D. G. MAYO AND M. KRUSE

PRINCIPLES OF INFERENCE AND THEIR CONSEQUENCES

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 ... [Edwards *et al.*, 1963, p. 193].

1 INTRODUCTION

What do data tell us about hypotheses or claims? When do data provide good evidence for or a good test of a hypothesis? These are key questions for a philosophical account of evidence and inference, and in answering them, philosophers of science have often appealed to formal accounts of probabilistic and statistical inference. In so doing, it is obvious that the answer will depend on the principles of inference embodied in one or another statistical account. If inference is by way of Bayes' theorem, then two data sets license different inferences only by registering differently in the Bayesian algorithm. If inference is by way of error statistical methods (e.g., Neyman and Pearson methods), as are commonly used in applications of statistics in science, then two data sets license different inferences or hypotheses if they register differences in the error probabilistic properties of the methods.

The principles embodied in Bayesian as opposed to error statistical methods lead to conflicting appraisals of the evidential import of data, and it is this conflict that is the pivot point around which the main disputes in the philosophy of statistics revolve. The differences between the consequences of these conflicting principles, we propose, are sufficiently serious as to justify supposing that one "cannot be just a little bit Bayesian" [Mayo, 1996], at least when it comes to a philosophical account of inference, but rather must choose between fundamentally incompatible packages of evidence, inference, and testing. In the remainder of this section we will sketch the set of issues that seems to us to serve most powerfully to bring out this incompatibility.

EXAMPLE 1 (ESP Cards). The conflict shows up most clearly with respect to the features of the *data generation process* that are regarded as relevant for assessing evidence. To jump right into the crux of the matter, we can consider a familiar type of example: To test a subject's ability, say, to predict draws from a deck of five ESP cards, he must demonstrate a success rate that would be very improbable if he were merely guessing. Supposing that after a long series of trials, our subject

D. Corfield and J. Williamson (eds.), Foundations of Bayesianism, 381–403.

attains a "statistically significant" result, the question arises: Would it be relevant to your evaluation of the evidence if you learned that he had planned all along to keep running trials *until* reaching such an improbable result? Would you find it relevant to learn that, having failed to score a sufficiently high success rate after 10 trials, he went on to 20 trials, and on and on until finally, say on trial number 1,030, he attained a result that would apparently occur only 5% of the time by chance?

A plan for when to stop an experiment is called a *stopping rule*. So our question is whether you would find knowledge of the subject's stopping rule relevant in assessing the evidence for his ESP ability. If your answer is *yes*, then you are in sync with principles from standard error statistics (e.g., significance testing and confidence interval estimation). Interestingly enough, however, this intuition conflicts with the principles of inference espoused by other popular philosophies of inference, i.e., the Bayesian and Likelihoodist accounts. In particular, it conflicts with the *likelihood principle* (LP). According to the LP, the fact that our subject planned to persist until he got the desired success rate, the fact that he *tried and tried again*, can make *no* difference to the evidential import of the data: the data should be interpreted in just the same way as if he had decided from the start that the experiment would consist of exactly 1,030 trials.

This challenge to the widely held supposition that stopping rules alter the import of data was L. J. Savage's central message to a forum of statisticians in 1959:

> 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....
>
> These truths 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, p. 18].

Savage rightly took this conflict as having very serious consequences for the foundations of statistics:

> In view of the likelihood principle, all of [the] classical statistical ideas come under new scrutiny, and must, I believe, be abandoned or seriously modified [Savage, 1962, pp. 17–18].

This conflict corresponds to two contrasting principles on what is required by an account of inference, "*evidential-relationship*" (E-R) principles and "*testing*".

2 EVIDENTIAL-RELATIONSHIP VS. (ERROR STATISTICAL) TESTING ACCOUNTS

In what we are calling E-R accounts, the evidential bearing of data on hypotheses is determined by a measure of support, probability, confirmation, credibility or the like to hypotheses given data. Testing approaches, in contrast, do not seek to assign measures of support or probability to hypotheses, but rather to specify methods by which data can be used to test hypotheses. Probabilistic considerations arise to characterize the probativeness, reliability, or severity of given tests, and specific inferences they license.

