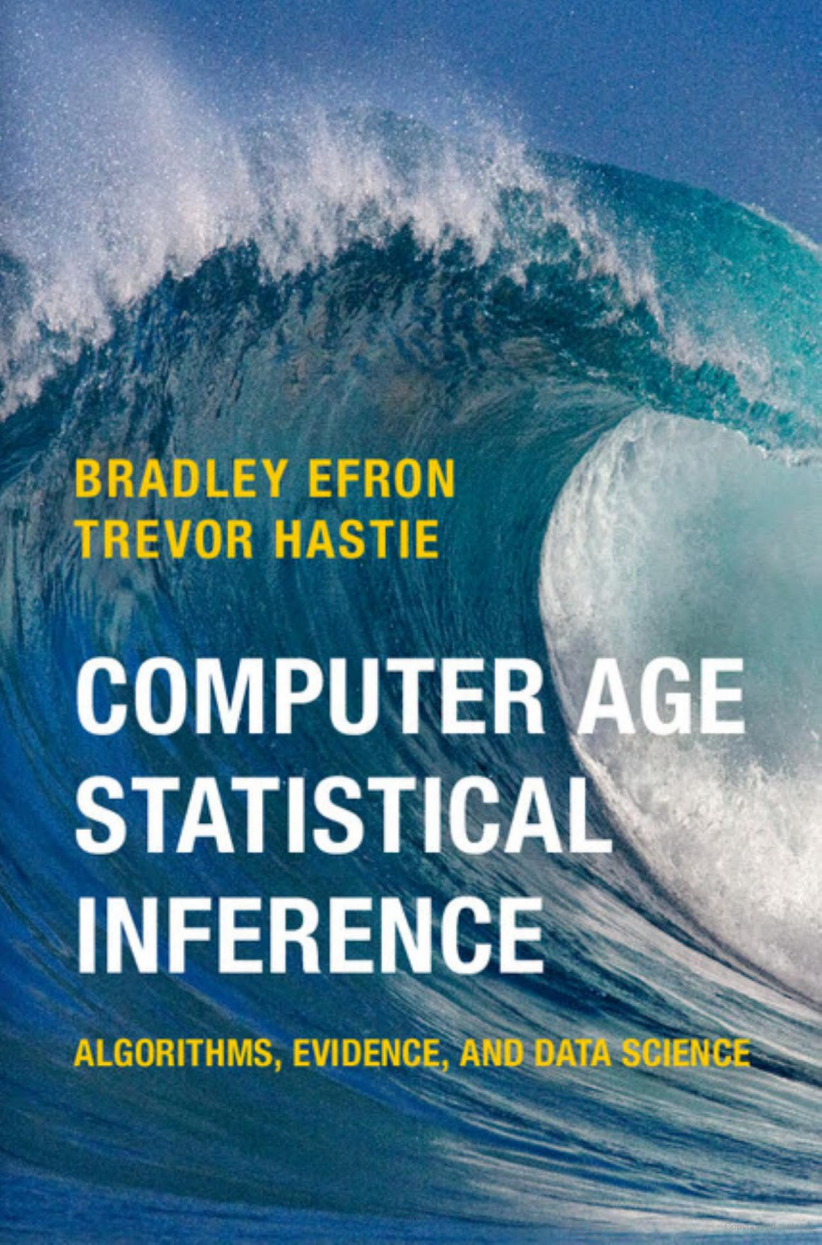
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE STATISTICAL INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE

