10.5).
- An independent Poisson model was assumed for the counts,

$$y_k \overset{\text{ind}}{\sim} \text{Poi}(\mu_k) \qquad \text{for } k = 1, 2, \ldots, K. \tag{10.54}$$

- The parametric model of degree "df" assumed that the μ_k values were described by an exponential polynomial of degree df in the $x_{(k)}$ values,

$$\log(\mu_k) = \sum_{j=0}^{\text{df}} \beta_j x_{(k)}^j. \tag{10.55}$$

- The MLE $\hat{\boldsymbol{\beta}} = (\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_{\text{df}})$ in model (10.54)–(10.55) was found.[5]
- The plotted curves in Figure 10.5 trace the MLE values $\hat\mu_k$,

$$\log(\hat\mu_k) = \sum_{j=0}^{\text{df}} \hat\beta_j x_{(k)}^j. \tag{10.56}$$

How accurate are the curves? Parametric bootstraps were used to assess their standard errors. That is, Poisson resamples were generated according to

$$y_k^* \overset{\text{ind}}{\sim} \text{Poi}(\hat\mu_k) \qquad \text{for } k = 1, 2, \ldots, K, \tag{10.57}$$

and bootstrap MLE values $\hat\mu_k^*$ calculated as above, but now based on count vector $\boldsymbol{y}^*$ rather than $\boldsymbol{y}$. All of this was done $B = 200$ times, yielding bootstrap standard errors (10.16).

The results appear in Table 10.2, showing $\widehat{\text{se}}_{\text{boot}}$ for df = 2, 3, ..., 7

[5] A single R command, `glm(y~poly(x,df),family=poisson)` accomplishes this.

Table 10.2 *Bootstrap estimates of standard error for the* `gfr` *density. Poisson regression models* (10.54)–(10.55)*,* df $= 2, 3, \ldots, 7$*, as in Figure 10.5; each* $B = 200$ *bootstrap replications; nonparametric standard errors based on binomial bin counts.*

	Degrees of freedom						Nonparametric standard error
`gfr`	2	3	4	5	6	7	
20.5	.28	.07	.13	.13	.12	.05	.00
29.5	.65	.57	.57	.66	.74	1.11	1.72
38.5	1.05	1.39	1.33	1.52	1.72	1.73	2.77
47.5	1.47	1.91	2.12	1.93	2.15	2.39	4.25
56.5	1.57	1.60	1.79	1.93	1.87	2.28	4.35
65.5	1.15	1.10	1.07	1.31	1.34	1.27	1.72
74.5	.76	.61	.62	.68	.81	.71	1.72
83.5	.40	.30	.40	.38	.49	.68	1.72
92.5	.13	.20	.29	.29	.34	.46	.00

degrees of freedom evaluated at nine values of `gfr`. Variability generally increases with increasing df, as expected. Choosing a "best" model is a compromise between standard error and possible definitional bias as suggested by Figure 10.5, with perhaps df $= 3$ or 4, the winner.

If we kept increasing the degrees of freedom, eventually (at df $= 32$) we would exactly match the bar heights y_k in the histogram. At this point the parametric bootstrap would merge into the nonparametric bootstrap. "Nonparametric" is another name for "very highly parameterized." The huge sample sizes associated with modern applications have encouraged nonparametric methods, on the sometimes mistaken ground that estimation efficiency is no longer of concern. It is costly here, as the "nonparametric" column of Table 10.2 shows.[6]

Figure 10.6 returns to the student score eigenratio calculations of Figure 10.2. The solid histogram shows 2000 parametric bootstrap replications (10.49), with $f_{\hat{\mu}}$ the five-dimensional bivariate normal distribution $\mathcal{N}_5(\bar{x}, \hat{\Sigma})$. Here $\bar{x}$ and $\hat{\Sigma}$ are the usual MLE estimates for the expectation vector and covariance matrix based on the 22 five-component student score vectors. It is narrower than the corresponding nonparametric bootstrap histogram, with $\widehat{\text{se}}_{\text{boot}} = 0.070$ compared with the nonparametric estimate

[6] These are the binomial standard errors $[y_k(n - y_k)/n]^{1/2}$, $n = 211$. The nonparametric results look much more competitive when estimating cdf's rather than densities.

