CHAPTER 16

ORTHODOX METHODS: HISTORICAL BACKGROUND

"With all this confounded trafficking in hypotheses about invisible connections with all manner of inconceivable properties, which have checked progress for so many years, I believe it to be most important to open people's eyes to the number of superfluous hypotheses they are making, and would rather exaggerate the opposite view, if need be, than proceed along these false lines." — H. von Helmholtz (1868).

This Chapter and Chapter 13 are concerned with the history of the subject rather than its present status. There is a complex and fascinating history before 1900, recounted by Stigler (1986), but we are concerned now with more recent developments. In the period from about 1900 to 1970, one school of thought dominated the field so completely that it has come to be called "orthodox statistics". It is necessary for us to understand it, because it is what most working statisticians active today were taught, and its ideas are still being taught, and advocated vigorously, in many textbooks and Universities.

In Chapter 17 we want to examine the "orthodox" statistical practice thus developed and compare its technical performance with that of the "probability as logic" approach expounded here. But first, to understand this weird course of events we need to know something about the problems faced then, the sociology that evolved to deal with them, the roles and personalities of the principal figures, and the general attitude toward scientific inference that orthodoxy represents.

The Early Problems

As we note repeatedly, the beginnings of scientific inference were laid in the 18'th and 19'th Centuries out of the needs of astronomy and geodesy. The principal figures were Daniel Bernoulli, Laplace, Gauss, Legendre, Poisson and others, whom we would describe today as mathematical physicists.

Transitions in the dominant mode of thinking take place slowly over a few decades, the working lifetime of one generation. But the beginning of our period, 1900, marks roughly the time when non–physicists moved in and proceeded to take over the field with quite different ideas. The end, 1970, marks roughly the time when those ideas in turn came under serious, concerted attack in our present "Bayesian Revolution".

During this period, as we analyzed in Chapter 10, the non–physicists thought that probability theory was a physical theory of "chance" or "randomness", with no relation to logic, while "Statistical Inference" was thought to be an entirely different field, based on entirely different principles. But, having abandoned the principles of probability theory, it seemed that they could not agree on what those new principles were; or even on whether the reasoning of statistical inference was deductive or inductive.

The first problems, dating back to the 18'th Century, were of course of the very simplest kind, estimating one or more location parameters θ from data $D = \{x_1 \dots x_n\}$ with sampling distributions of the form $p(x|\theta) = f(x-\theta)$. However, in practice this was not a serious limitation, because even a pure scale parameter problem becomes approximately a location parameter one if the quantities involved are already known rather accurately, as is generally the case in astronomy and geodesy.

Thus if the sampling distribution has the functional form $f(x/\sigma)$, and x and σ are already known to be about equal to x_0 and σ_0, we are really making inferences about the small corrections $q \equiv x - x_0$ and $\delta \equiv \sigma - \sigma_0$. Expanding in powers of δ and keeping only the linear term, we have

$$\frac{x}{\sigma} = \frac{x_0 + q}{\sigma_0 + \delta} = \frac{1}{\sigma_0}(x - \theta + \cdots)$$

where $\theta \equiv x_0\delta/\sigma_0$. Thus we may define a new sampling distribution function

$$h(x - \theta) \equiv f(x/\sigma)$$

and we are considering an approximately location parameter problem after all. In this way, almost any problem can be linearized into a location parameter one if the quantities involved are already known to fairly good accuracy. The 18'th Century astronomers took good advantage of this, as we should also.

Only toward the end of the 19'th Century did practice advance to the problem of estimating simultaneously both a location and scale parameter θ, σ from a sampling distribution of the form

$$p(x|\theta,\sigma) = f\left(\frac{x-\theta}{\sigma}\right)\frac{1}{\sigma}$$

and to the marvelous developments by Galton associated with the bivariate gaussian distribution, which we studied in Chapter 7. Virtually all of the development of orthodox statistics was concerned with these three problems or their reverbalizations in hypothesis testing form, and most of it only with the first. But even that seemingly trivial problem had the power to generate fundamental differences of opinion and fierce controversy over matters of principle.

Sociology of Orthodox Statistics

During the aforementioned period, the average worker in physics, chemistry, biology, medicine, economics with a need to analyze data could hardly be expected to understand theoretical principles that did not exist, and so the approved methods of data analysis were conveyed to him in many different, unrelated *ad hoc* recipes in "cookbooks" which in effect told one to "*Do this* $\cdots$ *then do that* $\cdots$ *and don't ask why*."

R. A. Fisher's *Statistical Methods for Research Workers* was the most influential of these cookbooks. In going through 13 editions in the period 1925–1960 it acquired such an authority over scientific practice that researchers in some fields such as medical testing found it impossible to get their work published if they failed to follow Fisher's recipes to the letter.

Fisher's recipes include Maximum Likelihood Parameter Estimation (MLE), Analysis of Variance (ANOVA), fiducial distributions, randomized design of experiments, and a great variety of significance tests, which make up the bulk of his book. The rival Neyman–Pearson school of thought offered unbiased estimators, confidence intervals, and hypothesis testing. The combined collection of the *ad hoc* recipes of the two schools came to be known as orthodox statistics, although arguments raged back and forth between them over fine details of their respective ideologies. It was just the absence of any unifying principles that perpetuated this division; there was no criterion acceptable to all for resolving differences of opinion.

