6

Empirical Bayes

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

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

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

6.1 Robbins' Formula

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

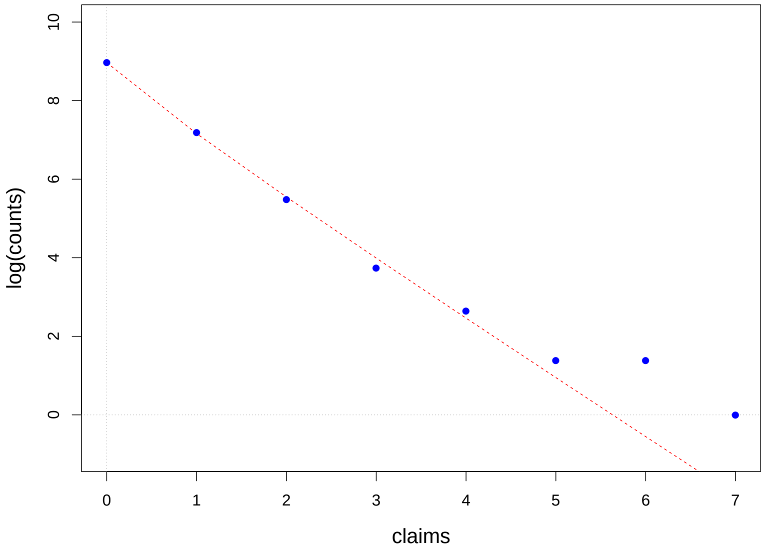

Figure 6.1 Auto accident data; log(counts) vs claims for 9461 auto insurance policies. The dashed line is a gamma MLE fit.

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

6.2 The Missing-Species Problem

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

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

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

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

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

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

The entries in Table 6.2 are

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

the number of species trapped exactly x times each.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

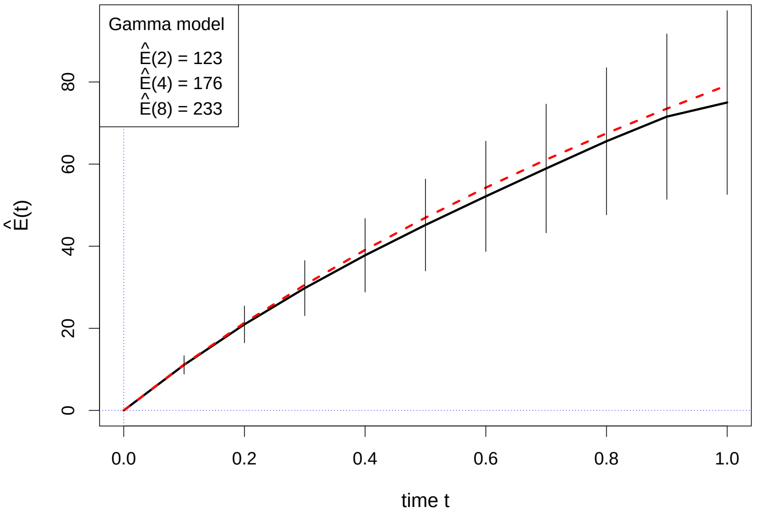

Figure 6.2 Butterfly data; expected number of new species in t units of additional trapping time. Nonparametric fit (solid) ± 1 standard deviation; gamma model (dashed).

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

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

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

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

Employing formulas (6.19) and (6.21) gives

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

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

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

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

The numerator has expectation

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

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

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

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

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

For the Shakespeare data,

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

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

All of this seems to live in the rarefied world of mathematical abstraction, but in fact some previously unknown Shakespearean work *might* have been discovered in 1985. A short poem, "Shall I die?," was found in the archives of the Bodleian Library and, controversially, attributed to Shakespeare by some but not all experts.

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

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

and a prediction from (6.19) of

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

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

6.3 A Medical Example

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

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

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

"positive" meaning malignant. The ratios

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

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

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

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

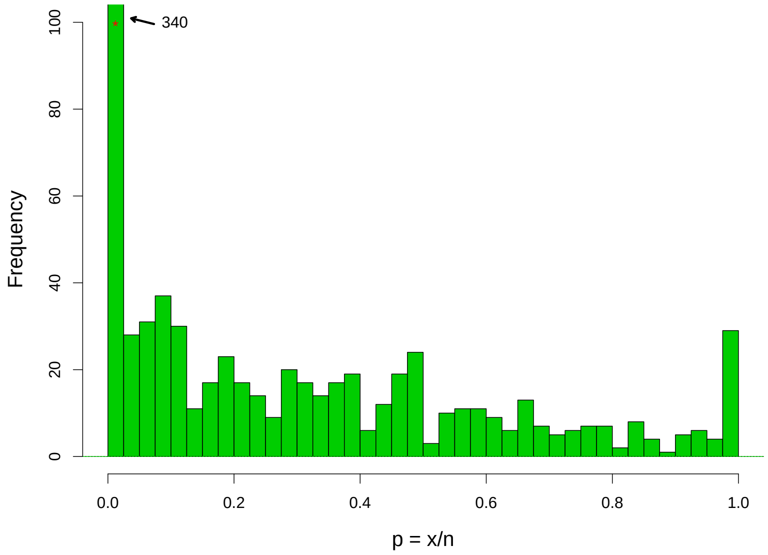

Figure 6.3 Nodes study; ratio $p = x/n$ for 844 patients; $n =$ number of nodes removed, $x =$ number positive.

