CHAPTER TEN

Why You Cannot Be Just a Little Bit Bayesian

To understand how radical the likelihood principle must appear to many objectivists, note first that in accepting this principle one renounces all desire to make his estimates unbiased. An even more radical consequence of the likelihood principle is the thesis of the innocuousness of [rules to stop the experiment].

—Bruno de Finetti, *Probability, Induction and Statistics: The Art of Guessing,* p. 170

The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proved or disproven.

—Ward Edwards, Harold Lindman, and Leonard Savage, "Bayesian Statistical Inference for Psychological Research," p. 193

It would indeed be strange if the information to be extracted from a body of data concerning the relative merits of two hypotheses should depend not only on the data and the hypotheses, but also on the purely external question of the generation of the hypotheses.

—A. W. F. Edwards, *Likelihood: An Account of the Statistical Concept of* Likelihood *and Its Application to Scientific Inference,* p. 30

The likelihood principle implies . . . the irrelevance of predesignation, of whether an hypothesis was thought of beforehand or was introduced to explain known effects.

—Roger Rosenkrantz, *Inference, Method and Decision: Towards a Bayesian Philosophy of Science,* p. 122

In appraising methodological rules for scientific inference the normative epistemologist needs to assess how well the rules promote a given experimental aim. It is entirely reasonable to expect that philosophical accounts of hypothesis testing or confirmation should have something to say in such a metamethodological assessment. I listed this task in chapter 3 as the third way in which accounts of hypothesis testing may be applied in philosophy of science. A danger persists, however, that the view of testing appealed to in such a metamethodological appraisal already embodies principles at cross-purposes with the aim underlying the account to be appraised. This situation exists, I have already alleged, in Bayesian appraisals of the use-novelty (UN) requirement. In this chapter I shall give a full-blown justification for my allegation—but that will be only my first stopping point on the way to a further destination. Catching the Bayesians in this misdemeanor uncovers a pervasive illicitness in the Bayesian Way of performing a methodological critique. The problem, in a nutshell, is this: the underlying rationale of a number of methodological rules is the aim of reliability or severity in the sense I have been advocating, yet that aim runs counter to the aim reflected in Bayesian principles. In section 10.3 I will explicitly take up an even more far reaching outgrowth of this recognition, which explains the title of this chapter.

The intent of the title is not to suggest that all Bayesians are radical subjectivists or strict Bayesians—but somewhat the opposite. What I am arguing is that insofar as one accepts inference according to Bayes's theorem, one is also buying into distinctive principles of relevant evidence, hence criteria for inferences, hence grounds for judging methodological rules. The key issue is the question of the relevance of error probabilities. Accepting minimal Bayesian principles compels renouncing standard error probability principles and their informal counterparts (e.g., severity and reliability in our sense).

The conflict between these two sets of principles is familiar to philosophers of statistics:

> It seems that the divisions in statistics result almost completely from differences in attitude to the question of whether operating characteristics of data analysis procedures are important or not. (Kempthorne 1972, 190)

Error probabilities are examples of operating characteristics of procedures, and they are the linchpin of error statistics. The difference in attitude reflects different principles for interpreting data. Given an observed outcome *x* the error statistician finds it relevant—indeed essential—to consider the other outcomes that could have resulted from the procedure that issued in data *x*. Those considerations are needed to calculate error probabilities. Bayesian inference—although it comes in many different forms—must hold to the likelihood principle, and this leads to the irrelevance of such calculations. James Berger and Robert Wolpert, in their monograph *The Likelihood Principle* (which they abbreviate as LP), assert that

> the philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with the observed x, while frequentist analyses involve averages over possible observations. . . . Enough direct conflicts have been . . . seen to justify viewing the LP as revolutionary from a frequentist perspective. (Berger and Wolpert 1988, 65–66)

I will be making use of their work in section 10.3.

Despite these direct conflicts, particular error statistical procedures often correspond to procedures Bayesians would countenance, albeit with differences in interpretation and in justification. This apparent overlapping of procedures is regarded by some as belittling the significance of the "philosophical" differences between Bayesians and error statisticians on matters of interpretation and justification. Even in the apparent eclecticism of statistical practice, however, the issues of interpretation and justification do not go away; and when it comes to utilizing statistical ideas in philosophy of science, these issues are paramount—although they have generally been overlooked. Since philosophy of statistics, in its formal guise, tends to occupy a rather separate niche in philosophy of science, it is not surprising to find that most philosophers of science are unaware that there are two major conflicting principles of confirmation, support, or testing. Nor is it obvious that this conflict should be of any particular concern to philosophers.

Is it possible that a conflict that, strictly speaking, emanates from two opposed formal statistical schools could shed any light upon the problems still facing philosophers of science? Is it possible that even as the logical empiricist ways are being replaced with "postpositivist" ones that a fundamental principle of evidence and evidential appraisal is still, unknowingly, retained? Is it possible that a good deal of the debate about methodological principles is rooted in the opposition between two principles, made explicit in theories of statistics? The answer to all these questions, I believe, is yes.

First I will illustrate how the novelty debate is skewed when seen through Bayesian glasses. Then we will arrive, finally, at the heart of the conflict between error probability principles and the likelihood principle.

10.1 Novelty and Severity through Bayesian Glasses

Let us begin with the flaw in Bayesian critiques of arguments for the UN requirement. While not immediately obvious—at least it does not seem to have been recognized—the flaw is not difficult to spot, having the results from the previous two chapters under our belts.

In chapter 8, recall, the UN requirement was found to reflect the desire to ensure that evidence counts as good grounds for *H* only to the extent that it may be seen to constitute a good test of *H*, meaning that the evidence stems from a procedure with a low probability of erroneously passing hypothesis *H*—that is, one with high severity. I then set out to evaluate how well the UN requirement accorded with the aim of severe tests. I showed that while a test that violates the UN requirement is assured of passing the hypothesis it tests, it does not follow that it was assured of doing so *whether or not that hypothesis is false*. In short, I showed the argument for requiring UN to be unsound by showing that violating UN need not lead to violating severity, despite the fact that the *intended* aim of use-novelty is severity.

The Bayesian appraisal of accounts of novelty takes a very different tack. To the Bayesian, it has been said, all things are Bayesian, and the Bayesian appraisal of the UN requirement is a perfect illustration of this. The Bayesian appraises the UN requirement according to whether it has a rationale from *its own* vantage point of what counts as good support for a hypothesis. Running the UN requirement through the Bayesian machinery means asking whether satisfying UN is necessary for Bayesian support.[1] Howson and Urbach (1989) make this Bayesian strategy very clear. They note that although "the Bayesian theory of support is certainly inconsistent with" the UN requirement,

> there are arguments for the view, and these both sound convincing and also number among their subscribers many if not most contemporary philosophers of science. We shall examine these arguments now and show that their plausibility vanishes on closer inspection. (Howson and Urbach 1989, 276)

They get the plausibility to vanish only by changing the argument—at least as it is offered by the non-Bayesians they consider (e.g., Giere, Glymour, Worrall). They change the argument by making the "closer inspection" consist of an examination through a Bayesian magnifying glass—through the Bayesian rule for support. I am not criticizing their

[1] That the Bayesian asks about support rather than about good and bad tests does not impede this analysis.

Bayesian scrutiny of the arguments offered by other Bayesians (e.g., Redhead). I concur that attempts at Bayesian justifications of UN will not wash. But aside from these few Bayesian exceptions, the "many if not most contemporary philosophers of science" to whom Howson and Urbach refer are not giving Bayesian arguments for requiring UN. Here is where the inappropriateness comes in.

Whose Rule of Support?

According to Howson and Urbach,

> attempts to show that data which hypotheses have been deliberately designed to entail, as opposed to independently predicting, do not support those hypotheses fail. On the contrary, *the condition for support,* that $\frac{P(e \mid \text{not-}H)}{P(e \mid H)}$ be small, may be perfectly well satisfied in many such cases. (Howson and Urbach, 1989, 279; emphasis added; I replace their *h* with *H*)

The condition for support? So confidently do Bayesians speak of "the condition for support" that the UN proponent may forget to ask whether this was the intended condition when thinking that UN is required for a good test. If it is not, then the Bayesian criticism fails to make a dent in the argument for UN. In fact, it is not.

First, let us be clear about the origin of this condition for support. It comes from the Bayesian condition that for evidence *e* to provide support for hypothesis *H* the posterior probability of *H* given *e* must be higher than the probability of *H* prior to *e*. That is, the posterior probability of *H* must exceed the prior probability of *H:*

> *Bayesian rule for support* (*first form*)*: e* supports *H* if $P(H \mid e)$ is greater than $P(H)$.

Although it is not immediately obvious, this rule is equivalent to the requirement that *e* be more probable under *H* than under not-*H*. That is,

> *Bayesian rule for support* (*second form*)*: e* supports hypothesis *H* if *e* is more probable under *H* than under not-*H*.

Equivalently,[2]

[2] To see how the second form of the Bayesian rule of support falls out from its first form, consider Bayes's theorem and then calculate the ratio of $P(H \mid e)$ and $P(H)$:

$$P(H \mid e) = \frac{P(e \mid H)\ P(H)}{P(e \mid H)\ P(H) + P(e \mid \text{not-}H)\ P(\text{not-}H).}$$

e supports *H* if $P(e \mid H)$ is greater than $P(e \mid \text{not-}H)$.

We have arrived at the rule for support to which Howson and Urbach refer in the above passage. Let us get a little fancier. Let us abbreviate the *Bayesian ratio of support*[3] for *H* as *BR:*

$$BR\text{:}\ \frac{P(e \mid \text{not-}H)}{P(e \mid H)}.$$

We can then write what Howson and Urbach refer to as "the condition for support" as the condition: *e* supports *H* if the Bayesian ratio BR is less than 1. The smaller the BR is, the greater the support for *H.* This condition does not require the posterior probability to be high, just that it be higher than the prior.

For the case where *H* is constructed to fit *e,* Howson and Urbach suppose that *H* entails *e* (hence $P(e \mid H) = 1$), so in considering their discussion I will too. In that case, the Bayesian rule of support becomes extremely simple:

> *Bayesian rule of support where* $P(e \mid H) = 1$: *e* supports hypothesis *H* if $P(e \mid \text{not-}H)$ is less than 1.

It is now easy to see, when the argument for the UN requirement is scrutinized through Bayesian glasses, why the proponent of the UN requirement *appears* to be claiming that violating UN leads to $P(e \mid \text{not-}H)$ being 1. Why? Because that is what it would mean for *a Bayesian* to assert that no support accrues (when *H* entails *e*).