Computer Age Statistical Inference

Algorithms, Evidence, and Data Science

BRADLEY EFRON
Stanford University, California

TREVOR HASTIE
Stanford University, California

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107149892

First published 2016
8th printing 2018

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-14989-2 Hardback

Contents

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

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

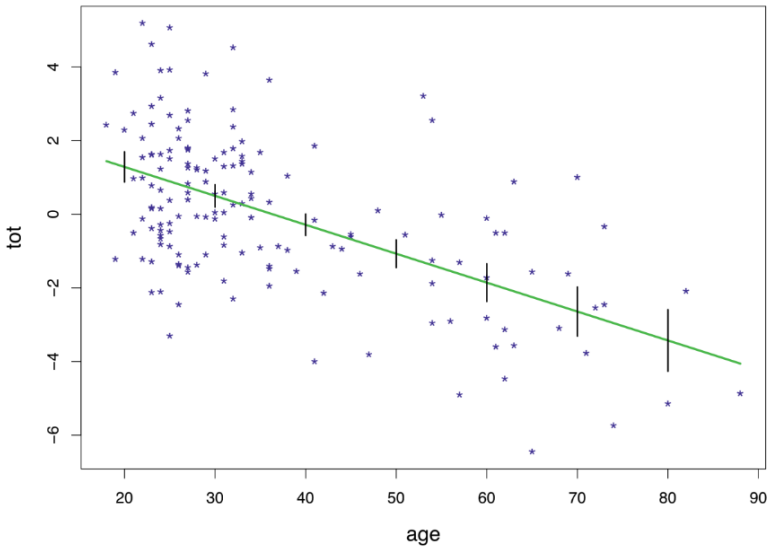

Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ±2 standard errors at selected values of `age`.

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

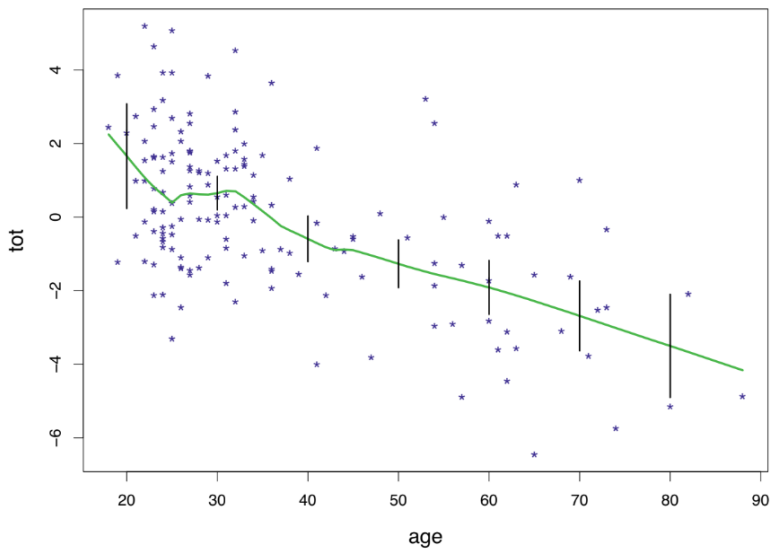

Figure 1.2 Local polynomial `lowess(x,y,1/3)` fit to the kidney-fitness data, with ±2 bootstrap standard deviations.

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.

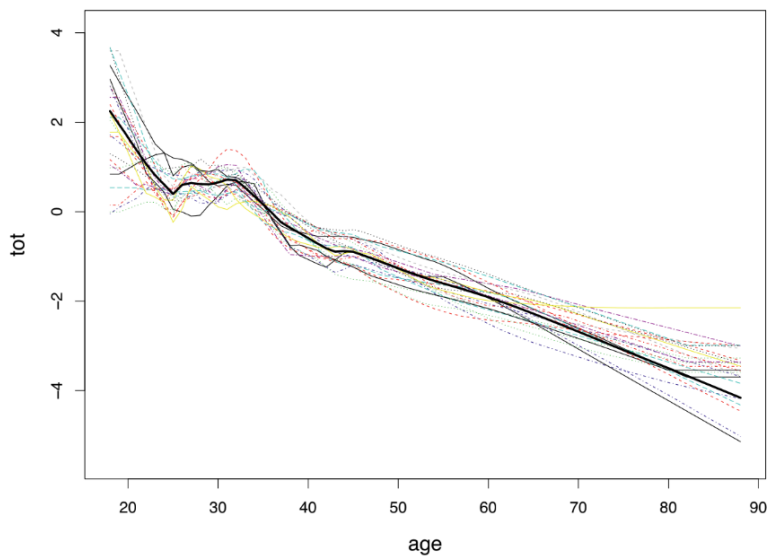

Figure 1.3 25 bootstrap replications of `lowess(x,y,1/3)`.

