17

Principles and pathology of orthodox statistics

> The development of our theory beyond this point, as a practical statistical theory, involves . . . all the complexities of the use, either of Bayes' law on the one hand, or of those terminological tricks in the theory of likelihood on the other, which seem to avoid the necessity for the use of Bayes' law, but which in reality transfer the responsibility for its use to the working statistician, or the person who ultimately employs his results.
>
> *Norbert Wiener (1948)*

To the best of our knowledge, Norbert Wiener never actually applied Bayes' theorem in a published work; yet he perceived the logical necessity of its use as soon as one builds beyond the sampling distributions involved in his own statistical work. In the present chapter we examine some of the consequences of failing to use Bayesian methods in some very simple problems, where the paradoxes of Chapter 15 never arise.

In Chapter 16 we noted that the orthodox objections to Bayesian methods were always philosophical or ideological in nature, never examining the actual results that they give, and we expressed astonishment that mathematically competent persons would use such arguments. In order to give a fair comparison, we need to adopt the opposite tactic here, and concentrate on the demonstrable facts that orthodoxians never mention. Since Bayesian methods have been so egregiously misrepresented in the orthodox literature throughout our lifetimes, we must lean over backwards to avoid misrepresenting orthodox methods now; whenever an orthodox method does yield a satisfactory result in some problem, we shall acknowledge that fact, and we shall not deplore its use merely on ideological grounds. On the other hand, when a common orthodox procedure leads to a result that insults our intelligence, we shall not hesitate to complain about it.

Our present goal is to understand the following. *In what circumstances, and in what ways, do the orthodox results differ from the Bayesian results*? *What are the pragmatic consequences of this in real applications*? The theorems of Richard Cox provide all the ideology we need, and all of our pragmatic comparisons only confirm, in many different contexts, what those theorems lead us to expect.

17.1 Information loss

It is not easy to cover all this ground, because orthodox statistics is not a coherent body of theory that could be confirmed or refuted by a single analysis. It is a loose collection of independent *ad hoc* devices, invented and advocated by many different people on many different intuitive grounds; and they are often in sharp disagreement with each other.

But one can see generally, once and for all, when and why orthodox methods, quite aside from their failure to use prior information, must also waste some of the information in the data. Consider estimation of a parameter θ from a data set $D \equiv \{x_1, \ldots, x_n\}$ represented by a point in R^n. Orthodoxy requires us to choose a single estimator $b(D) \equiv b(x_1, \ldots, x_n)$ *before we have seen the data*, and then use only $b(x)$ for the estimation. Now, specifying the observed numerical value of $b(x)$ locates the sample on a manifold (subspace of R^n) of dimension $(n - 1)$. Specifying the actual data set D tells us that, and also where on the manifold we are. If position on the manifold is irrelevant to θ, then $b(D)$ is a sufficient statistic for θ, and unless there are further circumstances, such as highly cogent prior information, the orthodox method will be satisfactory pragmatically whatever its proclaimed rationale. Otherwise, specifying D conveys additional information about θ that is not conveyed by specifying the statistic $b(D)$.

Put differently, given the actual data set D, all estimators that the orthodoxian might have chosen $\{b_1, b_2, \ldots\}$ are known, so Bayes' theorem has available for its use simultaneously all the information contained in the class of all possible estimators. If there is no sufficient statistic, it is able to choose the optimal estimator *for the present data set*.

If the estimator is not a sufficient statistic, its sampling distribution is irrelevant for us, because with different data sets we shall use different estimators. We saw this in some detail, from different viewpoints, in Chapters 8 and 13. The same considerations apply to hypothesis testing; the Bayesian procedure has available all the relevant information in the data, but an orthodox procedure based on a single statistic does not unless it is a sufficient statistic. If it is not sufficient, then we expect that the Bayesian procedure will be superior (in the sense of more accurate or more reliable) because it is extracting more information from the data. Once one understands this, it is easy to produce any number of examples which demonstrate it.

From the Neyman–Pearson camp of orthodoxy we have the devices of unbiased estimators, confidence intervals, and hypothesis tests which amount to a kind of decision theory. This line of thought was adopted more or less faithfully in the works of Herbert Simon (1977) in economics, Erich Lehmann (1986) in hypothesis testing, and David Middleton (1960) in electrical engineering.

From the Fisherian (sometimes called the piscatorial) camp, there are the principles of maximum likelihood, analysis of variance, randomization in design of experiments, and a mass of specialized 'tail area' significance tests. Fortunately, the underlying logic is the same in all such significance tests, so they need not be analyzed separately. Adoption of these methods has been almost mandatory in biology and medical testing. Also, Fisher advocated fiducial probability, which most statisticians rejected, and conditioning on ancillary

statistics, which we discussed in Chapter 8, and showed that it is mathematically equivalent to applying Bayes' theorem without prior information.

17.2 Unbiased estimators

Given a sampling distribution $p(x|\alpha)$ with some parameter α and a data set comprising n observations $D \equiv \{x_1, \ldots, x_n\}$, there are various orthodox principles for estimating α, in the particular use of an unbiased estimator, and maximum likelihood. In the former we choose some function of the observations $\beta(D) = \beta(x_1, \ldots, x_n)$ as our 'estimator'. The Neyman–Pearson school holds that it should be 'unbiased', meaning that its expectation over the sampling distribution is equal to the true value of α:

$$\langle\beta\rangle = E(\beta) = \int dx_1 \cdots dx_n \, \beta(x_1, \ldots, x_n) p(x_1 \cdots x_n|\alpha) = \alpha. \tag{17.1}$$

As noted in Chapter 13, Eq. (13.20), the expected square of the error, over the sampling distribution, is the sum of two positive terms,

$$\langle(\beta - \alpha)^2\rangle = (\langle\beta\rangle - \alpha)^2 + \mathrm{var}(\beta), \tag{17.2}$$

where what the orthodoxian calls the 'sampling variance of β' (more correctly, the variance *of the sampling distribution for* β) is $\mathrm{var}(\beta) = \langle\beta^2\rangle - \langle\beta\rangle^2$. At present, we are not after mathematical pathology of the kind discussed in Chapter 15 and Appendix B, but rather *logical* pathology – due to conceptual errors in the basic formulation of a problem – which persists even when all the mathematics is well-behaved. So we suppose that the first two moments of that sampling distribution, $\langle\beta\rangle, \langle\beta^2\rangle$, exist for all the estimators to be considered. If we introduce a fourth moment $\langle\beta^4\rangle$, we are automatically supposing that it exists also; this is the general mathematical policy advocated in Appendix B. Then an unbiased estimator has, indeed, the merit that it makes one of the terms of (17.2) disappear. But it does not follow that this choice minimizes the expected square of the error; let us examine this more closely.

What is the relative importance of removing bias and minimizing the variance? From (17.2) it would appear that they are of equal importance; there is no advantage in decreasing one if in so doing we increase the other more than enough to compensate. Yet that is what the orthodox statistician usually does! As the most common specific example, Cramér (1946, p. 351) considers the problem of estimating the variance μ_2 of a sampling distribution $p(x_1|\mu_2)$:

$$\mu_2 = \langle x_1^2\rangle - \langle x_1\rangle^2 = \langle x_1^2\rangle \tag{17.3}$$

from n independent observations $\{x_1, \ldots, x_n\}$. We assume, in (17.3) and in what follows, that $\langle x_1\rangle = 0$, since a trivial change of variables would in any event accomplish this. An elementary calculation shows that the sample variance (now correctly called the variance *of the sample* because it expresses the variability of the data within the sample, and does

not make reference to any probability distribution)

$$m_2 \equiv \overline{x^2} - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2 \tag{17.4}$$

has expectation, over the sampling distribution $p(x_1 \cdots x_n|\mu_2) = p(x_1|\mu_2)\cdots p(x_n|\mu_2)$, of

$$\langle m_2 \rangle = \frac{n-1}{n}\mu_2 \tag{17.5}$$

and thus, as an estimator of μ_2, it has a negative bias. So, goes the argument, we should correct this by using the unbiased estimator

$$M_2 \equiv \frac{n}{n-1} m_2. \tag{17.6}$$

Indeed, this has seemed so imperative that in most of the orthodox literature, the term 'sample variance' is *defined* as M_2 rather than m_2.

Now, of course, the only thing that really matters here is the *total* error of our estimate; the particular way in which you or I separate error into two abstractions labeled 'bias' and 'variance' has nothing to do with the actual quality of the estimate. So, let's look at the full mean-square error criterion (17.2) with the choices $\beta = m_2$ and $\beta = M_2$. Replacement of m_2 by M_2 removes a term $(\langle m_2 \rangle - \mu_2)^2 = \mu_2^2/n^2$, but it also increases the term $\mathrm{var}(m_2)$ by a factor $[n/(n-1)]^2$, so it seems obvious that, at least for large n, this has made things worse instead of better. More specifically, suppose we replace m_2 by the estimator

$$\beta \equiv c m_2. \tag{17.7}$$

What is the best choice of c by orthodox criteria? The expected quadratic loss (17.2) is now

$$\begin{aligned} \langle (c m_2 - \mu_2)^2 \rangle &= c^2 \langle m_2^2 \rangle - 2c \langle m_2 \rangle \mu_2 + \mu_2^2 \\ &= \langle (m_2 - \mu_2)^2 \rangle - \langle m_2^2 \rangle (\hat{c} - 1)^2 + \langle m_2^2 \rangle (c - \hat{c})^2, \end{aligned} \tag{17.8}$$

where

$$\hat{c} \equiv \frac{\mu_2 \langle m_2 \rangle}{\langle m_2^2 \rangle}. \tag{17.9}$$

Evidently, the best estimator in the class (17.7) is the one with $c = \hat{c}$, and the term $-\langle m_2^2 \rangle (\hat{c} - 1)^2$ in (17.8) represents the decrease in mean-square error obtainable by using $\hat{\beta} \equiv \hat{c} m_2$ instead of m_2. Another short calculation shows that

$$\langle m_2^2 \rangle = n^{-3}(n-1)[(n^2 - 2n + 3)\mu_2^2 + (n-1)\mu_4], \tag{17.10}$$

where

$$\mu_4 \equiv \langle (x_1 - \langle x_1 \rangle)^4 \rangle = \langle x_1^4 \rangle \tag{17.11}$$

is the fourth central moment of $p(x_1|\mu_2)$. We must understand $n > 1$ in all this, for if $n = 1$, we have $m_2 = 0$; in sampling theory, a single observation gives no information about the variance μ_2.[1]

From (17.5) and (17.10) we then find that $\hat{c}$ depends on the second and fourth moments of the sampling distribution:

$$\hat{c} = \frac{n^2}{n^2 - 2n + 3 + (n-1)K}, \tag{17.12}$$

where $K \equiv \mu_4/\mu_2^2 \geq 1$. We see that $\hat{c}$ is a monotonic decreasing function of K; so, if $K \geq 2$, (17.12) shows that $\hat{c} < 1$ for all n, instead of removing the bias in (17.5) we should always *increase* it!

In the case of a Gaussian distribution, $p(x|\mu_2) \propto \exp\{-x^2/2\mu_2\}$, we find $K = 3$. We will seldom have $K < 3$, for that would imply that $p(x|\mu_2)$ cuts off even more rapidly than Gaussian for large x. If $K = 3$, (17.12) reduces to

$$\hat{c} = \frac{n}{n+1}, \tag{17.13}$$

which, by comparison with (17.6), says that rather than removing the bias we should approximately double it, in order to minimize the mean-square sampling error.

How much better is the estimator $\hat{\beta} = \hat{c}m_2$ than M_2? In the Gaussian case the mean-square error of the estimator $\hat{\beta}$ is

$$\langle(\hat{\beta} - \mu_2)^2\rangle = \frac{2\mu_2^2}{n+1}. \tag{17.14}$$

The unbiased estimator M_2 corresponds to the choice

$$c = \frac{n}{n-1} \tag{17.15}$$

and thus to the mean-square error

$$\langle(M_2 - \mu_2)^2\rangle = \mu_2^2\left[\frac{2}{n+1} + \frac{2}{n}\right], \tag{17.16}$$

which is over twice the amount incurred by use of $\hat{\beta}$.[2] Most sampling distributions that arise in practice, if not Gaussian, have wider tails than Gaussian, so that $K > 3$; in this case the difference will be even greater.

Up to this point, it may have seemed that we are quibbling over a very small thing – changes in the estimator of one or two parts out of n. But now we see that the difference between (17.14) and (17.16) is not at all trivial. For example, *with the unbiased estimator M_2 you will need $n = 203$ observations in order to get as small a mean-square sampling*

[1] In Bayesian theory a single observation could give information about μ_2 if μ_2 is correlated, in the joint prior probability $p(\mu_2\theta|I)$, with some other parameter θ in the problem about which a single observation does give information; that is, $p(\mu\theta|I) \neq p(\mu|I)p(\theta|I)$. This kind of indirect information transfer can be helpful in problems where we have cogent prior information but only sparse data.

[2] Editor's footnote: It appears that Jaynes indavertantly calculated this expectation using $\langle m_2^2\rangle(c - c^*)$ rather than $\langle M_2^2\rangle(c - c^2)^2$ and so arrived at (17.16) rather than $\langle(M_2 - \mu_2)^2\rangle = 2\mu_2^2/(n-1)$.

error as the biased estimator $\hat{\beta}$ gives you with only 100 observations. This is typical of the way orthodox methods waste information; in this example we have, in effect, thrown away half of our data whatever the value of n, and therefore wasted half the work expended in acquiring the data.

R. A. Fisher, who often thought in terms of information, perceived this long ago; but modern orthodox practitioners seem never to perceive it, because they continue to fantasize about frequencies, and do not think in terms of information at all.[3] There is a work on econometrics (Valavanis, 1959, p. 60) where the author attaches such great importance to removing bias that he advocates throwing away not just half the data but practically all of them, if necessary, to achieve this.

Why do orthodoxians put such exaggerated emphasis on bias? We suspect that the main reason is simply that they are caught in a psychosemantic trap of their own making. When we call the quantity $(\langle\beta\rangle - \alpha)$ the 'bias', that makes it sound like something awfully reprehensible, which we must get rid of at all costs. If it had been called instead the 'component of error orthogonal to the variance', as suggested by the Pythagorean form of (17.2), it would have been clear to all that these two contributions to the error are on an equal footing; it is folly to decrease one at the expense of increasing the other. This is just the price one pays for choosing a technical terminology that carries an emotional load, implying value judgments; orthodoxy falls constantly into this tactical error.