Recall that $P(e \mid \text{not-}H)$ is our friend the Bayesian catchall factor (section 4.3). Calculating the Bayesian catchall factor (except where not-*H* is a point hypothesis) requires prior probability assignments to

Then the ratio,

$$\frac{P(H \mid e)}{P(H)} = \frac{P(e \mid H)}{P(e \mid H)P(H) + P(e \mid \text{not-}H)\,P(\text{not-}H)} = \frac{1}{P(H) + \dfrac{P(e \mid \text{not-}H)}{P(e \mid H)}\,P(\text{not-}H)}.$$

This exceeds 1 just so long as the denominator is less than 1. And remembering that $P(H) = 1 - P(\text{not-}H)$, it is seen that this occurs whenever $P(e \mid \text{not-}H) < P(e \mid H)$.

[3] The BR is often called the likelihood ratio but this is misleading since the hypotheses involved can be disjunctions requiring averaging over prior probabilities. What I am calling the Bayesian ratio of support is also called the Bayes's factor against *H,* but I did not want to confuse the BR with what I call the Bayesian factor on the catchall.

all alternatives to *H*. But the non-Bayesian refuses to employ prior probabilities. Nevertheless, this does not stop Howson and Urbach from using this critique against non-Bayesian arguments for use-novelty. From their Bayesian perspective, all that is needed to vitiate arguments for requiring UN is that UN is not required for Bayesian support. For this it suffices to show that even when *H* is use-constructed, an agent can assign the Bayesian catchall factor a value less than 1. And that is what they do.

Howson and Urbach most specifically consider Giere's position, which we are already familiar with, that evidence used to construct a hypothesis has "no chance of refuting it." In this connection, they consider Giere's discussion of Gregor Mendel. A constraint on Mendel's model, says Giere, was to fit the evidence of the two-to-one ratio of tall to dwarf plants. Thus, following the same pattern of argument articulated in chapter 8, he considers that such a fit was assured even if Mendel's hypothesis had been false. Through Bayesian glasses, Giere looks to be claiming that the Bayesian catchall factor is assured to be 1, and this Howson and Urbach deny:

> It is far from self-evident that Mendel's data would not be improbable were his own explanation of them to be false; indeed, as that was the *only* explanation which seemed plausible to Mendel, its falsity would presumably render those data, were they assumed to be still conjectural, relatively improbable as far as he was concerned. And this, as we have seen, is sufficient for a Bayesian to be able to explain the undoubted fact that Mendel himself took his data to be strongly confirmatory of his model. *Giere has not justified his thesis—nor indeed could he—that* $P(\mathrm{e} \mid not\text{-}\mathrm{H}) = 1$ *when* H *has been designed to explain* e. (Howson and Urbach 1989, 277; emphasis added. To be consistent with my notation, I capitalize their lowercase *h* and use "not-*H*" rather than ~*h*.)

So it suffices to vitiate Giere's argument, according to Howson and Urbach, that *as far as a given agent is concerned* there is no other plausible explanation of the evidence. In fact, Bayesian support will accrue (however small) so long as the subjective Bayesian catchall factor is less than 1.[4] Thus it suffices for support that the agent believes there to be at least one alternative to *H* that does not make the evidence certain! Since there is vast latitude to what agents can believe, their dismissal of Giere's argument is simplicity itself.

[4] There are really two mistaken assumptions in this Bayesian critique of Giere: that the probability of concern is the Bayesian factor on the catchall and that subjective probabilities are relevant.

But it has nothing to do with Giere's argument. Admittedly, Giere is careless in stating what he regards as the intended rationale for UN, but, having acknowledged that he is a non-Bayesian, it is odd that Howson and Urbach suppose that a subjective Bayesian analysis has relevance. They disregard Giere's warning that

> it is crucial to remember that the probabilities involved are physical probabilities inherent in the actual scientific process itself. If one slips into thinking in terms of probability relations among hypotheses, or between evidence and hypotheses, one will necessarily misunderstand this account of the nature of empirical testing. In particular, one must not imagine that to estimate the probability of [a failing result] one must be able to calculate the probability of this result as a weighted average of its probabilities relative to all possible alternative theories. No such probabilities are involved. (Giere 1983, 282–83)

And yet that is precisely how Howson and Urbach construe the probabilistic claim in Giere's argument.

Now it is true that Giere's argument for the necessity of UN is unsound (as I argued at length in chapter 8), but not for the reason the Bayesian alleges. The difference is altogether crucial. The proponent of the UN requirement is not at all claiming that violating UN precludes Bayesian support (whether subjectively or objectively interpreted)! The concern, rather, is with violating the severity requirement. And Bayesian support is easy to obtain even where severity is violated.

Low or Minimally Severe Tests Can Satisfy the Bayesian Requirement for Support

One way of making this point clearer is to consider a restricted version of the UN requirement, which *does* hold:

> *Restricted UN requirement:* Data used to arrive at and test a use-constructed hypothesis cannot count as a good test of that hypothesis if there is a high probability for passing some such hypothesis, even if it is false.

This is, of course, just an instance of the severity requirement: any test that lacks severity is a poor one, and in some cases violating UN makes it easier to carry out strategies that hinder severity. Equivalently, the restricted UN requirement says that a use-constructed hypothesis is poorly supported or poorly tested by evidence if the use-construction rule is one of the unreliable ones. In the extreme case of a gellerized rule, there is no test, and so no genuine support at all.

The existence of highly unreliable use-construction procedures is the reason that many are led to uphold the UN requirement and eschew "double counting" of data. Admittedly, as we saw in chapter 8, it is really only the restricted UN requirement that is warranted. Nevertheless, Howson and Urbach's Bayesian argument against the general UN requirement goes through just as well for this restricted UN requirement. Another way to put this is that even when severity is low or zero, the condition for Bayesian support can be satisfied. Hence finding Bayesian support still available simply cuts no ice with an error-severity person. Satisfying Bayesian support is not sufficient for severity.

This is obvious where calculating the Bayesian ratio (BR) is subjective, that is, where the reason the Bayesian support ratio is small is that the agent simply believes evidence *e* to be incredible under alternatives to *H*. However, minimally severe tests can muster Bayesian support even if the BR is determined by objective likelihoods from a probability model.

Maximally Likely Alternatives Again

We have already seen several examples that would show this. Recall our discussion of maximally likely alternatives and the problem of underdetermination in section 6.5. Let H_0 be the null or test hypothesis, and *H* an alternative hypothesis. The Bayesian condition of support for *H* is satisfied so long as *H* makes *e* more probable than does H_0—so long as the Bayesian ratio BR is less than 1. Here the BR equals

$$\frac{P(e \mid H_0)}{P(e \mid H)}.$$

But one can always find such an H (so long as H_0 does not give *e* probability 1). One simply uses evidence *e* to construct or select an *H* that perfectly fits the evidence *e*. Support would thereby accrue to *H*, even when the restricted UN requirement would be violated. The extreme example of gellerization could illustrate, but so could less artificial examples such as those from "hunting for statistical significance," discussed in chapter 9 (especially in section 9.2).[5]

[5] To see how the gellerized process in example 6.1 would do to make this point, note that the severe testing theorist would describe the process this way: Observe the outcome *e*, find a hypothesis $G(e)$ that makes *e* maximally likely, and then deem $G(e)$ supported by *e*. (Whether one goes on to measure the support is irrelevant.) Although the particular hypothesis erected to perfectly fit the data will vary in different trials, for every data set some such alternative may be found. Therefore, supporting the maximally likely hypothesis constructed is assured, even if that hypothesis is false.

Teddy Seidenfeld clued me in to a nifty real-life example from sports that offers a different kind of illustration. A person who wants to show he has a system for

Hunting Again

To see how the Bayesian rule accords support to a hypothesis that an error theorist would consider poorly tested, we can recall the strategy of hunting used in the study of infant training discussed by Kish (1970) and detailed in section 9.2. For simplicity, consider that a single infant training experience is of interest, early weaning. Suppose that the procedure is to search among 100 factors for 1 that is highly correlated with having been subjected to early weaning. Say that such a correlation is found between early weaning and a tendency toward shyness in older children. Outcome e is the difference between the proportions of the early weaners and the late weaners who are or claim to be shy. Hypothesis H_0 asserts that the observed correlation between early weaning and shyness is spurious, or due to chance. The procedure will test hypothesis H_0 only upon finding an e that is very improbable given that H_0 is true—so the numerator of the Bayesian ratio is small. For example, the difference sought may be required to be at least 2 standard deviations (corresponding to the .05 "computed" level of significance). The alternative H may assert that the correlation is real and in the direction observed:

> *H:* Early weaning is correlated with shyness in young children.

Hypothesis H is deliberately chosen or constructed so that the denominator of the BR, $P(e \mid H)$, is high (perhaps maximal). The evidence is far more probable given hypothesis H than given the null hypothesis H_0.

Notice the similarity between the improbability of the particular e given the chance hypothesis H_0 and the small "computed significance level" in the last chapter. The probability of this particular outcome e—a high correlation between early weaning and shyness—is very low given the null hypothesis H_0. The alternative—*by design*—makes this observed correlation highly probable. So the Bayesian rule of support is satisfied—indeed, it is well satisfied, since the ratio is not just less than 1 but very small.

Such an example reveals in no uncertain terms the mistake that has gone unchecked in Bayesian reconstructions of non-Bayesian arguments. The mistake stems from the fact that satisfying Bayesian sup-

predicting the winning slate in a football series offers this "test": If the list of winners he sends you at the start of the season turns out to be correct, then it is regarded as good support that his system really works. However, each possible permutation of winners is sent to a different sports fan, so his system is assured of acquiring high support (by someone) even if his system has no predictive ability beyond mere guessing.

port is not sufficient for satisfying severity. That a methodological rule is not required for Bayesian support does not license inferring that the rule is not required for non-Bayesian measures of support—such as those based on error probabilities.

Satisfying Severity versus Satisfying Bayesian Support

It will be useful to recapitulate the distinguishing feature of a severity calculation by way of the Bayesian rule of support (that *e* supports *H* if the BR is small). In appraising such a rule, the error statistician is concerned with its behavior under repetitions. With respect to the case being discussed, the severity criterion requires asking how often the rule would award support to hypotheses about the effects of infant training on personality, even if they are false and there is no real difference in personality traits among those subjected to the different infant training.[6] We can answer this question by viewing the Bayesian ratio BR as a statistic, a function of the data. For each trial of the experiment the BR takes some value. So long as the BR is less than 1, support accrues to the use-constructed alternative *H*, but we are imagining that the rule is even more demanding: the BR must be very small. Severe testers want to know how often such strong support would accrue for some nonchance hypothesis (between infant training and personality) or other, even if H_0 is true. And they want to know this even after the result is in and a particular value for the BR has been calculated. With 100 different tests, it is highly probable that at least one 2-standard-deviation difference would be found, even if all the null hypotheses were true.[7] The severity of the test that *H* passes is 1 minus this, and thus is very small, practically 0. Although the probability of *any particular* statistically significant result is low, the probability of some high correlation *or other* is high.