The difference between E-R and testing approaches is most dramatically revealed by the fact that two data sets $\boldsymbol{x}$ and $\boldsymbol{y}$ may have exactly the same evidential relationship to hypothesis H, on a given E-R measure, yet warrant very different inferences on testing accounts because $\boldsymbol{x}$ and $\boldsymbol{y}$ arose from tests with different characteristics. In particular, the two tests may differ in the frequency with which they would lead to erroneous inferences (e.g., passing a false or failing a true hypothesis). That is, the tests may have different *error probabilities*. We will refer to the testing philosophy as the *error-statistical approach*.

In *statistical* approaches to evidence, the main E-R measure is given by the probability conferred on $\boldsymbol{x}$ under the assumption that H is correct, $P(\boldsymbol{x}; H)$, i.e., the *likelihood* of H with respect to $\boldsymbol{x}$.[1] The LP, informally speaking, asserts that the evidential import of $\boldsymbol{x}$ on any two hypotheses, H and H', is given by the ratio of the likelihoods of H and H' with respect to $\boldsymbol{x}$.

To get a quick handle on the connection between the LP and stopping rules, suppose $\boldsymbol{x}$ arose from a procedure where it was decided in advance to take just n observations (i.e., n was *predesignated*), and $\boldsymbol{y}$ arose from our ESP subject's 'try and try again' procedure, which just happened to stop at trial n (*sequential* sampling). If, for every hypothesis H, $P(\boldsymbol{x}; H) = P(\boldsymbol{y}; H)$, then according to the LP it can make no difference to the inference which procedure was used. So, the fact that the subject stopped along the way to see if his success rate was sufficiently far from what is expected under chance makes no difference to "what the data are saying" about the hypotheses. This sentiment is quite clear in a seminal paper by Edwards, Lindman and Savage:

> 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 *et al.*, 1963, pp. 238–239].

[1] Note that $P(\boldsymbol{x}; H)$ is not a conditional probability usually written as $P(\boldsymbol{x}|H)$ because that would involve assigning prior probabilities to H — something outside the standard error statistical approach. The way to read $P(\boldsymbol{x}; H)$ is "The probability that X takes value $\boldsymbol{x}$ *according to* statistical hypothesis H." Any statistical hypothesis H must assign probabilities to the different experimental outcomes.

This is called the *irrelevance of the stopping rule* or the *Stopping Rule Principle* (SRP), and is an implication of the LP.[2]

To the holder of the LP, the intuition is that the stopping rule is irrelevant, and it is a virtue of the LP that it accords with this intuition. To the error statistician the situation is exactly the reverse. For her, the stopping rule is relevant because the persistent experimenter is more likely to find data in favor of H, even if H is false, than one who fixed the sample size in advance. Peter Armitage, in his comments to Savage at the 1959 forum, put it thus:

> 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 seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling [Armitage, 1962, p. 72], (emphasis added).

It is easy enough to dismiss long-run frequencies as irrelevant to interpreting given evidence, and thereby deny Armitage's concern, but we think that would miss the real epistemological rationale underlying Armitage's argument.[3] Granting that textbooks on "frequency methods" do not adequately supply the rationale, we propose to remedy this situation. Holders of the error statistical philosophy, as we see it, insist that data only provide genuine or reliable evidence for H if H survives a *severe test*. The severity of the test, as probe of H — e.g., the hypothesis that our ESP subject does better than chance — depends upon the test's ability to find that H is false when it is (i.e., when null hypothesis H_0 is true). H is *not* being put to a stringent test when a researcher allows trying and trying again until the data are far enough from H_0 to reject it in favor of H. This conception of tests provides the link between a test's error probabilities and what is required for a warranted inference based on the test. It lets us understand Armitage as saying that one would be misled if one could not take into account that two plans for generating data correspond to tests with different abilities to uncover errors of concern.

In the 40 years since this forum, the conflict between Bayesian and "classical" or error statistics has remained, and the problems it poses for evidence and inference are unresolved. Indeed, in the past decade, as Bayesian statistics has grown in acceptance among philosophers, the crux of this debate seems to have been largely forgotten. We think it needs to be revived.

---

[2]There are certain exceptions (the stopping rule may be "informative"), but Bayesians do not regard the examples we consider as falling under this qualification. See section 6.1.

[3]This dismissal is the basis of Howson and Urbach's response to Gillies' [1990] criticism of them.

3 THE LIKELIHOOD PRINCIPLE (LP)

The LP is typically stated with reference to two experiments considering the same set of statistical hypotheses H_i about a particular parameter, μ, such as the probability of success (on a Bernoulli trial) or the mean value of some characteristic.