Whenever a real scientific problem arose that was not covered by the published recipes, the scientist was expected to consult a professional statistician for advice on how to analyze his data, and often on how to gather it as well. There developed a statistician–client relationship rather like the doctor–patient one, and for the same reason. If there are simple unifying principles (as there are today in the theory we are expounding), then it is easy to learn them and apply them to whatever problem one has; each scientist can become his own statistician. But in the absence of unifying principles, the collection of all the empirical, logically unrelated procedures that a data analyst might need, like the collection of all the logically unrelated medicines and treatments that a sick patient might need, was too large for anyone but a dedicated professional to learn.

Undoubtedly, this arrangement served a useful purpose at the time, in bringing about a semblance of order into the way scientists analyzed and interpreted their data and published their conclusions. It was workable as long as scientific problems were simple enough so that the cookbook procedures could be applied and made some intuitive sense, even though they were not derived from any first principles. Then, had the proponents of orthodox methods behaved with the professional standards of a good doctor (who notes that some treatments have been found to be effective, but admits frankly that the real cause of a disorder is not known and welcomes further research to supply the missing knowledge) there could be no criticism of the arrangement.

But that is not how they behaved; they adopted a militant attitude, each defending his own little bailiwick against intrusion and opposing every attempt to find the missing unifying principles of inference. R. A. Fisher (1956) and M. G. Kendall (1963) attacked Neyman and Wald for seeking unifying principles in decision theory. R. A. Fisher (numerous articles), H. Cramér (1946), R. von Mises (1951), J. Neyman (1952), Wm. Feller (1950) – and even the putative Bayesian L. J. Savage (1954, 1980) – accused Laplace and Jeffreys of committing metaphysical nonsense for thinking that probability theory was an extension of logic, and seeking the unifying principles of inference on that basis.

We are at a loss to explain how they could have felt such a certainty about this, since they were all quite competent mathematically and presumably understood perfectly well what does and what does not constitute a proof. Yet they did not examine the consistency of probability theory as logic (as R. T. Cox did); nor did they examine its qualitative correspondence with common sense (as G. Pólya did). They did not even deign to take note of how it works out in practice (as H. Jeffreys had shown so abundantly in works which were there for their inspection). In fact, they offered no demonstrative arguments or factual evidence at all in support of their position; they merely repeated ideological slogans about "subjectivity" and "objectivity" which were quite irrelevant to the issues of rational thinking, logical consistency, and useful results.

We are equally helpless to explain why James Bernoulli and John Maynard Keynes (who expounded essentially the same views as did Laplace and Jeffreys) escaped their scorn. Evidently, the course of events must have had something to do with personalities; let us examine a few of them.

Ronald Fisher, Harold Jeffreys, and Jerzy Neyman

Sir Ronald Aylmer Fisher (1890 – 1962) was by far the dominant personality in this field in the period 1925 – 1960. A personal account of his life is given by his daughter (Joan Fisher Box, 1978). On the technical side, he had a deep intuitive understanding and produced a steady stream of important research in genetics. Sir Harold Jeffreys (1891 – 1989) working in geophysics, wielded no such influence, and for most of his life found himself the object of scorn and derision from the Fisherian camp.

Fisher's early fame (1915–1925) rested on his mathematical ability: given data $D \equiv \{x_1 \dots x_n\}$ to which we assign a multivariate gaussian sampling probability $p(D|\theta)$ with parameters $\theta \equiv \{\theta_1 \dots \theta_m\}$, how shall we best estimate those parameters from the data? Probability theory as logic considers it obvious that in any problem of inference we are always to calculate the probability of whatever is unknown and of interest, conditional on whatever is known and relevant; in this case, $p(\theta|D, I)$.

But the orthodox view rejects this on the grounds that $p(\theta|D, I)$ is meaningless because it is not a frequency; θ is not a 'random variable', only an unknown constant. Instead, we are to choose some function of the data $f(D)$ as our "estimator" of θ. The merits of any proposed estimator are to be determined solely from its sampling distribution $p(f|\theta)$. The data are always supposed to be obtained by 'drawing from a population' urn–wise, and $p(f|\theta)$ is always supposed to be a limiting frequency in many repetitions of that draw. A good estimator is one whose sampling distribution is strongly concentrated in a small neighborhood of the true value of θ.

But as we noted in Chapter 13, orthodoxy, having no general theoretical principles for constructing the 'best' estimator, must in every new problem guess various functions $f(D)$ on grounds of intuitive judgment, and then test them by determining their sampling distributions, to see how concentrated they are near the true value. Thus calculation of sampling distributions for estimators is the crucially important part of orthodox statistics; without it one has no grounds for choosing an estimator.

Now the sampling distribution for some complicated function of the data, such as the sample correlation coefficient, can become quite a difficult mathematical problem; but Fisher was very good at this, and found many of these sampling distributions for the first time. Technical details of these derivations, in more modern language and notation, may be found in Feinberg & Hinkley (1989).

Many writers have wondered how Fisher was able to acquire the multidimensional space intuition that enabled him to solve these problems. We would point out that just before starting to produce those results, Fisher spent a year (1912–1913) as assistant to the theoretical physicist Sir James Jeans, who was then preparing the second edition of his book on kinetic theory and worked daily on calculations with high–dimensional multivariate gaussian distributions (called Maxwellian velocity distributions).

But nobody seemed to notice that Jeffreys was able to bypass Fisher's calculations and derive his parameter estimates in a few lines of the most elementary algebra. For Jeffreys, using probability theory as logic, in the absence of any cogent and detailed prior information, the best estimators were always determined by the likelihood function, which can be written down by inspection of $p(D|\theta)$. This automatically constructed the optimal estimator for him, with no need for intuitive judgment and without ever calculating a sampling distribution for an estimator. Fisher's difficult calculations calling for all that space intuition, although interesting as mathematical results in their own right, were quite unnecessary for the actual conduct of inference.