To answer the severity question, the error statistician needs to consider something that from the Bayesian standpoint is irrelevant—the behavior of the statistic (in this case the BR) in a series of (real or hypothetical) repetitions. This is the experimental distribution of statistic BR. (Precisely why it is irrelevant for the Bayesian will be taken up in section 10.3.) Such considerations, in the view of the error statistician, are necessary to scrutinize the Bayesian procedure of assessing support by calculating the BR.

[6] Since this is an example of the type where the hypothesis selected for testing can vary, the severity criterion becomes SC with hypothesis construction (defined in section 6.6).

[7] Applying the calculation discussed in chapter 9, the probability is about .99.

Sum-Up

We have shown in this section that the Bayesian rule of support is unreliable in the sense that it allows support to accrue to hypotheses, with high probability, even if the hypotheses are false. True, merely satisfying the rule of support does not say that the posterior probability of *H* is high, nor that the increment in the posterior is large. It says merely that *some* support accrues to H; it may, depending on the prior probabilities, be tiny. *But that is the Bayesian condition for support that Howson and Urbach use to denounce arguments for the UN requirement.* Hence it is appropriate for us to consider it in questioning their denunciation. Moreover, it makes sense for them to consider the minimal requirement for support because that is what they regard as being challenged by the claim that UN is necessary. What I have shown is that their criticism is unsound because satisfying Bayesian support does not entail satisfying severity. And all I claim to be doing just now is discounting this Bayesian criticism.

It may be objected that I am evaluating the Bayesian criterion of support from an error probability (e.g., severity) stance. That is exactly right. It is entirely appropriate to do this in answering Bayesian critiques of the novelty requirement, because the aim of novelty is severity.

Novelty through Bayesian Glasses

Since the UN requirement reflects a concern about error-severity, it is easy to see why the Bayesian concludes that the novelty requirement will not hold up. The point is made informally by remembering what was said about Bayesian philosophers being descendants of holders of "logical theories" of confirmation. Like the logical theorists before them, those who assess the import of evidence by way of the Bayesian ratio regard as irrelevant how the hypothesis was generated. (There are also prior probabilities, of course, but these are separate from the import of the evidence for support.) The Bayesian support ratio BR, which is how the import of the evidence comes through for a Bayesian, is unaffected by the manner of hypothesis generation. Hence Howson and Urbach's (1989) assertion that "the Bayesian theory of support is certainly inconsistent with" the novelty requirement (p. 276).

Not all Bayesians deny the UN requirement, but both supporters and detractors share the assumption that what has to be shown or denied is its bearing on the Bayesian measure of support (the Bayesian support ratio). To be fair, the reason for this is not just the Bayesian tendency to appraise all principles of evidence according to the Bayesian formula. In this case, non-Bayesian UN proponents have unconsciously opened up their argument to Bayesian scrutiny by making ambiguous statements about the grounds for requiring UN. (We saw this in discussing Worrall and Giere in section 8.3.) What UN proponents have failed to see or failed to state unequivocally is that the raison d'être for the use-novelty principle has to do with severity—in our non-Bayesian sense. To think that severity matters is to think that test results cannot be appraised without considering the error properties of the entire procedure from which the results arose—the very properties that Bayesians are happy to declare irrelevant.

Bayesian Ways to Make Novelty Matter

If violating novelty is not relevant to Bayesians, then they are faced with the problem of accounting for scientific cases where it does seem to have mattered. It is open to the Bayesian to propose the classic Bayesian move. Any scientific appraisal that you think turned on the mode of hypothesis construction (and a concern with the corresponding lack of severity or reliability), the Bayesian may allege, can be reconstructed as having turned on some difference in prior probabilities. Indeed, any time there is a difference in the appraisal of two hypotheses that entail (or equally fit) evidence, the Bayesian *must* locate the source of the difference in the priors. There is no place else to locate the difference in the Bayesian algorithm. Granted, with given assumptions about one's prior probabilities in hypotheses, some, though not all, violations of severity can be made to correspond to tests that are poor or comparatively poor on Bayesian grounds. Even so, the Bayesian reconstruction incorrectly locates the actual rationale for disparaging these tests.

Various attempts in which, through just the right assumptions and prior probabilities, use-novel hypotheses receive higher Bayesian support than use-constructed ones include Campbell and Vinci 1983; Howson and Urbach 1989; Maher 1988, 1993c; Redhead 1986; and Rosenkrantz 1977. Regrettably, my remarks on them must be brief. These attempts, if they are not guilty of mistakes about probabilities, err in one or both of two ways: either (1) they violate the likelihood principle[8] or (2) they fail to capture the actual epistemological rationale for why certain violations of use-novelty are taken as problematic and why they should be.

[8] The arguments between Maher and his critics Howson and Franklin (1991) really turn on this issue, although the debate has not been framed in these terms. Maher needs to appeal to the idea of the reliability of the method by which hypotheses are generated. Such an idea finds its home in error-statistical not Bayesian accounts.

My basis for (2) is that the problem caused by unreliable use-construction procedures is not a problem about prior degrees of belief in hypotheses. To underscore this point, consider a single hypothesis so that the problem cannot be traced to prior probability assignments. A hypothesis that might do is one we saw in the example with which I began our discussion of novelty—the bombing of the World Trade Center. Let hypothesis *H* assert that group X drives into the garage at the given time and explodes the bomb. When advanced before the bombing, *H* passes a (relatively) severe test. The probability that the before-the-fact description of the bombing would have fit the actual facts so well if *H* were false is low. In contrast, hypothesis *H* does not pass a severe test when advanced after the details of the bombing have been reported. Being able to come up with a hypothesis that fits the reported occurrence is precisely the sort of move that is open to anyone who wants to assign credit for the bombing, even though the alleged group had nothing to do with it. At the stage that we disparage the after-the-fact calls claiming responsibility for the bombing, nothing has been done to rule out this error.

More generally, the difference in the evidential import of two pieces of evidence, both of which fit hypotheses equally well, is located in a difference in the reliability of test processes. This is where the error statistician locates it. The Bayesian desiring to make out a difference locates its source elsewhere (depending on which attempt one considers). For this reason, error statistical criteria match our intuitions about differential support better than Bayesian ones.

10.2 The Old Evidence Problem

The position that how or when a hypothesis is generated is irrelevant to a Bayesian may seem puzzling in the light of what has been said about Bayesians having an "old evidence" problem. The puzzle results because, whereas for most Bayesians novelty *never* matters to support, for a few others it *always* does! These other Bayesians are said to have an old evidence problem.

An account is said to have an old evidence problem if it has the consequence that old or known data fail to count as evidence in support of a hypothesis (it requires temporal novelty). As critics of temporal novelty show, this conflicts with many cases where known evidence is regarded as providing excellent support for hypotheses.[9]

[9] We have shown (in chapters 8 and 9) that contrary to what some have claimed, Neyman-Pearson statistics does not have an old evidence problem.

Why is the subjective Bayesian supposed to have an old evidence problem?

The allegation, brought to the forefront by Glymour (1980), goes like this: if probability is a measure of degree of belief, then if an agent already knows that *e* has occurred, the agent must assign $P(e)$ the value 1. Hence $P(e \mid H)$ is assigned a value of 1. But this means no Bayesian support accrues from *e*. For if $P(e) = P(e \mid H) = 1$, then $P(H \mid e) = P(H)$. The Bayesian condition for support is not met.

Another way of phrasing the problem is that if evidence *e* is known and so assigned a probability of 1 by an agent, then the agent also assigns a probability of 1 to the Bayesian catchall factor, that is, $P(e \mid \text{not-}H) = 1$. So, the BR equals 1, and no Bayesian support accrues to *H*.

How do subjective Bayesians respond to the charge that they have an old evidence problem? The standard subjective Bayesian response is given by Howson and Urbach (1989) and by Howson (1989).

> It has been argued (e.g., by Glymour . . .) that since e is, by assumption, known at the time h is formulated, its probability must be 1, so that P(e | ~h) = 1 also. . . . But is it? If Glymour is right then P(e | ~h) would be 1 even if h had not been constructed to explain e, if e is a known fact, and so e would not support h in this case either. But Glymour is not right. . . . The Bayesian interprets P(e | ~h) as how likely you think e *would be* were h to be false. . . . On this construction the value of P(e | ~h) is *independent* of whether h was or was not constructed to explain e. (Howson 1989, 386)

But many people—Bayesians included—are not too clear about how this "would be" probability is supposed to work.

Consider the known evidence in the Brownian motion example (chapter 7). Brownian motion was known before formulation of the Einstein-Smoluchowski theory. To assess the support that this phenomenon affords this theory, the subjectivist imagines Perrin or some other agent asking something like this: How probable would I regard the phenomenon of Brownian motion, were the Einstein-Smoluchowski theory false? If the agent thinks that *H* is the only plausible explanation of Brownian motion (*e*), he or she may well think the occurrence of the phenomenon very improbable in a world in which *H* were false. So the agent may well assign a small, and not a maximal, value to $P(e \mid \text{not-}H)$. If we ask how the agent figures out this probability, Howson and Urbach will say (recall chapter 3), *it is not our job.* "We are under no obligation to legislate concerning the methods people adopt for assigning prior probabilities" (Howson and Urbach 1989, 271).

It is pretty clear that an agent would not want to assign a probability of 1 to known evidence because the evidence is *always* known by the time Bayes's theorem is applied. Glymour's "mistake," according to Howson and Urbach, is in supposing that probability assignments are to be relative to the totality of current knowledge. In fact, "they should have been relativised to current knowledge minus *e*" (p. 271):

> Once *e* has become known . . . the probabilities $P(e \mid h)$, $P(h)$, and $P(e)$ . . . are relativised to the counter-factual knowledge state in which you still do not know *e*. (P. 271)

In their view, the support that *e* gives to *H* is to be computed by considering how knowledge of *e* would alter one's degree of belief in *H*, on the supposition that one did not yet know *e*.[10]

Counterfactual Degrees of Belief

The need to consider an agent's counterfactual knowledge state while still keeping assignments coherent is fraught with problems, many of which Glymour discusses (e.g., Glymour 1980, 87–91). Does one try to go back in time, and if so how far back? Were the counterfactual knowledge view taken seriously, says Glymour,

> we should have to condemn a great mass of scientific judgments on the grounds that those making them had not studied the history of science with sufficient closeness to make a judgment as to what their degrees of belief would have been in relevant historical periods. (Glymour 1980, 91)

What if we stay in the present but imagine subtracting out the knowledge of *e*? That seems silly and hardly an easier feat. In this view, for Einstein to assess the evidential bearing of the perihelion of Mercury on the relativistic theory of the gravitational law, he needs to imagine whether there would be an increase in his assignment of probability to the law had he not already known of the perihelion phenomenon. However, as Earman (1992) nicely points out, Einstein developed the theory hoping to account for just this phenomenon; had he not known about it, perhaps he would not have developed the theory. "And if someone else had formulated the theory, Einstein might not have taken it seriously enough to assign it a nonzero prior" (p. 123). In any

[10] Paul Horwich has a somewhat different tactic for dealing with this problem. He seems to allow that the probability of known evidence *e* equals 1, but maintains that a Bayesian should assess how much *e* would alter our degree of belief assignment to *H* relative to "our epistemic state prior to the discovery" of *e*, when its probability was not yet 1 (Horwich 1982, 53).

event, a scientist does not actually assess the import of the evidence *e* in hand by imagining the probability change that would obtain if *e* were unknown to him.

