CHAPTER 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 numerical results that they give in real problems, 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. And, 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: *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.

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. So, to understand the performance of a dozen such devices may require a dozen different analyses.

But one can see quite generally, once and for all, when and why orthodox methods must waste information. Consider estimation of a parameter θ from a data set $D \equiv \{x_1 \cdots x_n\}$ represented by a point in R^n. Orthodoxy requires us to choose a single estimator $b(D) \equiv b(x_1 \cdots x_n)$ *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 technical problems like nuisance parameters – 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 $b(D)$.

But it seems that Fisher never did appreciate the further conclusion that follows from this. Given the actual data set D, of course, all estimators that the orthodoxian might have chosen $\{b_1, b_2, \cdots\}$ are known. The Bayesian procedure chooses the estimate *after* seeing the data, and so has the benefit of the extra information contained in the specific data set. Therefore it is able to choose the optimal estimator *for that data set.* In effect Bayes' theorem has available for its use simultaneously all the information contained in the class of all possible estimators.†

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 takes into account all the relevant information in the data, but an orthodox procedure based on a single statistic often fails to do so. Then we expect that, whenever an orthodox procedure is not based on a sufficient statistic or conditioned on an ancillary statistic, the Bayesian procedure will be superior (in the sense of more accurate or more reliable) because its extra information restricts further the range of possibilities compatible with the estimator $b(D)$.

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 has been adopted more or less faithfully in the works of Herbert Simon in Economics, Erich Lehmann in hypothesis testing, and David Middleton 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, but which does not seem to be used appreciably in applied statistics.‡

Unbiased Estimators

Given a sampling distribution $p(x|\alpha)$ with some parameter α and a data set comprising n observations $D \equiv \{x_1 \cdots x_n\}$, there are various orthodox principles for estimating α, in particular use of an unbiased estimator, and maximum likelihood. In the former we choose some function of the observations $\beta(D) = \beta(x_1 \cdots x_n)$ 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:

$$\langle\beta\rangle = E(\beta) = \int \beta(x_1 \cdots x_n)\, p(x_1 \cdots x_n|\alpha)\, dx_1 \cdots dx_n = \alpha \tag{17–1}$$

As noted in Chap. 13, Eq. (13–20), the expected square of the error, over the sampling distribution, is the sum of two positive terms

† Perhaps it is now clearer why we have described orthodox and Bayesian methods as 'pre–data' and 'post–data' inferences.

‡ We find an interesting consistency here: the Fisherian methods that have been widely adopted are the ones whose results often disagree strongly with Bayesian results; the ones that have met with almost no use are just the ones that, when they are applicable, necessarily agree closely – often exactly – with the Bayesian ones.

$$\langle(\beta - \alpha)^2\rangle = (\langle\beta\rangle - \alpha)^2 + var(\beta) \tag{17–2}$$

where what the orthodoxian calls the "sampling variance of β" (more correctly, the variance *of the sampling distribution for* β) is $var(\beta) = \langle\beta^2\rangle - \langle\beta\rangle^2$. 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 of those terms 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 $\{x_1 \cdots x_n\}$. 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 labelled "bias" and "variance" has no effect on 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 $var(m_2)$ 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? The expected quadratic loss (17–2) is now

$$\begin{aligned}\langle (cm_2-\mu_2)^2\rangle &= c^2\,\langle m_2^2\rangle - 2c\langle m_2\rangle\,\mu_2 + \mu_2^2 \\ &= \langle (m_2-\mu_2)^2\rangle - \langle m_2^2\rangle(\hat{c}-1)^2 + \langle m_2^2\rangle (c-\hat{c})^2\end{aligned} \tag{17–8}$$

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 at all about the variance of $p(x_1|\mu_2)$.*

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 $K\equiv \mu_4/\mu_2^2 \ge 1$ (from a previous remark, we are assuming now that $p(x_i|\mu_2)$ has moments up to fourth order at least). We see that $\hat{c}$ is a monotonic decreasing function of K; so if $K\ge 2$, (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 $\hat{\beta}$ than M_2? In the Gaussian case the mean square error of the estimator $\hat{\beta}$ is

$$\langle (\hat{\beta}-\mu_2)^2\rangle = \frac{2}{n+1}\,\mu_2^2\,. \tag{17–14}$$

The unbiased estimator M_2 corresponds to the choice

---

* 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 important in problems where we have cogent prior information but only sparse data.

$$c = \frac{n}{n-1} \tag{17–15}$$

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 $\hat\beta$. 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 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.

