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

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

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

Perhaps the most familiar case is the normal family

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

by rewriting (3.5) as

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

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

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

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

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

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

3.1 Two Examples

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

In this situation μ, the unknown parameter (or "state of nature") is either *Identical* or *Fraternal* with prior probability 1/3 or 2/3; X, the possible sonogram results for twin births, is either *Same Sex* or *Different Sexes*, and $x =$ *Same Sex* was observed. (We can ignore sex since that does not affect the calculation.) A crucial fact is that identical twins are always same-sex while fraternals have probability 0.5 of same or different, so *Same Sex* in the sonogram is twice as likely if the twins are Identical. Applying Bayes' rule in ratio form (3.8) answers the physicist's question:

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

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

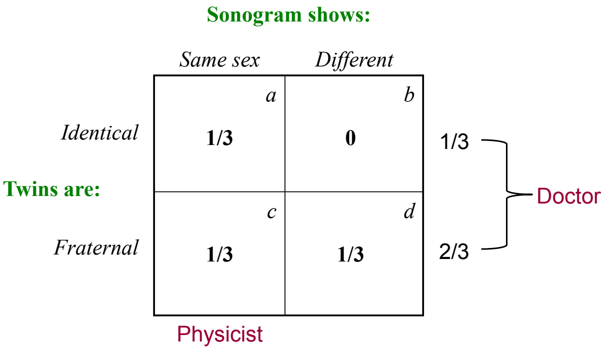

Figure 3.1 Analyzing the twins problem.

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

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

[4] They turned out to be Fraternal.

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

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

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

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
`mechanics`	7	44	49	59	34	46	0	32	49	52	44	36	42	5	22	18	41	48	31	42	46	63
`vectors`	51	69	41	70	42	40	40	45	57	64	61	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 $\hat\theta = 0.498$,

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

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.

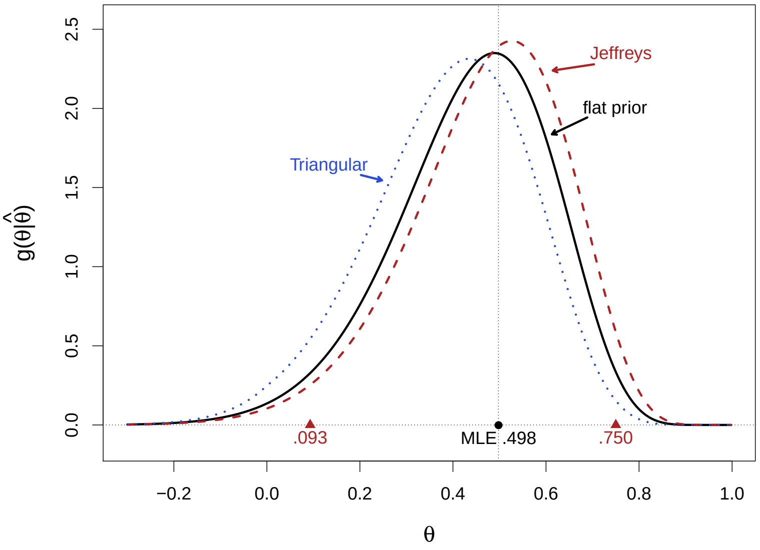

Figure 3.2 Student scores data; posterior density of correlation θ for three possible priors.

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,

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

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

Jeffreys' prior,

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

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

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

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

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

3.2 Uninformative Prior Distributions

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

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

Venn (of the Venn diagram) in the 1860s, and Fisher in the 1920s, attacking the routine use of Bayes' theorem, pointed out that Laplace's principle could not be applied consistently. In the student correlation example, for instance, a uniform prior distribution for θ would not be uniform if we changed parameters to $\gamma = e^{\theta}$; posterior probabilities such as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Multiparameter probability families, Chapter 4, make everything more difficult. Suppose, for instance, the statistician observes 10 independent versions of the normal model (3.2), with possibly different values of μ,

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

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

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

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

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

3.3 Flaws in Frequentist Inference

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

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

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

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

A Bayesian statistician comes to the meter reader's rescue. For any prior density $g(\mu)$, the posterior density $g(\mu|\boldsymbol{x}) = g(\mu) f_\mu(\boldsymbol{x})/f(\boldsymbol{x})$, where $\boldsymbol{x}$ is the vector of 12 measurements, depends only on the data $\boldsymbol{x}$ actually observed, and *not on other potential data sets* $\boldsymbol{X}$ *that might have been seen*. The flat Jeffreys' prior $g(\mu)$ = constant yields posterior expectation $\bar{x} = 92$ for μ, irrespective of whether or not the glitch would have affected readings above 100.

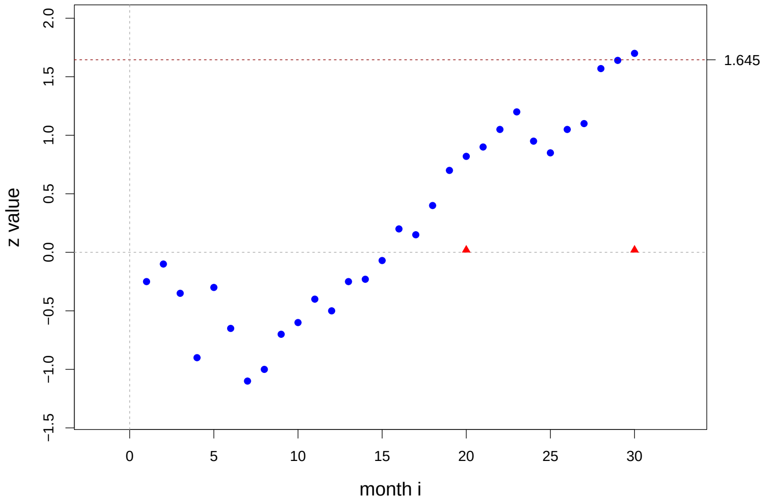

Figure 3.3 Z-values against null hypothesis $\mu = 0$ for months 1 through 30.