The problem of assigning counterfactual degrees of belief also makes it difficult to carry out Bayesian reconstructions of scientific episodes, upon which Bayesianism's usefulness to the philosophy of science depends. The Bayesian "solution" to Duhem's problem (recall chapter 4) requires the philosopher to assign probabilities so as to reflect the beliefs actually held by given scientists. But are philosophers in a position to follow Dorling's suggestion that we consider the betting odds a typical scientist would have been willing to place on *e*, assuming that *e* had not yet been discovered? (Dorling 1979, 182). Even if one could arrive at such a counterfactual probability assignment, the deeper question remains: why is it relevant to either making or scrutinizing a scientific inference from evidence *e?*

Take the example of the World Trade Center bombing. In the "imagine the evidence is not known yet" view, to assess the support to give a hypothesis *H* about the group responsible for the bombing—where this hypothesis is called in after the fact—I must consider how the evidence of the explosion would have altered my belief in *H* if I did not yet know of the explosion. I am to be just as impressed when nonnovel evidence fits a hypothesis as I am when novel evidence fits it. The reliability of the use-construction method does not enter. At the end of 1994 I enjoyed fitting the data on profitable stocks over the past twelve months into a highly profitable strategy for buying and selling during that time. Would the Bayesian assess how well these data support my system in the *same way* had I advanced it at the end of 1993? Yet isn't that what pretending you didn't yet know the data would seem to countenance?

The most satisfactory Bayesian way around the old evidence problem, in my view, is for the Bayesian to restrict the probability assignments to specific statistical models of experiments and generic outcomes of those experiments. For example, the probability of heads on a toss of a fair coin is one-half, independent of anyone knowing the outcome. Presumably, this is the "objective" Bayesian solution. This would considerably limit the scope of the Bayesian account in philosophy of science, and the problem of prior probabilities of hypotheses would remain.

Although there seems to be no single agreed-upon solution to the old evidence problem, some such solution is assumed by the generally accepted Bayesian position on novelty described in section 10.1: novelty (temporal and use-novelty)—in and of itself—never matters to support.

Garber's and Jeffrey's Way

The Bayesian rejection of the UN requirement may be put this way: an argument showing that violating UN precludes Bayesian support would also show that known evidence in general precludes Bayesian support (leading to the old evidence problem). There are a few Bayesians, however, who want to accept the UN requirement, yet get around the old evidence problem—even if this requires changing the probability axioms.

This way around the old evidence problem, developed by Daniel Garber, Clark Glymour, Richard Jeffrey, and others, reflects the idea that even old evidence can be made to support *H,* as long as the fact that *H* entails *e* is new. This attempt requires altering the axioms of the probability calculus relative to one's knowledge. For example, instead of stipulating that all tautologies have probability 1, the assignment would be relativized to *knowing* it to be a tautology. (Essentially the same idea was originally developed by I. J. Good, but for a different reason.) The problems that such attempts raise for Bayesians will not be gone into here.[11] They make confirmation even more a matter of subjective, psychological beliefs than traditional subjective Bayesianism.

Note mainly that accepting the Garber-Jeffrey way out of the old evidence problem takes us right back to the problem of use-novelty. With a use-constructed hypothesis *H* it *is* known that *H* entails or otherwise fits *e,* so it is not saved from the chopping block that spares cases in which the entailment is unknown. Those who appeal to such accounts to solve the old evidence problem *start out assuming* the position that violating UN always precludes support. Take Garber:

> Suppose that S constructed h *specifically* to account for e, and knew, from the start, that it would. It should not *add* anything to the credibility of h that it accounts for the evidence that S knew all along it would account for. (Garber 1983, 104)

There is no attempt to justify this on Bayesian principles. It is hard to see how it could be so justified since, as Bayesians are quick to show, the UN requirement violates the likelihood principle. Moreover, as I argued in chapter 8, the position that violating UN precludes support is untenable: it would rule out many cases with excellent and even maximal support (as in estimation techniques). It seems hardly worth changing the probability axioms only to be left upholding this position.

Regardless of how Bayesians settle this family quarrel about how

[11] See for example, Miller 1987, 305–8 and Earman 1992.

best to deal with old evidence, the problem only furthers my criticism of Bayesian critiques of methodological principles: methodological principles are often based on non-Bayesian ideas about error probabilities, and these ideas run counter to a fundamental Bayesian principle, the likelihood principle. If we are to use ideas from statistics to obtain a philosophical understanding of reasoning in science—something I heartily endorse—then we need to be very clear on the fundamental differences between Bayesian and error-probability approaches. That is the purpose of the next section.

10.3 The Likelihood Principle (LP) and Stopping Rules

> One of the claims [of the Bayesian approach] is that the experiment matters little, what matters is the likelihood function after experimentation. . . . It tends to undo what classical statisticians have been preaching for many years: think about your experiment, design it as best you can to answer specific questions, take all sorts of precautions against selection bias and your subconscious prejudices. (LeCam 1977, 158)

Why does embracing the Bayesian position tend to undo what classical statisticians have been preaching? Because Bayesian and classical statisticians view the task of statistical inference very differently.

In chapter 3 I contrasted these two conceptions of statistical inference by distinguishing evidential-relationship or E-R approaches from testing approaches, and explained why E-R approaches in general, and the Bayesian Way in particular, have appealed more to philosophers than classical testing approaches. The E-R view is modeled on deductive logic, only with probabilities. In the E-R view, the task of a theory of statistics is to say, for given evidence and hypotheses, how well the evidence confirms or supports hypotheses (whether absolutely or comparatively). There is, I suppose, a certain confidence and cleanness to this conception that is absent from the error-statistician's view of things. Error statisticians eschew grand and unified schemes for relating their beliefs, preferring a hodgepodge of methods that are truly ampliative. Error statisticians appeal to statistical tools as protection from the many ways they know they can be misled by data as well as by their own beliefs and desires. The value of statistical tools for them is to develop strategies that capitalize on their knowledge of mistakes: strategies for collecting data, for efficiently checking an assortment of errors, and for communicating results in a form that promotes their extension by others.

Given the difference in aims, it is not surprising that information relevant to the Bayesian task is very different from that relevant to the task of the error statistician. In this section I want to sharpen and make more rigorous what I have already said about this distinction.

The different positions staked out by error statisticians and by those who accept the likelihood principle (e.g., Bayesians), it should by now be clear, are not only of concern to philosophers of statistics. This opposition, I have been urging, while crystallized in formal statistical principles, is implicated, if only informally or implicitly, in a cluster of disputes in philosophy of science. Overlooking this distinction in underlying principles, we saw in section 10.1, has permitted what are essentially question-begging appraisals of methodological rules to go unchallenged. Further, the secret to solving a number of problems about evidence, I hold, lies in utilizing—formally or informally—the error probabilities of the procedures generating the evidence. It was the appeal to severity (an error probability), for example, that allowed distinguishing among the well-testedness of hypotheses that fit the data equally well (the alternative hypothesis objection, chapter 6).

Having been reminded of the philosophy of science ramifications of the dispute between opposing statistical philosophies, I want to revisit that dispute, but this time I want to go deeper into its core. Whereas in thinking of the key difference between Bayesians and error statisticians one most often thinks of the former's willingness to assign prior probabilities to hypotheses and the latter's insistence upon methods that do not require such assignments, the difference I now want to concentrate on is more fundamental. In this place, more than any other, one can see the chasm that divides the Bayesian from the error-statistician.

Stopping Rules

Let me begin with a question, as I did with hunting for statistical significance in chapter 9. The situation now resembles that case, but instead of hunting for a statistically significant property, we will imagine that the researchers have an effect they would like to demonstrate, and that they plan to keep experimenting until the data differ statistically significantly, say at the .05 level, from the null hypothesis of "no effect." (In other words, the researchers keep going until they get a 2-standard-deviation difference. .05 is the *computed* or "nominal" level of significance.) We can call this a "try and try again" procedure. The effect may be anything one would like to consider—that a subject can do better than chance at guessing an ESP card, that one treatment does better than some other (with regard to some symptom), that the discrepancy from some parameter value is real—or any of the other kinds of examples we have considered. You are presented, say, with the statistically significant data ultimately arrived at. The question is whether it is relevant to your appraisal of the effect that the data resulted from the try and try again procedure.

For a simple example, imagine a subject of an ESP experiment, Zoltan. During each trial of the experiment Zoltan must predict ahead of time the next ESP card in a deck of cards. Suppose that after a long series of trials Zoltan scores a relative frequency of successful predictions that exceeds the relative frequency expected by chance alone by an amount sufficient to attain a .05 significance level. Would it be relevant to your evaluation of the evidence if you learned that Zoltan had planned all along to keep running trials for as long as it took to reach the (computed) significance level of .05? Would you find it relevant to learn that after, say, 10 trials, having failed to rack up enough successes to reach the .05 level of statistical significance, Zoltan went on to 20 trials, and failing yet again he went to 30 trials, and then 10 more, and on and on until, say on trial 1,007, he finally attained a statistically significant result?

A plan for when to stop an experiment is called a *stopping rule*. So my question is whether you would find knowledge of the stopping rule relevant in assessing the evidence from a statistical test. If you would you are in good company, for that is how standard error statistics answers the question. From the Bayesian point of view, however, you are incoherent!