Chernoff and Moses (1959) give a more forceful example showing how an unbiased estimate may be far from what we want. A company is laying a telephone cable across San Francisco Bay. They cannot know in advance exactly how much cable will be needed, and so they must estimate. If they overestimate, the loss will be proportional to the amount of excess cable to be disposed of; but if they underestimate, and the cable end falls into the water, the result may be financial disaster. Use of an unbiased estimate here could only be described as foolhardy; this shows why a Wald-type decision theory is needed to fully express rational behavior.

Another reason for such an undue emphasis on bias is a belief that if we draw N successive samples of n observations each and calculate the estimators $\beta_1, \ldots, \beta_N$, the average $\overline{\beta} = N^{-1}\sum \beta_i$ of these estimates will converge in probability to $\langle\beta\rangle$ as $N \to \infty$, and thus an unbiased estimator will, on sufficiently prolonged sampling, give an arbitrarily accurate estimate of α. Such a belief is almost never justified, even for the fairly well-controlled measurements of the physicist or engineer, not only because of unknown systematic error, but because successive measurements lack the logical independence required for these limit theorems to apply.

In such uncontrolled situations as economics, the situation is far worse; there is, in principle, no such thing as 'asymptotic sampling properties' because the 'population' is always finite, and it changes uncontrollably in a finite time. The attempt to use only sampling

[3] Note that this difficulty does not arise in the Bayesian approach, in spite of a mathematical similarity. Again choosing any function $\beta(x_1, \ldots, x_n)$ of the data as an estimator, and letting the brackets $\langle\ \rangle$ stand now for expectations over the posterior pdf for α, we have the expected square of the error of $\langle(\beta - \alpha)^2\rangle = (\beta - \langle\alpha\rangle)^2 + \mathrm{var}(\alpha)$, rather like (17.2). But now changing the estimator β does not change $\mathrm{var}(\alpha) = (\langle\alpha^2\rangle - \langle\alpha\rangle^2)$, and so, by this criterion, the optimal estimator over the class of *all* estimators is always $\beta = \langle\alpha\rangle$.

distributions – always interpreted as limiting frequencies – in such a situation forces one to expend virtually all his efforts on irrelevant fantasies. What is relevant to inference is not any imagined (that is, nonobserved) frequencies, but *the actual state of knowledge that we have about the real situation*. To reject that state of knowledge – or any human information – on the grounds that it is 'subjective' is to destroy any possibility of finding useful results; for human information is all we have.[4]

Even if we accept these limit theorems uncritically, and believe faithfully that our sampling probabilities are also the limiting frequencies, unbiased estimators are not the only ones which approach perfect accuracy with indefinitely prolonged sampling. Many biased estimators approach the true value of α in this limit, *and do it more rapidly*. Our $\hat{\beta}$ is an example. Furthermore, asymptotic behavior of an estimator is not really relevant, because the real problem is always to do the best we can with a finite data set; therefore the important question is not *whether* an estimator tends to the true value, but *how rapidly* it does so.

Long ago, R. A. Fisher disposed of the unbiased estimate by a different argument that we noted in Chapter 6, Eq. (6.94). The criterion of bias is not really meaningful, because it is not invariant under a change of parameters; the square of an unbiased estimate of α is not an unbiased estimate of α^2. With higher powers α^k, the difference in conclusions can become arbitrarily large, and nothing in the formulation of a problem tells us which choice of k is 'right'. Thus, if you and I happen to choose k differently, the criterion of an unbiased estimate will lead us to different conclusions about α from the same data. However, many orthodoxians simply ignore these ambiguities (although they can hardly be unaware of them) and continue to use unbiased estimators whenever they can, aware that they are violating a rather basic principle of rationality, but unaware that they are also wasting information.[5]

Note, however, that, after all this argument, nothing in the above entitles us to conclude that $\hat{\beta}$ is the best estimator of μ_2 by the criterion of mean-square sampling error! We have considered only the restricted class of estimators (17.7) constructed by multiplying the sample variance (17.4) by some preassigned number; we can say only that $\hat{\beta}$ is the best one in that class. The question whether some other function of the sample values, not a multiple of (17.4), might be still better by the criterion of mean-square sampling error, remains completely open. That the orthodox approach to parameter estimation does not tell us how to find the best estimator, but only how to compare different intuitive guesses, was noted in Chapter 13 following Eq. (13.21); and we showed that the difficulty is overcome

[4] 'Objectivity' in inference consists, then, in carefully considering all the information we have about the real situation, and carefully avoiding fantasies about situations that do not actually exist. It seems to us that this should have been obvious to orthodoxians from the start, since it was obvious already to ancient writers such as Herodotus (*c* 500 BC) in his discussion of the policy decisions of the Persian kings.

[5] We noted in Chapter 6, Eqs. (6.94)–(6.98) that the Bayesian criterion of the posterior expectation has potentially the same ambiguity; different definitions of parameters will lead to different conclusions if we continue to use the criterion of posterior expectation after a parameter transformation. Curiously, this problem did not arise with Laplace's original criterion of posterior median and quartiles. But these were not entirely correct applications of Bayesian theory. When we completed the theoretical apparatus with the decision theory of Chapter 13, a transformation of parameters was accompanied by a corresponding transformation of the loss function, with the result that our final substantive conclusions are now invariant under arbitrary parameter redefinitions.

by a slight reformulation of the problem, which leads inexorably to the Bayesian algorithm as the one which accomplishes what we really want.

Exercise 17.1. Try to extend sampling theory to deal with the many questions left unanswered by the orthodox literature and the above discussion. Is there a general theory of optimal sampling theory estimators for finite samples? If so, does bias play any role in it? We know already, from the analysis in Chapter 13, that this cannot be a variational theory; but it seems conceivable that a theory somewhat like dynamic programming might exist. In particular, can you find an orthodox estimator that is better than $\hat{\beta}$ by the mean-square error criterion? Or can you prove that $\hat{\beta}$ cannot be improved upon within sampling theory?

In contrast to the difficulty of these questions in sampling theory, we have noted above and in Chapter 13 that the Bayesian procedure automatically constructs the optimal estimator for any data set and loss function, whether or not a sufficient statistic exists; and it leads at once to a simple variational proof of its optimality not within any restricted class, but with respect to *all* estimators. And it does this without making any reference to the notion of bias, which plays no role in Bayesian theory.

17.3 Pathology of an unbiased estimate

On closer examination, an even more disturbing feature of unbiased estimates appears. Consider the Poisson sampling distribution: the probability that, in one time unit, we observe n events, or 'counts', is

$$p(n|l) = \exp(-l)\frac{l^n}{n!}, \qquad n = 0, 1, 2, \ldots, \tag{17.17}$$

in which the parameter l is the sampling expectation of n, $\langle n \rangle = l$. Then what function $f(n)$ gives an unbiased estimate of l? Evidently, the choice $f(n) = n$ will achieve this; to prove that it is unique, note that the requirement $\langle f(n) \rangle = l$ is

$$\sum_{n=0}^{\infty} \exp(-l)\,\frac{l^n}{n!} f(n) = l, \tag{17.18}$$

and, from the formula for coefficients of a Taylor series, this requires

$$f(n) = \frac{\mathrm{d}^n}{\mathrm{d}l^n}\{l\exp(l)\}\Big|_{l=0} = n. \tag{17.19}$$

A reasonable result. But suppose we want an unbiased estimator of some function $g(l)$; by the same reasoning, the unique solution is

$$f(n) = \frac{\mathrm{d}^n}{\mathrm{d}l^n}\Big\{\exp(l)g(l)\Big\}\Big|_{l=0}. \tag{17.20}$$

Thus the only unbiased estimator of l^2 is

$$f(n) = \begin{cases} 0 & n = 0, 1 \\ n(n-1) & n > 1, \end{cases} \tag{17.21}$$

which is absurd for $n = 1$. Likewise, the only unbiased estimator of l^3 is absurd for $n = 1, 2$; and so on. Here the unbiased estimator does violence to elementary logic; if we observe $n = 2$, we are advised to estimate $l^3 = 0$; but if l^3 were zero, it would be impossible to observe $n = 2$! The only unbiased estimator of $\exp\{-l\}$ is

$$f(n) = \begin{cases} 1 & n = 0 \\ 0 & n > 0, \end{cases} \tag{17.22}$$

which is absurd for all positive n. An unbiased estimator for $(1/l)$ does not exist; it is mathematically pathological. Unbiased estimators can stand in conflict with deductive logic not just for a few data sets, but for all data sets. And if they can generate such pathology even in such a simple problem as this, what horrors await us in more complicated problems?

The remedy

In contrast, with uniform prior the Bayesian posterior mean estimate of any function $g(l)$ is

$$\langle g(l) \rangle = \frac{1}{n!} \int_0^\infty \mathrm{d}l \, \exp(-l) l^n g(l), \tag{17.23}$$

which is readily verified to be mathematically well-behaved and intuitively reasonable for all the above examples. The Bayes estimate of $(1/l)$ is just $(1/n)$; no pathology here. It is at first surprising that the Bayes estimate of $\exp\{-l\}$ is

$$f(n) = 2^{-(n+1)}. \tag{17.24}$$

Why would it not be just $\exp\{-n\}$? To see why, note the following points.

(1) The posterior distribution for l is skewed; the posterior probability that $l > n$ is

$$P(l > n) = \int_n^\infty \mathrm{d}l \, \exp\{-l\} \frac{l^n}{n!} = \exp(-n) \sum_{m=0}^{n} \frac{n^m}{m!}. \tag{17.25}$$

This decreases monotonically from 1 at $n = 0$ to 1/2 as $n \to \infty$. Thus, given n, the parameter l is always more likely to be greater than n than less.

(2) The posterior distribution for l is proportional to $\exp\{-l\} l^n$, which is concentrated mostly in the interval $(n \pm \sqrt{n})$. But $\exp\{-l\}$ is so rapidly varying that, in calculating its expectation, most of the contribution to the integral $\int \mathrm{d}l \, \exp\{-2l\} l^n$ comes from the region $(n/2 \pm \sqrt{n}/2)$; so $\exp\{-n/2\}$ would be closer to the correct estimator than $\exp\{-n\}$. Both of these circumstances affect the numerical value, in such as way that (17.24) finally emerges as the balance between these opposing tendencies. This is still another example where Bayes' theorem detects a genuinely complicated

situation and automatically corrects for it, but in such a slick, efficient way that one is unaware of what is happening.

Exercise 17.2. Consider the truncated Poisson distribution:

$$p(n|l) = \left[\frac{1}{\exp(l) - 1}\right] \frac{l^n}{n!}, \qquad n = 1, 2, \ldots. \tag{17.26}$$

Show that the unbiased estimator of l is now absurd for $n = 1$, and the unbiased estimator of $\exp(-l)$ is absurd for all even n and queer for all odd n.

Many other examples are known in which the attempt to find unbiased estimates leads to similar pathologies; several were noted by the orthodoxians Kendall and Stuart (1961). But their anti-Bayesian indoctrination was so strong that they would not deign to examine the corresponding Bayesian results; and so they never did learn that in all their cases Bayesian methods overcome the difficulty easily.[6]

17.4 The fundamental inequality of the sampling variance

A famous inequality, variously associated with the names Cramér, Rao, Darmois, Frechét and others, finds a lower bound to the sampling variance that can be achieved for any estimator – or, indeed, any statistic – with a continuous sampling distribution. Although the result is nearly trivial mathematically, it is important because it is almost the only bit of connected theory that orthodoxy has to guide it. An extensive discussion with examples is given by Cramér (1946, Chap. 32). Denote a data set of n observations by $x \equiv \{x_1, \ldots, x_n\}$ and integration over the sample space by $\int \mathrm{d}x\,(\quad)$. With a sampling distribution $p(x|\alpha)$ containing a parameter α, let

$$u(x, \alpha) \equiv \frac{\partial \log p(x|\alpha)}{\partial \alpha}. \tag{17.27}$$

Mathematically, the result we seek is just the Schwartz inequality: given two functions $f(x)$, $g(x)$ defined on the sample space, write $(f, g) \equiv \int \mathrm{d}x\, f(x)g(x)$. Then $(f, g)^2 \leq (f, f)\,(g, g)$ with equality if and only if $f(x) = qg(x)$, where q is a constant independent of x, although it may depend on α.[7] Now make the choices

$$f(x) \equiv u(x, \alpha)\sqrt{p(x|\alpha)}, \qquad g(x) \equiv [\beta(x) - \langle\beta\rangle]\sqrt{p(x|\alpha)}. \tag{17.28}$$

We find that $(f, g) = \langle\beta u\rangle - \langle\beta\rangle\langle u\rangle = \langle\beta u\rangle$, since $\langle u\rangle = \int \mathrm{d}x\, u(x, \alpha)p(x|\alpha) = \partial/\partial\alpha\,[\int \mathrm{d}x\, p(x|\alpha)] = 0$. Likewise, $(f, f) = \mathrm{var}(u)$, and $(g, g) = \mathrm{var}(\beta)$, so the Schwartz

[6] Maurice Kendall could have learned this in five minutes from Harold Jeffreys, whom he saw almost daily for years because they were both Fellows of St John's College, Cambridge, and ate at the same high table.