Indeed, R. A. Fisher perceived this long ago, remarking that a procedure that loses half the information in the data, wastes half of the work expended in acquiring the data. But modern orthodox practitioners seem never to perceive this, because they continue to fantasize about frequencies, and do not think in terms of information at all.† A fantastic example appeared in a work on econometrics (Valavanis, 1959, p. 60) where the author attached such great importance to removing bias that he advocated throwing away not just half the data but practically all them, if necessary, to achieve this.

Why do they do this? Why do orthodoxians put such exaggerated emphasis on bias? We suspect that the main reason is simply that they are caught in a psycho–semantic 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 & 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 $\beta_1 \cdots \beta_N$, the average $\overline{\beta} = N^{-1} \sum \beta_i$

---

† Note that this difficulty does not arise in the Bayesian approach in spite of a mathematical similarity. Again choosing any function $\beta(x_1 \cdots x_n)$ 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 $\langle(\beta - \alpha)^2\rangle = (\beta - \langle\alpha\rangle)^2 + var(\alpha)$, rather like (17–2). But now changing the estimator β does not change $var(\alpha) = (\langle\alpha^2\rangle - \langle\alpha\rangle^2)$, and so by this criterion, the optimal estimator over the class of *all* estimators is always $\beta = \langle\alpha\rangle$.

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 distributions, always interpreted as frequencies, in such a situation forces one to expend virtually all his efforts on irrelevant fantasies. What is relevant to inference is not any non–existent 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.‡

But even if we accept these limit theorems, 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 gave in Chap. 6, Eq. (6–90). 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 parameters is "right". However, many orthodoxians simply ignore these arguments (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.

But note 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 Chap. 13 following Eq. (13–21); and we showed that the difficulty is overcome by a slight reformulation of the problem, which leads inexorably to the Bayesian algorithm as the one which accomplishes what we really want.

‡ "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 (ca. 500 B.C.) in his discussion of the policy decisions of the Persian kings.

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.

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

$$p(n|\lambda) = e^{-\lambda}\,\frac{\lambda^n}{n!}\,, \qquad\qquad n = 0, 1, 2, \cdots \tag{17–17}$$

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

$$\sum_{n=0}^{\infty} e^{-\lambda}\,\frac{\lambda^n}{n!}\,f(n) = \lambda \tag{17–18}$$

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

$$f(n) = \frac{d^n}{d\lambda^n}(\lambda e^{\lambda})\Big|_{\lambda=0} = n \tag{17–19}$$

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

$$f(n) = \frac{d^n}{d\lambda^n}\,[e^{\lambda} g(\lambda)]\Big|_{\lambda=0} \tag{17–20}$$

Thus the only unbiased estimator of λ^2 is

$$f(n) = \left\{ \begin{array}{cc} 0, & n = 0, 1 \\ n(n-1), & n > 1 \end{array} \right\} \tag{17–21}$$

which is absurd for $n = 1$. Likewise, the only unbiased estimator of λ^3 is absurd for $n = 1, 2$. Here the unbiased estimator does violence to elementary logic; if we observe $n = 2$, we are advised to estimate $\lambda = 0$; but if λ were zero, it would be impossible to observe $n = 2$! An unbiased estimator for $1/\lambda$ does not exist, and the only unbiased estimator of $e^{-\lambda}$ is

$$f(n) = \left\{ \begin{array}{cc} 1, & n = 0 \\ 0, & n > 0 \end{array} \right\} \tag{17–22}$$

which is absurd for all positive n. Unbiased estimators can stand in conflict with deductive logic not just for a few data sets, but for all data sets.