The Likelihood Principle

Having alluded more than once to the likelihood principle (LP), I will now say more specifically what it asserts.[12] The LP is regarded as having been articulated by non-Bayesian statisticians, principally George Barnard (1947) and R. A. Fisher (1956). But, as it is *their* principle now, I will let the Bayesians do the talking.[13]

In their classic piece, Edwards, Lindman, and Savage (1963) spell out the LP as follows. They consider two experiments involving the same set of hypotheses H_1 up to H_n. Let D be an outcome from the first

[12] There is also something called the "weak likelihood principle," but since that is not in dispute between Bayesians and error statisticians I will not discuss it. Richard Miller (1987) uses the term to mean something different. What he has in mind is a principle sometimes called the law of likelihood (e.g., by Hacking; noted in section 6.6). The formal likelihood principle should not be confused with these other notions.

[13] I do not think there are more than a handful of non-Bayesian ("likelihoodists") who still accept the LP.

experiment and D' from the second. They ask, "Just when are D and D' thus evidentially equivalent, or of the same import?" (p. 237). Their answer is when, for some positive constant k,

$$P(D' \mid H_i) = kP(D \mid H_i)$$

for each i. That is, D and D' are evidentially equivalent when the likelihood of H_i given D is a multiple of the likelihood of H_i given D'. That is because the posterior probabilities of the hypotheses come out the same, as the interested reader can check.

A reminder: $P(D \mid H)$ is called the likelihood of H, but for a non-Bayesian this can be calculated only where H is a simple statistical hypothesis—not a disjunction. That is why, for example, where we considered $P(D \mid \text{not-}H)$, we gave it a different name (the Bayesian catchall factor). In discussing the LP, however, the Bayesian often wishes to identify a conflict between Bayesian and non-Bayesian treatments of evidence. To demonstrate this conflict the Bayesian has to consider only examples in which $P(D \mid H_i)$ is calculable for a non-Bayesian, that is, where these likelihoods are calculated the same way for Bayesians and non-Bayesians. I will also maintain this restriction.

To this end, the LP is often stated with reference to hypotheses about a particular parameter μ, such as the probability of success (on a Binomial trial) or the mean value of some characteristic. L. J. Savage (1962) states it this way, where x and y (rather than D' and D) now refer to the two results:

> According to Bayes's theorem, $P(x \mid \mu)$ . . . constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, if y is the datum of some other experiment, and if it happens that $P(x \mid \mu)$ and $P(y \mid \mu)$ are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ. (P. 17; I substitute P for his Pr and μ for his λ)

It is a short step from this reasoning to see the conflict with classical or "orthodox" theory. As Lindley (1976) puts it,

> we see that in calculating [the posterior], our inference about μ, the only contribution of the data is through the likelihood function. . . . In particular, if we have two pieces of data x_1 and x_2 with the same likelihood function . . . the inferences about μ from the two data sets should be the same. *This is not usually true in the orthodox theory, and its falsity in that theory is an example of its incoherence.* (P. 361; emphasis added. I replace his q with μ)[14]

[14] Where the likelihoods are proportional for the hypotheses under consideration they are sometimes said to be the same likelihood function. That is how

Savage's Message at the 1959 Forum

It was this conflict that was uppermost in Savage's mind when in 1959 he led a forum attended by several leaders in statistics. He declared:

> In view of the likelihood principle, all of these classical statistical ideas come under new scrutiny, and must, I believe, be abandoned or seriously modified. (Savage 1962, 18)

Attendees at this forum included P. Armitage, I. J. Good, G. Barnard, M. S. Bartlett, E. S. Pearson, D. Lindley, D. R. Cox, and others, representing a mixture of statistical schools. Savage announced to this distinguished group that all the classical statistical notions—all the notions under "error statistics"—significance levels and tests, confidence levels and interval estimates, criteria based on error probabilities—all are suspect. They are suspect because they come into conflict with the LP.

The conflict is most pronounced, Savage explains, on the relevance of stopping rules. While it is widely held that the import of the evidence depends on the stopping rule in examples like the one above, in fact, Savage warns, this violates the LP. The LP tells you that it can make no difference to the import of evidence whether the experimenter had planned to "try and try again" until a (computed) .05 significant result is achieved, or whether the experimenter had planned to run just one experiment, with some fixed sample size, and let the chips fall where they may. Let us refer to the former, try and try again plan, as the *optional stopping plan,* and the latter, prespecified plan as the *fixed sample size plan*. Why, according to the LP, does a result have exactly the same thing to say about μ when generated through optional stopping as when generated through a fixed sample size plan? Because the probabilities of the results from the two experiments (given μ)—i.e., the likelihoods—are proportional to each other.

Zoltan's 1,007 trials—whether by optional stopping or fixed sample size—consist of a string of *k* successes and 1,007 − k failures. It can be pictured as a string of 1,007 *s*s and *f*s, such as

s,s,f,s,f,f,f,s,f,f,s,f,s,. . . . .

This string is the outcome *x*. The hypothesis of interest is a hypothesized value for μ—the probability of success on each trial. The posterior probability accorded to μ with *either* experimental plan is a function of the prior probability and the likelihood, $P(x \mid \mu)$. And in both cases,

Lindley is using "same likelihood function" here. It will be less confusing to just say that their likelihoods are proportional.

$P(x \mid \mu) = \mu^k(1 - \mu)^{1007-k}$. That is, the data x enter into the Bayesian computation the same way whether they arose from the optional stopping plan or the fixed sample size plan.

> In general, suppose that you collect data of any kind whatsoever—not necessarily Bernoullian, nor identically distributed, nor independent of each other . . . —stopping only when the data thus far collected satisfy some criterion of a sort that is sure to be satisfied sooner or later, then the import of the sequence of n data actually observed will be exactly the same as it would be had you planned to take exactly n observations in the first place. (Edwards, Lindman, and Savage 1963, 238–39)

This is called the *irrelevance of the stopping rule.* Those who accept the LP hold to the irrelevance of the stopping rule.[15]

How then, in contrast, do error statisticians render stopping rules relevant? By operating with a different notion of relevant evidence. In their view, it *is* relevant to what the data are saying about the population parameter μ to learn that the result in front of them—x—came from the try and try again (optional stopping) method. Mathematically, this corresponds to the fact that x does not enter the error statistician's computations by itself but always by considering error properties of the experimental procedure from which x arose. Information about stopping rules does not show up in likelihoods, but it sure shows up in a procedure's error probabilities.

Edwards, Lindman, and Savage, quite rightly, regard this difference in attitude on the relevance of stopping rules as a central point of incompatibility between the two approaches. That is why it is so important for us. To the holder of the LP, the irrelevance of the stopping rule is a point in its favor, but to the error statistician the situation is exactly the reverse. P. Armitage (1962), the most forthright error statistician at the 1959 Savage forum, puts it plainly:

> I think it is quite clear that likelihood ratios, and therefore posterior probabilities, do not depend on a stopping rule. Professor Savage, Dr Cox and Mr Lindley take this necessarily as a point in favour of the use of Bayesian methods. My own feeling goes the other way. I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then "Thou shalt be misled if thou dost not know that." If so, prior probability methods

[15] There are certain exceptions where the stopping rule may be "informative," but I keep to examples that Bayesians do not regard as falling under this qualification.

> seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling. (P. 72)

The error statistician wants to take the method of sampling into account because, as was known in 1959, the try and try again method allows experimenters to attain as small a level of significance as they choose (and thereby reject the null hypothesis at that level), even though the null hypothesis is true.[16] If allowed to go on long enough, the probability of such an erroneous rejection is one! So the *actual* or *overall* significance level is not .05 but 1!

Optional Stopping Leads to High or Maximal Overall Significance Levels

Just as, in chapter 9, we calculated the actual significance level to be the probability of hunting down *some statistically significant factor or other* given that none are really correlated, here we calculate the actual or overall significance level as the probability of finding a statistically significant difference from a fixed null hypothesis *at some stopping point or other* up to the point at which one is actually found. The overall significance level accumulates.

We need to be extra careful with the term *statistically significant difference* in the optional stopping case. Here, one keeps taking more and more samples until the observed difference is *computed* to be statistically significant, until it is, say, 2 standard deviations away from the null hypothesis. The computed significance level with an optional stopping plan refers to the significance level that would be calculated under a fixed sample size plan—.05. Say it took *k* tries to achieve a difference computed to be .05 statistically significant. The *actual* or overall significance level is the probability that out of *k* tries at least one would be computed to be .05 statistically significant, even if the null hypothesis is true.

Unlike the case of hunting, there is a substantial literature on how to run and calculate overall significance levels—error probabilities—for tests with different stopping rules. These kinds of tests are called *sequential.* Sequential tests have long been part of the error-statistician's tool box. One reason they are so useful is that often it is estimated that a smaller number of samples is required with a sequential than with a fixed sample size test. Medical trials, especially, are often deliberately designed as sequential. Armitage, as it happens, is a leader in the devel-

[16] Feller (1940) is the first to show this explicitly. Other early discussions of this result include Anscombe 1954 and Robbins 1952. The result is also implicit in Good 1956 and Lindley 1957.

opment of sequential trials, having devoted whole books to their use and interpretation within the error statistical framework.

Example 10.1: Armitage: The Effect of Repeated Significance Tests

In his *Sequential Medical Trials,* Armitage (1975) discusses the effect of repeated significance tests. He illustrates with a common example from medicine. Suppose that each patient scores in some numerical fashion the effectiveness of two different treatments, say two types of painkillers A and B. Drug A is administered one week, drug B on a different week. The recorded observation on each patient is the difference between the two scores. Imagine that a new significance test is performed after each patient's scores are obtained, with a view toward finding a difference (in either direction) computed to be statistically significant at the .05 level. The null hypothesis assumes that drugs A and B are equally effective. By the time 30 patients are sampled, the probability of computing a statistically significant difference even though the null hypothesis is true is around .3—not .05. So with 30 patients, the actual probability of rejecting the null hypothesis erroneously is not .05 but around .3. We would say that the *calculated* (or fixed sample size) significance level is .05, but that the actual or *overall* level is .3. Armitage gives the overall significance level for each number of patients (based on the standard test of the difference between means called the *t*-test), as shown in table 10.1.