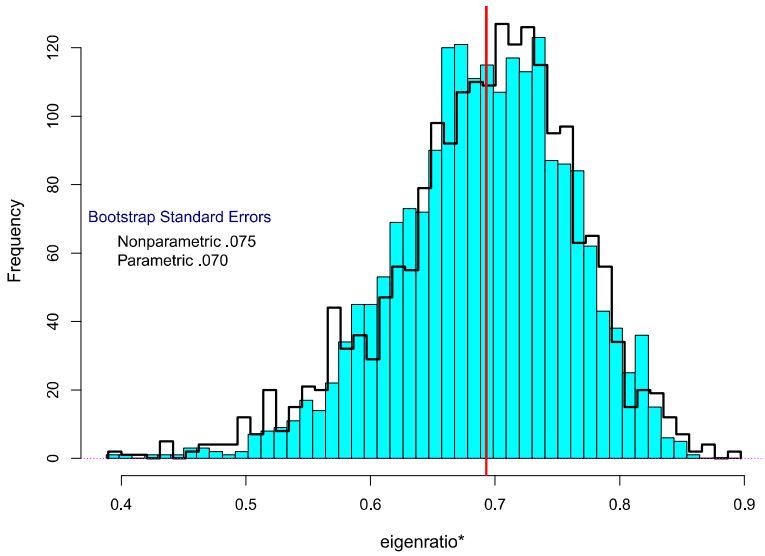

Figure 10.6 Eigenratio example, student score data. *Solid histogram* $B = 2000$ parametric bootstrap replications $\hat{\theta}^*$ from the five-dimensional normal MLE; *line histogram* the 2000 nonparametric replications of Figure 10.2. MLE $\hat{\theta} = .693$ is vertical red line.

0.075. (Note the different histogram bin limits from Figure 10.2, changing the details of the nonparametric histogram.)

Parametric families act as *regularizers*, smoothing out the raw data and de-emphasizing outliers. In fact the student score data is not a good candidate for normal modeling, having at least one notable outlier,[7] casting doubt on the smaller estimate of standard error.

The classical statistician could only imagine a mathematical device that given any statistic $\hat{\theta} = s(\boldsymbol{x})$ would produce a formula for its standard error, as formula (1.2) does for $\bar{x}$. The electronic computer *is* such a device. As harnessed by the bootstrap, it automatically produces a numerical estimate of standard error (though not a formula), with no further cleverness required. Chapter 11 discusses a more ambitious substitution of computer power for mathematical analysis: the bootstrap computation of confidence intervals.

[7] As revealed by examining scatterplots of the five variates taken two at a time. Fast and painless plotting is another advantage for twenty-first-century data analysts.

10.5 Influence Functions and Robust Estimation

The sample mean played a dominant role in classical statistics for reasons heavily weighted toward mathematical tractibility. Beginning in the 1960s, an important counter-movement, *robust estimation*, aimed to improve upon the statistical properties of the mean. A central element of that theory, the *influence function*, is closely related to the jackknife and infinitesimal jackknife estimates of standard error.

We will only consider the case where $\mathcal{X}$, the sample space, is an interval of the real line. The unknown probability distribution F yielding the iid sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ in (10.2) is now the cdf of a density function $f(x)$ on $\mathcal{X}$. A parameter of interest, i.e., a function of F, is to be estimated by the plug-in principle, $\hat{\theta} = T(\hat{F})$, where, as in Section 10.2, $\hat{F}$ is the empirical probability distribution putting probability $1/n$ on each sample point x_i. For the mean,

$$\theta = T(F) = \int_{\mathcal{X}} x f(x)\, dx \quad \text{and} \quad \hat{\theta} = T\left(\hat{F}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{10.58}$$

(In Riemann–Stieltjes notation, $\theta = \int x dF(x)$ and $\hat{\theta} = \int x d\hat{F}(x)$.)

The influence function of $T(F)$, evaluated at point x in $\mathcal{X}$, is defined to be

$$\mathrm{IF}(x) = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F + \epsilon\delta_x\right) - T(F)}{\epsilon}, \tag{10.59}$$

where δ_x is the "one-point probability distribution" putting probability 1 on x. In words, $\mathrm{IF}(x)$ measures the differential effect of modifying F by putting additional probability on x. For the mean $\theta = \int x f(x) dx$ we calculate that

$$\mathrm{IF}(x) = x - \theta. \tag{10.60}$$

†5 A fundamental theorem† says that $\hat{\theta} = T(\hat{F})$ is approximately

$$\hat{\theta} \doteq \theta + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(x_i), \tag{10.61}$$

with the approximation becoming exact as n goes to infinity. This implies that $\hat{\theta} - \theta$ is, approximately, the mean of the n iid variates $\mathrm{IF}(x_i)$, and that the variance of $\hat{\theta}$ is approximately