> According to Bayes' theorem, $P(x;\mu)\ldots$ 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 $P(x;\mu)$ and $P(y;\mu)$ 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 $\mu\ldots$ I, and others, call this important principle the likelihood principle. The function $P(x;\mu)$ — rather this function together with all others that result from it by multiplication by a positive constant — is called the likelihood [Savage, 1962, p. 17]. (We substitute his $\Pr(x|\lambda)$ with $P(x;\mu)$.)

The *likelihood function* gives the probability (or density) of a *given* observed value of the sample under the different values of the unknown parameter(s) such as μ. More explicitly, writing the n-fold sample $(X_1, X_2, \ldots, X_n)$ as $\boldsymbol{X}$, the likelihood function is defined as the probability (or density) of $\{x = (x_1, x_2, \ldots, x_n)\}$ — arising from the *joint distribution* of the random variables making up the sample X — under the different values of the parameter(s) μ.

Even granting that two experiments may have different error probabilities over a series of applications, for a holder of the LP, once the data are in hand, only the actual likelihoods matter:

> The Likelihood Principle. In making inferences or decisions about μ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about μ if they are proportional to each other (as functions of μ) [Berger, 1985, p. 28].

That is, the LP asserts that:

> If two data sets x and y have likelihood functions which are (a) functions of the same parameter(s) μ and (b) proportional to each other, then x and y contain the same experimental information about μ.[4]

[4] We think this captures the generally agreed upon meaning of the LP although statements may be found that seem stronger. For example, Pratt, Raiffa, and Schlaifer characterize the LP in the following way:

> If, in a given situation, two random variables are observable, and if the value x of the first and the value y of the second give rise to the same likelihood function, then observing the value x of the first and observing the value y of the second are equivalent in the sense that *they should give the same inference, analysis, conclusion, decision, action, or anything else* ([Pratt *et al.*, 1995, p. 542]; emphasis added).

4 STATISTICAL SIGNIFICANCE LEVELS: TESTING PARAMETERS OF BERNOULLI TRIALS

The error statistical approach is *not* consistent with the LP because the error statistical calculations upon which its inferences are based depend on more than the likelihood function. This can be seen by considering Neyman–Pearson statistical significance testing.

Significance testing requires identifying a statistical hypothesis H_0 that will constitute the *test* or *null hypothesis* and an alternative set of hypotheses reflecting the discrepancy from H_0 being probed. A canonical example is where $X := (X_1, X_2, \ldots, X_n)$ is a random sample from the Bernoulli distribution with parameter μ, the probability of success at each trial. In a familiar "coin tossing" situation, we test $H_0 : \mu = 0.5$, (the coin is "fair") against the claim that $J : \mu > 0.5$.

Once a null hypothesis is selected, we define a test statistic, i.e., a characteristic of the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ that we are interested in such as $\bar{X}$, the proportion of successes in n Bernoulli trials. Then, we define a measure of *fit* or *distance* between the test statistic, and the value of the test statistic expected under H_0 (in the direction of some alternative hypothesis J). For example, in testing hypothesis $H_0 : \mu = 0.5$, a sensible distance measure $d(\boldsymbol{X}; H_0)$ is the (positive) difference between $\bar{X}$ and the expected proportion of successes under H_0, 0.5, in standard deviation units:

$$d(\mathbf{X}; H_0) = \frac{(\bar{X} - 0.5)}{\sigma_{\bar{X}}}.$$

Our distance measure may also be set out in terms of the likelihoods of H_0 as against different alternatives in J. A result $\boldsymbol{x}$ is further from H_0 to the extent that H_0 is less likely than members of J, given $\boldsymbol{x}$. This distance measure, which we may write as $d'(\boldsymbol{X}, H)$ gives us a likelihood ratio (LR). That is:

$$d'(\mathbf{X}, H_0) = LR = \frac{P(x; H_0)}{P(x; J)}$$

In the case of composite hypotheses we take the maximum value of the likelihood.[5]

No matter which distance measure is used, the key feature of the test is based on considering not just the one value of d that happened to occur, but *all* of the possible values. That is, d is itself a statistic that can take on different values in repeated trials of the experimental procedure generating the data. This probability distribution is called the *sampling* distribution of the distance statistic, and is what allows calculating error probabilities, one of which is the *statistical significance level* (SL):

[5]Otherwise, one would need to have prior probability assignments for each hypothesis within the composite alternative. Some strict likelihoodists, who do not use prior probabilities, regard this likelihood as undefined (e.g., [Edwards, 1992; Royall, 1997]).