[7] Proof: $\int \mathrm{d}x\,[f(x) - qg(x)]^2 \geq 0$ for all constants q, in particular for the value $q = (f, g)/(g, g)$ which minimizes the integral. Then we have equality if and only if $f(x) - qg(x) = 0$. Note that this remains true whatever the range of integration; it need not be the entire sample space.

inequality reduces to

$$\langle \beta u \rangle \leq \sqrt{\mathrm{var}(\beta)\mathrm{var}(u)}. \tag{17.29}$$

But $\langle \beta u \rangle = \int \mathrm{d}x\, \beta \partial p(x|a)/\partial\alpha = \mathrm{d}\langle\beta\rangle/\mathrm{d}\alpha = 1 + b'(\alpha)$, where $b(\alpha) \equiv (\langle\beta\rangle - \alpha)$ is the bias of the estimator. Thus the famous inequality sought is

$$\mathrm{var}(\beta) \geq \frac{[1 + b'(\alpha)]^2}{\int \mathrm{d}\alpha\, (\partial \log p(x|\alpha)/\partial\alpha)^2 p(x|\alpha)}. \tag{17.30}$$

Now, substituting (17.27) into the necessary and sufficient condition for equality ($f = qg$) and making a change of parameters ($\alpha \to l$), where l is defined by $q(\alpha) = -\partial l/\partial\alpha$, we have

$$\frac{\partial \log p(x|\alpha)}{\partial\alpha} = -l'(\alpha)[\beta(x) - \langle\beta\rangle], \tag{17.31}$$

and, integrating over α, the condition for equality becomes

$$\log p(x|a) = -l(\alpha)\beta(x) + \int \mathrm{d}l\, \langle\beta\rangle + \text{const.} \tag{17.32}$$

To put this into more familiar notation, note that the integral in (17.32) is a function of α; let us call it $-\log Z(\alpha)$, defining the function $Z(\alpha)$. Likewise, the constant of integration in (17.32) is independent of α but may depend on x; so call it $\log m(x)$, defining the function $m(x)$. With these changes of notation, the necessary and sufficient condition for equality in (17.30) becomes

$$p(x|\alpha) = \frac{m(x)}{Z(l)} \exp\{-l(\alpha)\beta(x)\}. \tag{17.33}$$

But we recognize this as just the distribution that we found in Chapter 11, produced by the maximum entropy principle with a constraint fixing $\langle\beta(x)\rangle$. In (17.33) the denominator $Z(l)$ is evidently a normalizing constant, therefore equal to

$$Z(l) = \int \mathrm{d}x\, m(x) \exp\{-l\beta(x)\}, \tag{17.34}$$

whereupon the constraint is just

$$\langle\beta\rangle = -\frac{\partial \log(Z)}{\partial l}, \tag{17.35}$$

which is identical with (11.60). This generalizes at once to the case where α, β are vectors of any dimensionality, the exponent becoming $\{-\sum l_i(\alpha)\ \beta_i(x)\}$ as in (11.43); so we are just rediscovering the maximum entropy formalism of Chapter 11!

These results teach us something very important about the basic unity and mutual consistency of several principles that had seemed, up till now, distinct from each other. We noted in Chapter 14 that the notion of sufficiency, which was always associated with the notion of information, is in fact definable in terms of Shannon's information measure of entropy. Long ago, the Pitman–Koopman theorem (Koopman, 1936; Pitman, 1936) proved

that the condition for existence of a sufficient statistic is just that the sampling distribution be of the functional form (17.32). Therefore, if we use the maximum entropy principle to assign sampling distributions, this automatically generates the distributions with the most desirable properties from the standpoint of inference in either sampling theory (because the sampling variance of an estimator is then the minimum possible value) or Bayesian theory (because then in applying Bayes' theorem we need only calculate a single function of the data).

Indeed, if we think of a maximum entropy distribution as a sampling distribution parameterized by the Lagrange multipliers l_j, we find that the sufficient statistics are precisely the data images of the constraints that were used in defining that distribution. Thus, the maximum entropy distribution generated from the set of constraints fixing $\{\langle\beta_1(x)\rangle, \langle\beta_2(x)\rangle, \ldots, \langle\beta_k(x)\rangle\}$ as expectations over the probability distribution, has k sufficient statistics which are just $\{\beta_1(x), \ldots, \beta_k(x)\}$, in which now x is the observed data set. This is proved in Jaynes (1978, Eq. B82); we leave it as an exercise for the reader to reconstruct the proof.

If the sampling distribution does not have the form (17.33) or its generalization, there are two possibilities. Firstly, if the sampling distribution is continuous in α, then the lower bound (17.28) cannot be attained, and there seems to be no theory to determine the correct lower bound, much less to construct an estimator that achieves it. Then if $\langle\beta\rangle$ is unbiased, the ratio of the minimum possible variance, right-hand side of (17.30), to the actual var (β) was called by Fisher the *efficiency* of the estimator β, and an estimator with efficiency of one was called an *efficient estimator*. Nowadays it is usually called an 'unbiased minimum variance' (UMV) estimator.[8]

Secondly, if $p(x|\alpha)$ has discontinuities, Cramér (1946, p. 485) finds that there are estimators that actually achieve a lower variance than (17.28). But how is this possible, since the Schwartz inequality does not seem to admit to any exceptions? We consider this a mathematical error, for reasons explained in Appendix B (had Cramér approached a discontinuous function as the limit of a sequence of continuous ones, another term, a delta-function, would have appeared in the limit, which just accounts for the discrepancy and makes the inequality (17.30) correct whether $p(x|\alpha)$ is continuous or discontinuous). This is a typical case where failure to recognize the necessary role of delta-functions in analysis leads one into errors.

17.5 Periodicity: the weather in Central Park

A common problem, important in economics, meteorology, geophysics, astronomy and many other fields, is to decide whether certain data taken over time provide evidence for a periodic behavior. Any clearly discernible periodic component (in births, diseases, rainfall, temperature, business cycles, stock market, crop yields, incidence of earthquakes, brightness of a star) provides an evident basis for improved prediction of future behavior,

[8] Note that the notion of efficiency is even more parameter-dependent than that of an unbiased estimate; if an efficient estimator of α exists, then an efficient estimator of α^2 does not.

on the presumption (that is, inductive reasoning) that periodicities observed in the past are likely to continue in the future. But even apart from prediction, the principle for analyzing the data for evidence of periodicity in the past is still controversial: is it a problem of significance tests, or one of parameter estimation? Different schools of thought come to opposite conclusions from the same data.

Consider an example from the recent literature of orthodox reasoning and procedure here; this will also provide an easy introduction to Bayesian spectrum analysis. Bloomfield (1976, p. 110) gives a graph showing mean January temperatures observed over about 100 years in Central Park, New York. The presence of a periodicity of roughly 20 years with a peak-to-peak amplitude of about 4 °F is perfectly evident to the eye, since the irregular 'noise' is only about 0.5 °F. Yet Bloomfield, applying an orthodox significance test introduced by Fisher, concludes that there is no significant evidence for any periodicity!

17.5.1 The folly of pre-filtering data

In trying to understand this we note first that the data of Bloomfield's graph have been 'pre-filtered' by taking a 10 year moving average. What effect does this have on the evidence for periodicity? Let the original raw data be $D = \{y_1, \ldots, y_n\}$ and consider the discrete Fourier transform

$$Y(\omega) \equiv \sum_{t=1}^{n} y_t \exp\{\mathrm{i}\omega t\}. \tag{17.36}$$

This is well-defined for continuous values of ω and is periodic: $Y(\omega) = Y(\omega + 2\pi)$. Therefore there is no loss of information if we confine the frequency to $|\omega| < \pi$. But even that is more than necessary; the values of $Y(\omega)$ at any n consecutive and discrete 'Nyquist' frequencies[9]

$$\omega_k \equiv 2\pi k/n, \quad 0 \le k < n, \tag{17.37}$$

already contain all the information in the data, for by the orthogonality $n^{-1}\sum_k \exp\{\mathrm{i}\omega_k(s - t)\} = \delta_{st}$, the data can be recovered from them by the Fourier inversion:

$$\frac{1}{n}\sum_{k=1}^{n} Y(\omega_k)\exp\{-\mathrm{i}\omega_k t\} = y_t, \qquad 1 \le t \le n. \tag{17.38}$$

Suppose the data were replaced with an m year moving average over past values, with weighting coefficient of w_s for lag s:

$$z_t \equiv \sum_{s=0}^{m-1} y_{t-s} w_s. \tag{17.39}$$

[9] Harry Nyquist was a mathematician at the Bell Telephone Laboratories who, in the 1920s, discovered a great deal of the fundamental physics and information theory involved in electrical communication. The work of Claude Shannon is a continuation, 20 years later, of some of Nyquist's pioneering work. All of it is still valid and indispensable in modern electronic technology. In Chapter 7 we have already considered the fundamental, irreducible 'Nyquist noise' in electrical circuits due to random thermal motion of electrons.

The new Fourier transform would be, after some algebra,[10]

$$Z(\omega) = \sum_{t=1}^{n} z_t \exp\{\mathrm{i}\omega t\} = W(\omega)Y(\omega), \tag{17.40}$$

where

$$W(\omega) \equiv \sum_{s=0}^{m-1} w_s \exp\{\mathrm{i}\omega s\} \tag{17.41}$$

is the Fourier transform of the weighting coefficients. This is just the convolution theorem of Fourier theory. Thus, taking any moving average of the data merely multiplies its Fourier transform by a known function. In particular, for uniform weighting

$$w_s = \frac{1}{m}, \qquad 0 \le s < m, \tag{17.42}$$

we have

$$W(\omega) = \frac{1}{m}\sum_{s=0}^{m-1} \exp\{-\mathrm{i}\omega s\} = \exp\left\{-\mathrm{i}\frac{\omega}{2}(m-1)\right\}\left[\frac{\sin(m\omega/2)}{m\sin(\omega/2)}\right]. \tag{17.43}$$

In the case $m = 10$ we find, for a 10 year and 20 year periodicity, respectively,

$$W(2\pi/10) = 0; \qquad W(2\pi/20) = 0.639\exp\{-9\pi\mathrm{i}/20\}. \tag{17.44}$$

Thus, taking a 10 year moving average of any time series data represents an irreversible loss of information; it completely wipes out any evidence for a 10 year periodicity, and reduces the amplitude of a 20 year periodicity by a factor 0.639, while shifting its phase by $9\pi/20 = 1.41$ radian. In addition, the magnitude of $W(\omega)$ is decreasing at $\omega = 2\pi/20$ so the apparent frequency is shifted; the peak in $Z(\omega)$ occurs at a lower frequency than the true peak in $Y(\omega)$. We conclude that the original data had a periodicity of roughly 20 years with a peak-to-peak amplitude of about $4/0.639 = 6.3\,^{\circ}$F, even more obvious to the eye and nearly 90 degrees out of phase with the periodicity visible in Bloomfield's graph; and the true frequency is somewhat higher than one would estimate from the graph. Taking the moving average has severely mutilated and distorted the information in the data.

At several places we warn against the common practice of pre-filtering data in this way before analyzing them. The only thing it can possibly accomplish is the cosmetic one of making the graph of the data look prettier to the eye. But if the data are to be analyzed

[10] At this point, many authors get involved in a semantic hangup over exactly what one means by the term 'm year moving average' for a series of finite length. If we have only y_t for $t > 0$, then it seems to many that the m year moving average (17.39) could start only at $t = m$. But then they find that their formulas are not exact, but require small 'end-effect' correction terms of order m/n. We avoid this by a slight change in definitions. Consider the original time series $\{y_t\}$ augmented by 'zero-padding'; we define $y_t \equiv 0$ when $t < 1$ or $t > n$, and likewise the weighting coefficients are defined to be zero when $s < 0$ or $s \ge m$. Then we may understand the above sums over t, s to be over $(-\infty, +\infty)$, and the first few terms $(z_1, \ldots, z_{m-1})$, although averages over m years of the padded data, are actually averages over less than m years of nonzero data. The differences are numerically negligible when $m \ll n$, but we gain the advantage that the simple formulas (17.36)–(17.42) with sums taken instead over $\pm\infty$ and t in (17.39) allowed to take all positive values, are all exact as they stand, without our having to bother with messy correction terms. Furthermore, it is evident that failure to do this means that some of the information in the first m and last m data values is lost. This particular definition of the term 'moving average' for a finite series (which was basically arbitrary anyway) is thus the one appropriate to the subject.

by a computer, this does not help in any way; it only throws away or distorts some of the information that the computer could have extracted from the original, unaltered data. It renders the filtered data completely useless for certain purposes. For all we know, there might have been a strong periodicity of about 10 years in the original data, corresponding to the well known 11 year periodicity in sunspot numbers; but, if so, taking a 10 year moving average has wiped out the evidence for it.

The periodogram of the data is then the power spectral density:

$$P(\omega) \equiv \frac{1}{n}|Y(\omega)|^2 = \frac{1}{n}\sum_{t,s} y_t y_s \exp\{\mathrm{i}\omega(t-s)\}. \tag{17.45}$$

Note that $P(0) = (\sum y_t)^2/n = n\overline{y}^2$ determines the mean value of the data, while the average of the periodogram at the Nyquist frequencies is the mean-square value of the data:

$$P(\omega_k)_{\mathrm{av}} = \frac{1}{n}\sum_{k=1}^{n} P(\omega_k) = \overline{y^2}. \tag{17.46}$$

Fisher's proposed test statistic for a periodicity is the ratio of peak/mean of the periodogram:

$$q = \frac{P(\omega_k)_{\max}}{P(\omega_k)_{\mathrm{av}}}, \tag{17.47}$$

and one computes its sampling distribution $p(q|H_0)$ conditional on the null hypothesis H_0 that the data are Gaussian white noise. Having observed the value q_0 from our data, we find the so-called 'P-value', which is the sampling probability, conditional on H_0, that chance alone would have produced a ratio as great or greater:

$$P \equiv p(q > q_0|H_0) = \int_{q_0}^{\infty} \mathrm{d}q\; p(q|H_0), \tag{17.48}$$

and if $P > 0.05$ the evidence for periodicity is rejected as 'not significant at the 5% level'.[11]

This test looks only at probabilities conditional on the 'null hypothesis' that there is no periodic term. It takes no note of probabilities of the data conditional on the hypothesis that a periodicity is present; or on any prior information indicating whether it is reasonable to expect a periodicity! We commented on this kind of reasoning in Chapter 5; how can one test any hypothesis rationally if he fails to specify (1) the hypothesis to be tested; (2) the alternatives against which it is to be tested; and (3) the prior information that we bring to the problem? Until we have done that much, we have not asked any definite, well-posed question.

Equally puzzling, how can one expect to find evidence for a phenomenon that is real, if he starts with all the cards stacked overwhelmingly against it? The only hypothesis H_0 that this test considers is one which assumes that the totality of the data are part of a

[11] This is a typical orthodox 'tail area' significance test; we discussed such tests in Chapter 9, and noted that the orthodox chi-squared test has serious shortcomings, but there is a similar Bayesian ψ-test that is exact and is free of those difficulties. Many other Bayesian ψ-tests can be set up, which test some hypothesis H against a specified class C of alternatives. But now we note a different way of looking at this situation that is generally more useful: a significance test can often be replaced by a parameter estimation problem that is simpler and more informative.

'stationary Gaussian random process' *without* any periodic component. According to that H_0, the appearance of anything resembling a sine wave would be purely a matter of chance; even if the noise conspires, by chance, to resemble one cycle of a sine wave, it would still be only pure chance – equally unlikely according to the orthodox sampling distribution – that would make it resemble a second cycle of that wave; and so on.