$$\mathrm{var}\left\{\hat{\theta}\right\} \doteq \frac{1}{n}\,\mathrm{var}\left\{\mathrm{IF}(x)\right\}, \tag{10.62}$$

var{IF(x)} being the variance of IF(x) for any one draw of x from F. For the sample mean, using (10.60) in (10.62) gives the familiar equality

$$\text{var}\{\bar{x}\} = \frac{1}{n}\,\text{var}\{x\}. \tag{10.63}$$

The sample mean suffers from an *unbounded* influence function (10.60), which grows ever larger as x moves farther from θ. This makes $\bar{x}$ unstable against heavy-tailed densities such as the Cauchy (4.39). Robust estimation theory seeks estimators $\hat{\theta}$ of bounded influence, that do well against heavy-tailed densities without giving up too much efficiency against light-tailed densities such as the normal. Of particular interest have been the trimmed mean and its close cousin the winsorized mean.

Let $x^{(\alpha)}$ denote the 100αth percentile of distribution F, satisfying $F(x^{(\alpha)}) = \alpha$ or equivalently

$$\alpha = \int_{-\infty}^{x^{(\alpha)}} f(x)\,dx. \tag{10.64}$$

The *αth trimmed mean of F*, $\theta_{\text{trim}}(\alpha)$, is defined as

$$\theta_{\text{trim}}(\alpha) = \frac{1}{1-2\alpha}\int_{x^{(\alpha)}}^{x^{(1-\alpha)}} xf(x)\,dx, \tag{10.65}$$

the mean of the central $1-2\alpha$ portion of F, trimming off the lower and upper α portions. This is not the same as the *αth winsorized mean* $\theta_{\text{wins}}(\alpha)$,

$$\theta_{\text{wins}}(\alpha) = \int_{\mathcal{X}} W(x)f(x)\,dx, \tag{10.66}$$

where

$$W(x) = \begin{cases} x^{(\alpha)} & \text{if } x \le x^{(\alpha)} \\ x & \text{if } x^{(\alpha)} \le x \le x^{(1-\alpha)} \\ x^{(1-\alpha)} & \text{if } x \ge x^{(1-\alpha)}; \end{cases} \tag{10.67}$$

$\theta_{\text{trim}}(\alpha)$ removes the outer portions of F, while $\theta_{\text{wins}}(\alpha)$ moves them into $x^{(\alpha)}$ or $x^{(1-\alpha)}$. In practice, empirical versions $\hat{\theta}_{\text{trim}}(\alpha)$ and $\hat{\theta}_{\text{wins}}(\alpha)$ are used, substituting the empirical density $\hat{f}$, with probability $1/n$ at each x_i, for f.

There turns out to be an interesting relationship between the two: the influence function of $\theta_{\text{trim}}(\alpha)$ is a function of $\theta_{\text{wins}}(\alpha)$,

$$\text{IF}_\alpha(x) = \frac{W(x) - \theta_{\text{wins}}(\alpha)}{1-2\alpha}. \tag{10.68}$$

This is pictured in Figure 10.7, where we have plotted empirical influence

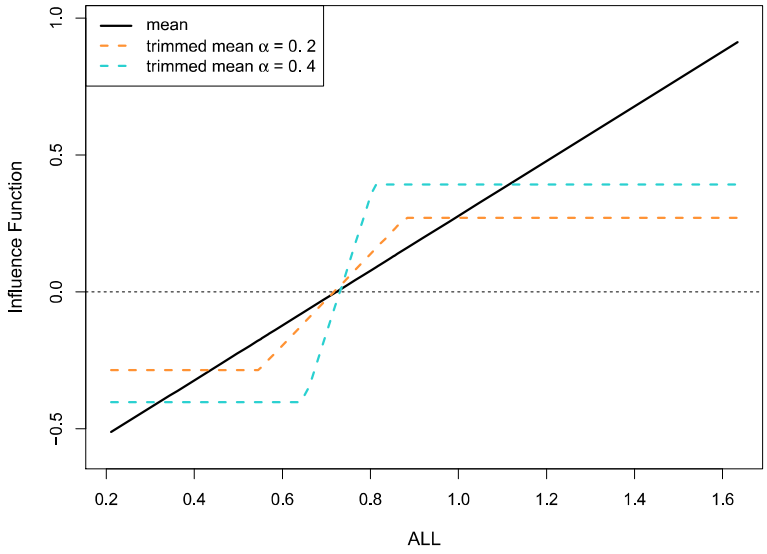

Figure 10.7 Empirical influence functions for the 47 leukemia **ALL** scores of Figure 1.4. The two dashed curves are $\mathrm{IF}_\alpha(x)$ for the trimmed means (10.68), for $\alpha = 0.2$ and $\alpha = 0.4$. The solid curve is $\mathrm{IF}(x)$ for the sample mean $\bar{x}$ (10.60).