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

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

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

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

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

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

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

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

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

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

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).

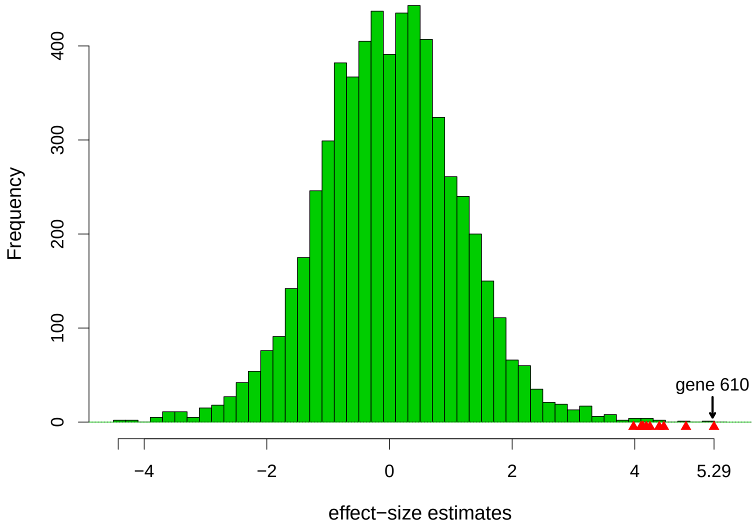

Figure 3.4 Unbiased effect-size estimates for 6033 genes, prostate cancer study. The estimate for gene 610 is $x_{610} = 5.29$. What is its effect size?

The lenient nature of Bayesian inference can look less benign in multiparameter settings. Figure 3.4 concerns a prostate cancer study comparing 52 patients with 50 healthy controls. Each man had his genetic activity measured for a panel of $N = 6033$ genes. A statistic x was computed for each gene,[5] comparing the patients with controls, say† †4

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

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

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

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

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

3.4 A Bayesian/Frequentist Comparison List

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

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

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

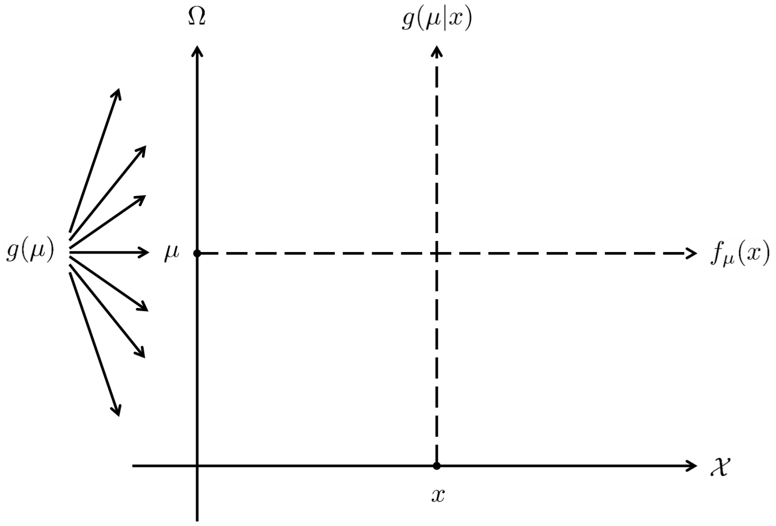

Figure 3.5 Bayesian inference proceeds vertically, given x; frequentist inference proceeds horizontally, given μ.

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

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

- Having chosen $g(\mu)$, only a single probability distribution $g(\mu|x)$ is in play for Bayesians. Frequentists, by contrast, must struggle to balance the behavior of $t(x)$ over a family of possible distributions, since μ in Figure 3.5 is unknown. The growing popularity of Bayesian applications (usually begun with uninformative priors) reflects their simplicity of application and interpretation.
- The simplicity argument cuts both ways. The Bayesian essentially bets it all on the choice of his or her prior being correct, or at least not harmful. Frequentism takes a more defensive posture, hoping to do well, or at least not poorly, whatever μ might be.
- A Bayesian analysis answers *all* possible questions at once, for example, estimating $E\{\text{gfr}\}$ or $\Pr\{\text{gfr} < 40\}$ or anything else relating to Figure 2.1. Frequentism focuses on the problem at hand, requiring different estimators for different questions. This is more work, but allows for more intense inspection of particular problems. In situation (2.9) for example, estimators of the form
$$\sum (x_i - \bar{x})^2/(n-c) \tag{3.29}$$
might be investigated for different choices of the constant c, hoping to reduce expected mean-squared error.
- The simplicity of the Bayesian approach is especially appealing in dynamic contexts, where data arrives sequentially and updating one's beliefs is a natural practice. Bayes' rule was used to devastating effect before the 2012 US presidential election, updating sequential polling results to correctly predict the outcome in all 50 states. Bayes' theorem is an excellent tool in general for combining statistical evidence from disparate sources, the closest frequentist analog being maximum likelihood estimation.
- In the absence of genuine prior information, a whiff of subjectivity[6] hangs over Bayesian results, even those based on uninformative priors. Classical frequentism claimed for itself the high ground of scientific objectivity, especially in contentious areas such as drug testing and approval, where skeptics as well as friends hang on the statistical details.

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

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

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

3.5 Notes and Details

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

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

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

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

satisfies

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

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

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

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

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

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