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

$$\langle g(\lambda)\rangle = \frac{1}{n!}\int_0^\infty e^{-\lambda}\lambda^n g(\lambda)d\lambda \tag{17–23}$$

which is readily verified to be mathematically well–behaved and intuitively reasonable for all the above examples. The Bayes estimate of $1/\lambda$ is just $1/n$. It is at first surprising that the Bayes estimate of $e^{-\lambda}$ is

$$f(n) = 2^{-(n+1)}\,. \tag{17–24}$$

Why would it not be just e^{-n}? To see why, note that the posterior distribution for λ is not symmetric but strongly skewed for small n; the posterior probability that $\lambda > n$ is

$$P = \int_n^\infty e^{-\lambda}\frac{\lambda^n}{n!}d\lambda = \frac{e^{-n}}{n!}\sum_{m=0}^{n}\binom{n}{m}n^m\int_0^\infty e^{-x}x^{n-m}dx = e^{-n}\sum_{m=0}^{n}\frac{n^m}{m!} \tag{17–25}$$

This decreases monotonically from 1 at $n = 0$ to $1/2$ as $n \to \infty$. Thus given n, the parameter λ is always more likely to be greater than n than less, so e^{-n} would systematically overestimate $e^{-\lambda}$. Bayes' theorem automatically detects this and corrects for it.

Exercise 17.2 Consider the truncated Poisson distribution:

$$p(n|\lambda) = \frac{1}{e^\lambda - 1}\frac{\lambda^n}{n!}\,, \qquad n = 1, 2, \cdots$$

Show that the unbiased estimator of λ is now absurd for $n = 1$, and the unbiased estimator of $e^{-\lambda}$ 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 & Stuart (1961). But their anti–Bayesian indoctrination (from Fisher) was so strong that they would not deign to examine the corresponding Bayesian results; and so they failed to learn that in all cases Bayesian methods overcome the difficulty effortlessly. Maurice Kendall could have learned this in five minutes from Harold Jeffreys, whom he saw almost daily because they were both Fellows of St. John's College, Cambridge and ate at the same high table.

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, 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.

Let us consider in detail 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^o Fahrenheit is perfectly evident to the eye, since the irregular 'noise' is only about 0.5^o. Yet Bloomfield, applying an orthodox significance test introduced by Fisher, concludes that there is no significant evidence for any periodicity!

The folly of pre–filtering data: In trying to understand this we note first that the data of Bloomfields'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 $D = \{y_1 \cdots y_n\}$ and consider the discrete fourier transform

$$Y(\omega) \equiv \sum_{t=1}^{n} y_t \, e^{i\omega t} \tag{17–26}$$

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‡

$$\omega_k \equiv 2\pi k/n, \quad 0 \le k < n \tag{17–27}$$

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

$$\frac{1}{n}\sum_{k=1}^{n} Y(\omega_k)\, e^{-i\omega_k t} = y_t, \qquad 1 \le t \le n\,. \tag{17–28}$$

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

$$z_t \equiv \sum_{s=0}^{m-1} y_{t-s}\, w_s \tag{17–29}$$

The new fourier transform would be, after some algebra,*

‡ Harry Nyquist was a mathematician at the Bell Telephone Laboratories who in the 1920's 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 indispensible in modern electronic technology. In Chapter 7 we have already considered the fundamental, irreducible "Nyquist noise" in electrical circuits due to thermal motion of electrons.

* At this point, many authors get involved in an annoying little 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 (17–29) 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 $(z_1, \cdots, z_{m-1})$, 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 $m << n$, but we gain the advantage that the simple formulas (17–26)–(17–32) with sums taken instead over $\pm\infty$ and t in (17–29) 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.

$$Z(\omega) = \sum_{t=1}^{n} z_t \, e^{i\omega t} = W(\omega)\, Y(\omega) \tag{17–30}$$

where

$$W(\omega) \equiv \sum_{s=0}^{m-1} w_s \, e^{i\omega s} \tag{17–31}$$

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:

$$w_s = \frac{1}{m}, \qquad 0 \le s < m \tag{17–32}$$

we have

$$W(\omega) = \frac{1}{m} \sum_{s=0}^{m-1} e^{-i\omega s} = \exp[-i\frac{\omega}{2}\,(m-1)] \left(\frac{\sin\, m\frac{\omega}{2}}{m\, \sin\, \frac{\omega}{2}} \right) . \tag{17–33}$$

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

$$W(2\pi/10) = 0\,; \qquad\qquad W(2\pi/20) = 0.639\, \exp[-9\pi i/20]\,. \tag{17–34}$$

Thus, taking a ten–year moving average of any time series data represents an irreversible loss of information; it completely wipes out any evidence for a ten–year periodicity, and reduces the amplitude of a twenty–year periodicity by a factor .639 while shifting its phase by $9\pi/20 = 1.41$ radians. We conclude that the original data had a periodicity of roughly 20 years with a peak–to–peak amplitude of about $4/.639 = 6.3^o$ F, even more obvious to the eye and nearly 90 degrees out of phase with the periodicity visible in Bloomfield's graph.