Before we rule out the null (or "mere chance") hypothesis and argue that the result is indicative of a genuine difference, we want to be able to sustain a reliable argument from error. We want to be able to say that our procedure would probably have ruled in favor of the "mere chance" explanation, were that the case. But the procedure of trying and trying again cannot be said to have a good chance of ruling in favor of the null hypothesis—even if the null is true. With enough significance tests, the try and try again procedure will almost never pass the null hypothesis even if it is true. In their useful booklet on statistics for doctors, Bjorn Andersen and Per Holm (1984) provide a humorous analogy for this unfairness toward the null hypothesis:

> The procedure might be compared with new rules for determination of the world championship in heavy-weight boxing: Only the reigning champion is allowed to strike. The fight is over, whenever the contender is out for the count of 10. The contender (like H_0) has little chance of winning, no matter how "good" he is. (P. 57)

A Funny Thing Happened at the 1959 Savage Forum

Now Savage knows all about the effect of optional stopping—he knows all about how the try and try again method ensures reaching

TABLE 10.1 The Effect of Repeated Significance Tests (the "Try and Try Again" Method) (Armitage 1975, p. 29)

Number of patients (differences in scores) *n*	Probability of a 0.05 "significant" result at or before this stage, given the null hypothesis is true
2	0.05
10	0.09
20	0.26
30	0.29
40	0.31
50	0.33
100	0.39
infinity	1.00

statistical significance. In his opening remarks at the 1959 forum, Savage rehearses how "the persistent experimenter can arrive at data that nominally reject any null hypothesis at any significance level, when the null hypothesis is in fact true" (Savage 1962, 18). Because the persistent experimenter is thereby assured of rejecting a perfectly true null hypothesis, the standard error statistician denies that such a rejection provides genuine evidence against the null. But Savage audaciously declares that the lesson to draw from the optional stopping effect is just *the reverse* of the one the error statistician draws. The problem is not with the data arrived at by a procedure of trying and trying again, the problem is with significance levels!

> These truths [about the optional stopping effect] are usually misinterpreted to suggest that the data of such a persistent experimenter are worthless or at least need special interpretation. . . . The likelihood principle, however, affirms that the experimenter's intention to persist does not change the import of his experience. (Savage 1962, 18)

I shall come to the business of the relevance of "intentions" in a moment. Savage's argument is this: if calculating the significance level is altered (by the stopping rule), then there must be something wrong with significance levels because likelihoods are unaffected. According to the LP, says Savage, "optional stopping is no sin," so the problem must lie with the use of significance levels (1964, 185). But why should we accept the likelihood principle?

Yes, the LP follows from Bayes's theorem, but significance tests are non-Bayesian techniques. Apparently, the LP is regarded by some as so intrinsically plausible that it seems any sensible account of inference should obey it. Bayesians do not seem to think any argument is necessary for this principle, and rest content with echoing Savage's declaration in 1959: "I can scarcely believe that some people resist an idea so patently right" (1962, 76). However much Savage deserves reverence, that is still no argument. Ironically, what prompted Savage's famous declaration as to the patent rightness of the LP was a heretical confession by George Barnard. Barnard—the statistician whose arguments Savage claims (p. 76) convinced *him* (in 1952) of the *irrelevance* of optional stopping—had just announced to the forum that he had changed his mind!

Explaining why he now thinks that stopping rules do matter, Barnard describes an example quite like the one with which I began this section:[17]

> Suppose somebody sets out to demonstrate the existence of extrasensory perception and says "I am going to go on until I get a one in ten thousand significance level." Knowing that this is what he is setting out to do would lead you to adopt a different test criterion. What you would look at would not be the ratio of successes obtained, but how long it took him to obtain it. And you would have a very simple test of significance which said if it took you so long to achieve this increase in the score above the chance fraction, this is not at all strong evidence for E.S.P., it is very weak evidence. (Barnard 1962, 75)

By altering the test criteria accordingly, Barnard continues, one would avoid misinterpreting the evidence.[18] That is just what error statisticians recommend—thereby making them incoherent from the Bayesian standpoint.

The Argument from Intentions

Startled by this turnabout, Savage reminds Barnard of the persuasive argument he himself urged (in 1952) against the relevance of stopping rules.

> The argument then was this: The design of a sequential experiment is, in the last analysis, what the experimenter actually intended to do. His intention is locked up inside his head. (Savage 1962, 76)

The experimenter's intentions about when to stop sampling are locked up in his head, and it seems absurd for intentions to influence what

[17] The suggestion Barnard made to the forum was that stopping rules matter when you do not have explicit alternatives. He himself was a likelihoodist and not a Bayesian, although he came to give that up as well. (See, for example, Barnard 1972.)

[18] In practice, the alteration is generally to lower the computed or nominal significance level sufficiently so that the overall significance level is still .05. Armitage and others have done extensive work on this for a variety of types of sequential trials.

the data have to say. Since significance levels take stopping rules into account, significance levels let experimenter's intentions count. In their joint paper, Edwards, Lindman, and Savage remark:

> The irrelevance of stopping rules is one respect in which Bayesian procedures are more objective than classical ones. Classical procedures . . . insist that the intentions of the experimenter are crucial to the interpretation of data. (Edwards, Lindman, and Savage 1963, 239)

Although Savage (1962, 76) declared himself uncomfortable with the argument from intentions, it is repeated again and again by followers of Savage. Howson and Urbach think it substantiates some rather dire conclusions about significance tests:

> A significance test inference, therefore, depends not only on the outcome that a trial produced, but also on the outcomes that it could have produced but did not. And the latter are determined by certain private intentions of the experimenter, embodying his stopping rule. It seems to us that this fact precludes a significance test delivering any kind of judgment about empirical support. . . . For scientists would not normally regard such personal intentions as proper influences on the support which data give to a hypothesis. (Howson and Urbach 1989, 171)

In their view, apparently, to take account of the experimenter's sampling plan is to take personal intentions into account and is unscientific, while the properly scientific way of assessing evidential support is to ask for the agent's personal degrees of belief in hypotheses.

In fact, the whole insinuation that to regard optional stopping as relevant is to make private intentions relevant is fallacious. Any and all aspects of what goes into specifying an experiment could be said to reflect intentions—sample size, space of hypotheses, prediction to test, and so on—but it does not mean that paying attention to those specifications is tantamount to paying attention to the experimenter's intentions. Yet Howson and Urbach are pretty plainly arguing that since a significance test's error probabilities are determined by the experimenter's personal intentions and since intentions should not matter to support, a test's error probabilities (e.g., significance levels) do not or should not be relevant to support. Are Bayesians just committing a gross fallacy here?

They are, but they cannot see it. They have got Bayesian glasses on, and they will not take them off. Through Bayesian glasses, there is no place in the inference scheme to record the effect of the sampling plan—at least not once the data are in hand. So, they view it as locked inside someone's head.

Recall hunting for a statistically significant difference in the infant training example again (chapter 9). Suppose the hunter reports the single factor found to be statistically significant out of 20 that are checked (e.g., late weaning and left-handedness). We have this one statistically significant result before us, but where, one might ask, is the fact that it was the single factor found significant in a hunting expedition through 20 factors? Is it locked up in the experimenter's head? Not if he or she is an honest hunter, nor if the one scrutinizing the result is an error statistician. But that means having an eye for error probabilities—being able to see, in particular, that the actual probability of erroneously declaring statistical significance in this case is not .05 but over .6. The Bayesian glasses have a substantial blind spot here.

Likewise with optional stopping. If one is wearing Bayesian glasses, that is, if one adheres to the LP, then two experiments that give the same (i.e., proportional) likelihoods to a hypothesis have the same evidential bearing on the hypothesis. If one is wearing Bayesian glasses, then, once the data are available, one cannot make out any difference between that data having arisen from a try and try again method or from a (nonsequential) experiment where the subject declares ahead of time, "If I have not shown statistical significance in exactly *n* trials then conclude I have not shown the effect." One cannot see the difference because the likelihoods are unchanged. One may well know there is a difference in sampling plans, but that just means one knows they had different intentions, and that cannot possibly make a difference to the meaning of evidence. That, at any rate, is the way things look through Bayesian glasses. That is the way things look to anyone peering at evidence through the LP. Ian Hacking, in his Likelihood testing period, also gives the argument from intentions:[19]

> Can testing depend on hidden intentions? Surely not; hence optional stopping should not matter after all. (Hacking 1965, 109)

Examples of philosophers espousing the argument from intentions could easily be multiplied.

Notice a certain similarity with justifying why novelty should matter. If a violation of novelty is nothing more than that the experimenter intended to find a way to account for the data, then it looks as if propo-

[19] That account, developed in Hacking 1965, was based on Hacking's likelihood rule of support noted in section 6.6.

nents of novelty appeal to the psychological intentions of the investigator. Once the aim of novelty is recognized to be severity, violating novelty shows up as a problem (when it *is* one) with a test's severity; and the effect on severity, whether formally or informally calculated, shows up in a procedure's error probabilities. In precisely the same way, the error statistician has a perfectly nonpsychologistic way of taking account of the impact of stopping rules, as well as other aspects of experimental plans. The impact is on the error probabilities (operating characteristics) of a procedure.[20]

In the optional stopping plan, the difference in the test procedure clearly shows up in the difference in the set of possible experimental outcomes. Certain outcomes possible in the fixed sample size (nonsequential) version of the test are no longer possible.[21] If the stopping rule is open-ended, then the possible outcomes do not contain any that fail to reject the null hypothesis!

It might be asked: But does the difference in error probabilities corresponding to a difference in sampling plans correspond to any real difference in the experiments? Absolutely. The researchers really did something different in the try and try again scheme and, quoting Armitage, "thou shalt be misled" if you do not know this. It is not just that incorrectly reporting a test's error probabilities incorrectly reports what happened in obtaining a result, it also incorrectly reports *what should be expected to happen* (with various probabilities) in subsequent experiments on the phenomenon of interest. It must be remembered that every error statistical inference includes a statement about future experiments, whether or not they will be carried out. With an incorrect report of a test's error probabilities, an experimenter seeking to check or repeat the previous results would be misled. The reported error probabilities would not be close to those that would actually be found in such repetitions.

I think enough has been said to banish the common allegation that letting stopping plans matter is tantamount to letting intentions matter. As such, we can reject the argument that the LP must be embraced on pain of unjustly letting intentions enter into the appraisal of evidence.

[20] The only time it seems unwarranted to draw a distinction is if the experimenter stops after the first test because a statistically significant result is achieved on the first try. But in that case the difference between the computed and the overall significance level is extremely small, and should make no difference in a "nonautomatic" use of tests. I discuss what is wrong with automatic or recipelike uses of tests in chapter 11.

[21] The sample space differs but because the likelihoods are proportional, the difference cancels out for a holder of the LP.

But, if threats will not win us over, the Bayesian tempts with the goodies that await those who accept the LP.

Bayesian Freedom, Bayesian Magic

A big selling point for adopting the LP, and with it the irrelevance of stopping rules, is that it frees us to do things that are sinful and forbidden to an error statistician.

> This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson). . . . Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money or patience. . . . Classical statisticians . . . have frowned on [this]. (Edwards, Lindman, and Savage 1963, 239)

Breaking loose from the grip imposed by error probabilistic requirements returns to us an appealing freedom.