In almost every application one can think of, our prior knowledge about the real world tells us that in speaking of 'periodicity' we have in mind some systematic physical influence that repeats itself; indeed, our interest in it *is due entirely to the fact that we expect it to repeat*.[12] Thus we expect to see some periodicity in the weather because we know that this is affected by periodic astronomical phenomena; the rotation of the Earth on its axis, its yearly orbital motion about the Sun, and the observed periodicity in sunspot numbers, which affect atmospheric conditions on the Earth. So the hypothesis H_1 that we want to test for is quite unrelated to the hypothesis H_0 that is used in Fisher's test.[13]

This is the kind of logic that underlies all orthodox significance tests. In order to argue for an hypothesis H_1 that some effect exists, one does it indirectly: invent a 'null hypothesis' H_0 that denies any such effect, then argue against H_0 in a way that makes no reference to H_1 at all (that is, using only probabilities conditional on H_0). To see how far this procedure takes us from elementary logic, suppose we decide that the effect exists; that is, we reject H_0. Surely, we must also reject probabilities conditional on H_0; but then what was the logical justification for the decision? Orthodox logic saws off its own limb.[14]

Harold Jeffreys (1939, p. 316) expressed his astonishment at such limb-sawing reasoning by looking at a different side of it: 'An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it.'

Thus, if we say that there is a periodicity in temperature, we mean by this that there is some periodic physical influence at work, the nature of which may not be known with certainty, but about which we could make some reasonable conjectures. For example, the aforementioned periodicity in solar activity, already known to occur by the 11 year periodic variation in sunspot numbers (which many believe, with good reason, to be a rectified

[12] It is not necessary for successful prediction that the physical cause of the periodicity be actually understood; in ancient India records of eclipses were maintained carefully over centuries. From these observations they 'got the rhythm of it' and were able to predict future eclipses very accurately, although they had no conception of their causes.

[13] If an apparent periodicity were only a momentary artifact of the noise as supposed by H_0, we would not consider it a real periodicity at all, and would not want our statistical test to take any note of it. But, unfortunately, it is always possible for noise artifacts to appear momentarily real to any test one can devise. The remedy is to check whether the apparent effect is reproducible; a noise artifact will in all probability never occur again in the same way. A physicist can, almost always, use this remedy easily; an economist usually cannot.

[14] An historical study has suggested that the culprit who started this kind of reasoning was not a statistician, but the physicist Arthur Schuster (1897), who invented the periodogram for the purpose of refuting some claims of periodicity in earthquakes in Japan. Never thinking in terms of information, he achieved his preconceived goal by the simple device of analyzing the data in a way that *threw away* the information about that periodicity! But then this was taken up by many others, including Fisher, Feller, Blackman and Tukey, and Bloomfield. Nevertheless, we shall see that the periodogram does contain basic information that Schuster and his followers failed to recognize. They thought that the information was contained in the *sampling distribution* for the periodogram; whereas the analysis given here shows that it was actually contained in the *shape* of the periodogram.

22 year periodicity), would cause a periodic variation in the number of charged particles entering our atmosphere (the reality of this is shown by the observed periodic variations in the *aurora borealis*), varying the ion concentration and therefore the number of raindrop condensation centers. This would cause periodic variations in the cloud cover, and hence in the temperature and rainfall, which might be very different in different locations on the Earth because of prevailing atmospheric circulation patterns.

We do not mean to say that we firmly believe this mechanism to be the dominant one; only that it is a conceivable one, which does not violate any known laws of physics, but whose magnitude is difficult to estimate theoretically. But already, this prior information prepares us not to be surprised by a periodic variation in temperature in Central Park somewhat like that observed[15] and leads us to conjecture that the July temperatures might give even better evidence for periodicity.

Once a data set has given mild evidence for such a periodicity, its reality could be definitely confirmed or refuted by other observations, correlating other data (astronomical, atmospheric electricity, fish populations, etc.) with weather data at many different locations. A person trained only in orthodox statistics would not hesitate to consider all these phenomena 'independent'; a scientist with some prior knowledge of astrophysics and meteorology would not consider them independent at all.

If editors of scientific journals refuse to publish that first mild evidence on the grounds that it is not significant *in itself* by an orthodox significance test at the 5% level, the confirmatory observations will, in all probability, never be made; a potentially important discovery could be delayed by a century. Physicists and engineers have been largely spared from such fiascos because they hardly ever took orthodox teachings seriously anyway; but others working in economics, artificial intelligence, biology, or medical research who, in the past, allowed themselves to be cowed by Fisher's authority, have not been so fortunate.

Contrast our position just stated with that of Feller (II, p 76–77), who delivers another polemic against what he calls the 'old wrong way'. Suppose the data are expanded in sinusoids:

$$y_t = \sum_{j=1}^{n} (A_j \cos \omega_j t + B_j \sin \omega_j t). \tag{17.49}$$

We can always approximate y_t this way. Then it seems that A_j, B_j must be 'random variables' if the $\{y_t\}$ are. Feller warns us against that old wrong way: fit such a series to the data with well-chosen frequencies $\{\omega_1, \ldots, \omega_n\}$ and assume all A_j, $B_j \sim N(0, \sigma)$. If one of the $R_j^2 = A_j^2 + B_j^2$ is big, conclude that there is a true period. He writes of this:

For a time it was fashionable to introduce models of this form and to detect 'hidden periodicities' for sunspots, wheat prices, poetic creativity, etc. Such hidden periodicities used to be discovered as easily as witches in medieval times, but even strong faith must be fortified by a statistical

[15] One who was also aware of the roughly 20 year periodicity in crop yields, well known to Kansas wheat farmers for a century, would be even less surprised.

test. A particularly large amplitude R_j is observed; One wishes to prove that this cannot be due to chance and hence that ω_j is a true period. To test this conjecture one asks whether the large observed value of R is plausibly compatible with the hypothesis that all n components play the same role.

Apparently, Feller did not even believe in the sunspot periodicity, which no responsible scientist has doubted for over a century; the evidence for it is so overwhelming that nobody needs a 'statistical test' to see it. He states that the usual procedure was to assume the A_j, B_j iid normal $N(0, \sigma)$.[16] Then the R_j^2 are held to be independent with an exponential distribution with expectation $2\sigma^2$. 'If an observed value R_j^2 deviated 'significantly' from this predicted expectation it was customary to jump to the conclusion that the hypothesis of equal weights was untenable, and R_j represented a 'hidden periodicity'. At this point, Feller detects that we are using the wrong sampling distribution:

> The fallacy of this reasoning was exposed by R. A. Fisher who pointed out that the maximum among n independent observations does not obey the same probability distribution as each variable taken separately. The error of treating the worst case statistically as if it had been chosen at random is still common in medical statistics, but the reason for discussing the matter here is the surprising and amusing connection of Fisher's test of significance with covering theorems.

Feller then states that the quantities

$$V_j = \frac{R_j^2}{\sum R_i^2}, \qquad 1 \le j \le n, \tag{17.50}$$

are 'distributed' as the lengths of the n segments into which the interval (0,1) is partitioned by a random distribution of $n - 1$ points. The probability that all $V_j < a$ is then given by a covering theorem noted by Feller.

Of course, our position is that both Feller's 'old wrong' and 'new right' sampling distributions are irrelevant to the inference; the two quantities that are relevant (the prior information that expresses our knowledge of the phenomenon and the likelihood function that expresses the evidence of the data) are not even mentioned.

In any event, the bottom line of this discussion is that Fisher's test fails to detect the perfectly evident 20 year periodicity in the New York Central Park January temperatures. But this is not the only case where simple visual examination of the data is a more powerful tool for inference than the principles taught in orthodox textbooks. Crow, Davis and Maxfield (1960) present applications of the orthodox F-test and t-test which we examine in Jaynes (1976) with the conclusions that (1) the eyeball is a more reliable indicator of an effect than an orthodox equal-tails test, and (2) the Bayesian test confirms quantitatively what the eyeball sees qualitatively. This is also relevant to the notions of domination and admissibility discussed elsewhere.

[16] The abbreviation 'iid' is orthodox jargon standing for 'independently and identically distributed'. For us, this is another form of the mind projection fallacy. In the real world, each individual coefficient A_j, B_j is a definite, fixed quantity that is known from the data; it is not 'distributed' at all!

17.6 A Bayesian analysis

Now we examine a Bayesian analysis of these same data, and for pedagogical reasons we want to explain its rationale in some detail. There may be various different Bayesian treatments of data for periodicity, corresponding to different information about the phenomenon, expressed by different choices of a model, and different prior information concerning the parameters in a model. Our Bayesian model is as follows. We consider it possible that the temperature data have a periodic component due to some systematic physical influence on the weather:

$$A\cos\omega t + B\sin\omega t, \tag{17.51}$$

where, as noted, we may suppose $|\omega| \leq \pi$ (with yearly data it does not make sense to consider periods shorter than a year). In addition the data are contaminated with variable components e_t that we call 'irregular' because we cannot control them or predict them and therefore cannot make allowance for them. This could be because we do not know their real causes or because, although we know the causes, we lack the data on initial conditions that would enable predictions.[17] Then, as explained in Chapter 7, it will almost always do justice to the real prior information that we have to assign a Gaussian sampling distribution with parameters (μ, σ) to the irregulars. There is hardly any real problem in which we would have the detailed prior information that would justify any more structured sampling distribution.

Thus μ is the 'nominal true mean temperature' not known in advance; we can estimate it from the data very easily (intuition can see already that the mean value of the data $\overline{y}$ is about as good an estimate of μ that we can make from the information we have); but it is not of present interest and so we treat it as a nuisance parameter. We do not know σ in advance either, although we can easily estimate it too from the data. But that is not our present interest, and so we shall let σ also be a nuisance parameter to be integrated out as explained in Chapter 7. Our model equation for the data is then

$$y_t = \mu + A\cos\omega t + B\sin\omega t + e_t, \qquad 1 \leq t \leq n, \tag{17.52}$$

and our sampling distribution for the irregular component is

$$p(e_1 \cdots e_n|\mu\sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_t e_t^2\right\}. \tag{17.53}$$

Then the sampling (density) distribution for the data is

$$p(y_1 \cdots y_n|\mu\sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{Q}{2\sigma^2}\right\} \tag{17.54}$$

[17] In meteorology, although the principles of thermodynamics and hydrodynamics that determine the weather are well-understood, weather data taken on a 50 mile grid are grossly inadequate to predict the weather 24 hours in advance. Partial differential equations require an enormous amount of information on initial conditions to determine anything like a unique solution.

with the quadratic form

$$Q(A, B, \omega) \equiv \sum (y_t - \mu - A\cos\omega t - B\sin\omega t)^2 \tag{17.55}$$

or

$$\begin{aligned} Q = n\Big[\ & \overline{y^2} - 2\bar{y}\mu + \mu^2 - 2A\overline{y_t\cos\omega t} - 2B\overline{y_t\sin\omega t} + 2\mu A\overline{\cos\omega t} \\ & + 2\mu B\overline{\sin\omega t} + 2AB\overline{\cos\omega t\sin\omega t} + A^2\overline{\cos^2\omega t} + B^2\overline{\sin^2\omega t}\ \Big], \end{aligned} \tag{17.56}$$

where all the overbar symbols denote sample averages over t. A great deal of detail has suddenly appeared that was not present in the orthodox treatment; but now all of this detail is actually *relevant to the inference*. In any nontrivial Bayesian solution we may encounter much analytical detail because every possible contingency allowed by our information is being taken into account (as is required by our basic desiderata in Chapters 1 and 2). Most of this detail is not perceived at all by orthodox principles, and it would be difficult to handle by paper-and-pencil calculation.

In practice, a Bayesian learns to recognize that much of this detail actually makes a negligible difference to the final conclusions, and so we can almost always make such good approximations that we can do the special calculation needed for our present purpose with pencil and paper after all. But, fortunately, masses of details are no deterrent to a computer, which can happily grind out the exact solution.[18] Now, in the present problem, (A, B, ω) are the interesting parameters that we want to estimate, while (μ, σ) are nuisance parameters to be eliminated. We see that of the nine sums in (17.56), four involve the data y_t; and since this is the only place where the data appear, these four sums are the jointly sufficient statistics for all the five parameters in the problem. The other five sums can be evaluated analytically once and for all, before we have the data.

Now, what is our prior information? Surely, we knew in advance that A, B must be less than 200 °F. If there were a temperature variation that large, New York City would not exist; there would have been a panic evacuation of that area long before, by anyone who happened to wander into it and survived long enough to escape. Thus the empirical fact that New York City *exists* is highly cogent information relevant to the question being asked; it is already sufficient to ensure proper priors for (A, B) in the Bayesian calculation. Also, we have no prior information about the phase $\theta = \tan^{-1}(B/A)$ of any periodicity, which we express by a uniform prior over θ.

We could cite various other bits of relevant prior information, but we know already from the results found in Chapter 6, Exercise 6.6, that, unless we have prior information that reduces the possible range to something like 30 °F, it will make a numerically negligible difference in the conclusions (a strictly nil difference if we report our conclusions only to three decimal digits). So let us see what Bayesian inference gives with just this. By an

[18] Indeed, the exact general solution is often easier to program than is any particular special case of it or approximation to it, because one need not go into the details that make the case special. And the program for the exact solution has the merit of being crash-proof if written to prevent underflow or overflow (for approximations will almost surely break down for some data sets, but the exact solution – with proper priors – must always exist for every possible data set.

argument essentially the same as the Herschel derivation of the Gaussian distribution in Chapter 7, we may assign a joint prior

$$p(AB|I) = \frac{1}{2\pi\delta^2}\exp\left\{-\frac{A^2+B^2}{2\delta^2}\right\}, \tag{17.57}$$

where δ is of the order of magnitude of 100 °F; we anticipate that its exact numerical value can have no visible effect on our conclusions (nevertheless, such a proper prior may be essential to prevent computer crashes).