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 by a computer, this does not help in any way; it only throws away some of the information that the computer could have extracted from the original, unmutilated data. It renders the filtered data completely useless for certain purposes. For all we know, there might have been a strong periodicity of about ten years in the original data; but taking a ten–year moving average has wiped out the evidence for it.‡

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

$$P(\omega) \equiv \frac{1}{n}\, |Y(\omega)|^2 = \frac{1}{n} \sum_{t,s} y_t \, y_s \, e^{i\omega(t-s)}\,. \tag{17–35}$$

† We hasten to add that Fisher and Bloomfield are not guilty of this; but it is practiced egregiously by others such as Blackman and Tukey (1958).

‡ This data prefiltering is the one–dimensional version of the practice of 'apodization' in optics. But as we have noted elsewhere (Jaynes, 1988) this throws away highly cogent information about the fine details in the image, which a computer could have extracted, leading to much better resolution than that apparent to the eye, if one had refrained from apodization. The term 'apodization' means literally 'removing the foot'. It is singularly well–chosen; one who commits apodization is, in effect, shooting himself in the foot.

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:

$$P(\omega_k)_{av} = \frac{1}{n}\sum_{k=1}^{n} P(\omega_k) = \overline{y^2}\,. \tag{17–36}$$

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

$$q = \frac{P(\omega_k)_{max}}{P(\omega_k)_{av}} \tag{17–37}$$

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:

$$P \equiv p(q > q_0|H_0) = \int_{q_0}^{\infty} p(q|H_0)\,dq \tag{17–38}$$

and if $P > 0.05$ the evidence for periodicity is rejected as "not significant at the 5% level". This is a typical orthodox "tail area" significance test.*

But 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 '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.

But 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 it we expect it to repeat.* 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.†

---

* The choice of the 5% significance level is, of course, only an arbitrary convention; yet it has been adopted so religiously that it has become almost mandatory. Anyone who failed to use it would be considered queer by many of his colleagues.

† If an apparent periodicity were only a momentary artifact of the noise as suposed 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

But 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.‡

Harold Jeffreys (1939, p. 316) expressed his astonishment at such 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. The same applies to all the current significance tests based on P–values."

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 22–year periodicity), causes a periodic variation in the number of charged particles entering our atmosphere (indicated by 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‡ and leads us to conjecture that the July temperatures (the record of which presumably still exists) 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.

---

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.

‡ An historical study has suggested that the culprit who started this kind of reasoning was not any statistician, but the physicist Arthur Schuster (1897), who invented the periodogram for the purpose of refuting some claims of periodicity in earthquakes in Japan. He achieved his preconceived goal by the simple device of analyzing the data in a way that threw away the information about that periodicity; and then this was taken up by many others. Nevertheless, we shall see that the periodogram does contain basic information that Schuster, and Blackman & Tukey, failed to recognize. They thought that the information was contained in the sampling distribution of the periodogram; whereas it was actually contained in the shape of the periodogram.

‡ 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.

But 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, biology, or medical research who in the past allowed themselves to 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 expanded in sinusoids:

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

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 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 $N(0,\sigma)$,* 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 (1929) 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."

He then states that the quantities

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

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 given by the covering theorem of W. L. Stevens (I, 9.9).

---

* 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! Quite generally, orthodoxy tries to draw inferences from imaginary data sets that one thinks might have been seen, but were not. The pragmatic consequences of this nonsense are probably the most dangerous error in orthodox reasoning. In our closing comments we shall note why orthodox ideology forces one to it.

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 by Fisher, Feller, or Bloomfield, so they are in no position to draw inferences about periodicity.

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 & 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 below.

A Bayesian analysis. Now we examine a Bayesian analysis of these same data, and for pedagogical reasons we want to explain its rationale in great 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. Our Bayesian model is: we consider it possible that the data have a periodic component due to some systematic physical influence on the weather:

$$A \cos \omega t + B \sin \omega t \tag{17–39}$$

where as noted, we may suppose $|\omega| \le \pi$ (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.† 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

$$y_t = A \cos \omega t + B \sin \omega t + e_t, \qquad 1 \le t \le n \tag{17–40}$$

and our sampling distribution for the irregular component is

$$p(e_1 \cdots e_n | \mu, \sigma, I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_t (e_t - \mu)^2\right] \tag{17–41}$$

† In meteorology, although the laws 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.