Fisher's later dominance of the field derives less from his technical work than from his flamboyant personal style and the worldly power that went with his official position, in charge of the work and destinies of many students and subordinates. For 14 years (1919–1933) he was at the Rothamsted agricultural research facility with an increasing number of assistants and visiting students, then holder of the Chair of Eugenics at University College, London, and finally in 1943 Balfour Professor of Genetics at Cambridge where he also became President of Caius College. He was elected Fellow of the Royal Society in 1929, and was knighted in 1952.

Within his field of geophysics, Harold Jeffreys also showed an outstandingly high competence, was elected Fellow of the Royal Society in 1925, became Plumian Professor of Astronomy at Cambridge in 1946, and was knighted in 1953. The treatise on mathematical physics by Sir Harold and Lady Jeffreys (1946) was for many years the standard textbook in the field, with far more advanced and useful mathematics than is contained in all of Fisher's works. But Jeffreys remained all his life as a Fellow of St. John's College, Cambridge, working quietly and modestly, and hardly visible outside his field of geophysics; he had only one doctoral student in probability theory (V. S. Huzurbazar).

In sharp contrast Fisher, possessed of a colossal, overbearing ego, thrashed about in the field, attacking the work of everyone else† with equal ferocity. Somehow, early in life Fisher's mind became captured by the dogma that by "probability" one is allowed to mean only limiting frequency in a

† For the record (we shall not go into them deeply): we consider Fisher's criticisms of Karl Pearson on grounds of maximum likelihood vs. moment fitting and the proper number of degrees of freedom in Chi–squared, and of Jerzy Neyman on grounds of confidence intervals, unbiased estimators, and the meaning of

random experiment (although he usually stated this as the ratio of two infinite numbers, rather than the limit of a ratio of finite numbers) and that any other meaning is metaphysical nonsense, unworthy of a scientist. Conceivably, this view might have come from the philosopher John Venn (an earlier President of Caius College, Cambridge, where Fisher was an undergraduate in 1909–1912). In a very influential work which went through three editions, Venn ridiculed Laplace's conception of probability theory as logic; and Fisher's early work sounds very much like this.

However, we see a weakening of resolve in Fisher's final book (1956) where he actually defends Laplace against the criticisms of Venn, and suggests that Venn did not understand mathematics well enough to comprehend what Laplace was saying. His criticisms of Jeffreys are now much toned down. Noting this, some have opined that, were Fisher alive today, he would now be a Bayesian.[†]

In both science and art every creative person must, at the beginning of his career, do battle with an Establishment that, not comprehending the new ideas, is more concerned with putting him down than with understanding his message. Karl Pearson (1857–1936), as Editor of *Biometrika*, performed that 'service' for Fisher in his early attempts at publication, and Fisher never forgave him for this. But curiously, in his last book Fisher's attacks against Pearson are, if anything, more violent and personal than ever before. This is hard to understand, for by 1956 the battle was long since won; Pearson had been dead for twenty years, and it was universally recognized that in all their disputes Fisher had been in the right. Why should the bitterness remain thirty years after it had ceased to be relevant? This tells us much about Fisher's personality.

Fisher's articles are most easily found today in two "collected works" (Fisher, 1952, 1974). The ones on the principles of inference have an interesting characteristic pattern. They start with a paragraph or two of polemical denunciation of Jeffreys' use of Bayes' theorem (at that time called *inverse probability*). Then he formulates a problem, sees the correct solution intuitively, and does the requisite calculations in a very efficient, competent way. But just at the point where one more step of the logical argument would have forced him to see that he was only rediscovering, in his own way, the results of applying Bayes' theorem, the article comes to an abrupt end.

Harold Jeffreys (1939) was able to derive all the same results far more easily, by direct use of probability theory as logic, and this automatically yielded additional information about their range of validity and how to generalize them, that Fisher never did get. But whenever Jeffreys tried to point this out, he was buried under an avalanche of criticism which simply ignored his mathematical demonstrations and substantive results and attacked his ideology. His perceived sin was that he did not require a probability to be also a frequency, and so admitted the notion of probability of an hypothesis. Nobody seemed to perceive the fact that this broader conception of probability was just what was giving him those computational advantages.

---

significance levels, to be justified on grounds of technical fact. It is perhaps a measure of Fisher's influence that the two disputes where we think that Fisher was in the wrong – the one with W. S. Gossett over randomization and the one with Jeffreys on the whole meaning and philosophy of inference, are still of serious concern today.

[†] But against this supposition is the fact that in the last year of his life Fisher published an article (Fisher, 1962) examining the possibilities of Bayesian methods, but *with the prior probabilities to be determined experimentally*!! This shows that he never accepted – and probably never comprehended – the position of Jeffreys about the meaning and function of a prior probability. Anything obtained experimentally would be, for Jeffreys and for us, part of the data; the prior probabilities represent instead whatever additional information we bring to the problem, exclusive of the data. In his final pronouncement on this, Fisher would still leave us no way to take that prior information into account, although we saw in Chapter 6 how crucially important even qualitative prior information can be in a real problem. It seems to us likely that Fisher never faced a problem in which we had cogent prior information that cannot be expressed by choice of a model; for Jeffreys and for us, that is a common situation.