functions (plugging in $\hat{F}$ for F in definition (10.59)) relating to the 47 leukemia **ALL** scores of Figure 1.4: $\mathrm{IF}_{0.2}(x)$ and $\mathrm{IF}_{0.4}(x)$ are plotted, along with $\mathrm{IF}_0(x)$ (10.60), that is, for the mean.

Table 10.3 *Trimmed means and their bootstrap standard deviations for the 47 leukemia* **ALL** *scores of Figure 1.4; $B = 1000$ bootstrap replications for each trim value. The last column gives empirical influence function estimates of the standard error, which are also the infinitesimal jackknife estimates* (10.41)*. These fail for the median.*

	Trim	Trimmed mean	Bootstrap sd	(IFse)
Mean	.0	.752	.040	(.040)
	.1	.729	.038	(.034)
	.2	.720	.035	(.034)
	.3	.725	.044	(.044)
	.4	.734	.047	(.054)
Median	.5	.733	.053	

The upper panel of Figure 1.4 shows a moderately heavy right tail for the **ALL** distribution. Would it be more efficient to estimate the center of the distribution with a trimmed mean rather than $\bar{x}$? The bootstrap provides an answer: $\widehat{\text{se}}_{\text{boot}}$ (10.16) was calculated for $\bar{x}$ and $\hat{\theta}_{\text{trim}}(\alpha)$, $\alpha = 0.1, 0.2, 0.3, 0.4$, and 0.5, the last being the sample median. It appears that $\hat{\theta}_{\text{trim}}(0.2)$ is moderately better than $\bar{x}$. This brings up an important question discussed in Chapter 20: if we use something like Table 10.3 to select an estimator, how does the selection process affect the accuracy of the resulting estimate?

We might also use the square root of formula (10.62) to estimate the standard errors of the various estimators, plugging in the empirical influence function for $\text{IF}(x)$. This turns out to be the same as using the infinitesimal jackknife (10.41). These appear in the last column of Table 10.3. Predictably, this approach fails for the sample median, whose influence function is a square wave, sharply discontinuous at the median θ,

$$\text{IF}(x) = \pm 1 \big/ (2f(\theta)) . \tag{10.69}$$

Robust estimation offers a nice illustration of statistical progress in the computer age. Trimmed means go far back into the classical era. Influence functions are an insightful inferential tool for understanding the tradeoffs in trimmed mean estimation. And finally the bootstrap allows easy assessment of the accuracy of robust estimation, including some more elaborate ones not discussed here.

10.6 Notes and Details

Quenouille (1956) introduced what is now called the jackknife estimate of bias. Tukey (1958) realized that Quenouille-type calculations could be repurposed for nonparametric standard-error estimation, inventing formula (10.6) and naming it "the jackknife," as a rough and ready tool. Miller's important 1964 paper, "A trustworthy jackknife," asked when formula (10.6) could be trusted. (Not for the median.)

The bootstrap (Efron, 1979) began as an attempt to better understand the jackknife's successes and failures. Its name celebrates Baron Munchausen's success in pulling himself up by his own bootstraps from the bottom of a lake. Burgeoning computer power soon overcame the bootstrap's main drawback, prodigous amounts of calculation, propelling it into general use. Meanwhile, 1000+ theoretical papers were published asking when the bootstrap itself could be trusted. (Most but not all of the time in common practice.)

A main reference for the chapter is Efron's 1982 monograph *The Jackknife, the Bootstrap and Other Resampling Plans*. Its Chapter 6 shows the equality of three nonparametric standard error estimates: Jaeckel's (1972) infinitesimal jackknife (10.41); the empirical influence function estimate, based on (10.62); and what is known as the nonparametric delta method.

Bootstrap Packages

Various bootstrap packages in **R** are available on the CRAN contributed-packages web site, **bootstrap** being an ambitious one. Algorithm 10.1 shows a simple **R** program for nonparametric bootstrapping. Aside from bookkeeping, it's only a few lines long.