Now the most general application of Bayes' theorem for this problem would proceed as follows. We first find the joint posterior distribution for all five parameters:

$$p(AB\omega\mu\sigma|DI) = p(AB\omega\mu\sigma|I)\frac{p(D|AB\omega\mu\sigma I)}{p(D|I)}. \tag{17.58}$$

Then integrate out the nuisance parameters:

$$p(AB\omega|DI) = \int \mathrm{d}\mu \int \mathrm{d}\sigma\; p(AB\omega\mu\sigma|DI). \tag{17.59}$$

But this is a far more general calculation than we need for present purposes; it is prepared to take into account arbitrary correlations in the prior probabilities. Indeed, we can always factor the prior thus:

$$p(AB\omega\mu\sigma|I) = p(AB\omega|I)p(\mu\sigma|AB\omega I); \tag{17.60}$$

and so the most general solution appears formally simpler:

$$p(AB\omega|DI) = Cp(AB\omega|I)L^*(A, B, \omega), \tag{17.61}$$

where C is a normalization constant, and L^* is the quasi-likelihood

$$L^*(A, B, \omega) \equiv \int \mathrm{d}\mu \int \mathrm{d}\sigma\; p(\mu\sigma|AB\omega I)p(D|AB\omega\mu\sigma I). \tag{17.62}$$

In (17.61) the nuisance parameters are already out of sight. But in our present problem, evidently knowledge of the parameters (A, B, ω) of the systematic periodicity would tell us nothing about the parameters (μ, σ) of the irregulars; so the prior for the latter is just

$$p(\mu\sigma|AB\omega I) = p(\mu\sigma|I), \tag{17.63}$$

so what is our prior information about (μ, σ)? Surely we knew also, for the same 'panic evacuation' reason, that neither of these parameters could be as large as 200 °F. And we know that σ could not be as small as 10^{-6} °F, because, after all, our data are taken with a real thermometer, and no meteorologist's thermometer can be read to that accuracy (if it could, it would not give reproducible readings to that accuracy). We could just as well ignore that practical consideration and argue that σ could not be as small as 10^{-20} °F because the concept of temperature is not defined, in statistical mechanics, to that accuracy. Numerically, it will make no difference at all in our final conclusions, but it is still conceivable that a proper prior may be needed to avoid computer crashes in all contingencies. So, to be on the

safe side, we assign the prior Gaussian in μ, because it is a location parameter, a truncated Jeffreys prior for σ, because we have seen in Chapter 12 that the Jeffreys prior is uniquely determined as the only completely uninformative prior for a scale parameter:

$$p(\mu\sigma|I) \propto \frac{1}{\sigma\sqrt{2\pi\alpha^2}} \exp\{-\mu^2/2\alpha^2\}, \qquad a \leq \sigma \leq b, \tag{17.64}$$

in which α and b are also of the order of 100 °F, while $a \simeq 10^{-6}$; we are only playing it extremely safe in the expectation that most of this care will prove in the end to have been unnecessary.

Our quasi-likelihood is then

$$L^*(A, B, \omega) = \int_{-\infty}^{\infty} \mathrm{d}\mu \, \exp\{-\mu^2/2\alpha^2\} \int_a^b \frac{\mathrm{d}\sigma}{\sigma^{n+1}} \exp\{-Q/2\sigma^2\}. \tag{17.65}$$

But now it is evident that the finite limits on σ are unnecessary; for if $n > 0$ the integral over σ converges both at zero and infinity, and

$$\int_0^{\infty} \frac{\mathrm{d}\sigma}{\sigma^{n+1}} \exp\{-Q/2\sigma^2\} = \frac{1}{2}\frac{(n/2-1)!}{(Q/2)^{n/2}}, \tag{17.66}$$

and the integral of this over μ is also guaranteed to converge. For tactical reasons, let us do the integration over μ first. We begin by rewriting Q as

$$Q = n[s^2 + (\mu - \overline{d})^2]. \tag{17.67}$$

Editor's Exercise 17.3(a) The equation for Q is formally identical to (7.29); however, as written, none of the quantities were defined by Jaynes. Show that s^2 may be written as

$$s^2 \equiv \overline{d^2} - \overline{d}^2, \tag{17.68}$$

where $\overline{d}$ and $\overline{d^2}$ are the mean and mean-square of an effective data defined as

$$d_i = y_i - A\cos(\omega t_i) - B\sin(\omega t). \tag{17.69}$$

(b) Evaluate the integral over u and σ to obtain the marginal $p(AB\omega|DI)$.

(c) Unfortunately, the $p(AB\omega|DI)$ does not summarize all of the information in the data concerning frequency estimation, to do that we need $p(\omega|DI)$; derive it in closed form.

(d) The posterior probability $p(\omega|DI)$ makes the implicit assumption that a resonance is present and so will estimate the frequency regardless of whether or not such a resonance exists. How would you use probability theory and the results derived so far to determine if a resonance is present?[19]

[19] For an example of such a signal detection statistic, see my article: Bretthorst, G. L. (1990), *J. Mag. Resonance* **88,** 571–595.

17.7 The folly of randomization

Many writers introduce randomized methods by the example of 'Monte Carlo integration'. Let a function $y = f(x)$ have its domain of existence in the unit square $0 \le x, y \le 1$; we wish to compute numerically the integral

$$\theta \equiv \int_0^1 \mathrm{d}x\, f(x). \tag{17.70}$$

Perhaps this is too complicated analytically, or perhaps $f(x)$ was only empirically determined; we do not have it in analytical form. Then let us just choose n points at random (x, y) in the unit square and determine for each whether it lies below the graph of $f(x)$; that is, whether $y \le f(x)$. Let the number of such points be r; then we estimate the integral as $(\theta)_{\text{est}} = r/n$ and as $n \to \infty$ we might expect this to approach the correct Riemannian integral; but how accurate is it? Always, one would suppose independent binomial sampling: the sampling distribution for r is taken to be

$$p(r|n\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \tag{17.71}$$

which has (mean) $\pm$ (standard deviation) of

$$\theta \pm \sqrt{\frac{\theta(1-\theta)}{n}}, \tag{17.72}$$

and if the width of the sampling distribution is held to indicate the accuracy of our estimate, one would think it reasonable to assign a probable error to $(\theta)_{\text{est}}$ given by

$$(\theta)_{\text{est}} = \frac{r}{n} \pm \sqrt{\frac{r(n-r)}{n^3}}. \tag{17.73}$$

For example, suppose the true θ is 1/2, and $n = 100$. Then, having observed $r = 43$ we would get the estimate of

$$(\theta)_{\text{est}} = 0.43 \pm \sqrt{\frac{0.43 \times 0.57}{n}} = 0.43 \pm 0.05, \tag{17.74}$$

or an accuracy of about 11.5%. But the trouble with such methods is that they improve only as $1/\sqrt{n}$.

Now let's take our n sampling points in a nonrandomized way on a uniform grid: divide the unit square into $\sqrt{n}$ steps each way, take one sampling point at each grid point, and again count how many (r) are below the curve. The maximum error we can make in each step is

$$[\text{error in determining } f(x)] \times [\text{width of step}] = \frac{1}{2\sqrt{n}} \times \frac{1}{\sqrt{n}} = \frac{1}{2n}. \tag{17.75}$$

Therefore the maximum possible error in the integral is

$$[\text{number of steps}] \times [\text{maximum error in each step}] = \frac{1}{2\sqrt{n}}. \tag{17.76}$$

So if $\theta \simeq 0.5$, the *probable error* in the Monte Carlo method is about equal to the *maximum possible error* in the uniform grid sampling method. But the probable error in the uniform grid method is much less than this: the central limit theorem tells us that, with a rectangular distribution of error probability in each step, the expected square of the error in determining $f(x)$ in that step is

$$\sqrt{n}\int_0^{\sqrt{n}} \mathrm{d}x\,\left(x - \frac{1}{2\sqrt{n}}\right)^2 = \frac{1}{12n^2}. \tag{17.77}$$

If the errors in different steps are independent, the expected square of the total error is

$$(\text{mean square error per step}) \times (\text{number of steps}) = \frac{1}{12n^{3/2}}, \tag{17.78}$$

and the probable error in the integral is about

$$\pm\frac{1}{\sqrt{12}n^{3/4}}. \tag{17.79}$$

Thus, if $n = 100$ and $\theta \simeq 0.5$, the Monte Carlo method gives a probable error of about 0.05, the uniform grid sampling gives 0.00913, less than one-fifth as much. With $n = 1000$, the Monte Carlo probable error is 0.0158, the uniform grid probable error is 0.00162, about one-tenth as much. The uniform grid calculation at $n = 100$ points yields the same probable error as does the Monte Carlo method at $n = 3000$ points. This corresponds rather nicely to the italicized statement following (17.16). Another example is given by Royall and Cumberland (1981); this is particularly cogent because the authors are not Bayesian and did not start out with the intention of exposing the folly of randomization, but did so anyway.

17.8 Fisher: common sense at Rothamsted

From the study of several such examples, we propose as a general principle: *Whenever there is a randomized way of doing something, there is a nonrandomized way that yields better results from the same data, but requires more thinking*. Perhaps this principle does not have quite the status of a theorem, but we are confident that, whenever one is willing to do the required extra thinking, it will be confirmed.

17.8.1 The Bayesian safety device

We note that Bayesian methods are not only more powerful than orthodox ones; they are also safer (i.e. they have automatic built-in safety devices that prevent them from misleading us with the over-optimistic or over-pessimistic conclusions that orthodox methods can produce). It is important to understand why this is true. In parameter estimation, for example, whether or not there is a sufficient statistic, the log-likelihood function is

$$\log L(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha) = n\overline{\log p(x_i|\alpha)}, \tag{17.80}$$

in which we see the average of the log-likelihoods over each individual data point. The log-likelihood is always spread out over the full range of variability of the data, so if we happen to get a very bad (spread out) data set, no good estimate is possible and Bayes' theorem warns us about this by returning a wide posterior distribution. With a location parameter $p(x|\alpha) = h(x - \alpha)$ and an uninformative prior, the width of the posterior distribution for α is essentially $(R + W)$

$$\text{(range of the data)} + \text{(width of individual likelihoods)}. \tag{17.81}$$

If we happen to get a very good (sharply concentrated) data set, a more accurate estimate of α is possible and Bayes' theorem takes advantage of this, returning a posterior distribution whose width approaches a lower bound determined by that of the single-point likelihood $L_i(\alpha) = p(x_i|\alpha)$ and the amount n of data.

In the orthodox method, the accuracy claim is essentially the width of the sampling distribution for whatever estimator β we have chosen to use. But this takes no note of the range of the data! Orthodox estimation based on a single statistic will claim just the same accuracy whether the data range is large or small. Far worse, that accuracy expresses entirely the variability of the estimator *over other data sets that we think might have been obtained but were not*. But again this concentrates attention on an irrelevancy, while ignoring what is relevant; unobserved data sets are only a figment of our imagination. Surely, if we are only imagining them, we are free to imagine anything we please. That is, given two proposed conjectures about unobserved data, what is the test by which we could decide which one is correct?

In spite of its mathematical triviality, we stress the fundamental importance of (17.80) for demonstrating the inner mechanism of Bayes' theorem. It clarifies several other questions often raised about Bayesian methods. We note one of the most important.

17.9 Missing data

This is a problem that does not exist for us; Bayesian methods work by the same algorithm whatever data we have. For example, in estimating a parameter θ from a data set $D \equiv \{x_i\}$, where the indices i refer to the times $\{t_i\}$ of observation and take on values in some set T, the data affect the result through the likelihood function $L(\theta)$, given by

$$\log L(\theta) = \sum_{i \in T} \log p(x_i|\theta), \tag{17.82}$$

where the sum is over whatever data values we have. The point is that, whether the times $\{t_i\}$ are consecutive and equally spaced, or completely irregular with large gaps, makes no difference; probability theory as logic tells us that (17.82) yields the optimal inference, which captures all the evidence in the data set that we happen to have. One can write a single computer program which, once and for all, accepts whatever data (that is, whatever set of numbers $\{x_i; t_i\}$) we give it, and proceeds to do the correct calculation for that data set.

In contrast, note what happens in orthodox statistics, where estimation is obliged to proceed through the sampling distribution of some 'statistic' $\theta^*(x_i)$. If any data are missing from the set T which was assumed in setting up the problem, one has two ways of dealing with this. Firstly, the theoretically correct procedure would recognize that this not only changes the sampling distribution for any statistic; it requires one to go back to the beginning and reconsider the whole problem. This can get us into a horrendous situation – every different kind of missing data or extra data can oblige us to define a new sample space, choose a new statistic θ^{**} and calculate a new sampling distribution $p(\theta^{**}|\theta)$.

Alternatively, one can invent a new *ad hockery* and try to estimate the missing data values from the ones we have, and use these as if they were real data. Obviously, this procedure is not only unjustified logically, it is highly ambiguous because that estimation could be made in many different ways. These difficulties are seen at first hand in Little and Rubin (1987).

The missing data problem was so cumbersome in orthodox statistics that some who saw the light and moved over into the Bayesian camp failed to perceive that they had left this problem behind them. Instead of applying such simple rules as (17.82) directly, which would have led them immediately to the correct solution whatever the data, they proceeded out of force of habit to follow the orthodox custom and invent new *ad hoc* devices like the one just noted, as 'corrections' to the Bayesian or maximum entropy methods, thus grotesquely mutilating them and getting a worse inference by a bigger computation. To those accustomed to orthodox difficulties, the power and simplicity of the Bayesian method in this application seems unbelievable; and one must think long and hard to understand how it is possible.

As a more general comment, there is a simple strategy that will serve in almost all of these Bayes/orthodox comparisons: 'magnification', as demonstrated for the chi-squared test in Chapter 9. When we find a quantitative difference in the orthodox and Bayesian conclusions, it may appear at first glance so small that our common sense is unable to judge the issue. But then we can usually find some extreme problem in which the small difference is magnified into a large one, or preferably to a qualitative difference in the conclusions. Our common sense will then tell us very clearly which procedure is giving reasonable results, and which one is not. Indeed, it is often possible to magnify to the point where one procedure is yielding an obvious violation of deductive reasoning or pathology like that noted above for an unbiased estimate. Now we examine another very important example where we can compare orthodox and Bayesian results by magnification.

17.10 Trend and seasonality in time series

The observed time series generated by the real world seldom appear to be 'stationary' but exhibit more complicated behavior. In most series, particularly demographic or economic data, trend is the most common form of nonstationarity. Many economic time series are so dominated by trend (due, for example, to steadily rising population, inflation, or technological advances) that any attempt to study other regularities, such as cyclical fluctuations or settling back after response to a shock – and, particularly, correlations between different

time series – can be more misleading than helpful until we have a safe way of dealing with trend.

The story is told – perhaps apocryphal – of a researcher who announced the discovery of a strong positive correlation between membership in the Church of England and incidence of suicide in England, and concluded that it would be safer to keep away from the Church. The true explanation was, of course, that the population of England was growing steadily, so membership in the Church, incidence of suicides – and almost any other demographic variable – were all growing together. False correlations of this type have led many to nonsensical conclusions because of the almost universal tendency to jump to the conclusion that correlation implies causation.

The problem of contaminated data has been with us from the very beginning. We noted in Chapter 9 how Edmund Halley (1693) was obliged to deal with it in compiling the first tables of mortality. The real key to dealing with it is recognition of the useful functional role of nuisance parameters in probability theory.