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.

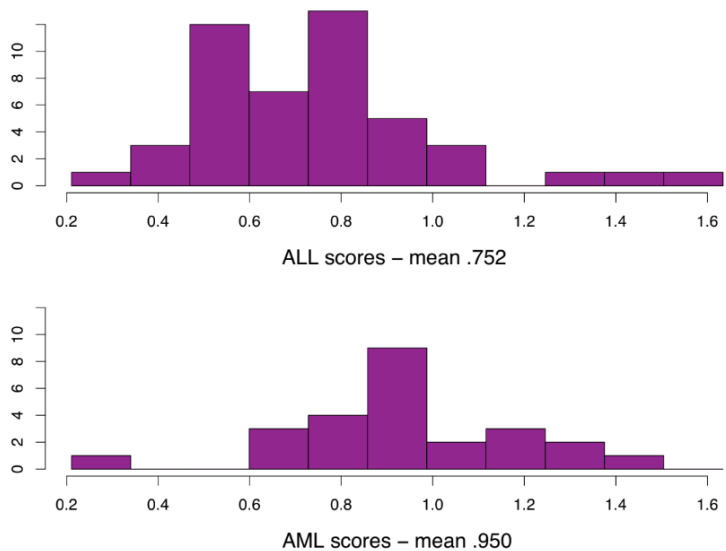

Figure 1.4 Scores for gene 136, leukemia data. Top `ALL` ($n = 47$), bottom `AML` ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

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

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

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

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

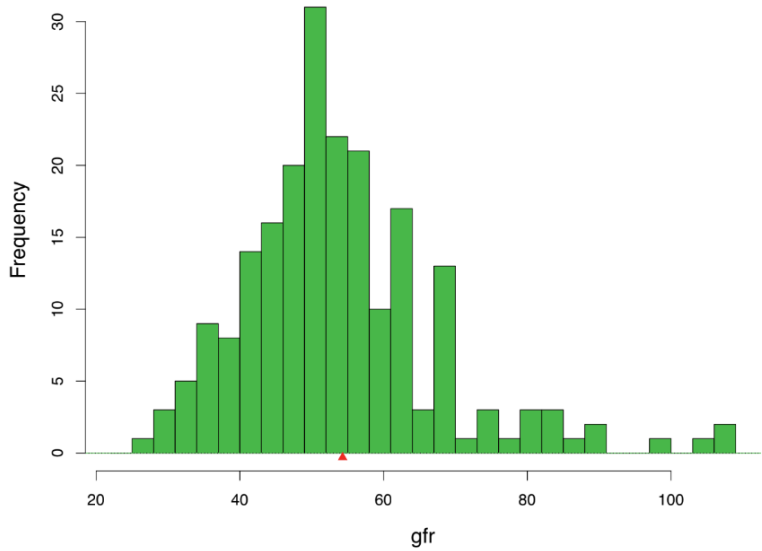

Figure 2.1 Glomerular filtration rates for 211 kidney patients; mean 54.25, standard error .95.

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 $\text{var}_F(X)$, the variance of a single X drawn from F,

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

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

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

Plugging formula (2.9) into (2.8) gives $\widehat{\text{se}}$ (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

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

with $\widehat{\text{se}}$ 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 $[p(1-p)/n]^{1/2}$.