LeCam, a leading error statistician (cited at the start of section 10.3) hits the nail on the head:

> It is characteristic of [Bayesian approaches] . . . that they . . . tend to treat experiments and fortuitous observations alike. In fact, the main reason for their periodic return to fashion seems to be that they claim to hold the magic which permits [us] to draw conclusions from whatever data and whatever features one happens to notice. (LeCam 1977, 145)

In contrast, the error probability assurances go out the window if you are allowed to change the experiment as you go along. Repeated tests of significance (or sequential trials) are permitted, are even desirable for the error statistician; but a penalty must be paid for perseverance—for optional stopping. Before-trial planning stipulates how to select a small enough significance level to be on the lookout for at each trial[22] so that the overall significance level is still low. That is what Armitage's work on sequential clinical trials is all about.

But the Bayesian pays no penalty, or so it seems. I. J. Good, a veteran Bayesian, often puts it this way:

> Given the likelihood, the inferences that can be drawn from the observations would, for example, be unaffected if the statistician arbi-

[22] This is the level that would be required to be reached on any given significance test so as to stop the trials. Setting it small enough ensures that the probability of an erroneous rejection of the null is still small in sequential trials.

> trarily and falsely claimed that he had a train to catch, although he really had decided to stop sampling because his favorite hypothesis was ahead of the game. . . . On the other hand, the "Fisherian" tail-area method for significance testing violates the likelihood principle because the statistician who is prepared to pretend he has a train to catch (optional stopping of sampling) can reach arbitrarily high significance levels, given enough time, even when the null hypothesis is true. (Good 1983, 36)

"Arbitrarily high significance levels" means significance levels as *small* as one wants. Elsewhere in Good's *Good Thinking:*

> The way I usually express this "paradox" is that a Fisherian [but not a Bayesian] can cheat by pretending he has a train to catch like a gambler who leaves the table when he is ahead. (Good 1983, 135)

As often as my distinguished colleague presents this point, I remain baffled as to its lesson about who is allowed to cheat. The significance tester—as Good well knows—does not allow reaching arbitrarily high (meaning small) significance levels through optional stopping. The significance tester is not allowed to change the sample size at will, stopping just because he is ahead. When error statisticians perform sequential tests, the *overall* (and not the computed) significance level must be reported. To the error statistician, what would be cheating would be to report the significance level you persevered to attain, say .05, *just as if* the test were the ordinary nonsequential sort.

Good's point seems to be this: Error statisticians are forced to fret about a consideration the Bayesian is free to ignore. Wearing our error probability glasses—glasses that compel us to see how certain procedures alter error probability characteristics of tests—we are forced to say, with Armitage, that "Thou shalt be misled if thou dost not know that" the data resulted from the try and try again stopping rule. To avoid having a high probability of following false leads, the error statistician must scrupulously follow a specified experimental plan. But that is because we hold that error probabilities of the procedure alter what the data are saying—whereas Bayesians do not. The Bayesian is permitted the luxury of optional stopping and has nothing to worry about. The Bayesians hold the magic.

Or is it voodoo statistics?

Armitage's Example

To some, the magic is accomplished by smoke and mirrors and wearing Bayesian glasses. At the 1959 forum, Armitage, building on his earlier remarks, went on to say that

> [Savage] remarked that, using conventional significance tests, if you go on long enough you can be sure of achieving any level of significance; does not the same sort of result happen with Bayesian methods? The departure of the mean by two standard errors corresponds to the ordinary five per cent level. It also corresponds to the null hypothesis being at the five per cent point of the posterior distribution. Does it not follow that by going on sufficiently long one can be sure of getting the null value arbitrarily far into the tail of the posterior distribution? (Armitage 1962, 72)

Armitage's point can be simply put as follows. In many cases, rejecting a null hypothesis H_0, say at level of significance .05, corresponds to a result that would lead a Bayesian to assign a low (e.g., .05) posterior probability to H_0. This occurs with so-called uniform or uninformative priors. (That this is so is often touted by Bayesians as a point in their favor: Whereas the most an error statistician can say is that this procedure has a low [.05] probability of erroneously rejecting the null, the Bayesian, thanks to his prior probability assignment, can assign the low probability to the specific hypothesis H_0.)[23] Hence, Armitage reasons, if error statisticians—if they go on sampling long enough—are assured of reaching a .05 significant result, even though H_0 is true, then Bayesians—if they go on sampling long enough—would be assured of reaching a low (.05) posterior probability in H_0, even though H_0 is true. (The assurance here is with high probability or, in the limit, with probability one.) That is:

1. In certain cases, rejecting a null hypothesis H_0, say at level of significance .05, corresponds to a result that would lead a Bayesian to assign a low (e.g., .05) posterior probability to H_0.
2. If one is allowed to go on sampling long enough (i.e., the try and try again procedure), then, even if H_0 is true, one is assured of achieving a .05 statistically significant difference from the null hypothesis H_0.
3. Therefore, if one is allowed to go on sampling long enough, then, in the cases described in (1), one is assured of reaching a low posterior probability in H_0, even though H_0 is true.

Now the error statistician is not allowed to go on trying and trying, at least not without paying a penalty. The penalty, we said, is that the overall significance level—in the extreme case 1—must be reported. The stopping rule matters. But Bayesians are free! They are allowed to go on sampling and the stopping rule does not alter the likelihoods,

[23] See, for example, DeGroot 1973.

hence the posterior is just the same as if the case were nonsequential. It follows that, in going on long enough, a Bayesian is assured of assigning a low posterior probability to H_0 even though H_0 is true.

So, who is allowed to mislead?

Although Savage wants to deny Armitage's implication, he appears to grant it, though fuzzily, and skips to a different sort of example. While I think there are problems with this different example as well (see Rosenkrantz 1977, 199), I want to keep to Armitage's kind of example.[24]

Armitage's example goes like this. The null hypothesis H_0 is an assertion about a population parameter μ. As in example 10.1, μ might measure the mean difference in the effectiveness of two drug treatments. H_0 asserts that the treatments are equally effective, that is, that μ equals 0.

H_0: The treatments are equally effective: μ equals 0.

The experiment records $\overline{X}$ which, in this case, is the mean difference in scores accorded to the two drug treatments in a sample of n patients. The sample size n, however, is not fixed but is determined by a stopping rule. The stopping rule—an example of a try and try again plan—is to keep taking more samples until H_0 is rejected at the .05 level, *by the usual prespecified significance test:*

(1) *Stopping rule:* Keep sampling until H_0 can be rejected at the .05 level.

That is, the stopping rule is to keep sampling until $\overline{X}$ is 2 standard deviations away from 0 (the hypothesized value of μ in H_0) in either direction. The standard deviation here is the standard deviation of the statistic $\overline{X}$, but for simplicity, let us just abbreviate it as s.d.[25] So we have (letting $|\overline{X}|$ be the absolute value of $\overline{X}$)

(1) *Stopping rule:* Keep sampling until $|\overline{X}| \geq$ *2 s.d.*

Following this stopping rule, one is assured of achieving a .05 significant difference even if H_0 is true. But with a so-called uninformative

[24] The example Savage skips to involves comparing two simple hypotheses. Rather than lending plausibility to Savage's cause, Rosenkrantz (while himself a Bayesian) thinks it shows that Savage goes too far in ignoring the stopping rule. Rosenkrantz's analysis has been questioned by Seidenfeld (1979b, n. 4).

[25] This would more properly be written as s.d.($\overline{x}$). When the standard deviation is estimated, as is most often the case, it is called the standard error, but it seems simpler to stick with a single term. Armitage takes the random variable X to be Normally distributed with mean μ and standard deviation 1. In that case, s.d.($\overline{x}$) is $n^{-1/2}$.

or diffuse prior probability assignment to μ, such an occurrence would correspond to assigning a low posterior probability to H_0. Hence, following this stopping rule, the Bayesian would be assured of assigning a low probability to H_0 even though H_0 is true. This is Armitage's argument. No satisfactory answer has been forthcoming, nor is there one. Armitage is right.

Berger and Wolpert

To my knowledge, there are only a handful of Bayesians (or other holders of the LP) who have specifically addressed Armitage's example: most just accept Savage's dismissal of it.[26] Berger and Wolpert (1988), in their interesting monograph, show themselves to be as ardent a pair of proponents of the LP as it is likely to have. Still, even they concede Armitage's point. But, as they want to retain the LP, some defensive moves are called for.[27]

Since Berger and Wolpert treat Armitage's example in terms of confidence intervals, I will too. Recall example 8.2. Given some result, one forms an interval within which the parameter of interest, μ, is hypothesized to lie. As is standard, we can use a lowercase $\bar{x}$ to represent the observed value of the random variable $\bar{X}$. (This is easier to read than $\bar{X}_{obs}$ here.) The standard 95 percent confidence interval takes this form

(2) Estimate that μ is within 2 standard deviations of the observed mean $\bar{x}$.

(I use 2 rather than the exact value of 1.96.) That is, the 95 percent confidence interval is

(2) Estimate that μ equals $\bar{x} \pm 2$ s.d.

Berger and Wolpert agree that a Bayesian would use this interval in the usual fixed sample size case, adding:

> Of course, he would not interpret confidence in the frequency sense, but instead would (probably) use a posterior Bayesian viewpoint with the noninformative prior density. (Berger and Wolpert 1988, 80)

[26] No one, to my knowledge, has identified the flaw in Savage's use of "the simple general formula" on page 73 of Savage 1962.

[27] In a forthcoming paper, "Reasoning to a Foregone Conclusion," Kadane, Schervish, and Seidenfeld set out mathematical conditions under which Bayesians are and are not allowed to reason "to a foregone conclusion" erroneously.

Whereas the frequentist says only that this particular estimate was generated by a procedure with a 95 percent probability of correctly including the value of μ, the Bayesian can assign the particular estimate a posterior probability.[28] In particular, with the so-called noninformative prior, they would assign it a .95 posterior probability. We can now see how Armitage's argument goes through.

Berger and Wolpert continue: "Suppose now that the experimenter has an interest in seeing that $\mu = 0$ is not in the confidence interval. He could then use the stopping rule" (1) above (ibid.; I replace θ with μ). Let us rewrite the stopping rule to relate directly to confidence intervals. The null hypothesis H_0 asserts that $\mu = 0$. Rejecting H_0—finding $\bar{x}$ statistically significant from 0—is equivalent to 0 *not* being included in the interval estimate formed with $\bar{x}$. Hence assuring that $\mu = 0$ is *not* in the 95 percent confidence interval is equivalent to assuring that the null hypothesis H_0 is rejected at the .05 level. So the stopping rule in (1)—keep sampling until H_0 is rejected—stated in terms of confidence intervals is

> (1) Keep sampling until the 95 percent confidence interval formed excludes 0.

So the Bayesian experimenter interested in keeping 0 out of the interval is free to use stopping rule (1). At the same time, Berger and Wolpert concede, "the [Bayesian] conditionalist, being bound to ignore the stopping rule, will still use (2) as his confidence interval, but this can *never* contain zero" (ibid., 81).

