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 $\boldsymbol{x}$.

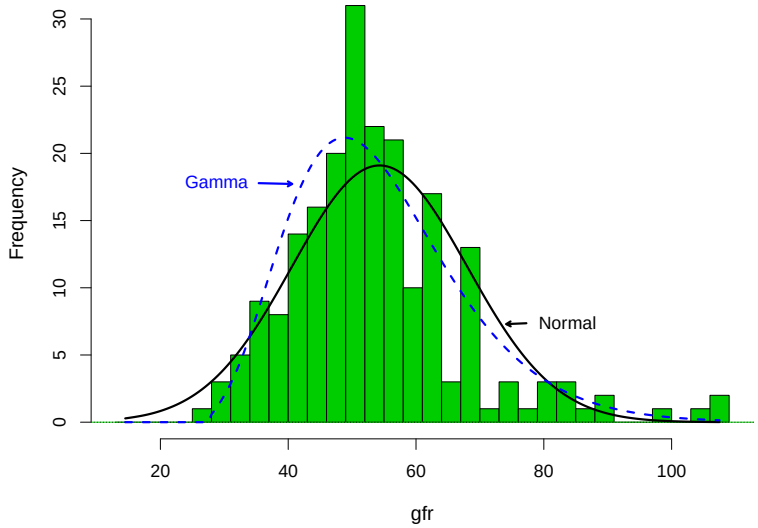

Figure 4.1 Glomerular filtration data of Figure 2.1 and two maximum-likelihood density estimates, normal (solid black), and gamma (dashed blue).

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

$$f_\mu(x) = \frac{(x-\lambda)^{\nu-1}}{\sigma^2\Gamma(\nu)}\, e^{-\frac{x-\lambda}{\sigma}} \qquad \text{(for } x \geq \lambda, 0 \text{ otherwise).} \tag{4.7}$$

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

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

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

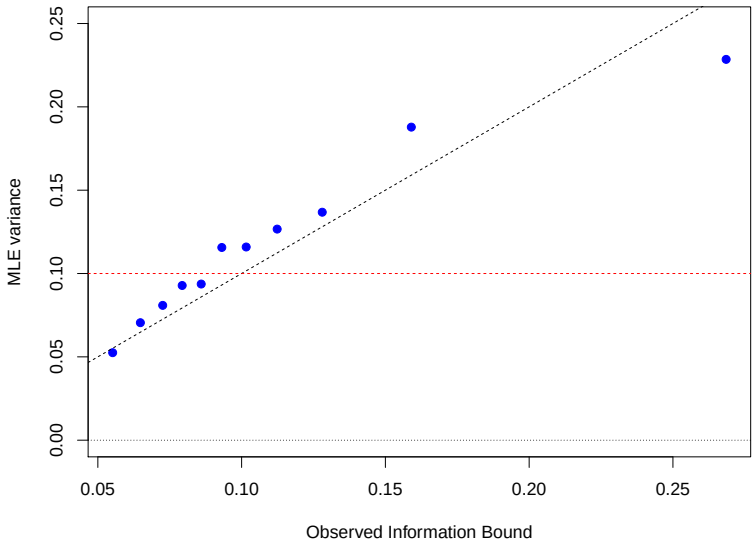

Figure 4.2 Conditional variance of MLE for Cauchy samples of size 20, plotted versus the observed information bound $1/I(\boldsymbol{x})$. Observed information bounds are grouped by quantile intervals for variance calculations (in percentages): (0–5), (5–15), . . . , (85–95), (95–100). The broken red 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 3.13, 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,

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

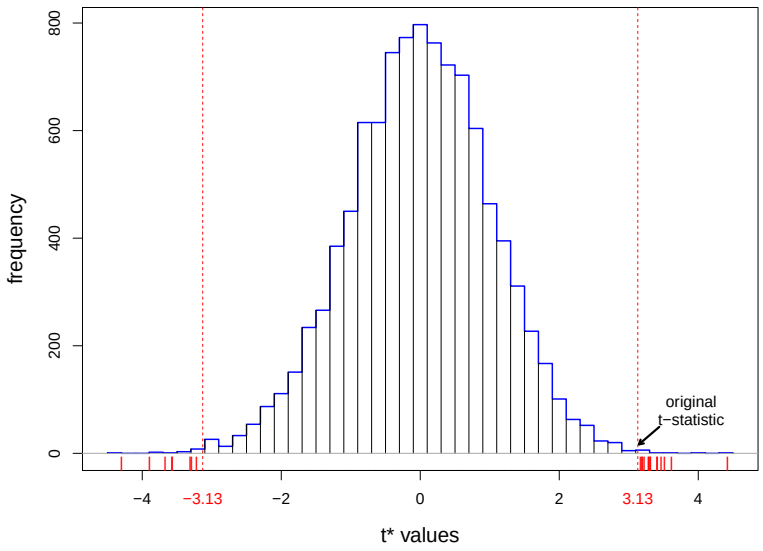

Figure 4.3 10,000 permutation t^*-values for testing **ALL** vs **AML**, for gene 136 in the **leukemia** data of Figure 1.3. Of these, 22 t^*-values (red ticks) exceeded in absolute value the observed t-statistic 3.13, giving permutation significance level 0.0022.

Figure 4.3 shows the histogram of $B = 10{,}000$ t_i^* values for the gene 136 data in Figure 1.3: 22 of these exceeded $t = 3.13$ in absolute value, yielding significance level 0.0022 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.

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

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

Here $\boldsymbol{\mathcal{X}}$ 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.