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]

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

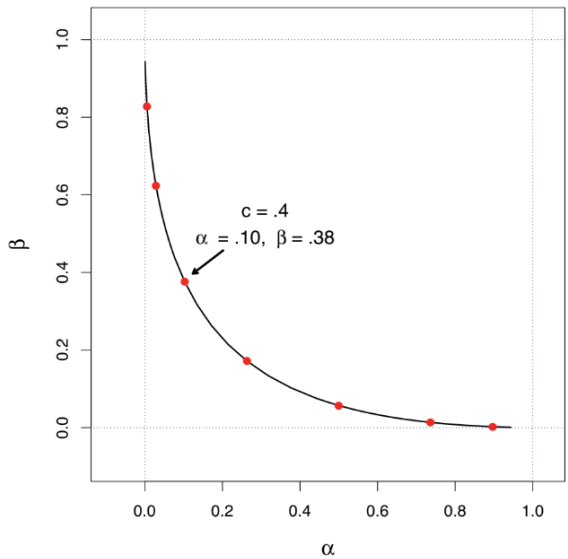

Figure 2.2 Neyman–Pearson alpha–beta curve for $f_0 \sim \mathcal{N}(0, 1)$, $f_1 \sim \mathcal{N}(.5, 1)$, and sample size $n = 10$. Red dots correspond to cutoffs $c = .8, .6, .4, \ldots, -.4$.

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."

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

$$\frac{g(\text{Identical} \mid \text{Same})}{g(\text{Fraternal} \mid \text{Same})} = \frac{g(\text{Identical})}{g(\text{Fraternal})} \cdot \frac{f_{\text{Identical}}(\text{Same})}{f_{\text{Fraternal}}(\text{Same})} = \frac{1/3}{2/3} \cdot \frac{1}{1/2} = 1. \tag{3.9}$$

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.

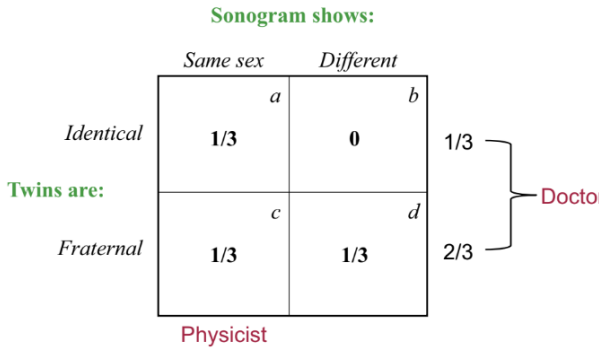

Figure 3.1 Analyzing the twins problem.

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.

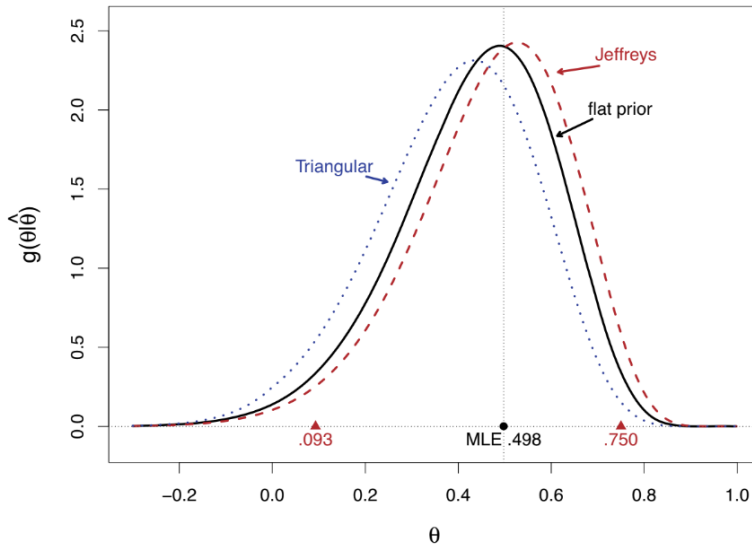

Figure 3.2 Student scores data; posterior density of correlation θ for three possible priors.