Likewise, today many time series are so dominated by cyclic fluctuations (seasonal effects in economic data, periodicity in weather, hum in electrical circuits, synchronized growth in bacteriology, vibrations in helicopter blades) that the attempt to extract an underlying 'signal', such as a long-term trend from a short run of data, is frustrated. We want to contrast how orthodox statistics and probability theory as logic deal with the problem of extracting the information one wants, in spite of such data contaminations.

17.10.1 Orthodox methods

The traditional procedures do not apply probability theory to this problem; instead, one resorts to inventing the same kind of intuitive *ad hoc* devices that we have noted so often before. The usual ones are called 'detrending' and 'seasonal adjustment' in the economic literature, 'filtering' in the electrical engineering literature. Like all such *ad hockeries* not derived from first principles, they capture enough of the truth to be usable in some problems; but they can generate disaster in others.

The almost universal detrending procedure in economics is to suppose the data (or the logarithm of the data) to be $y(t) = x(t) + Bt + e(t)$, composed additively of the component of interest $x(t)$, a linear 'trend' Bt, and a 'random error' or 'noise' $e(t)$. We estimate the trend component, subtract it from the data, and proceed to analyze the resulting 'detrended data' for other effects. However, many writers have noted that detrending may introduce spurious artifacts that distort the evidence for other effects. Detrending may even destroy the relevance of the data for our purposes, and we saw in the scenario of the weather in Central Park that filtering of data can also do this.

Similarly, the traditional way of dealing with seasonal effects is to produce 'seasonally adjusted' data, in which one subtracts an estimate of the seasonal component from the true data, then tries to analyze the adjusted data for other effects. Indeed, most of the economic time series data one can obtain have been rendered nearly useless because they have been seasonally adjusted in an irreversible way that has destroyed information which probability

theory could have extracted from the raw data. We think it imperative that this be recognized, and that researchers be able to obtain the true data – free of detrending, seasonal adjustment, pre-filtering, smoothing, or any other destructive mutilation of the information in the data.

Electrical engineers would think instead in terms of Fourier analysis and resort to 'high-pass filters' and 'band-rejection filters' to deal with trend and seasonality. Again, the philosophy is to produce a new time series (the output of the filter) which represents in some sense an estimate of what the real series would be if the contaminating effect were absent. Then choice of the 'best' physically realizable filter is a difficult and basically indeterminate problem; fortunately, intuition has been able to invent filters good enough to be usable if one knows in advance what kind of contamination will occur.

17.10.2 The Bayesian method

The direct application of probability theory as logic leads us to an entirely different philosophy; always, the correct procedure is to calculate the probability of whatever is unknown and of interest, conditional on whatever is known. This means that we do not seek to remove the trend or seasonal component from the data: that is fundamentally impossible because there is no way to know the 'true' trend or seasonal term. Any assumption about them is necessarily in some degree arbitrary, and is therefore almost certain to inject false information into the detrended or seasonally adjusted series. Rather, we seek to remove the effect of trend or seasonality from our *final conclusions*, taking into account all the relevant information we have, while leaving the actual data intact. We develop the Bayesian procedure for this and compare it in detail with the conventional one.

Firstly, we analyze the simplest possible nontrivial model, which can be solved completely from either point of view and will enable us to understand the exact relationship between the two procedures. Having this understanding, the extension to the most complicated multivariate case will be an easy mathematical generalization – essentially, just promotion of numbers to matrices while retaining the same formal equations.

Suppose the model consists of only a single sinusoid and a linear trend: $y(t) = A\sin\omega t + Bt + e(t)$, where A is the amplitude of interest to be estimated and B is the unknown trend rate. If the data are monthly economic data and the sinusoid represents a yearly seasonal effect, then ω will be $2\pi/12 = 0.524$ months^{-1}. But, for example, if we are trying to detect a cycle with a period of 20 years, ω will be $0.524/20 = 0.00262$. Estimation of an unknown ω from such data is the very important problem of spectrum analysis, considered in the scenario of the weather in Central Park. But for the present we consider the case where ω is known (usually, because we know that the seasonality has a period of one year for unchanging astronomical reasons). Writing for brevity $s_t = s(t) \equiv \sin(\omega t)$, our model equation is then

$$y(t) = As(t) + Bt + e(t), \tag{17.83}$$

and the available data $y \equiv (y_1, \ldots, y_N)$ are values of this at N equally spaced times $t = 1, 2, \ldots, N$. Assigning – for reasons already explained sufficiently – the noise components

e_i an independent Gaussian prior probability density function $e_t \sim N(0, \sigma)$ with variance σ^2, the sampling pdf for the data is

$$p(y|AB\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left\{-\frac{N}{2\sigma^2} Q(A, B)\right\}, \tag{17.84}$$

and, as in any Gaussian calculation, the first task is to rearrange the quadratic form

$$Q(A, B) \equiv \frac{1}{N} \sum_t (y_t - As_t - Bt)^2 = \overline{y^2} + A^2\,\overline{s^2} + B^2\overline{t^2} - 2A\overline{sy} - 2B\overline{ty} + 2AB\overline{st}, \tag{17.85}$$

where

$$\overline{y^2} \equiv \frac{1}{N}\sum_{t=1}^{N} y_t^2, \qquad \overline{sy} \equiv \frac{1}{N}\sum_{t=1}^{N} s_t y_t, \tag{17.86}$$

etc. denote averages over the data sample. Three of these averages, $(\overline{s^2}, \overline{t^2}, \overline{st})$ are determined by the 'design of the experiment' and can be known before one has the data. In fact, we have nearly

$$\overline{s^2} \simeq 1/2, \qquad \overline{t^2} \simeq N^2/3 \tag{17.87}$$

with errors of relative order $O(1/N)$. But $\overline{st}$ is highly variable; it is certainly less than $N/2$, since that could be achieved only if $s(t) = 1$ at every sampling point. Generally, $\overline{st}$ is much less than this, of the order $\overline{st} \simeq 1/\omega$ due to near cancellation of positive and negative terms.

The other three averages $(\overline{y^2}, \overline{sy}, \overline{ty})$ depend on the data, and, since they are the only terms containing the data, they are the jointly sufficient statistics for our problem, to be calculated as soon as one has the data.

Suppose that it is the seasonal amplitude A that we wish to estimate, while the trend rate B is the nuisance parameter that contaminates our data. We want to make its effects disappear, as far as is possible. We shall do this by finding the joint posterior distribution for A and B,

$$p(AB|DI), \tag{17.88}$$

and integrating out B to get the marginal posterior distribution for A,

$$p(A|DI) = \int \mathrm{d}B\; p(AB|DI). \tag{17.89}$$

This is the quantity that tells us everything the data D and prior information I have to say about A, whatever the value of B; this would be called 'Bayesian detrending'. Conversely, if we wanted to estimate B, then A would be the nuisance parameter, and we would integrate it out of (17.88) to get the marginal posterior distribution $p(B|DI)$; and this would be called 'Bayesian seasonal adjustment'.

In the limit of diffuse priors for A and B (i.e. their prior densities do not vary appreciably over the region of high likelihood), the appropriate integration formula for (17.89) is

$$\int_{-\infty}^{\infty} \mathrm{d}B \, \exp\left\{-\frac{N\,Q(A,B)}{2\sigma^2}\right\} = (\text{const.}) \times \exp\left\{-\frac{N}{2\sigma^2}\left[\frac{\left(\overline{s^2}\right)\left(\overline{t^2}\right) - \left(\overline{st}\right)^2}{\overline{t^2}}\right]\left(A-\hat{A}\right)^2\right\}, \tag{17.90}$$

where

$$\hat{A} \equiv \frac{\left(\overline{t^2}\right)\left(\overline{sy}\right) - \left(\overline{st}\right)\left(\overline{ty}\right)}{\left(\overline{s^2}\right)\left(\overline{t^2}\right) - \left(\overline{st}\right)^2} \tag{17.91}$$

and the (const.) is independent of A. Thus the marginal posterior distribution for A is proportional to (17.90), and the Bayesian posterior (mean) $\pm$ (standard deviation) estimate of A, regardless of the value of B, is

$$(A)_{\text{est}} = \hat{A} \pm \sigma \sqrt{\frac{\overline{t^2}}{N\left[\left(\overline{s^2}\right)\left(\overline{t^2}\right) - \left(\overline{st}\right)^2\right]}} = \hat{A} \pm \frac{\sigma}{\sqrt{N\overline{s^2}(1-r^2)}}, \tag{17.92}$$

where

$$r \equiv \frac{\overline{st}}{\sqrt{\overline{s^2}\,\overline{t^2}}} \tag{17.93}$$

is the correlation coefficient of s and t.

Some orthodox writers have railed against this process of integrating out nuisance parameters – in spite of the fact that it is uniquely determined by the rules of probability theory as the correct procedure – on the usual grounds that the probability of a parameter is meaningless because a parameter is not a 'random variable'. Even worse, in the integration we introduced a prior that they consider arbitrary (although for us it represents our real state of prior information which is clearly relevant to the inference, but is ignored by orthodoxy). But, independently of all such philosophical hangups, we can examine the facts of actual performance of the Bayesian and orthodox procedures.

The integration of a nuisance parameter may be related to the detrending procedure as follows. The joint posterior pdf may be factored into marginal and conditional pdfs in two different ways:

$$p(AB|DI) = p(A|DI)p(B|ADI), \tag{17.94}$$

or, equally well,

$$p(AB|DI) = p(A|BDI)p(B|DI). \tag{17.95}$$

From (17.94) we see that (17.89) follows at once, and from (17.95) we see that (17.89) can be written as

$$p(A|DI) = \int \mathrm{d}B\; p(A|BDI)p(B|DI). \tag{17.96}$$

Thus the marginal pdf for A is a weighted average of the conditional pdfs that we would have if B were known:

$$p(A|BDI). \tag{17.97}$$

But if B is known, then (17.97), in its dependence on A, is just (17.84) with B held fixed. This is, from (17.85),

$$p(A|BDI) \propto \exp\left\{-\frac{N\overline{s^2}}{2\sigma^2}(A - A^*)^2\right\}, \tag{17.98}$$

where

$$A^* \equiv \frac{\overline{sy} - B\overline{st}}{\overline{s^2}}. \tag{17.99}$$

This is just the estimate that one would make by ordinary least squares fitting of $As(t)$ to the detrended data $y(t)_{\mathrm{det}} \equiv y(t) - Bt$

$$A^* = \frac{(\overline{sy})_{\mathrm{det}}}{\left(\overline{s^2}\right)}. \tag{17.100}$$

That is, A^* is the estimate the orthodoxian would make if he estimated the trend rate to be B. Of course, if his estimate of B were exactly correct, then he would indeed find the best estimate possible; but any error in his estimate of the trend rate will bias his estimate of A.

The Bayesian estimate of A obtained from (17.96) does not assume any particular trend rate B; it is a weighted average over all possible values that the trend rate might have, weighted according to their respective posterior probabilities. Thus, if the trend rate is very well determined by the data, so that the probability $p(B|DI)$ in (17.96) has a single very sharp peak at $B = B^*$, then the Bayesian and orthodoxian will be in essential agreement on the estimate of A if the orthodoxian also happens to estimate B as B^*. If the trend rate is not well determined by the data, then the Bayes estimate is a more conservative one that takes into account all possible values that B might have, while the orthodox estimate can vary widely.

Although an orthodoxian might accept what we have done as mathematically consistent, this argument would not convince him of the superiority of the Bayesian estimate (implicitly based, as usual, on a quadratic loss function), because he judges estimates by a different criterion. So let us compare them more closely.

17.10.3 Comparison of Bayesian and orthodox estimates

Having found a Bayesian estimator, which theorems demonstrate to be optimal by the Bayesian criterion of performance, nothing prevents us from examining its performance from the orthodox sampling theory viewpoint and comparing it with orthodox estimates. We introduce a useful method of doing this, which makes it clear what the two methods are doing. Let A_0 and B_0 be the unknown true values of the parameters, and let us describe the situation as it would appear to one who already knew A_0 and B_0, but not what data we shall find. As he would know, but we would not, our data vector will in fact be

$$y_t = A_0 s_t + B_0 t + e_t, \tag{17.101}$$

and we shall calculate the statistic

$$\overline{sy} = A_0 \overline{s^2} + B_0 \overline{st} + \overline{se} \tag{17.102}$$

in which the first two terms are fixed (i.e. independent of the noise) and only the last varies with different noise samples. Similarly, he knows that we shall find the statistic

$$\overline{ty} = A_0 \overline{st} + B_0 \overline{t^2} + \overline{te}. \tag{17.103}$$

Although $\overline{sy}$ and $\overline{ty}$ are known to us from the data, we cannot solve (17.102) and (17.103) for $(A_0,\ B_0)$ because $\overline{se}$ and $\overline{te}$ are unknown. We are obliged to continue using probability theory to get the best possible estimates of $A_0,\ B_0$. Substituting (17.102) and (17.103) into (17.91), we find that the Bayes estimate that we shall get reduces to

$$(\hat{A})_{\text{Bayes}} = A_0 + \frac{\left(\overline{t^2}\right)(\overline{se}) - (\overline{st})(\overline{te})}{\left(\overline{s^2}\right)\left(\overline{t^2}\right) - (\overline{st})^2}, \tag{17.104}$$

which is independent of the true trend rate, B_0 having cancelled out. Therefore the Bayesian estimate does indeed eliminate the effect of trend entirely from our conclusions; one could hardly do so more completely than that. But the unknown error vector e must, necessarily, produce some error in our estimate of A, and (17.104) tells us exactly how much.

On the other hand, if the orthodoxian uses the conventional ordinary least squares estimator (17.100) from detrended data $[y_t - \hat{B}t]$ based on any estimate $\hat{B}$, he will find instead

$$(\hat{A})_{\text{orthodox}} = A_0 + \frac{\overline{se} + (B_0 - \hat{B})\overline{st}}{\overline{s^2}}, \tag{17.105}$$

and any error in the trend rate estimate $\hat{B}$ contributes to the error in his estimate of the seasonal component. If, as is the usual practice, one uses the ordinary least squares estimate

of the trend from the original data,

$$\hat{B} = \frac{(\overline{ty})}{(\overline{t^2})}, \tag{17.106}$$

(17.105) becomes

$$(\hat{A})_{\text{orthodox}} = A_0 + \frac{(\overline{t^2})(\overline{se}) - (\overline{ty})(\overline{st}) + B_0(\overline{t^2})(\overline{st})}{(\overline{t^2})(\overline{s^2})} = (1 - r^2)A_0 + \frac{(\overline{se})}{(\overline{s^2})}, \tag{17.107}$$

where we have again used (17.102) and (17.103). Thus (17.107) is also exactly independent of the true trend rate B_0. But orthodox teaching would hold that the estimator (17.107) has a negative bias, since $\overline{se}$ is, 'on the average', zero. One might wish to 'correct' for this by the same device as in (17.6): by multiplying by a suitable factor. But this is not obviously the best procedure. It is far from clear that the optimal estimator can be found merely by multiplying the ordinary least squares estimate by a constant.