Statistical Significance Level of the observed difference $d(\boldsymbol{x})$ (in testing H_0) = the probability of a difference as large as or larger than $d(\boldsymbol{x})$, under H_0.

In calculating the statistical significance level, one sums up the probabilities of outcomes as far as or further from H_0 as $\boldsymbol{x}$ is. The smaller the significance level, the further $\boldsymbol{x}$ is from what is expected under H_0: if it is very small, say 0.05 or 0.01, then the outcome is said to be statistically significant at these levels.

To highlight how an analysis by way of significance levels violates the LP, the most common example alludes to two different ways one could generate a series of n independent (Bernoulli) coin-tossing trials, with μ the probability of "heads" on each trial.

EXAMPLE 2 (Case 1: The Binomial Distribution). In the first case, it is decided in advance to carry out n flips, stop, and record the number of successes, which we can represent as random variable Z. Here, Z is a Binomial variable with parameters μ and n and the probability distribution of Z takes the form:

$$(1) \quad P_1(Z = z; \mu) = \binom{n}{z} \mu^z (1-\mu)^{n-z}$$

Suppose it is decided to observe $n = 12$ trials and the observed result is $Z = 9$ heads. The probability of this result, under the assumption that $\mu = \mu_0$ is:

$$P_1(Z = 9; H_0 : \mu = \mu_0) = \binom{12}{9} \mu_0^9 (1-\mu_0)^3$$

EXAMPLE 2 (Case 2: The Negative Binomial Distribution — A Case of Sequential Sampling). In case 2, by contrast, we are to consider that the experimenter was interested in the number of heads observed, Z, before obtaining r tails, for some fixed value r. In this sampling scheme the random variable Z follows the Negative Binomial distribution:

$$(2) \quad P_2(Z = z; \mu) = \binom{z+r-1}{z} \mu^z (1-\mu)^r.$$

This experiment can be viewed as conducting Bernoulli trials with the following stopping rule: Stop as soon as you get a total of r tails. We are next to imagine that r had been set in advance to 3, and it happens that 9 heads were observed before the third tail, thereby allowing the trials to terminate. We then have:

$$P_2(Z = 9, r = 3; H_0 : \mu = \mu_0) = \binom{11}{9} (\mu_0)^9 (1-\mu_0)^3.$$

In each of the two cases above, the data set consists of 9 heads and 3 tails. We see immediately that (1) and (2) differ only by a constant. So, a set of z heads and r tails in n=z+r Bernoulli trials defines the *same* likelihood whether by Binomial

sampling (n fixed) or Negative Binomial sampling (r fixed). In both cases, the likelihood of μ given $z = \mu^z(1-\mu)^{n-z}$. According to the LP, then, this difference between the two cases makes no difference to what the outcomes tell us about the various values of μ:

> If a Bernoulli process results in z successes in n trials, it has the likelihood function $\mu^z(1-\mu)^{n-z}$ and as far as inferences about μ are concerned, it is irrelevant whether either n or r was predetermined ([Pratt *et al.*, 1995, p. 542]. We replace their p with μ for consistency of notation).

Nevertheless, as the holder of the LP goes on to show, the *significance level* attained in case 1 differs from that of case 2, thereby showing that significance levels violate the LP. In particular, we have

(i) The statistical significance level for the Binomial (n fixed at 12)=

$$P_1(Z \geq 9; H_0 : \mu = 0.5) = P_1(z = 9 \text{ or } 10 \text{ or } 11 \text{ or } 12; \mu = 0.5) \approx 0.075$$

whereas

(ii) The significance level for the Negative Binomial (r fixed at 3)=

$$P_2(z = 9 \text{ or } 10 \text{ or } \ldots; \mu = 0.5) \approx 0.0325.$$

Thus, if the level of significance before rejecting H_0 were fixed at 0.05, we would reject H_0 if the observations were the result of Binomial trials, but we would *not* reject it if *those same observations* were the result of *Negative* Binomial trials.

5 THE OPTIONAL STOPPING EFFECT

Although the contrasting analysis demanded by the error statistician in considering the Binomial vs. the Negative Binomial (Example 2) was not very pronounced, the example we used at the opening of our paper points to much more extreme contrasts.