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

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

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

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

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

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

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

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

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

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

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

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

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

The maximum likelihood estimate of α is the maximizer

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

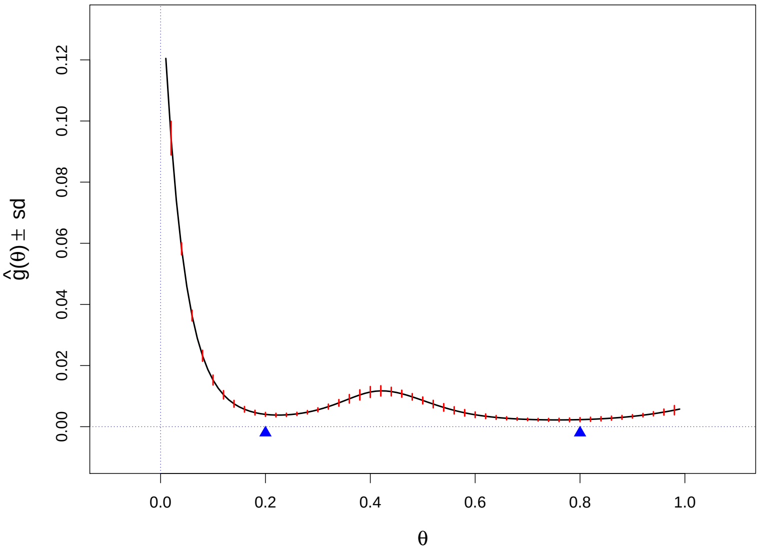

Figure 6.4 Estimated prior density $g(\theta)$ for the nodes study; 59% of patients have $\theta \leq 0.2$, 7% have $\theta \geq 0.8$.

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

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

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

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

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

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

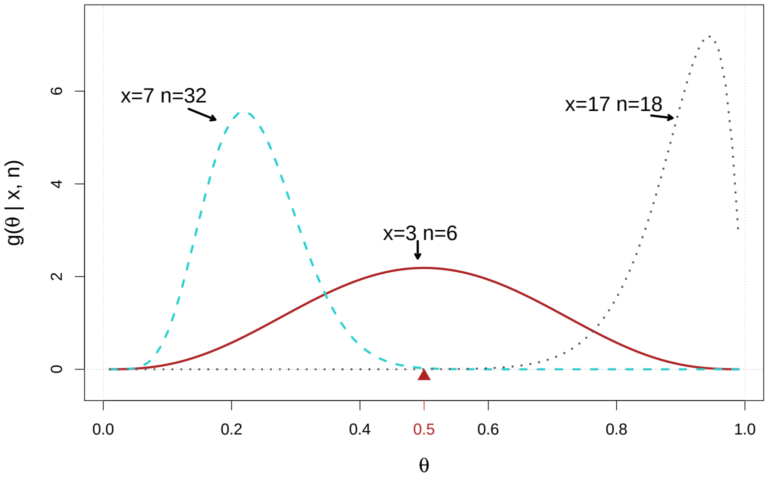

Figure 6.5 Empirical Bayes posterior densities of θ for three patients, given x = number of positive nodes, n = number of nodes.

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

6.4 Indirect Evidence 1

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

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

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

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

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

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

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

6.5 Notes and Details

Robbins (1956) introduced the term "empirical Bayes" as well as rule (6.7) as part of a general theory of empirical Bayes estimation. 1956 was also the publication year for Good and Toulmin's solution (6.19) to the missing-species problem. Good went out of his way to credit his famous Bletchley colleague Alan Turing for some of the ideas. The auto accident data is taken from Table 3.1 of Carlin and Louis (1996), who provide a more complete discussion. Empirical Bayes estimates such as 11430 in (6.25) do not depend on independence among the "species," but accuracies such as ±178 do; and similarly for the error bars in Figures 6.2 and 6.4.

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

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

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

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

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

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

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

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

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

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

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

$$y_x \overset{\text{ind}}{\sim} \text{Poi}(e_x), \qquad \text{for } x = 0, 1, 2, \dots, \tag{6.48}$$

in notation (6.17). Formula (6.19) and $\text{var}\{y_x\} = e_x$ then give

$$\text{var}\left\{\hat{E}(t)\right\} = \sum_{x \geq 1} e_x t^{2x}. \tag{6.49}$$

Substituting y_x for e_x produces (6.21). Section 11.5 of Efron (2010) shows that (6.49) is an upper bound on $\text{var}\{\hat{E}(t)\}$ if N is considered fixed rather than Poisson.

†3 [p. 81] *Formula* (6.23). Combining the case $x = 1$ in (6.17) with (6.15) yields

$$E(t) = \frac{e_1 \left[\int_0^\infty e^{-\theta} g(\theta)\, d\theta - \int_0^\infty e^{-\theta(1+t)} g(\theta)\, d\theta\right]}{\int_0^\infty \theta e^{-\theta} g(\theta)\, d\theta}. \tag{6.50}$$

Substituting the gamma prior (6.8) for $g(\theta)$, and using (6.45) three times, gives formula (6.23).