Likewise, having recognized what he would consider a shortcoming of (17.107), and perceiving that the Bayesian result (17.104) has at least the merit (from his viewpoint) of being unbiased, it still would not follow that the Bayesian solution is the best possible one. Indeed, one who has absorbed a strong anti-Bayesian indoctrination would, we suspect, reject any such suggestion and would say that we should be able to correct the defects of (17.107) by a little more careful thinking about the problem from the orthodox viewpoint. Let us try.

17.10.4 An improved orthodox estimate

Starting back at the beginning of the problem, orthodox reasoning proceeded as follows. If one had in mind only the seasonal term and was not aware of trend, one would be led to estimate the cyclic amplitude as

$$\hat{A}^{(0)} = \frac{(\overline{sy})}{(\overline{s^2})}, \tag{17.108}$$

the conventional regression solution. Many different lines of reasoning, including ordinary least squares fitting of the data to the sinusoid As_t, lead us to this result.

But then one realizes that (17.108) is not a very good estimate because it ignores the disturbing effect of trend. A better seasonal estimate could be made from the detrended data

$$(y_t)_{\text{det}} \equiv y_t - \hat{B}t, \tag{17.109}$$

where $\hat{B}$ is an estimate of the trend rate, and it seems natural to estimate it by the conventional regression rule

$$\hat{B}^{(0)} = \frac{(\overline{ty})}{(\overline{t^2})} \tag{17.110}$$

from ordinary least squares fitting of a straight line Bt to the data. Using the detrended data (17.109) in (17.108) yields the 'corrected' cyclic amplitude estimate

$$\hat{A}^{(1)} = \frac{\overline{sy} - \overline{st}\,\hat{B}^{(0)}}{\overline{s^2}} = \frac{(\overline{t^2})(\overline{sy}) - (\overline{st})(\overline{ty})}{(\overline{t^2})(\overline{s^2})} \tag{17.111}$$

which is the conventional orthodox result for the problem.

But now we see that this is not the end of the story; for A and B enter into the model on just the same footing. If it is true that we should estimate the cyclic amplitude A from detrended data $y_t - \hat{B}^{(0)}t$, surely it is equally true that we should estimate the trend rate B from the decyclized data $y_t - \hat{A}^{(0)}s_t$. Thus a better estimate of trend than (17.110) would be

$$\hat{B}^{(1)} = \frac{(\overline{ty}) - (\overline{st})\hat{A}^{(0)}}{(\overline{t^2})} = \frac{(\overline{s^2})(\overline{ty}) - (\overline{st})(\overline{sy})}{(\overline{t^2})(\overline{s^2})}, \tag{17.112}$$

where we used (17.108). But now, with this better estimate of trend, we can obtain a better estimate of the seasonal component than (17.111) by using (17.112):

$$\hat{A}^{(2)} = \frac{(\overline{sy}) - (\overline{st})\hat{B}^{(1)}}{(\overline{s^2})}. \tag{17.113}$$

This improved estimate of the seasonal amplitude will in turn enable us to achieve a still better estimate of trend

$$\hat{B}^{(2)} = \frac{(\overline{ty}) - (\overline{st})\hat{A}^{(1)}}{(\overline{t^2})} \tag{17.114}$$

... and so on, forever!

Therefore, the reasoning underlying the conventional detrending procedure, if applied consistently, does not stop at the conventional result (17.100). It leads us into an infinite sequence of back-and-forth revisions of our estimates, each set $[\hat{A}^{(n)}, \hat{B}^{(n)}]$ better than the previous $[\hat{A}^{(n-1)}, \hat{B}^{(n-1)}]$. Does this infinite sequence converge to a final 'best of all' set of estimates $[\hat{A}^{(\infty)}\hat{B}^{(\infty)}]$? If so, this is surely the optimal way of dealing with a nuisance parameter from the orthodox viewpoint. But can we calculate these final optimal estimates directly without going through the infinite sequence of updatings?

To answer this define the (2×1) vector of nth order estimates:

$$V_n \equiv \begin{pmatrix} \hat{A}^{(n)} \\ \hat{B}^{(n)} \end{pmatrix}. \tag{17.115}$$

Then the general recursion relation is, as we see from (17.111)–(17.113),

$$V_{n+1} = V_0 + M V_n, \tag{17.116}$$

where the matrix M is

$$M = \begin{pmatrix} 0 & -\dfrac{\overline{(st)}}{\overline{(s^2)}} \\ \dfrac{\overline{(st)}}{\overline{(t^2)}} & 0 \end{pmatrix}. \tag{17.117}$$

The solution of (17.116) is

$$V_n = (1 + M + M^2 + \cdots + M^n)V_0. \tag{17.118}$$

By the Schwartz inequality, $\overline{(st)}^2 \leq \overline{(s^2)}\,\overline{(t^2)}$, the eigenvalues of M are less than unity, so as $n \to \infty$ this infinite series sums to

$$V_\infty = (I - M)^{-1} V_0. \tag{17.119}$$

Now we find readily that

$$(I - M)^{-1} = \frac{1}{\overline{(s^2)}\,\overline{(t^2)} - \overline{(st)}^2} \begin{pmatrix} \overline{(t^2)}\,\overline{(s^2)} & -\overline{(t^2)}\,\overline{(st)} \\ -\overline{(s^2)}\,\overline{(st)} & \overline{(t^2)}\,\overline{(s^2)} \end{pmatrix}, \tag{17.120}$$

and so our final, best of all, estimate is

$$\hat{A}^{(\infty)} = \frac{\overline{(t^2)}\,\overline{(s^2)}\,\hat{A}^{(0)} - \overline{(t^2)}\,\overline{(st)}\,\hat{B}^{(0)}}{\overline{(s^2)}\,\overline{(t^2)} - \overline{(st)}^2} = \frac{\overline{(t^2)}\,\overline{(sy)} - \overline{(st)}\,\overline{(ty)}}{\overline{(s^2)}\,\overline{(t^2)} - \overline{(st)}^2}. \tag{17.121}$$

But this is precisely the Bayesian estimate that we calculated far more easily in (17.92)! Likewise, the final best possible orthodox estimate of trend rate is

$$\hat{B}^{(\infty)} = \frac{\overline{(s^2)}\,\overline{(ty)} - \overline{(ty)}\,\overline{(sy)}}{\overline{(s^2)}\,\overline{(t^2)} - \overline{(st)}^2}, \tag{17.122}$$

which is just the Bayesian estimate that we find by integrating out A as a nuisance parameter from (17.88).

This is another example of what we found in Chapter 13: if the orthodoxian will think his estimation problems through to the end, he will find himself obliged to use the Bayesian mathematical algorithm, even if his ideology still leads him to reject the Bayesian rationale

for it. Independently of all philosophical hangups, this mathematical form is determined by elementary requirements of rationality and consistency.

Now we see the relationship between the orthodox and Bayesian procedures in an entirely different light. The Bayesian procedure of integrating out a nuisance parameter is summing an infinite series of mutual updatings for us, and in such a slick way that, to the best of our knowledge, no orthodox writer has yet realized that this is what is happening. What we have just found is not limited to trend and seasonal parameters: it will generalize effortlessly to far more complex problems.

As we noted before (Jaynes, 1976) in many other cases, it is a common phenomenon that orthodox results, when improved to the maximum possible extent, become mathematically equivalent to the results that Bayesian methods give us far more easily. Indeed, it is one of the problems we have that Bayesian and maximum entropy methods are so easy that orthodoxians accuse us of trying to get something for nothing.

Thus, in the long run, attempts to evade the use of Bayes' theorem do not lead to different final results; they only make us work an order of magnitude harder to get them.

17.10.5 The orthodox criterion of performance

In our endeavor to understand this situation fully, let us examine it from a different viewpoint. According to orthodox theory, the accuracy of an estimation procedure is to be judged by the sampling distribution of the estimator, while in Bayesian theory it should be judged from the posterior pdf for the parameter. Let us compare these. For the orthodox analysis, note that in both (17.104) and (17.107) the terms containing the noise vector e combine to make a linear combination of the form

$$\overline{ge} \equiv \frac{1}{N}\sum_{t=1}^{N} g_t e_t. \tag{17.123}$$

Then over the sampling pdf for the noise we have

$$E(\overline{ge}) = \frac{1}{N}\sum_{t} g_t E(e_t) = 0 \tag{17.124}$$

$$E[(\overline{ge})^2] = \frac{1}{N}\sum g_t g_{t'} E(e_t e_{t'}) = \overline{g^2}\sigma^2, \tag{17.125}$$

since $E(e_t e_{t'}) = \sigma^2\delta(t, t')$. Thus, the sampling pdf would estimate this error term by (mean) $\pm$ (standard deviation):

$$(\overline{ge})_{\text{est}} = 0 \pm \sigma\sqrt{\overline{g^2}}. \tag{17.126}$$

For the Bayes estimator (17.104)

$$g_t = \frac{\left(\overline{t^2}\right)(s_t) - \left(\overline{st}\right)t}{\left(\overline{t^2}\right)\left(\overline{s^2}\right) - \left(\overline{st}\right)^2}, \tag{17.127}$$

and after some algebra we find

$$\overline{g^2} = \frac{\left(\overline{t^2}\right)\left[\left(\overline{s^2}\right)\left(\overline{t^2}\right) - \left(\overline{st}\right)^2\right]}{\left(\overline{s^2}\right)\left(1 - r^2\right)}, \tag{17.128}$$

where r is the correlation coefficient defined before. Thus the sampling distribution for the Bayes estimator (17.104) has (mean) $\pm$ (standard deviation) of

$$\tilde{A} \pm \sigma\sqrt{\hat{N}s^2(1 - r^2)}, \tag{17.129}$$

while for the orthodox estimator this is

$$(1 - r^2)\tilde{A} \pm \sigma\sqrt{\frac{1 - r^2}{N\left(\overline{s^2}\right)}}. \tag{17.130}$$

17.11 The general case

Having shown the nature of the Bayesian results from several different viewpoints, we now generalize them to a fairly wide class of useful problems. We assume that the N data are not necessarily uniformly spaced in time, but are taken at times in some set $\{t : t_1, \ldots, t_N\}$ that the noise probability distribution, although Gaussian, is not necessarily stationary or white (uncorrelated) and that the prior probabilities for the parameters are not necessarily independent. It turns out that the computer programs to take all this into account are not appreciably more difficult to write, if the most general analytical formulas are in view when we write them.

So now we have the model

$$y_{t_i} = T(t_i) + F(t_i) + e(t_i), \quad 1 \le i \le N, \tag{17.131}$$

in which we may write $y_i \equiv y(t_i)$, etc., with data $D = (y_1, \ldots, y_N)$, where $T(t)$ is the trend function, not necessarily linear, $F(t)$ is the periodic seasonal function, not necessarily sinusoidal, and $e(t)$ is the irregular component. To define our matrices we suppose $T(t)$ expanded in some linearly independent basis functions $\Phi_k(t)$ (for example, Legendre polynomials):

$$T(t) = \sum \gamma_k \Phi_k(t). \tag{17.132}$$

Similarly, $F(t)$ is expanded in sinusoids:

$$F(t) = \sum \left[A_k \cos(kt) + B_k \sin(kt)\right]. \tag{17.133}$$

The joint likelihood of all the parameters is

$$L(\gamma, A, B, \sigma) = p(D|\gamma AB\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{N}[y_i - T(t_i) - F(t_i)]^2\right\}. \tag{17.134}$$

The quadratic form may be written as

$$Q(\alpha_k, \gamma_j) \equiv \sum_{i=1}^{N} \left[y_i - \sum_{j=1}^{r} \gamma_j T_j(t_i) - \sum_{k=1}^{m} \alpha_k F_k(t_i) \right]^2, \tag{17.135}$$

where, in the seasonal adjustment problem, $m = 12$ and

$$\{\alpha_1, \ldots, \alpha_m\} = \{A_0, A_1, \ldots, A_6, B_1, B_2, \ldots, B_5\}. \tag{17.136}$$

Likewise,

$$F_k(t) = \begin{cases} \cos(k\omega t) & 0 \le k \le 6 \\ \sin([k-6]\omega t) & 7 \le k \le 12. \end{cases} \tag{17.137}$$

But if we combine α, γ into a single vector of dimension $n = m + r$:

$$q \equiv (\alpha_1, \ldots, \alpha_m, \gamma_1, \ldots, \gamma_r) \tag{17.138}$$

and define the function

$$G_k(t) = \begin{cases} F_k(t) & 1 \le k \le m \\ T_k(t) & m+1 \le k \le n, \end{cases} \tag{17.139}$$

then the model is in the more compact form

$$y(t) = \sum_{j=1}^{n} q_j G_j(t) + e(t), \tag{17.140}$$

and the data vector is

$$y_i = \sum_{j=1}^{n} q_j G_j(t_i) + e(t_i), \quad 1 \le i \le N \tag{17.141}$$

or

$$y = qG + e. \tag{17.142}$$

The 'noise' values $e = e(t_i)$ have the joint prior probability density

$$p(e_1 \cdots e_N) = \frac{\sqrt{\det K}}{(2\pi)^{N/2}} \exp\left\{ -\frac{1}{2} e^T K e \right\}, \tag{17.143}$$

where K^{-1} is the $(N \times N)$ noise prior covariance matrix. For 'stationary white noise', it reduces to

$$K^{-1} = \sigma^2 \delta_{ij}, \qquad 1 \le i, j \le N. \tag{17.144}$$

Given K and the parameters $\{q_j\}$, the sampling pdf for the data takes the form

$$p(y_1 \cdots y_N | q K I) = \frac{\sqrt{\det(K)}}{(2\pi)^{N/2}} \exp\left\{ -\frac{1}{2} (y - qG)^T K (y - qG) \right\}. \tag{17.145}$$

Likewise, a very general form of joint prior pdf for the parameters is

$$p(q_1 \cdots q_n|I) = \frac{\sqrt{\det(L)}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(q-q_0)^T L(q-q_0)\right\}, \tag{17.146}$$

where L^{-1} is the $(n \times n)$ prior covariance matrix and q_0 is the vector of prior estimates. Almost always we shall take L to be diagonal:

$$L_{ij} = \sigma_j^2 \delta_{ij}, \qquad 1 \le i, j \le n, \tag{17.147}$$

and q_0 to be zero. But the general formulas without these simplifying assumptions are readily found and programmed.