An example that has received considerable attention is of the type raised by Armitage at the "1959 Savage Forum." The null hypothesis H_0 is an assertion about a population parameter μ, the mean value of some quantity, say, a measure of the effectiveness of some medical treatment. The experiment involves taking a random sample of size n, $X_1, \ldots X_n$ and calculating its mean. Let $\bar{X}_n$ be the sample mean of the n observations of the X_is, where we know that each X_i is distributed Normally with *unknown* mean μ and *known* variance 1, i.e., $X_i \sim$ Normal$(\mu, 1)$. The null hypothesis asserts the treatment has no effect:

$$H_0 : \mu = 0.$$

The alternative hypothesis H_1 is the complex hypothesis consisting of all values of μ other than 0:

$$H_1 : \mu \neq 0.$$

As before, we are to consider two different stopping rules:

EXAMPLE 3 (Case 1: Test T-1 (fixed sample size)). In this case we take n samples, evaluate the distance between the observed mean, $\bar{x}_n$, and the mean hypothesized in H_0, namely 0, and then calculate the SL of this difference. For example, if $\bar{x}_n$ is 2 standard deviation units from 0, then the SL is approximately 0.05, regardless of the true value of the mean.[6] This is the *nominal (or computed) SL.*

A familiar test rule is to reject H_0 whenever the SL reaches some level, say 0.05. This test rule can be described as follows:

$$\text{Test T-1: Reject } H_0 \text{ at } SL = 0.05 \text{ iff } |\bar{X}_n| \geq 2/\sqrt{n}.$$

The standard deviation of $\bar{X}_n$ in this example is $1/\sqrt{n}$. We have

$$P(\text{Test T-1 rejects } H_0; H_0) = 0.05.$$

Since rejecting H_0 when H_0 is true is called the *type I error* of the test, we can also say

$$P(\text{Test T-1 commits a type I error}) = 0.05.$$

EXAMPLE 3 (Case 2: Test T-2 (Sequential testing)). In the second case sample size n is *not* fixed in advance. The stopping rule is:

(T-2) Keep sampling until $\bar{X}_n$ is 2 standard deviations away from 0 (the hypothesized value of μ in H_0) in either direction.

So we have

(T-2) Keep sampling until $|\bar{X}_n| \geq 2/\sqrt{n}$.

The difference between the two cases is that in T-2 the tests are applied *sequentially*. If we have not reached a 2 standard deviation difference after, say, the first 10 trials, we are to go on to take another 10 trials, and so on, as in the "try and try again" procedure of Example 1. The more generalized stopping rule T-2 for the Armitage example is:

$$\text{Keep sampling until } |\bar{X}_n| \geq k_\alpha/\sqrt{n}$$

where k_α is the number of standard deviations away from 0 that corresponds to a (nominal) SL of α. The probability that this rule will stop in a finite number oftrials is *one*, no matter what the true value of μ is; it is what is called a *proper* stopping rule.

[6] That is because we here have a two-sided test.

Table 1. The Effect of Repeated Significance Tests (the "Try and Try Again" Method)

Number of trials n	**Probability of rejecting H_0 with a result nominally significant at the 0.05 level at or before n trials, given H_0 is true**
1	0.05
2	0.083
10	0.193
20	0.238
30	0.280
40	0.303
50	0.320
60	0.334
80	0.357
100	0.375
200	0.425
500	0.487
750	0.512
1000	0.531
Infinity	1.000

Nominal SL vs. Actual SL

The probability that Test T-2 rejects H_0 even though H_0 is true — the probability it commits a type I error — *changes* according to how many sequential tests are run before we are allowed to stop. Because of this, there is a change in the *actual* significance level.

Suppose it takes 1000 trials to reach the 2-standard deviation difference. The SL for a 2-standard deviation difference, in Case 1, where n was fixed, would be 0.05, the *computed* or *nominal significance level*. But the actual probability of rejecting H_0 when it is true increases as n does, and so to calculate the actual SL, we need to calculate:

$$P(\text{Test T-2 stops and rejects } H_0 \text{ at or before } n{=}1000;\ H_0 \text{ is true}).$$

That is, the actual or overall significance level is the probability of finding a 0.05 *nominally* statistically significant difference from a fixed null hypothesis *at some stopping point or other* up to the point at which one is actually found. In other words, in sequential testing, the *actual* significance level accumulates, a fact reflected in Table 1.

