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. From our viewpoint, the necessity for this was established already in Chapter 2. In the present Chapter we are not concerned particularly with any abstract theory or with mathematical techniques; but rather with showing some of the pragmatic consequences of failing to use Bayesian methods in some very simple problems, where the mysteries of infinite sets never arise. These simple problems are just the ones that the "orthodox" school of thought believed that it had solved without using Bayesian methods; but in every case experience with the orthodox solution revealed defects that were corrected at once by use of Bayesian methods.

In Chapter 16 we noted that the orthodox objections to Bayesian methods were always ideological in nature, never examining the actual results that they give in real problems, and we expressed astonishment that mathematically competent people would use such arguments. In order to give a fair comparison, we need to adopt the opposite tactic here, and concentrate on examining the pragmatic results. Indeed, since Bayesian methods have been so egregiously misrepresented in the orthodox literature, we must lean over backwards to avoid misrepresenting orthodox methods now; whenever orthodox methods do yield satisfactory results in some problem, we shall want to acknowledge that fact, and we shall not deplore their use merely on ideological grounds (although our educational purpose requires that we explain what the orthodox ideology is). 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 on pragmatic grounds. The demonstrable facts, which amount to many different confirmations of what Cox's theorems led us to expect, are quite sufficient to make our point.

We have several independent *ad hoc* orthodox devices to consider. From the Neyman–Pearson camp of orthodoxy we have the principles of unbiased estimators, confidence intervals, and hypothesis tests which amount to a kind of decision theory. From the Fisherian camp there are the principles of maximum likelihood, fiducial probability, randomization in design of experiments, analysis of variance, and a mass of significance tests. Our present goal is to understand just what results these principles yield in practice, independently of whatever rationale they might have to the orthodoxian; in particular we want to know this: *In what circumstances, and in what ways, do the orthodox results differ from the Bayesian results*?

Unbiased Estimators

***************************** MORE HERE! **************************

Periodicity: The Weather in Central Park

A common problem, important in economics, meteorology, geophysics, and astronomy, is to decide whether certain data taken over time provide evidence for a periodic behavior. Any clearly discernible periodic component (in business cycles, stock market, rainfall, temperature, incidence of earthquakes, brightness of a star) provides an evident basis for improved prediction of future behavior, on the presumption that periodicities observed in the past will tend to continue in the future. But even the principle for analyzing the data to extract 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.

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 20 year period 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 rejects this as not significant by an orthodox significance test advocated by Fisher.

In reconsidering this we note first that the data of the graph have been mutilated by taking a 10 year moving average. We must first understand what effect this has on the evidence for periodicity. Let the original raw data be $D = \{y_1 \dots y_n\}$ and define the discrete fourier transform

$$Y(\omega) \equiv \sum_{t=1}^{n} y_t \, e^{i\omega t} \tag{17–b1}$$

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' $\omega_k \equiv 2\pi k/n$, $0 \le k < n$ already contain all the information in the data, for the data can be recovered from them by the fourier inversion:

$$\frac{1}{n} \sum_{k=1}^{n} Y(\omega_k) \, e^{-i\omega_k t} = y_t, \qquad 1 \le t \le n \tag{17–b2}$$

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

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

The new fourier transform would be, after some algebra,

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

where

[†] Many authors here get involved in an annoying little semantic complication about exactly what one means by an "m–year weighted average". If we have only y_t for $t > 0$, then it seems to many that an m–year moving average can start only with z_m. This leads to small "end effect" corrections of order m/n. We have avoided this by a slight reinterpretation. 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 sums over t, s to be over $(-\infty, +\infty)$, and the terms $(z_1, \cdots, z_{m-1})$ are actually weighted averages over less than m years. The differences are numerically negligible for large n, but we gain the advantage that the simple formulas (17–b1) – (17–b4) are all exact as they stand, without our having to bother with messy correction terms. This particular choice of definition of terms (which were basically arbitrary anyway) is thus the one appropriate to the subject. This same zero–padding technique is used extensively in dealing with fast fourier transforms, for the same reason.

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

is the fourier transform of the weighting coefficients. Thus taking any moving average of the data merely multiplies its fourier transform by a known function. In particular, for uniform weighting:

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

we have

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

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

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

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 twenty–year periodicity with a peak–to–peak amplitude of about $4/.639 = 6.3$ degrees F, even more obvious to the eye and nearly 90 degrees out of phase with the periodicity visible in the moving average 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 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 useless for certain purposes. For all we know, there might have been a strong ten–year periodicity in the original data; but taking a ten–year moving average has wiped out all evidence for it.[†]

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

$$P(\omega) \equiv \frac{1}{n}\, |Y(\omega)|^2 = \frac{1}{n} \sum_{t,s} y_t\, y_s\, e^{i\omega(t-s)} \tag{17–b3}$$

Note that $P(0) = (\sum y_t)^2/n = n\bar{y}^2$ 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:

$$P(\omega_k)_{av} = \frac{1}{n} \sum_{k=1}^{n} P(\omega_k) = \overline{y^2} \tag{17–b4}$$

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

$$q = \frac{P(\omega_k)_{max}}{P(\omega_k)_{av}} \tag{17–b5}$$

[†] This data filtering is the one–dimensional version of the practice of 'apodization' in optics, smearing out an image in a way that makes it look to the eye easier to resolve close objects; while actually throwing away highly cogent information about the fine details in the image, which a computer could have extracted, leading to much better resolution that that apparent to the eye, if one had refrained from apodization. As we have noted elsewhere (Jaynes, 1988) the process is singularly well–named; one who commits apodization is, quite literally, shooting himself in the foot.