(The term "conditionalist" comes from the fact that, for a holder of the LP, inference must be conditional on the actual $\bar{x}$ observed.)

Hence Berger and Wolpert allow that "the frequentist probability" that intervals formed by this procedure would include 0, even when 0 is the true value, equals zero! Equivalently, there is zero probability of accepting the hypothesis that $\mu = 0$, even when that hypothesis is true. In short, they find they cannot get around the conclusion that, despite the fact that μ *does* equal 0,

> the experimenter has thus succeeded in getting the conditionalist to perceive that $\mu \neq 0$, and has done so honestly. (Pp. 80–81)

Thus, they concede Armitage's point—the very point that Savage had denied or skirted.

Now for the defensive moves. Berger and Wolpert are at pains to

[28] It leads to a posterior distribution for μ equal to a normal distribution with mean $\overline{X}$ and a standard deviation equal to $n^{-1/2}$.

uphold the LP. In examples such as Armitage's, Berger and Wolpert maintain, the LP only "seems to allow the experimenter to mislead a [Bayesian] conditionalist. The 'misleading,' however, is solely from a frequentist viewpoint, and will not be of concern to a conditionalist" (ibid., 81). Bayesians remain unconcerned, presumably, because they are not in the business of calculating error frequencies.

Despite their professed lack of concern, Berger and Wolpert, like Savage, are plainly uncomfortable with Armitage's result. They leave off the example suggesting that in appraising the plausibility of the LP we should trust our intuitions in one of the other examples they offer—one where the LP gives the intuitively correct inference—"rather than in extremely complex situations such as [Armitage's example]" (ibid., 83). But the example we are to trust does not involve optional stopping,[29] and the confidence interval example is rather ordinary. Armitage tells us it is a standard situation in clinical trials.[30]

Examples analogous to Armitage's have been produced by others, notably Alan Birnbaum.[31] To the error statistician, such examples are counterexamples to adopting the LP:

> Thus it seems that the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations. (Birnbaum 1969, 128)

I shall come back to Birnbaum in chapter 11.

It should be emphasized that this problem exists even for so-called objective Bayesians (those who strive to determine objective prior probabilities). That is the reason I said it was the outgrowth of a differ-

[29] Perhaps the open-endedness of the stopping rule makes the case exotic, but less dramatic and still seriously troubling cases are generated with stopping rules only as high as 100, as Armitage shows in his examples with medical trials.

[30] One Bayesian ploy would be to insist that learning of the use of such a stopping rule would make the agent change his prior in such a way that the high posterior would be avoided. Not only is the kind of prior that leads to the trouble a commonly acceptable one, but such an admission would also conflict with the Bayesian insistence that once the evidence is at hand the likelihoods tell all. (As always, we are talking about the cases in which stopping rules are uninformative according to the LP.) Ironically, since error probabilities are not supposed to matter for a Bayesian, this ploy really would seem to appeal to the intentions of the investigator. Moreover, this Bayesian ploy depends on the agent reasoning that an experimenter using such a stopping rule probably thinks the null hypothesis is true, and so revising his prior accordingly. But it seems at least as plausible, if not more so, to suppose that an experimenter planning to go on until the null hypothesis is rejected really believes that the effect is real and that the null hypothesis is false.

[31] Birnbaum cites a similar result by Neyman from 1938, collected in Neyman 1952.

ence between Bayesians and error statisticians that runs even deeper than the use or nonuse of prior probabilities in hypotheses. It is the difference between the irrelevance and the relevance of error probabilities of procedures. If one is Bayesian enough to adhere to Bayesian coherency, hence to the LP, one is enough of a conditionalist to reject error probabilities and with them the familiar methods of standard error statistics.

That the standard methods conflict with the LP is readily accepted by leading Bayesians. Take Lindley:

> The most obvious violation of the likelihood principle occurs with the idea of a confidence interval, with its concept of repetition of the experiment. (Lindley 1976, 361)

Savage, discussing the

> "nice properties," exemplified by unbiasedness, stringency, minimum mean squared error, symmetry (or invariance), a given significance level, and so on (Savage 1964, 179)

declares that

> practically none of the "nice properties" respect the likelihood principle. (Ibid., 184)

That is why I think Berger and Wolpert's initial response to Armitage is the honest one from the Bayesian viewpoint; namely, that "the 'misleading' . . . is solely from a frequentist viewpoint." After all, it is only through frequentist considerations of error probabilities that Armitage's case is problematic, and those considerations violate the LP to which Bayesians adhere.

There, then, we have it. The reason Bayesians cannot be misled (in the case of optional stopping) is that they reject (as violating the LP) the frequentist viewpoint on which the error calculation depends! Anguish over a procedure's high probability of being wrong (in Armitage's example, as high as probability 1) is an error statistician's affliction. The Bayesian is not so afflicted. If I never check my bank account (and I always believe the correctness of my statement), then, in a sense, the bank can never mislead me.

The Relevance of Outcomes Other Than the One Observed

Let us explore a bit more why error probabilities violate the LP. The reason, in a nutshell, is that error probabilities ask what would happen for data sets other than the one actually observed. What is wrong with us error statisticians, from the Bayesian conditionalist perspective, is that we keep thinking that considerations of outcomes that could have resulted—outcomes other than the one that did result—are relevant for interpreting the evidential import of the data.

> Those who do not accept the likelihood principle believe that the probabilities of sequences that might have occurred, but did not, somehow affect the import of the sequence that did occur. (Edwards, Lindman, and Savage 1963, 238)

The error statistician has only one way of responding to this allegation. "Guilty as charged!" We remain steadfast no matter how leadingly Bayesians ask (echoing a line made famous by Harold Jeffreys), "What has what might have happened, but did not, got to do with inferences from the experiment?" (Lindley 1976, 361), and no matter how intimidating the rhetoric of prominent Bayesians is (e.g., E. T. Jaynes, an objective Bayesian):

> The question of how often a given situation would arise is utterly irrelevant to the question how we should reason when it *does* arise. I don't know how many times this simple fact will have to be pointed out before statisticians of "frequentist" persuasions will take note of it. (Jaynes 1976, 247)

What we error statisticians must rightly wonder is how many times we will have to point out that to us, reasoning from the result that did arise *is* crucially dependent upon how often it would arise. Lacking such information prevents us from ascertaining which inferences can be reliably drawn.

In criticizing the hunter, the error statistician notes, "But had this one not been statistically significant, it is very probable that you would have unearthed some other factor that was—even if none are really correlated." What would have happened is at the heart of the worry in the try and try again (optional-stopping) plan as well. The severity of a test is a measure of the relative frequency with which the test would lead to correctly failing (or not passing) a hypothesis in some sequence of applications. Virtually all the uses of statistical ideas in learning from error throughout this book depend critically on such considerations of "would have beens." What makes standard error statistical tools so useful for scientific inference is that their formal properties, error probabilities, enable learning about what would be expected if various errors exist—the key to experimental arguments from error. Yet these error properties and test criteria based on them are what the Bayesian is only too happy to declare irrelevant. As Lindley (1971) stresses,

> unbiased estimates . . . sampling distributions, significance levels, power, all depend on something more [than the likelihood function]—something that is irrelevant in Bayesian inference—namely the sample space. (Lindley 1971, 436)

In his admirably fair-minded work comparing different schools of inference, Vic Barnett explains why. In the Bayesian view,

> inferences are conditional on the realized value *x;* other values which *may* have occurred are regarded as irrelevant. . . . No consideration of the *sampling distribution* of a statistic is entertained; sample space averaging is ruled out. Thus there can be no consideration of the *bias* of an estimation procedure and this concept is totally disregarded. (Barnett 1982, 226)

Bayesian consistency requires rejecting the foundations of the error statistical methods, despite the widespread use of these methods throughout science.

This, then, is the bottom line. Our aims, our notions of relevant evidence, our criteria for judging satisfactory inference (the "nice properties"), our notions of probability (with a few exceptions) are strikingly different from those of the Bayesians. Quite aside from whether one accepts my position on the value of error statistical ideas, what cannot be denied are these differences. The lesson for metamethodology is this: Every critique of methodology from the Bayesian perspective must be seen as contingent upon accepting their aims in favor of error statistical ones. If the methodological rule in question turns out to concern promoting an error statistical aim (e.g., severity), then a Bayesian critique will be misleading if not just question-begging. The error statistician's conception of "being misled" is very different from that of the Bayesian: perhaps it is a gestalt switch that separates them. The philosopher seeking to apply ideas from statistics to the philosophy of science needs to decide whether to sign up for the LP (e.g., Bayesian) paradigm or the error statistical one, or perhaps something altogether different.

10.4 Some Anticipated Objections

Some might object that I am overlooking the ways in which some manage to be Bayesian while at least a little bit of an error statistician at the same time. A main way would be to use Bayesian ways and yet strive to assess the reliability of these methods in a genuine error probability sense. Such error statistical (or "robust") Bayesians, if I understand their position, seem to me to fall onto the error statistical side.[32] Nothing in the error probability approach prevents using Bayesian measures as measures of "fit" whose operating properties can be investigated. Such developments may well be part of "the historical process of development" of error statistical theory to which E. S. Pearson alludes (Pearson 1966d, 276).

Aside from such new and innovative hybrid approaches, am I not overlooking the eclecticism that exists in statistical practice? No. I began this chapter acknowledging that error statistical inferences often correspond to procedures Bayesians would countenance, albeit with differences in interpretation and justification. It is certainly open to error statisticians to apply Bayes's theorem with well-defined statistical hypotheses where standard prior probabilities have been found to work well. Likewise, Bayesians can and do appropriate standard (error-statistical) methods by giving them Bayesian justifications. One might regard I. J. Good's "Bayes-non-Bayes compromise" as a systematic attempt to appropriate error statistical methods in this way.[33] In practice, dabbling in one or the other of these methodologies even without being too clear on the justification is often innocuous. This is not the case when Bayesian principles are applied to philosophy of science.

I earlier outlined three main ways of applying a theory of statistics to philosophy of science: (1) to solve philosophical problems (e.g., Duhem's problem); (2) to model scientific inference; and (3) to carry out a metamethodological critique (e.g., appraise novelty requirements). For each of these applications, the differences of interpretation and justification called for by the Bayesian and error-statistical philosophies are serious and are ignored at our peril.

[32] This is not the case for Bayesians who are only willing to employ error-statistical methods if they can be given a subjective Bayesian interpretation, or who employ error probabilities disingenuously (e.g., because the customer wants or expects them).

[33] Good's compromise, as I understand it, remains fully Bayesian (see note 32). However, his program has brought forth a number of important relationships between error probability and Bayesian calculations.