While the *nominal* SL is 0.05, the *actual* SL for Case 2 is about 0.53: 53% of the time H_0 would be rejected even though it is true. More generally, applying

stopping rule T-2 would lead to an actual significance level that would differ from, and be greater than, α (unless it stopped at the first trial). If allowed to go on long enough, the probability of such an erroneous rejection is one![7]

By contrast, as Berger and Wolpert note:

> The SRP would imply, [in the Armitage example], that if the observation in Case 2 happened to have $n = k$, then the evidentiary content of the data would be the same as if the data had arisen from the fixed $[k]$ sample size experiment in Case 1 [Berger and Wolpert, 1988, p. 76].

So, in particular, if $n = 1000$, there would be no difference in "the evidentiary content of the data" from the two experiments.

Now holders of the LP do not deny that the actual significance levels differ dramatically, nor do error statisticians deny that alternative hypothesis $\mu = \bar{x}$ is more likely than the null hypothesis $\mu = 0$. Where the disputants *disagree* is with respect to what these facts *mean* for the evidential import of the data. Specifically, the error statistician's concern for the actual and not the nominal significance level in such cases leads her to infer that the stopping rule matters.

In contrast, the fact that the likelihood ratio is unaffected leads the proponent of the LP to infer that there is no difference in the evidential import, notwithstanding the difference in significance levels. Thus, according to the intuitions behind the LP, it is a virtue of a statistical account that it reflects this.

> 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) ... [Edwards *et al.*, 1963, p. 239].

We can grant some simplicity is lost, but that is because the error probability assurances are lost if one is allowed to change the experiment as one goes along, without reporting the altered significance level. Repeated tests of significance (or sequential trials) are permitted — are even *desirable* — in many situations. However, the error statistician requires that the interpretation of the resulting data reflect the fact that the error characteristics of a sequential test are different from those of a fixed-sample test. In effect, a *penalty* must be paid for perseverance. Before-trial planning stipulates how to select a small enough *nominal* significance level to compute at each trial so that the *actual* significance level is still low.[8] By contrast, since data $\boldsymbol{x}$ enter the Bayesian computation by means of the likelihood function, identical likelihood functions yield identical assignments of posterior probability or density — so no alteration is required with the two stopping rules, according to the LP.

[7] Feller [1940] is the first to show this explicitly.

[8] Medical trials, especially, are often deliberately designed as sequential. See [Armitage, 1975].

This leads to the question whether Bayesians are not thereby led into a situation analogous to the one that error statisticians would face were they to ignore the stopping rule.

EXAMPLE 3 (continued). Armitage continued his earlier remarks to Savage at the 1959 forum as follows:

> [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, p. 72]; (emphasis added).

That is, if we consider in Armitage's example the "uninformative" prior distribution of μ, uniform over $(-\infty, +\infty)$ and given that $\sigma^2 = 1$, then the posterior distribution for μ will be:

$$Normal\ (\bar{x}_n, 1/n).$$

The methods that Bayesians use to draw inferences about μ all depend on this posterior distribution in one way or another.[9] One common method of Bayesian inference involves using $\bar{x}$ to form an interval of μ values with highest posterior density, the "highest posterior density" (HPD) interval. In this case, the (approximate) 0.95 HPD interval will be $C_n(\bar{x}) = (\bar{x} - 2/\sqrt{n}, \bar{x} + 2/\sqrt{n})$. The Armitage stopping rule allows us to stop only when $|\bar{X}_n| > 2/\sqrt{n}$, and so that stopping rule *insures* that $\mu = 0$ is excluded from the HPD, even if $\mu = 0$ is true.

As even some advocates of the LP note, this looks very troubling for the Bayesian:

> The paradoxical feature of this example is that ... the experimenter can ensure that $C_n(\bar{x})$ *does not contain zero*; thus, as a classical confidence procedure, $\{C_n(\bar{x})\}$ will have zero coverage probability at $[\mu = 0]$ .... It thus seems that the experimenter can, through sneaky choice of the stopping rule, "fool" the Bayesian into believing that $[\mu]$ is not zero [Berger, 1985, p. 507].