Jerzy Neyman, whom we met in Chapter 14, also rejected Jeffreys' work on the same ideological grounds as did Fisher (but in turn had his own work rejected by Fisher). Neyman, too, directs scathing ridicule at Jeffreys, far beyond what would have been called for even if Neyman had been technically correct and Jeffreys wrong [for example, Neyman (1952, p. 11) becomes heated over a problem involving five balls in two urns, so simple that it would not be considered worthy of being a homework problem today, in which Jeffreys (1939, §7.02) is clearly in the right].

In view of all this, it is pleasant to be able to record that in the end Harold Jeffreys outlived his critics and the merit of his work, on both the theoretical and the pragmatic levels, was finally recognized. In the last years of his life he had the satisfaction of seeing Cambridge University – from the Cavendish Physics Laboratory on the North to the Molecular Biology Laboratory on the South – well populated with young scientists studying and applying his work and, with the new tool of computers, demonstrating its power for the current problems of science.

The exchanges between Fisher and Jeffreys over these issues in the British Journals of the 1930's were recalled recently by S. Geisser (1980) and D. Lane (1989), with many interesting details. But we want to add some additional comments to theirs, because a fellow physicist is in a better position to appreciate Jeffreys' motivations, highly relevant for the applications we are concerned with today.

Firstly, we need to recognize that a large part of their differences arose from the fact that Fisher and Jeffreys were occupied with different problems. Fisher studied biological problems where one had no prior information and no guiding theory (this was long before the days of the DNA helix) and the data taking was very much like drawing from Bernoulli's Urn. Jeffreys studied problems of geophysics where one had a great deal of cogent prior information and a highly developed guiding theory (all of Newtonian mechanics giving the theory of elasticity and seismic wave propagation, plus the principles of physical chemistry and thermodynamics); and the data taking procedure had no resemblance to drawing from an Urn. Fisher, in his cookbook, §1, defines statistics as *the study of populations*; Jeffreys devotes virtually all of his analysis to problems of inference where there is no population.

Late in life, Jerzy Neyman was able to perceive this difference. His biographer Constance Reid (1982, p. 229) quotes Neyman thus: "*The trouble is that what we* [statisticians] *call modern statistics was developed under strong pressure on the part of biologists. As a result, there is practically nothing done by us which is directly applicable to problems of astronomy.*"

Fisher advanced, very aggressively, the opposite view; that the methods which were successful in his biological problems must be also the general basis of all scientific inference. What Fisher was never able to see is that, from Jeffreys' viewpoint, Fisher's biological problems were trivial, both mathematically and conceptually. In his early Chapters, Jeffreys (1939) disposes of them in a few lines, obtaining Fisher's inference results far more easily than Fisher did, as the simplest possible applications of Bayes' theorem,* then goes on to more complex problems beyond the ambit of Fisher's methods. Jeffreys (1939, Chap. 7) then summarizes the comparisons with Fisher and Neyman in more general terms.

As science progressed to more and more complicated problems of inference, the shortcomings of the orthodox methods became more and more troublesome. Fisher would have been nearly helpless, and Neyman completely helpless, in a problem with many nuisance parameters but no sufficient or ancillary statistics. Accordingly, neither ever attempted to deal with what is actually the most common problem of inference faced by experimental scientists; linear regression with

* Of course, Fisher's randomized planting methods, which we think to be not actually wrong, but for reasons explained elsewhere in this work hopelessly inefficient, were not reproduced by Jeffreys; nor would he wish to. It appears to be a quite general principle that, whenever there is a randomized way of doing something, then there is a nonrandomized way that delivers better performance but requires more thought.

both variables subject to unknown error. For Bayesian methods [Zellner (1987), Bretthorst (1988)] these are minor technical details that do not deter one from finding the straightforward and useful solutions.

However, we recognize Fisher's high competence in the problems which concerned him. An honest man can maintain an ideology only as long as he confines himself to problems where its shortcomings are not evident. Had Fisher tried more complex problems, we think that he would have perceived the superior power of Jeffreys' methods rather quickly; the mathematics forces one to it, independently of all ideology. As noted, it may be that he started to see this toward the end of his life.

Secondly, we need to recognize the very different personalities and habits of scholarly conduct of the combatants. In any field, the most reliable and instantly recognizable sign of a fanatic is a lack of any sense of humor. Colleagues have reported their experiences at meetings, where Fisher could fly into a rage over some harmless remark that others would only smile at. Even his most ardent disciples (for example, M. Kendall, 1963) noted that the character defects that he attributed to others were easily discernible in Fisher himself; as one put it, "Whenever he paints a portrait, he paints a self–portrait."

Harold Jeffreys maintained his composure, never took these disputes personally and, even in his nineties when the present writer knew him, it was a delight to converse with him because he still retained a wry, slightly mischevious, sense of humor. The greatest theoretical physicists of the nineteenth and twentieth Centuries, James Clerk Maxwell and Albert Einstein, showed just the same personality trait, as testified by many who knew them.

Needless to say (since Fisher's methods were mathematically only special cases of those of Jeffreys), Fisher was never able to exhibit a specific problem in which his methods gave a satisfactory result and Jeffreys' methods did not. Therefore we see in Fisher's words almost no pointing to actual results in real problems. Usually Fisher's words convey only a spluttering exasperation at the gross ideological errors of Jeffreys and his failure to repent. His few attempts to address technical details only reveal his own misunderstandings of Jeffreys.