The joint posterior pdf for the parameters $\{q_j\}$ is then

$$p(q|yI) = \frac{\exp\{-Q/2\}}{\int \mathrm{d}q_1 \cdots \mathrm{d}q_n \, \exp\{-Q/2\}}, \tag{17.148}$$

where Q is the quadratic form

$$Q \equiv (y - Gq)^T K (y - Gq) + (q - q_0)^T L (q - q_0), \tag{17.149}$$

which we may expand into eight terms:

$$Q = y^T K y - y^T K G q - q^T G^T K y + q^T G^T K G q + q^T L q - q^T L q_0 - q_0^T L q + q_0^T L q_0. \tag{17.150}$$

We want to bring out the dependence on q by writing this in the form

$$Q = (q - \hat{q})^T M (q - \hat{q}) + Q_0, \tag{17.151}$$

where Q_0 is independent of q. Writing this out and comparing with (17.150), we have

$$\begin{aligned} M &= G^T K G + L, \\ M\hat{q} &= G^T K y + L q_0, \\ \hat{q}^T M \hat{q} + Q_0 &= y^T K y + q_0^T L q_0. \end{aligned} \tag{17.152}$$

Thus M, $\hat{q}$, and Q_0 are uniquely determined, because the equality of (17.150) and (17.151) must be an identity in q:

$$\hat{q} = M^{-1}\left[G^T K y + L q_0\right] \tag{17.153}$$

$$Q_0 = y^T K y + q_0^T L q_0 - \hat{q}^T M \hat{q}. \tag{17.154}$$

The denominator of (17.148) is then found using (17.151), with the final result

$$p(q_1 \cdots q_n|yKLI) = \frac{\sqrt{\det(M)}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(q-\hat{q})^T M (q-\hat{q})\right\}. \tag{17.155}$$

The components $q_1, \ldots, q_m$ are the seasonal amplitudes we wish to estimate, while $(q_{m+1}, \ldots, q_n)$ are the trend nuisance parameters to be eliminated. From (17.155) the

marginal pdf we want is

$$p(q_1 \cdots q_m \mid yKLI) = \int \mathrm{d}q_{m+1} \cdots \mathrm{d}q_n \; p(q_1 \cdots q_n | yKLI)$$
$$= \frac{\sqrt{\det(M)}}{(2\pi)^{n/2}} \frac{(2\pi)^{(n-m)/2}}{\sqrt{\det(W)}} \exp\left\{-\frac{1}{2}(u-\hat{u})^T U (u-\hat{u})\right\} \quad (17.156)$$
$$= \frac{\sqrt{\det(U)}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}(u-\hat{u})^T U (u-\hat{u})\right\}$$

where U, V, W, u are defined by ( ), ( ), ( ), ( ).

Editor's Exercise 17.4. Jaynes never defined U, V, W, and u. In (17.155) multiply out all of the terms in the exponent, obtain the appropriate sub-matrices, vectors, and scalars and then define each of these four quantities.

From the fact that the various probabilities are normalized, we see that

$$\det(M) = \det(W)\det(U), \quad (17.157)$$

a remarkable theorem not at all obvious from the definitions except in the case $V = 0$. This is another good example of the power of probabilistic reasoning to prove purely mathematical theorems.

Thus, the most general solution consists, computationally, of a string of elementary matrix operations and is readily programmed. To summarize the final computation rules:

K^{-1} is the $(N \times N)$ prior covariance matrix for the 'noise';
L^{-1} is the $(n \times n)$ prior covariance matrix for the parameters;
F is the $(N \times n)$ matrix of model functions.

Firstly, calculate the $(n \times n)$ matrix

$$M \equiv F^T K F + L \quad (17.158)$$

and decompose it into block form representing the interesting and uninteresting subspaces:

$$M = \begin{pmatrix} U_0 & V \\ V^T & W_0 \end{pmatrix}. \quad (17.159)$$

Then calculate the $(m \times m)$ and $(r \times r)$ renormalized matrices

$$U \equiv U_0 - V W_0^{-1} V^T \quad (17.160)$$

$$W \equiv W_0 - V^T U_0^{-1} V. \quad (17.161)$$

This much is determined by the definition of the model; the computer can work all this out in advance, before the data are known, and use the result on any number of data sets.

Now given y, the $(N \times 1)$ data vector and q_0, the $(n \times 1)$ vector of prior estimates, the computer should calculate the $(n \times 1)$ vector

$$\hat{q} = M^{-1}\left[F^T K y + L q_0\right] \tag{17.162}$$

of 'best' estimates of the parameters. Actually, the first m of them are the interesting ones wanted, and the remaining $r = n - m$ components are not needed unless one also wants an estimate of the trend function. Then we can use the following result.

The inverse M^{-1} can be written in the same block form as M:

$$M^{-1} = \begin{pmatrix} U^{-1} & -U_0 V W^{-1} \\ -W_0 V^T U^{-1} & W^{-1} \end{pmatrix}, \tag{17.163}$$

where, analogous to U,

$$W \equiv W_0 - V^T U_0^{-1} V. \tag{17.164}$$

Then F^T has the same block form with respect to its rows:

$$(F^T)_{ji} = \left[G_j(t_i) T_i(t_i)\right] \qquad \begin{matrix} 1 \le j \le m \\ 1 \le i \le N \\ (m+1) \le K \le n, \end{matrix} \tag{17.165}$$

where $G_j(t)$ are the seasonal sinusoids and $T_k(t)$ the trend functions.

Almost always, $q_0 = 0$, and so the 'interesting' seasonal amplitudes are given by

$$\hat{q} = RKy, \tag{17.166}$$

where R is the reduced $(m \times N)$ matrix

$$R \equiv U^{-1} G - U_0^{-1} V W^{-1} T \tag{17.167}$$

and U^{-1} is the joint posterior covariance matrix for the interesting parameters $\{q_1, \ldots, q_m\}$. Note that R and U^{-1} are also determined by the model, so the computer can calculate them once and for all before the data are available.

Editor's Exercise 17.5. Jaynes never finished this section, so we can only speculate as to what he would have put in here. So let's speculate. Firstly, look at (17.156): this is the joint posterior probability for all of the seasonal amplitudes, but the amplitudes are not the same thing as the seasonal component itself. The seasonal component is given by

$$S(t) = \sum_{k=1}^{m} q_k G_k(t) \tag{17.168}$$

and is a continuous function of time. Can (17.156) and (17.168) be used to compute $p(S(t)|yKLI)$, the joint posterior probability for the seasonal? In other words, can a simple change of variables plus marginalization over the remaining q's be used to compute $p(S(t))|yKLI)$? If not, how would you compute this joint posterior probability?

17.12 Comments

Let us try to summarize and understand the underlying technical reasons for the facts noted in the preceding two chapters. Sampling theory methods of inference were satisfactory for the relatively simple problems considered by R. A. Fisher in the 1930s. These problems had the features of:

(a) few parameters;
(b) presence of sufficient statistics;
(c) no important prior information;
(d) no nuisance parameters.

When all these conditions are met, and we have a reasonably large amount of data (say, $n \geq 30$), orthodox methods become essentially equivalent to the Bayesian ones, and it will make no pragmatic difference which ideology we prefer. But today we are faced with important problems in which some or all of these conditions are violated. Only Bayesian methods have the analytical apparatus capable of dealing with such problems without sacrificing much of the relevant information available to us. Bayesian methods are more powerful; if there is no sufficient statistic, they extract more information from the data for reasons explained at the beginning of this chapter. Also, they take note of possibly highly important prior information, and deal easily with nuisance parameters, turning them into an important asset.

Today one wonders how it is possible that orthodox logic continues to be taught in some places year after year and praised as 'objective', while Bayesians are charged with 'subjectivity'. Orthodoxians, preoccupied with fantasies about nonexistent data sets and, in principle, unobservable limiting frequencies – while ignoring relevant prior information – are in no position to charge anybody with 'subjectivity'. If there is no sufficient statistic, the orthodox accuracy claim based on a single 'statistic' simply ignores not only the prior information, but also all the evidence in the data that is relevant to that accuracy: hardly an 'objective' procedure. If there are ancillary statistics and the orthodoxian follows Fisher by conditioning on them, he obtains just the estimate that Bayes' theorem based on a noninformative prior would have given him by a shorter calculation. Bayes' theorem would have given also a defensible accuracy claim.

We shall illustrate this in later chapters with several examples, including interval estimation, dealing with trend, linear regression, detection of cycles, and prediction of time series. In all these cases, 'orthodox' methods can miss important evidence in the data; but they can also yield conclusions not justified by the evidence because they ignore highly cogent prior information. No case of such failure of Bayesian methods has been found; indeed, the optimality theorems well known in the Bayesian literature lead one to expect this from the start. Psychologically, however, practical examples seem to have more convincing power than do optimality theorems.

Historically, scientific inference has been dominated overwhelmingly by the case of univariate or bivariate Gaussian sampling distributions. This has produced a distorted picture of the field: the Gaussian case is the one in which 'orthodox', or 'sampling theory' methods do best, and the difference between pre-data and post-data procedures is the least. On the

basis of this limited evidence, orthodox theory (in the hands of Fisher) tried to claim general validity for its methods, and attacked Bayesian methods savagely without ever examining the results they give.

Even in the Gaussian case, there are important problems where sampling theory methods fail for technical reasons. An example is linear regression with both variables subject to error of unknown variance; indeed, this is perhaps the most common problem of inference faced by experimental scientists. Yet sampling theory is helpless to deal with it, because each new data point brings with it a new nuisance parameter. The orthodox statistical literature offers us no satisfactory way of dealing with this problem. See, for example, Kempthorne and Folks (1971), in which the (for them) necessity of deciding which quantities are 'random' and which are not, leads the authors to formulate *16* different linear regression models to describe what is only a single inference problem; then they find themselves helpless to deal with most of them, and give up with the statement that 'It is all very difficult.'

When we depart from the Gaussian case, we open up a Pandora's box of new anomalies, logical contradictions, absurd results, and technical difficulties beyond the means of sampling theory to handle. Several examples were noted already by the devout orthodoxians Kendall and Stuart (1961).

These examples show the fundamental error in supposing that the quality of an estimate can be judged merely from the sampling distribution of the estimator. This is true only in the simpler Gaussian cases for reasons of mathematical symmetry; in general, as Fisher noted, many different samples which all lead to the same estimator nevertheless determine the values of the parameters to very different accuracy because they have different configurations (ranges). But Fisher's remedy – conditioning on ancillary statistics – is seldom possible, and, when it is possible, we saw in Chapter 8 that it is mathematically equivalent to use of Bayes' theorem. In the case of the 'student' t-distribution this was shown already by Jeffreys in the 1930s. In Jaynes (1976) we demonstrate it in detail for the Cauchy distribution, which orthodoxy regards as 'pathological'.

What the orthodox literature invariably fails to recognize is that all of these difficulties are resolved effortlessly by the uniform application of the single Bayesian method. In fact, once the Bayesian analysis has shown us the correct answer, one can often study it, understand intuitively why it is right, and, with this deeper understanding, see how that answer might have been found by some *ad hoc* device acceptable to orthodoxy.

We shall illustrate this in later chapters by giving the solution to the aforementioned regression problem, and to some inference problems with the Cauchy sampling distribution. To the best of our knowledge, these solutions cannot be found in any of the orthodox statistical literature.

But we must note with sadness that, in much of the current Bayesian literature, very little of the orthodox baggage has been cast off. For example, it is rather typical to see a Bayesian article start with such phrases as: 'Let X be a random variable with density function $p(x|\theta)$, where the value of the parameter θ is unknown. Suppose this parametric family contains the true distribution of X ....' Or, one describes a uniform prior $p(\theta|I)$ by saying: 'θ is supposed uniformly distributed'. The analytical solutions thus obtained will doubtless be a

valid Bayesian result; but one is still clinging to the orthodox fiction of 'random variables' and 'true distributions'. θ is simply an unknown constant; it is not 'distributed' at all. What is 'distributed' is our state of knowledge about θ: again there is that persistent mind projection fallacy that contaminates all of probability theory, leading inexperienced readers far astray as to what we are doing. Equally bad, those who commit this fallacy seem unaware that this is restricting the application to a small fraction of the real situations where the solution might be useful. In the vast majority of real applications there are no 'random variables' (What defines 'randomness'?) and no 'true distribution' (What defines it? What test could we apply to decide whether some proposed distribution is or is not the 'true' one?); yet probability theory as logic applies to all of them.

Unlike orthodox tests, Bayesian posterior probabilities or odds ratios can tell us quantitatively how strong the evidence is for some effect taking into account *all* the evidence at hand, not merely the evidence of one data set.

L. J. Savage (1962, pp. 63–67) gives, by a tortuously long, closely reasoned argument using only sampling probabilities, a rationale for the Bayesian algorithm. The Bayesian argument expounded here, which he rejects as a 'necessary' view, derives the same conclusion, in greater generality, directly from first principles.

These comparisons show that in order to deal successfully with current real problems, it is essential to jettison tradition and authority, which have retarded progress throughout this century. It is deplorable that orthodox methods and terminology continue to be taught at all to young statisticians, economists, biologists, psychologists, and medical researchers; this has done serious damage in these fields for decades.

Yet everywhere we look there are glimmerings of hope. In physics, Bretthorst (1988) has treated the analysis of magnetic resonance data, extracting by Bayesian methods far more information from the data than was possible with the previous *ad hoc* Fourier analysis. In econometrics Prof. Arnold Zellner is the founder of a large, active, and growing school of Bayesian analysis which has given rise to a vast literature. In medical diagnosis the great physician Sir William Osler (1849–1919) noted long ago that:[20] *Medicine is a science of uncertainty and an art of probability*. In recent years several people have started to take this remark seriously. Lee Lusted (1968) gives worked-out examples, with flow charts and source code, of the Bayesian computer diagnoses of six important medical conditions, as well as a great deal of qualitative wisdom in medical testing.[21] Peter Cheeseman (1988) has been developing expert systems for medical diagnosis based on Bayesian principles.

[20] Quoted by Bean (1950, p. 125).

[21] Lusted later founded the Society for Medical Decision Making in 1978, and served as the first editor of its journal. At the time of his death in February 1994, he was retired but still serving as Adjunct Professor at the Stanford University Medical School, advising medical students in problems of decision analysis.