That is, (using the non-informative prior density) the use of the stopping rule in T-2 ensures the Bayesian will accord a high posterior probability to an interval that *excludes* the true value of μ. Rather than use the HPD intervals, the analogous

[9] That this uninformative prior results in posteriors that match the values calculated as error probabilities is often touted by Bayesians as a point in *their favor*. For example, where the most an error statistician can say is that a confidence interval estimator contains the true value of μ 95% of the time, the Bayesian, with his uniform prior, can assign .95 posterior probability to the specific interval obtained.

point can be made in reference to Bayesian hypothesis testing.[10] Nor can one just dismiss the issue by noting the obvious fact that the probability for any value of the continuous parameter is zero. Bayesians supply many procedures for inferences about continuous parameters, and the issue at hand arises for each of them. One procedure Bayesians supply is to calculate the posterior probability of a small interval around the null, (0-ε, 0+ε). With ε small enough, the likelihood is constant in a neighborhood of 0, so the posterior probability obtained from the Armitage stopping rule (T-2) will be very low for (0-ε, 0+ε), even if $\mu = 0$. And since T-2 is a proper stopping rule, such a low posterior for a true interval around 0 is assured.

In discussions of Armitage's example, most of the focus has been on ways to avoid this very extreme consequence–the guarantee (with probability 1) of arriving at an HPD interval that excludes the true value, $\mu = 0$, or a low posterior density to a true null hypothesis. For example, because the extreme consequence turns on using the (improper) uniform prior, some Bayesians have taken pains to show that this may be avoided with countably additive priors (e.g., [Kadane *et al.*, 1999]).[11]

Nevertheless, the most important consequence of the Armitage example is not so much the extreme cases (where one is guaranteed of strong evidence against the true null) but rather the fact that ignoring stopping rules can lead to a high probability of error, and that this high error probability is not reflected in the interpretation of data according to the LP. Even allowing that the Bayesians have ways to avoid the extreme cases, therefore, these gambits fail to show how to adhere to the LP and avoid a high probabilities of strong evidence against a true null.

To underscore this point, consider a modified version of T-2: the experimenter will make at most 1000 trials, but will stop before then if $\bar{X}_n$ falls more than 2 standard deviations from zero. This modified rule (while also proper) does not assure that when one stops one has $|\bar{X}_n| \geq 2/\sqrt{n}$. Nevertheless, were our experimenter to stop at the 1000th trial, the error probability is high enough (over 0.5) to be disturbing for an error statistician. (See Table 1.) So the error statistician would be troubled by any interpretation of the data that was not altered by dint of this high error probability (due to the stopping rule). Followers of the LP do not regard this stopping rule as altering the interpretation of the data — whatever final form of evidential appraisal or inference they favor. None of the discussions of the Armitage example address this consequence of the less extreme cases.

[10] HPDs are not invariant under one-one transformations of the parameter space [Berger, 1985, p. 144]. Some Bayesians find this a compelling reason to avoid HPDs altogether, but this method nevertheless is commonly used.

[11] One might propose that after the first observation, one could use the result to arrive at a new countably additive prior. But this altering of the prior so that the so-called "foregone conclusion" is avoided is not the Armitage example anymore, and so does not cut against that example which concerns an after-trial analysis of the data once one stops.

6 REACTIONS TO THE CONSEQUENCES OF THE LP

For the most part, holders of the LP have not shirked from but have *applauded* the fact that the inferential consequences of the LP conflict with those of error statistical principles. Indeed, those who promote Bayesianism over error statistical approaches often tout the fact that stopping rules (and other aspects of the data generation procedure) do not alter the Bayesian's inference.

At the same time, however, many Bayesians and other holders of the LP are plainly uncomfortable with the fact that the LP can lead to high error probabilities and attempt to deny or mitigate this consequence. We do not think that any existing attempts succeed. Before explaining why, we should emphasize that the consequences of the Armitage-style stopping rule example are not the only ways that adherence to the LP conflicts with the control of error probabilities. Because of this conflict, many have rejected the LP — including some who at first were most sympathetic, most notably Allan Birnbaum, who concluded that

> 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, p. 128].[12]

Therefore, in our view, a strategy to block high probabilities of erroneous interpretations as a result of stopping rules will not do unless it can be demonstrated that:

1. It is part of a complete account that blocks high probabilities of erroneous inferences (whatever the form of inference or evidential appraisal the account licenses.)