For example, Jeffreys (1932) gave a beautiful derivation of the $d\sigma/\sigma$ prior for a scale parameter, that we recall in Chapter 12. Given two observations x_1, x_2 from a Gaussian distribution, the predictive probability density for the third observation is

$$p(x_3|x_1, x_2, I) = \int d\mu \int d\sigma \, p(x_3|\mu, \sigma, I) \, p(\mu, \sigma|x_1, x_2, I)$$

If initially σ is completely unknown, then our estimates of σ ought to follow the data difference $|x_2 - x_1|$, with the result that the predictive probability for the third observation to lie between them ought to be $1/3$, independently of x_1 and x_2 (with independent sampling, every permutation of the three observations has the same probability). He shows that this will be true only for the $d\sigma/\sigma$ prior.

But Fisher (1933), failing to grasp the concept of a predictive distribution, takes this to be a statement about the sampling distribution $p(x_3|\mu, \sigma, I)$, which is an entirely different thing; and jumps to the conclusion that Jeffreys is guilty of a ridiculous elementary error. This launches him into seven pages of polemical attacks on all of Jeffreys' work, which display in detail his own total lack of comprehension of what Jeffreys was doing. All readers who want to understand the conceptual hangups that delayed the progress of this field for decades, should read this exchange very carefully.

But in Jeffreys' words there is no misunderstanding of Fisher, no heaping of scorn and no ideological sloganeering; only a bemused sense of humor at the whole business. The issue as Jeffreys saw it was not any error of Fisher's actual procedures on his particular biological problems, but the incompleteness of his methods for more general problems and the lack of any justification for his dogmatically asserted premises (in particular that one must conjure up some hypothetical infinite population from which the data are drawn, and that every probability must have an objectively 'true' value, independently of human information; Jeffreys' whole objective was to use probability to *represent* human information). Furthermore, Jeffreys always made his point quite gently.

For example, Jeffreys (1939, p. 325), perceiving what we noted above, writes of Fisher that, "In fact, in spite of his occasional denunciations of inverse probability, I think that he has succeeded better in making use of what it really says than many of its professed users have." As another example, in one of the exchanges Jeffreys complained that Fisher had "reduced his work to nonsense". In reply, Fisher pounced upon this, and wrote, gleefully: "I am not inclined to deny it." Geisser (1980) concludes that Jeffreys came off second best here; we see instead Jeffreys smiling at the fact that Fisher was deflected from the real issue and fell headlong into the little trap that Jeffreys had set for him.

Having said something of their differences, we should add that, as competent scientists, Fisher and Jeffreys were necessarily in close agreement on more basic things; in particular on the role of induction in science. Neyman, not a scientist but a mathematician, tried to claim that his methods were entirely deductive. For example, in Neyman (1952, p. 210), he states: "$\cdots$ *in the ordinary procedure of statistical estimation there is no phase corresponding to the description of 'inductive reasoning.'* $\cdots$ *all the reasoning is deductive and leads to certain formulae and their properties.*" But Neyman (1950) was willing so speak of inductive *behavior*.

Fisher and Jeffreys, aware that all scientific knowledge has been obtained by inductive *reasoning* from observed facts, naturally enough denied the claim of Neyman that inference does not use induction, and of the philosopher Karl Popper that induction was impossible. We discussed this claim at the end of Chapter 9. Jeffreys expressed himself on this more in private conversations than in public utterances; Fisher publicly likened Popper's and Neyman's strictures to political thought–control. As he put it (Fisher, 1956, p. 7): *"To one brought up in the free intellectual atmosphere of an earlier time there is something rather horrifying in the ideological movement represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world."*

Indeed, their reaction to Popper may be a repetition of what happened in the 18'th Century. Fisher (1956, p. 10), Stigler (1983), and Zabell (1989) present quite good evidence – which seems to us, in its totality, just short of proof – that Thomas Bayes had found his result as early as 1748, and the original motivation for this work was his annoyance at the claim of the 18'th Century philosopher David Hume of the impossibility of induction. We may conjecture that Bayes sought to give an explicit counter–example, but found it a bit more difficult than he had at first expected, and so delayed publishing it. This would give such a neat and natural explanation of so many otherwise puzzling facts that we rather expect that some future historian may uncover definitive proof of it.

Pre–data and Post–data Considerations

The basic pragmatic difference in the two approaches is in how they relate to the data; orthodox practice is limited at the outset to "pre–data" considerations. That is, it gives correct answers to questions of the form:

(A): Before you have seen the data, what data do you expect to get?

(B): If the as yet unknown data are used to estimate parameters by some known algorithm, how accurate do you expect the estimates to be?

(C): If the hypothesis being tested is in fact true, what is the probability that we shall get data indicating that it is true?

Of course, probability theory as logic automatically includes all sampling distribution calculations; so in problems where such questions are the ones of interest, we shall do the same calculations and reach the same numerical conclusions, with at worst a verbal disagreement over terminology.

But as we have stressed repeatedly, virtually all real problems of scientific inference are concerned with "post–data questions":

(A'): After we have seen the data, do we have any reason to be surprised by it?
(B'): After we have seen the data, what parameter estimates can we now make, and what accuracy are we entitled to claim?
(C'): What is the probability *conditional on the data*, that the hypothesis is true?

Orthodoxy is prevented from dealing with post–data questions by its different philosophy. The basic tenet that determines the form of orthodox statistics is that the reason why inference is needed lies not in mere human ignorance of the true causes operative, but in a "randomness" that is attributed instead to Nature herself; just what we call the Mind Projection Fallacy. This leads to the belief that probability statements can be made only about random variables and not about unknown fixed parameters. However, although the property of being 'random' is considered a real objective attribute of a variable, orthodoxy has never produced any definition of the term "random variable" that could actually be used in practice to decide whether some specific quantity, such as the number of beans in a can, is or is not "random".