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 †2 approximate standard deviation

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

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

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).

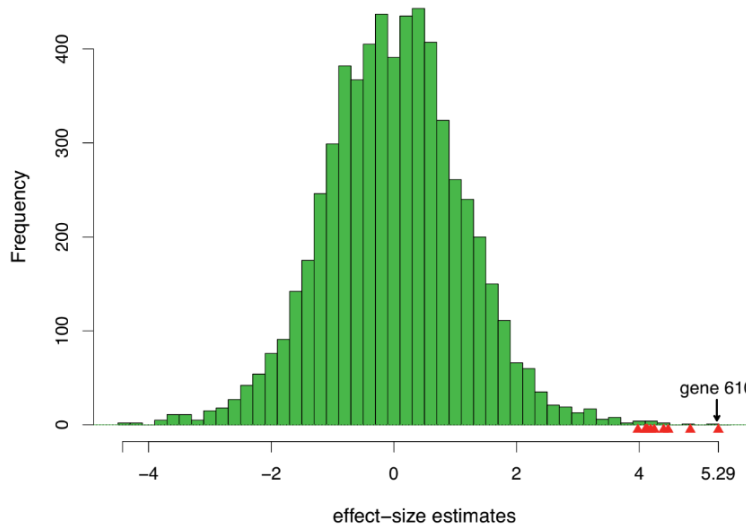

Figure 3.4 Unbiased effect-size estimates for 6033 genes, prostate cancer study. The estimate for gene 610 is $x_{610} = 5.29$. What is its effect size?

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 †5 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).

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.

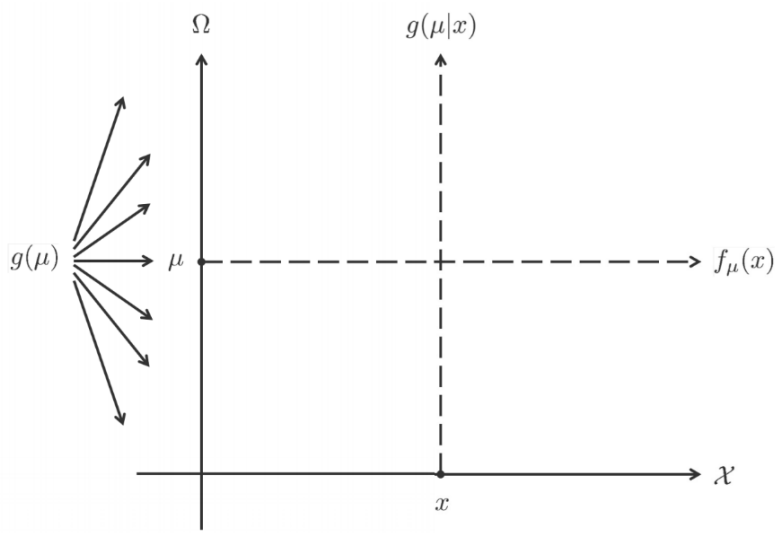

Figure 3.5 Bayesian inference proceeds vertically, given x; frequentist inference proceeds horizontally, given μ.

- 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 $\tilde{f}_{\tilde{\mu}}(x)$

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 $\boldsymbol{x}$.