Algorithm 10.1 R PROGRAM FOR THE NONPARAMETRIC BOOTSTRAP.

```
Boot <- function (x, B, func, ...){
 # x is data vector or matrix (with each row a case)
 # B is number of bootstrap replications
 # func is R function that inputs a data vector or
 # matrix and returns a numeric number or vector
 # ... other arguments for func
    x <- as.matrix(x)
    n <- nrow(x)
    f0=func(x,...) # get size of output
    fmat <- matrix(0,length(f0),B)
    for (b in 1:B) {
         i=sample(1:n, n, replace = TRUE)
         fmat[,b] <- func(x[i, ],...)
     }
    drop(fmat)
}
```

†1 [p. 158] *The jackknife standard error.* The 1982 monograph also contains Efron and Stein's (1981) result on the bias of the jackknife variance estimate, the square of formula (10.6): modulo certain sample size considerations, the expectation of the jackknife variance estimate is biased upward for the true variance.

For the sample mean $\bar{x}$, the jackknife yields exactly the usual variance estimate (1.2), $\sum_i (x_i - \bar{x})^2/(n(n-1))$, while the ideal bootstrap estimate ($B \to \infty$) gives

$$\sum_{i=1}^{n} (x_i - \bar{x})^2/n^2. \tag{10.70}$$

As with the jackknife, we could append a fudge factor to get perfect agreement with (1.2), but there is no real gain in doing so.

†2 [p. 161] *Bootstrap sample sizes.* Let $\widehat{\text{se}}_B$ indicate the bootstrap standard error estimate (10.16) based on B replications, and $\widehat{\text{se}}_\infty$ the "ideal bootstrap," $B \to \infty$. In any actual application, there are diminishing returns from increasing B past a certain point, because $\widehat{\text{se}}_\infty$ is itself a statistic whose value varies with the observed sample $\boldsymbol{x}$ (as in (10.70)), leaving an irreducible remainder of randomness in any standard error estimate. Section 6.4 of Efron and Tibshirani (1993) shows that $B = 200$ will almost always be plenty (for standard errors, but not for bootstrap confidence intervals, Chapter 11). Smaller numbers, 25 or even less, can still be quite useful in complicated situations where resampling is expensive. An early complaint, "Bootstrap estimates are random," is less often heard in an era of frequent and massive simulations.

†3 [p. 161] *The Bayesian bootstrap.* Rubin (1981) suggested the Bayesian bootstrap (10.44). Section 10.6 of Efron (1982) used (10.45)–(10.46) as an objective Bayes justification for what we will call the percentile-method bootstrap confidence intervals in Chapter 12.

†4 [p. 161] *Jackknife-after-bootstrap.* For the eigenratio example displayed in Figure 10.2, $B = 2000$ nonparametric bootstrap replications gave $\widehat{\text{se}}_{\text{boot}} = 0.075$. How accurate is this value? Bootstrapping the bootstrap seems like too much work, perhaps 200 times 2000 resamples. It turns out, though, that we can use the jackknife to estimate the variability of $\widehat{\text{se}}_{\text{boot}}$ based on just the original 2000 replications.

Now the deleted sample estimate in (10.6) is $\widehat{\text{se}}_{\text{boot}(i)}$. The key idea is to consider those bootstrap samples $\boldsymbol{x}^*$ (10.13), among the original 2000, that *do not include the point* x_i. About 37% of the original B samples will be in this subset. Section 19.4 of Efron and Tibshirani (1993) shows that applying definition (10.16) to this subset gives $\widehat{\text{se}}_{\text{boot}(i)}$. For the estimate of Figure 10.2, the jackknife-after-bootstrap calculations gave $\widehat{\text{se}}_{\text{jack}} = 0.022$ for $\widehat{\text{se}}_{\text{boot}} = 0.075$. In other words, 0.075 isn't very accurate, which is to be expected for the standard error of a complicated statistic estimated from only $n = 22$ observations. An infinitesimal jackknife version of this technique will play a major role in Chapter 20.

†5 [p. 174] *A fundamental theorem.* Tukey can justly be considered the founding father of robust statistics, his 1960 paper being especially influential. Huber's celebrated 1964 paper brought the subject into the realm of high-concept mathematical statistics. *Robust Statistics: The Approach Based on Influence Functions*, the 1986 book by Hampel *et al.*, conveys the breadth of a subject only lightly scratched in our Section 10.5. Hampel (1974)

introduced the influence function as a statistical tool. Boos and Serfling (1980) verified expression (10.62). Qualitative notions of robustness, more than specific theoretical results, have had a continuing influence on modern data analysis.