Then the sampling (density) distribution for the data is

$$p(y_1 \cdots y_n|\mu, \sigma, I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp[-Q/2\sigma^2] \tag{17–42}$$

with the quadratic form

$$Q(A, B, \omega) \equiv \sum (y_t - \mu - A\cos\omega t - B\sin\omega t)^2 \tag{17–43}$$

or,

$$\begin{aligned} Q = n[\overline{y^2} - 2\overline{y}\mu + \mu^2 - 2A\overline{y_t\cos\omega t} - 2B\overline{y_t\sin\omega t} + 2\mu A\overline{\cos\omega t} \\ + 2\mu B\overline{\sin\omega t} + 2AB\overline{\cos\omega t\sin\omega t} + A^2\overline{\cos^2\omega t} + B^2\overline{\sin^2\omega t}] \end{aligned} \tag{17–44}$$

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. 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 negligibly small difference for the final conclusions, and so we can almost always make approximations so good that we can do the special calculation needed for our present purpose with pencil and paper after all. But fortunately, details are no deterrent to a computer, which can happily grind out the exact solution.⋆ 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 (17–44), 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 100^o 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. 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 10^o F, it will make a numerically negligible difference in the conclusions (a strictly nil difference if we record our conclusions only to two or three decimal digits). So let us see what Bayesian inference gives with just this. By an argument essentially the same as the Herschel derivation of the gaussian distribution in Chapter 7, we may assign a joint prior

‡ This is just the expression of the fact that probability theory as logic is the *exact* system for inference; therefore it will seek out relentlessly every scrap of information that has any relevance at all to the question being asked.

⋆ 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.

$$p(A, B|I) = \frac{1}{2\pi\delta^2} \exp\left[-\frac{A^2 + B^2}{2\delta^2}\right] \tag{17–45}$$

where δ is of the order of magnitude of 100^o 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:

$$p(A, B, \omega, \mu, \sigma|D, I) = p(A, B, \omega, \mu, \sigma|I)\, \frac{p(D|A, B, \omega, \mu, \sigma, I)}{p(D|I)} \tag{17–46}$$

then integrate out the nuisance parameters:

$$p(A, B, \omega|D, I) = \int d\mu \int d\sigma\, p(A, B, \omega, \mu, \sigma|D, I) \tag{17–47}$$

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:

$$p(A, B, \omega, \mu, \sigma|I) = p(A, B, \omega|I)\, p(\mu, \sigma|A, B, \omega, I) \tag{17–48}$$

and thus the most general solution appears formally simpler:

$$p(A, B, \omega|D, I) = C\, p(A, B, \omega|I)\, L^*(A, B, \omega) \tag{17–49}$$

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

$$L^*(A, B, \omega) \equiv \int d\mu \int d\sigma\, p(\mu, \sigma|A, B, \omega, I)\, p(D|A, B, \omega, \mu, \sigma, I) \tag{17–50}$$

In (17–49) 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

$$p(\mu, \sigma|A, B, \omega, I) = p(\mu, \sigma|I) \tag{17–51}$$

so what is our prior information about (μ, σ)? Surely we know also, for the same "panic evacuation" reason, that neither of these parameters could be as large as 100^o F. And we know that σ could not be as small as 10^{-6} degrees 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 surely would not give reproducible readings to that accuracy over many years). We could just as well ignore that practical information and argue that σ could not be as small as 10^{-20} degrees F because 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:

$$p(\mu,\sigma|I) = \frac{1}{\sqrt{2\pi\alpha^2}}\exp[-\mu^2/2\alpha^2]\cdot\frac{1}{\sigma}\,, \qquad\qquad a\le\sigma\le b \tag{17–52}$$

in which α and b are also of the order of 100^o 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

$$L^*(A,B,\omega) = \int_{-\infty}^{\infty} d\mu\,\exp[-\mu^2/2\alpha^2]\int_a^b\frac{d\sigma}{\sigma^{n+1}}\exp[-Q/2\sigma^2] \tag{17–53}$$

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

$$\int_0^\infty \frac{d\sigma}{\sigma^{n+1}}\exp[-Q/2\sigma^2] = \frac{1}{2}\,\frac{(n/2-1)!}{(Q/2)^{n/2}} \tag{17–54}$$

and the integral of this over μ is also guaranteed to converge. But for tactical reasons, let us do the integration over μ first.

$$Q = n[s^2 - (\mu-\overline{y})^2$$

************** MUCH, MUCH MORE HERE! *****************

COMMENTS

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

(A) Few parameters
(B) No nuisance parameters
(C) No important prior information
(D) Presence of sufficient statistics.