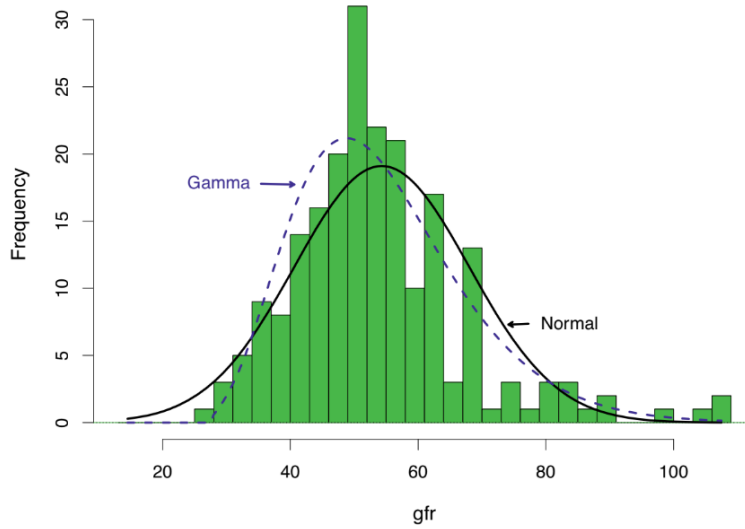

Figure 4.1 Glomerular filtration data of Figure 2.1 and two maximum-likelihood density estimates, normal (solid black), and gamma (dashed blue).

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 minimization, 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 $1/\mathcal{I}_\theta$,

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

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,

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

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

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

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

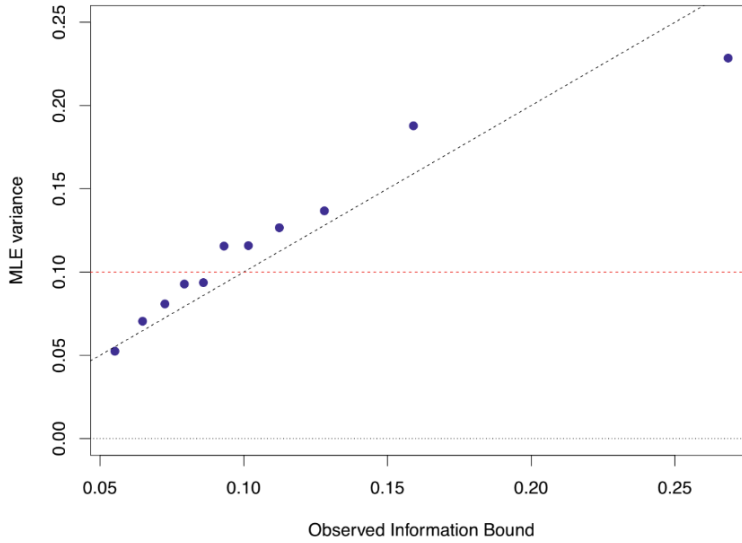

Figure 4.2 Conditional variance of MLE for Cauchy samples of size 20, plotted versus the observed information bound $1/I(\boldsymbol{x})$. Observed information bounds are grouped by quantile intervals for variance calculations (in percentages): (0–5), (5–15), . . . , (85–95), (95–100). The broken red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

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

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 3.13, with two-sided significance level 0.0025 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,

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

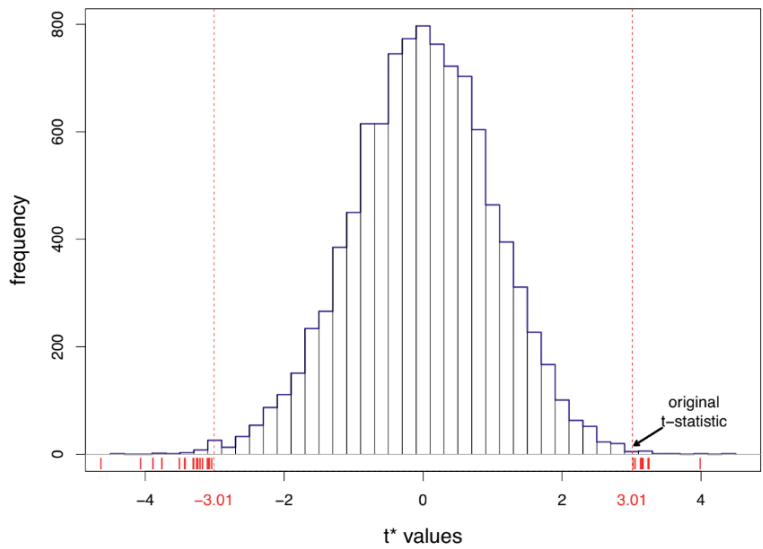

Figure 4.3 10,000 permutation t^*-values for testing **ALL** vs **AML**, for gene 136 in the **leukemia** data of Figure 1.3. Of these, 26 t^*-values (red ticks) exceeded in absolute value the observed t-statistic 3.01, giving permutation significance level 0.0026.