11

Bootstrap Confidence Intervals

The jackknife and the bootstrap represent a different use of modern computer power: rather than extending classical methodology—from ordinary least squares to generalized linear models, for example—they extend the reach of classical inference.

Chapter 10 focused on standard errors. Here we will take up a more ambitious inferential goal, the bootstrap automation of confidence intervals. The familiar *standard intervals*

$$\hat{\theta} \pm 1.96\,\widehat{\text{se}}, \tag{11.1}$$

for approximate 95% coverage, are immensely useful in practice but often not very accurate. If we observe $\hat{\theta} = 10$ from a Poisson model $\hat{\theta} \sim \text{Poi}(\theta)$, the standard 95% interval (3.8, 16.2) (using $\widehat{\text{se}} = \hat{\theta}^{1/2}$) is a mediocre approximation to the exact interval[1]

$$(5.1, 17.8). \tag{11.2}$$

Standard intervals (11.1) are symmetric around $\hat{\theta}$, this being their main weakness. Poisson distributions grow more variable as θ increases, which is why interval (11.2) extends farther to the right of $\hat{\theta} = 10$ than to the left. Correctly capturing such effects in an automatic way is the goal of bootstrap confidence interval theory.

11.1 Neyman's Construction for One-Parameter Problems

The student score data of Table 3.1 comprised $n = 22$ pairs,

$$x_i = (m_i, v_i), \qquad i = 1, 2, \ldots, 22, \tag{11.3}$$

[1] Using the Neyman construction of Section 11.1, as explained there; see also Table 11.2 in Section 11.4.

where m_i and v_i were student i's scores on the "mechanics" and "vectors" tests. The sample correlation coefficient $\hat{\theta}$ between m_i and v_i was computed to be

$$\hat{\theta} = 0.498. \tag{11.4}$$

Question: What can we infer about the true correlation θ between m and v? Figure 3.2 displayed three possible Bayesian answers. Confidence intervals provide the frequentist solution, by far the most popular in applied practice.

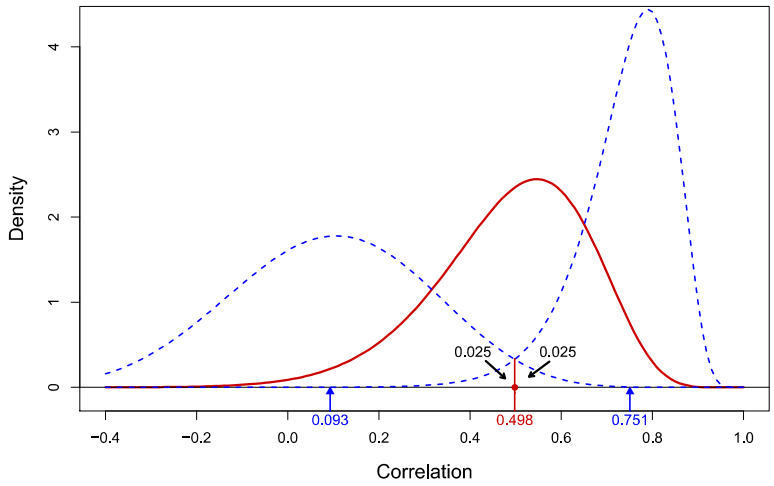

Figure 11.1 The solid curve is the normal correlation coefficient density $f_{\hat{\theta}}(r)$ (3.11) for $\hat{\theta} = 0.498$, the MLE estimate for the student score data; $\hat{\theta}(\text{lo}) = 0.093$ and $\hat{\theta}(\text{up}) = 0.751$ are the endpoints of the 95% confidence interval for θ, with corresponding densities shown by dashed curves. These yield tail areas 0.025 at $\hat{\theta}$ (11.6).

Suppose, first, that we assume a bivariate normal model (5.12) for the pairs (m_i, v_i). In that case the probability density $f_\theta(\hat{\theta})$ for sample correlation $\hat{\theta}$ given true correlation θ has known form (3.11). The solid curve in Figure 11.1 graphs f for $\theta = 0.498$, that is, for θ set equal to the observed value $\hat{\theta}$. In more careful notation, the curve graphs $f_{\hat{\theta}}(r)$ as a function of the dummy variable[2] r taking values in $[-1, 1]$.

[2] This is an example of a parametric bootstrap distribution (10.49), here with $\hat{\mu}$ being $\hat{\theta}$.