Therefore, although the question: "Which quantities are random?" is crucial for everything he does, we are unable to explain how the orthodoxian actually decides this; we can only observe what decisions he makes. For some reason, data are always considered random, almost everything else is nonrandom; but to the best of our knowledge, there is no principle in orthodox statistics which would have enabled one to predict this choice. Indeed, in a real situation the data are usually the only things that *are* definite and known, and almost everything else in the problem is unknown and only conjectured; so the opposite choice would seem more natural.

Orthodox theory does not admit the existence of prior or posterior probabilities for a fixed parameter or an hypothesis, because fixed parameters and hypotheses are not considered random variables. We want, then, to examine how orthodoxy manages to pass off the answer to a pre–data question as if it were the answer to a post–data one. Mostly this is possible because of mathematical accidents, such as symmetry in parameter and estimator.

The sampling distribution for an estimator

We have noted why a major part of the orthodox literature is devoted, necessarily, to calculating, approximating, and comparing sampling *pdf*'s for estimators; this is the only criterion orthodoxy has for judging estimators and in a new problem one may need to find sampling distributions for a half–dozen different estimators before deciding which one is best.

The sampling *pdf* for an estimator does not have the same importance in Bayesian analysis, because we do have the needed theoretical principles; if an estimator has been derived from Bayes' theorem and a specified loss function, then we know from perfectly general theorems that it is the optimal estimator for the problem as defined, whatever its sampling distribution may be. In fact, the sampling *pdf* for an estimator plays no functional role in post–data inference, and so we have no reason to mention it at all, unless pre–data considerations are of some interest (for example, in planning an experiment and deciding what kind of data to take and when to stop).

But in addition to this negative (non–functionality) reason, there is a stronger positive reason for diverting attention away from the sampling *pdf* for an estimator; it is not the proper *criterion* of the quality of an inference. Suppose a scientist is estimating a physical parameter α such as the mass of a planet. If the sampling *pdf* for the estimator is indeed equal to the long–run frequencies in many repetitions of the measurement, then its width would answer the pre–data question:

> Q1: *How much would the estimate of α vary over the class of all data sets that we might have got but did not?*

But this is not the relevant question for the scientist. His concern is with the post–data one:

> Q2: *How accurately is the value of α determined by the one data set D that we actually have?*

and according to probability theory as logic, the correct measure of this is the width of the posterior *pdf* for the parameter, not the sampling *pdf* for the estimator. Since this is a major bone of contention between the Orthodox and Bayesian schools of thought, let us understand why they can sometimes be the same, with resulting confusion of pre–data and post–data considerations. Later, we shall see some of the horrors that can arise when they are not the same.

Historically, since the time of Laplace, scientific inference has been dominated overwhelmingly by the case of Gaussian sampling distributions which have the aforementioned symmetry. Suppose we have a data set $D = \{y_1 \dots y_n\}$ with a sampling distribution

$$p(D|\mu, \sigma, I) \propto \exp\left[-\sum_i \frac{(y_i - \mu)^2}{2\sigma^2}\right] \qquad (16\text{–}1)$$

with σ known. Then the Bayesian posterior *pdf* for μ, with uniform prior, is

$$p(\mu|D, \sigma, I) \propto \exp\left[-\frac{n(\mu - \bar{y})^2}{2\sigma^2}\right] \qquad (16\text{–}2)$$

from which the post–data (mean $\pm$ standard deviation) estimate of μ is

$$(\mu)_{est} = \bar{y} \pm \frac{\sigma}{\sqrt{n}} \qquad (16\text{–}3)$$

which shows that the sample mean $\bar{y} \equiv n^{-1}\sum y_i$ is a sufficient statistic. Then if the orthodoxian decided to use $\bar{y}$ as an estimator of μ, he would find its sampling distribution to be

$$p(\bar{y}|\mu, \sigma, I) \propto \exp\left[-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right] \qquad (16\text{–}4)$$

and this would lead him to make the pre–data estimate

$$(\bar{y})_{est} = \mu \pm \frac{\sigma}{\sqrt{n}} \qquad (16\text{–}5)$$

But although (16–3) and (16–5) have entirely different meanings conceptually, they are mathematically so nearly identical that the Bayesian and Orthodoxian would make the same actual numerical estimate of μ and claim the same accuracy. In problems like this, which have sufficient statistics but no nuisance parameters, there is a mathematical symmetry (approximate or exact) which can make the answers to a pre–data question and a post–data question closely related if we have no very cogent prior information which would break that symmetry.

This accidental equivalence has produced a distorted picture of the field; the Gaussian case is the one in which orthodox methods do best because the symmetry is exact, and the difference between pre–data and post–data results is the least. On the basis of such limited evidence, orthodox theory tried to claim general validity for its methods.

But had the early experience referred instead to Cauchy distributions

$$p(y|\mu) = \frac{1}{\pi}\,\frac{1}{1+(y-\mu)^2} \tag{16–6}$$

the distinction could never have been missed because the answers to the pre–data and post–data questions are so different that common sense would never have accepted the answer to one as the answer to the other. In this case, with an uninformative prior the Bayesian posterior *pdf* is

$$p(\mu|D,I) \propto \prod_{i=1}^{n} \frac{1}{1+(\mu-y_i)^2} \tag{16–7}$$

which is still straightforward, although analytically inconvenient. Numerically, the (posterior mean $\pm$ standard deviation) or (posterior median $\pm$ interquartile) estimates are readily found by computer, but there is no sufficient statistic and therefore no good analytical solution.