When all these conditions are met and we have a reasonably large amount of data (say, $n \ge 30$), 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 in the data. They are more powerful (*i.e.*, if there is no sufficient statistic, Bayesian methods extract more information from the data because they make use of all the data, while an orthodox method will still use only one function of the data, namely some arbitrarily chosen 'estimator' $\beta(x_1\cdots x_n)$.

But at the same time Bayesian methods are 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). In parameter estimation, for example, whether or not there is a sufficient statistic, the log–likelihood function is

$$\log L(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha) = n\,\overline{\log p(x_i|\alpha)} \qquad (17\text{–sam})$$

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 and an uninformative prior, the width of the posterior distribution for α is essentially $(R + W)$

(range of the data) + (width of individual likelihoods)

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 $L_i(\alpha) = p(x_i|\alpha)$.

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! Whether the data range is large or small, orthodoxy will claim just the same accuracy for its estimate. Far worse, that accuracy expresses entirely the variability of the estimator *over other data sets that we might have obtained but did not*. But as noted, unobserved data sets are entirely a figment of our imagination, and so are irrelevant to the inference being made.

One wonders how is is possible that this orthodox logic continues to be taught year after year as 'objective', while charging Bayesians with 'subjectivity'. When we examine the rationale of their procedures, it is evident that orthodoxians are in no position to charge anybody with 'subjectivity'. If there is no sufficient statistic, the orthodox accuracy claim simply ignores all the evidence in the data that is relevant to the accuracy.

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 data. 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 exemples 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.

But even in the multivariate 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 & Folks (1971), in which the (for them) necessity of deciding which quantities are "random", and which are not, leads them to formulate sixteen different linear regression models to describe what is only a single inference problem; then they find themselves helpless to deal with most of them.

When we depart from the Gaussian case, we open up a Pandora's box of 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 & Stuart (1961)].

These 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; 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, it is mathematically equivalent to use of Bayes' theorem.

Unfortunately, what the orthodox literature fails to recognize is that all of these problems are solved 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 will illustrate this 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 uknown. Suppose this parametric family contains the true distribution of $X \cdots$." The analytical solutions thus obtained will doubtless be valid Bayesian results; but one is still clinging to the orthodox fiction of 'random variables' and 'true distributions', 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 there are no 'random variables' and no 'true distribution'; 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 rather long, closely reasoned argument using only sampling probabilities, a rationale for the Bayesian algorithm. The Bayesian argument expounded here in Chapter 4, which he rejects as a "necessary" view, yields the same conclusion, in greater generality, by three lines of elementary algebra.

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

Yet everywhere we look there are glimmerings of hope. For example, in medical diagnosis the great physician Sir William Osler (1849 – 1919) long ago noted that:[†] *Medicine is a science of uncertainty and an art of probability.* The book of Dr. Lee Lusted (1965) 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. Lusted later founded the Society for Medical Decision Making in 1978, and served as the first Editor of its journal, *Medical Decision Making*. 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. Dr. Peter Cheeseman has been developing Expert Systems for medical diagnosis based on Bayesian principles.[‡]

[†] Quoted by Wm. B. Bean (1950); p. 125

[‡] As noted in Jaynes (1990b) this aroused fierce opposition from those with an entrenched vested interest in

the old *ad hoc* principles; but in all their statistical training they had never seen a Bayesian solution, and did not understand what Bayesian methods are. Once the final results are in hand, such uninformed opposition melts away like an ice cube in a furnace. As we noted in Chapter 5 under "Evolution into Bayesianity", Cox's theorems show that Bayesian methods are uniquely determined by elementary requirements of consistency. Therefore, to deny that the human mind reasons according to Bayesian principles is to assert that it operates in a deliberately inconsistent way. Nobody could maintain such a position if he were aware of this.