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, remarkably close to the normal-theory significance level 0.0025. (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.

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),

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

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 $\text{Poi}(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}$	$\{0, 1, \ldots\}$	$\mu > 0$	μ μ
Binomial $\text{Bi}(n, \pi)$	$\frac{n!}{x!(n-x)!}\,\pi^x(1-\pi)^{n-x}$	$\{0, 1, \ldots, n\}$	$0 < \pi < 1$	$n\pi$ $n\pi(1-\pi)$
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.

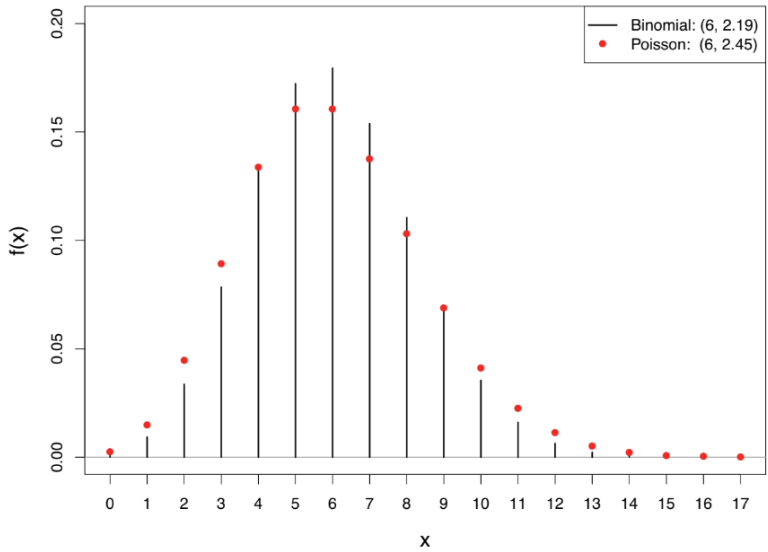

Figure 5.1 Comparison of the binomial distribution Bi(30, 0.2) (black lines) with the Poisson Poi(6) (red dots). In the legend we show the mean and standard deviation for each distribution.

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-

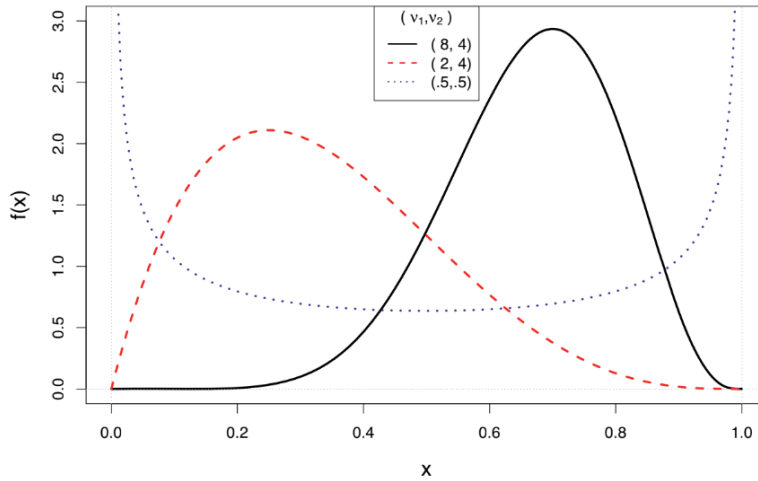

Figure 5.2 Three beta densities, with (ν_1, ν_2) indicated.

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.