But orthodoxy has never found any satisfactory estimator at all for this problem! If we try again to use the sample mean $\bar{y}$ as an estimator, we find to our dismay that its sampling *pdf* is

$$p(\bar{y}|\mu,I) \propto [1+(\bar{y}-\mu)^2]^{-1}$$

which is identical with (16–6); the mean of any number of observations is, according to orthodox criteria, no better than a single observation. Although Fisher noted that for large samples, the sample median tends to be more strongly concentrated near the true μ than does the sample mean, this gives no reason to think that it is the *best* estimator by orthodox criteria even in the limit of large samples, and the question remains open today.

We expect that both the Bayesian posterior mean and posterior median value estimators would prove to be considerably better, by orthodox criteria of performance, than any presently known orthodox estimator. Simple computer experiments would be able to confirm or refute this conjecture; but we doubt whether they will be done, because the question is of no interest to a Bayesian, while an indoctrinated orthodoxian would not voluntarily examine any Bayesian result.*

Pro–Causal and Anti–Causal Bias

A criticism of orthodox methods that we shall find in the next Chapter is not ideological, but that they have technical shortcomings (waste of information) which, in practice, all tend to bias our inferences in the same direction. The result is that, when we are testing for a new phenomenon, orthodoxy in effect considers it a calamity to give credence to a phenomenon that is not real, but is quite unconcerned about the consequences of failing to recognize a phenomenon that is real.

* For example, many years ago the writer attempted to publish an article demonstrating the superior performance of Bayesian estimation with a Cauchy distribution, in the small sample case which can be solved analytically – and had the work twice rejected. The Referee accused me of unfair tactics for bringing up the matter of the Cauchy distribution at all, because "$\cdots$ it is well known that the Cauchy distribution is a pathological, exceptional case." Thus did one orthodoxian insulate himself and the readers from the unpleasant truth that Bayesian analysis does not break down on this problem. To the best of our knowledge Bayesian analysis has no pathological, exceptional cases; a reasonable question always has a reasonable answer. Finally, after 13 years of struggling, we did manage to get that analysis published after all (Jaynes, 1976).

But to be fair, at this point we should keep in mind the historical state of affairs, and the far worse practices that the early workers in this field had to counteract. As we noted in Chapter 5, the uneducated mind always sees a causal relation – even where there is no conceivable physical mechanism for it – out of the most far–fetched coincidence.

Johannes Kepler (1571–1630) was obliged to waste much of his life casting horoscopes for his patron (and complained about it privately). No amount of evidence showing the futility of this seems to shake the belief in it; even today, more people make their living as astrologers than astronomers.

In the 18'th and 19'th Centuries, science was still awash with superstitious beliefs in causal influences that do not exist, and Laplace (1819) warned against this in terms that seem like platitudes today, although they made him enemies then. Our opening quotation from Helmholtz shows his exasperation at the fact that progress in physiology was made almost impossible by common belief in all kinds of causal influences for which there was no physical mechanism and no evidence. Louis Pasteur (1822–1895) spent much of his life trying to overcome the universal belief in spontaneous generation.

Although the state of public health was intolerable by present standards, hundreds of plants were credited with possessing miraculous medicinal properties; at the same time, tomatoes were believed to be poisonous. As late as 1910 it was still being reported as scientific fact that poison ivy plants emit an 'effluvium' which infects those who merely pass by them without actual contact, although the simplest controlled experiment would have disproved this at once.

Today, science has advanced far beyond this state of affairs, but common understanding has hardly progressed at all. On the package of a popular brand of rice, the cooking instructions tell us that we must use a closed vessel, because "the steam does the cooking". Since the steam does not come into contact with the rice, this seems to be on a par with the poison ivy myth. Surely, a controlled experiment would show that the *temperature of the water* does the cooking. But at least this myth does no harm.

Other spontaneously invented myths can do a great deal of harm. If we have a single unusually warm Summer, we are besieged with dire warnings that the Earth will soon be too hot to support life. Next year we will have an unusually cold Winter, and the same disaster mongers will be right there shouting about the imminent ice age. Both times they will receive the most full and sympathetic coverage by the news media, who with their short memory and in their belief that they are doing a public service, amplify a thousandfold the capacity of the disaster monger to do mischief, and encourage ever more irresponsible disaster–mongering as the surest way to get free personal publicity.

In 1991, some persons without the slightest conception of what either electricity or cancer are, needed only to hint that the weak 60Hz fields around home wiring or power lines are causing cancer; and the news media gave it instant credence and full prime time radio and TV coverage, throwing the uneducated public into a panic. They set up picket lines and protest marches to prevent installation of power lines where they were needed. The right of the public to be protected against the fraud of false advertising is recognized by all; so when will we have the right to be free of the fraud of sensationally false news reporting – which is also sold to us for profit?

To counter this universal tendency of the untrained mind to see causal relations and trends where none exist, responsible science *requires* a very skeptical attitude, which demands cogent evidence for an effect; particularly one which has captured the popular imagination. Thus we can easily understand and sympathize with the orthodox conservatism in accepting new effects.

But there is another side to this; skepticism can be carried too far. The orthodox bias against a real effect does help to hold irresponsibility in check, but today it is also preventing recognition of effects that *are* real and important. The history of science offers many examples of important discoveries that had their origin in the perception of someone who saw a small unexpected thing in his data, that an orthodox significance test would have dismissed as a random error.† The discovery of argon by Lord Rayleigh, and of cosmic rays by Victor Hess, are examples that come to mind immediately. Of course, they did not jump to sweeping conclusions from a single observation as do the disaster mongers; they used the single surprising observation rather to motivate a careful investigation that culminated in overwhelming evidence for the new phenomenon.