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 probability that chance alone would have produced a ratio as great or greater:

$$P \equiv p(q > q_0|H_0) = \int_{q_0}^{\infty} p(q|H_0)\, dq \tag{17–b6}$$

and if $P > 0.05$ the hypothesis of periodicity is "rejected as not significant at the 5% level".

Say something about P-values in general. Jeffreys, p. 316

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; and (2) the alternatives against which it is to be tested? Until we have done that, 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 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 assumption, 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 that would make it resemble a second cycle of that wave.

But in almost every application one can think of, we know perfectly well that if a periodicity is present, it is caused by some systematic influence that repeats itself; indeed, our interest in it *is due entirely to the fact that it will repeat.* The hypothesis that we want to test – to see whether the data give evidence for it – is crudely something like the opposite of the hypothesis that is assumed in Fisher's test.

But this is the peculiar logic that underlies all orthodox significance tests. In order to argue for an hypothesis H, do it indirectly: invent a "null hypothesis" H_0 that denies H, then argue against H_0. But of course, H_0 is not the direct denial $\overline{H}$ of Aristotelian logic; indeed, H is usually stated or implied to be a disjunction of many different hypotheses (in the present case, specifying the period, amplitude, and phase of the periodic term), while H_0 denies all of them while assuming things (gaussian noise) that H neither assumes nor denies. To see how far this procedure takes us from conventional logic, note the following difficulty: suppose 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: "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 there is a periodicity in temperature, we mean that there is some periodic physical influence at work, the nature of which is not known with certainty, but about which we could make some reasonable conjectures. For example, periodicity in solar activity, already known to occur by the periodically variable sunspot numbers, could conceivably cause 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 operating; only that it is a conceivable one, which does not violate any known laws of physics, but whose magnitude is difficult to estimate theoretically, and may or may not be sufficient to account for the data. But its presence or absence could be determined by other observations, correlating other astronomical and atmospheric electricity data with weather data at many different locations.

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". The result is that the procedure he recommends is too feeble to extract any information from the data.

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

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 what he calls the Old Wrong Way: Fit to the data with well-chosen frequencies $\{\omega_1 \ldots \omega_n\}$. If one of the $R_j^2 = A_j^2 + B_j^2$ is big, say there is a true period, assume all $A_j, B_j \sim N(0, \sigma)$. 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.

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 indepdendent 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 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."

The quantities

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

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).

Of course, our position is that this sampling distribution urged on us by both Fisher and Feller is quite irrelevant to the inference; the two quantities that are relevant (the prior information and the likelihood function) are not even mentioned by Fisher or Feller, so they are in no position to draw inferences about anything.

[‡] The abbreviation "*iid*"" is orthodox jargon standing for "Independently and Identically Distributed". For us, this is another form of the Mind Projection Fallacy; it considers the probablity distribution to be a real physical property of the A_j, B_j, in spite of the obvious fact that each individual coefficient is a definite, if unknown quantity; it is not "distributed" at all!

In any event, the bottom line of this discussion is that Fisher's test fails to detect the 20 year periodicity in the New York Central Park January temperatures, although that periodicity is perfectly obvious to the eye without any calculation.

But this is not the only case where ordinary common–sense examination of the data is a more powerful tool for inference than the principles taught in orthodox textbooks. Crow, Davis & Maxfield (1960) F–test - analyzed in Jaynes 1976 examples of t-test and F–test.

Now we examine a Bayesian analysis of these same data. Prior information: it is surely safe to say that 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.

Can We Achieve Bayesian – Orthodox Equivalence?

When one who thinks habitually in orthodox terms contemplates Bayesian methods with the thought of looking for a compromise, he seems to be interested only in finding the prior probabilities that would make the Bayesian results the same as the orthodox ones for, say, interval estimation. Then he intends that Bayesians are to use henceforth only that prior, so that all pragmatic reasons for contention disappear. The Bayesian, of course, has no such intentions; for one of the main advantages of Bayesian methods is that by use of varions prior probabilities they achieve pragmatic results far beyond those accessible to orthodox methods. We saw a striking example in the different results for inversion of the Urn sampling distribution, in Chapter 6.

But the idea of a prior that makes the procedures equivalent in final numerical results is useful to the Bayesian for just the opposite reason; it facilitates meaningful comparison of the Bayesian and orthodox results, making it easy to demonstrate the pragmatic superiority of the Bayesian results. Such a prior $p(\theta|I)$ would be, in a sense uninformative. But does it always exist?

D. J. Bartholomew (1970) made some conjectures to the effect that prior $p(\theta|I) \propto [I(\theta)]^{-1/2}$ where $I(\theta)$ is the Fisher information, would accomplish this approximately, but his calculations did not really establish this. However, the exact answer is obvious to a Bayesian who thinks in terms of information content rather than frequencies:

(1) The orthodox results depend on which estimator the orthodoxian has chosen to use; so the results cannot always be equivalent.

(2) The Bayesian procedure always extracts all of the relevant information from the data $x \equiv \{x_1 \cdots x_n\}$. The orthodox procedure does so only if the estimator is a sufficient statistic. Therefore, such a prior cannot exist if there is no sufficient statistic (or if the orthodoxian has chosen to use an estimator which is not a sufficient statistic, even though one exists).

REFERENCES

Bartholomew, D. J. (1970), "A Comparison of Frequentist and Bayesian Approaches to Inference with Prior Knowledge", in Symposium on Foundations of Statistical Inference, March 31 – April 9, University of Waterloo, Canada.