In other fields we must wonder how many important discoveries, particularly in medicine, have been prevented by Editorial policies which refuse to publish that necessary first evidence for some effect, because the one data set that the researcher was able to obtain did not quite achieve an arbitrarily imposed significance level in an orthodox test. This could well defeat the whole purpose of scientific publication; for the cumulative evidence of three or four such data sets might have yielded overwhelming evidence for the effect. Yet this evidence will never be found unless the first data set can manage to get published.

How can editors recognize that scientific discovery is not a one–step process, but a many–step one, without thereby releasing a new avalanche of irresponsible, sensational publicity seekers? The problem is genuinely difficult, and we do not pretend to know the full answer.

In the next Chapter we study some very instructive case histories of science gone wrong, when orthodox statistics was used to support either an unreasonable belief or an unreasonable disbelief in some phenomenon. In every case, a Bayesian analysis – taking into account all the evidence, not just the evidence of one data set – would have led to far more defensible conclusions; so editorial policies that required Bayesian standards of reasoning would go a long way toward solving this problem.

This orthodox bias against an effect is seen in the fact that Feller and others heap ridicule on "cycle hunters" as being irresponsible, seeing in phenomena such as weather and earthquakes periodicities that are not there. It is conceivable that there may be instances of this; but those who make the charge do not document specific examples which we can verify, and so we do not know of any. In economics, belief in business cycles goes in and out of style cyclically. Those who, like the economist Arthur Burns, merely look at a plot of the data, see the cycles at once. Those who, like Fisher, Feller, and Tukey (1958), use orthodox data analysis methods, do not find them.‡ Those who, like Bretthorst (1988) use probability theory as logic are taking into account more evidence than either of the above groups, and may or may not find them. More generally, the reason why some skeptics do not see real effects is that they use methods of data analysis which violate the likelihood principle, and therefore waste some of the information in the data.

What is Real; the Probability or the Phenomenon?

This Orthodox reluctance to see causal effects even when they are real, has another psychological danger because eventually it becomes extrapolated into a belief in the existence of "stochastic processes" in which no causes at all are operative, and probability itself is the only real physical

---

† Jeffreys (1939, p. 321) notes that there has never been a time in the history of gravitational theory when an orthodox significance test (which takes no note of alternatives) would not have rejected Newton's law and left us with no law at all. But nevertheless Newton's law did lead to constant improvements in the accuracy of our accounting of the motions of the moon and planets for Centuries, and it was only when an alternative (Einstein's law) had been stated fully enough to make very accurately known predictions of its own that a rational person could have ceased using Newton's law.

‡ Indeed, orthodox spectrum analysis was invented by Arthur Schuster (1897) for the specific purpose of refuting some claims of periodicities in earthquakes. As noted later, Schuster's periodogram is relevant to this problem; but the evidence lies in its shape rather than its sampling distribution. Very clear evidence for periodicity can be missed by those who do not understand this.

phenomenon. When the search for any causal relation whatever is deprecated and discouraged, scientific progress is brought to a standstill.

Belief in the existence of "stochastic processes" in the real world; (*i.e.* that the property of being "stochastic" rather than "deterministic" is a real physical property of a process, that exists independently of human information) is another example of the Mind Projection Fallacy: attributing one's own ignorance to Nature instead. The current literature of probability theory is full of claims to the effect that a 'Gaussian random process' is fully determined by its first and second moments. If it were made clear that this is only the defining property for an abstract mathematical model, there could be no objection to this; but it is always presented in verbiage that implies that one is describing an objectively true property of a real physical process. To one who believes such a thing literally, there could be no motivation to investigate the causes more deeply than noting the first and second moments, and so the real processes at work might never be discovered.

This is not only irrational because one is throwing away the very information that is essential to understand the physical process; if carried into practice it can have disastrous consequences. Indeed, there is no such thing as a "stochastic process" in the sense that the individual events have no specific causes. One who views human diseases or machine failures as "stochastic processes" as described in some orthodox textbooks, would be led thereby to think that in gathering statistics about them he is measuring the one controlling factor; the physically real "propensity" of a person to get a disease or a machine to fail; and that is the end of it.

Yet where our real interests are involved, such foolishness is usually displaced rather quickly. Every individual disease in every individual person has a definite cause; fortunately, Louis Pasteur understood this in the 19'th Century and our medical researchers understand it today. In medicine one does not merely collect statistics about the incidence of diseases; there are large organized research efforts to find their specific causes.

Likewise, every machine failure has a definite cause; after every airplane crash the Federal Aviation Officials arrive and, if necessary, spend months sifting through all the evidence trying to determine the exact cause. Only by this pursuit of each individual cause can the level of public health and the safety and reliability of our machines be improved.

COMMENTS

An important general conclusion is that in analyzing data – particularly when searching for new effects – scientists are obliged to find a very careful compromise between seeing too little and seeing too much. Only methods of inference which realize all the "resolving power" possible, by taking careful account of all the relevant prior information, all the previously obtained data, and all the information in the likelihood function, can steer a safe course between these dangers and yield justifiable conclusions. Throughout this work we adduce theoretical arguments and numerical examples showing why probability theory as logic cannot give us misleading conclusions unless we feed it false information or fail to give it all the relevant information we have.