2. It is not merely *ad hoc*. There must be a general rationale for the strategy that is also consistent with the LP.

6.1 Can the Stopping Rule Alter the Likelihood Function?

Upon first hearing of the Armitage example, one might assume that the stopping rule T-2 *must* make some kind of difference to the likelihood function. This is especially so for those inclined to dabble informally with likelihoods or Bayes' Theorem apart from any explicit mathematical definition of the likelihood function. We know of no formal statistical treatment of the Armitage example that has seriously claimed that the two stopping rules imply different likelihood functions. (Other types of strategies are proposed, which we will consider.) But these informal intuitions are important, especially for philosophers seeking an adequate account of statistical inference.

[12] See Birnbaum [1961; 1962; 1972], Giere [1977], as well as citations in [Barnard and Godambe, 1982] and [Bjornstad, 1992].

To begin with, it is worth noting that there are *other* kinds of situations in which stopping rules *will* imply different likelihood functions. These are known as *informative stopping rules*, an example of which is given by Edwards, Lindman, and Savage:

> A man who wanted to know how frequently lions watered at a certain pool was chased away by lions before he actually saw any of them watering there; in trying to conclude how many lions do water there he should remember why his observation was interrupted when it was [Edwards *et al.*, 1963, p. 239].

Although a more realistic example might seem more satisfactory, in fact, it is apparently very difficult to find a *realistic* stopping rule that is genuinely informative. (For a discussion, see [Berger and Wolpert, 1988, pp. 88–90].) As Edwards, *et al.*, then add: "We would not give a facetious example had we been able to think of a serious one." In any event, this issue is irrelevant for the Armitage-type example because T-2 is not an informative stopping rule. Although the probability of deciding to take more observations at each stage depends on $\boldsymbol{x}$, it does *not* depend on the parameter μ under test.[13]

Nevertheless, we are willing to address those who assume that an informal or subjectivist construal of probabilities gives them a legitimate way to alter the likelihood based on the stopping rule T-2. But to address them we need more than their intuitive hunch, they need to tell us in general how we are to calculate the likelihoods that will be needed, whether the account is purely likelihoodist (e.g., Royall [1992; 1997]) or Bayesian. Are we to substitute error probabilities in for likelihoods? Which ones? And how will this escape the Bayesian incoherence to which error probabilities such as significance levels are shown to lead?

To see that any such suggested alteration of likelihoods runs afoul of the LP, it must be remembered that the likelihood is a function of the observed $\boldsymbol{x}$:

> The philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with the observed $\boldsymbol{x}$, 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, pp. 65–66].

Once the data $\boldsymbol{x}$ are in hand, the holder of the LP insists on the "irrelevance of the sample space" — the irrelevance of the other outcomes that *could* have occurred but did not when drawing inferences from $\boldsymbol{x}$ (e.g., [Royall, 1997]). This is often

[13] As Berger and Wolpert [1988, p. 90] observe, the mere fact that the likelihood function depends on N, the number of observations until stopping, does *not* imply that the stopping rule is informative: "Very often N will carry information about [the parameter], but to be informative a stopping rule must carry information about [the parameter] additional to that available in [the sample $\boldsymbol{X}$], and this last will be rare in practice" (*ibid.*, 90). For further discussion of informative stopping rules, see [Roberts, 1967].

expressed by saying the holder of the LP is a *conditionalist*: for them inferences are always conditional on the *actual* value $\boldsymbol{x}$.

With respect to stopping rules, the conditionalist asks: Why should our interpretation of the data in front of us, $\boldsymbol{x}$, depend upon what would have happened *if* the trials were stopped earlier than they *actually* were stopped?

> 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 *et al.*, 1963, p. 238].

But altering the likelihood because of the stopping rule *is* to take into account the stopping plan, e.g., that if he hadn't gotten a significant result at 10 trials, he *would have* continued, and so on, thereby violating the LP. So anyone who thinks a subjectivist or informal construal of likelihoods gives them a legitimate way out, must be aware of this conflict with the conditionalist principle. Certainly this would put them at odds with leading subjective Bayesians who condemn error statisticians for just such a conflict:

> 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, 1993, p. 212].

Thus, the intuition that the stopping rule should somehow alter the likelihood is at odds with the most well-entrenched subjective Bayesian position and constitutes a shift toward the error statistical (or "frequentist") camp and away from the central philosophy of evidence behind the LP. According to the LP philosophy:

> [I]t seems very strange that a frequentist could not analyze a given set of data, such as $(x_1, \ldots, x_n)$ [in Armitage's example] if the stopping rule is not given. ... data should be able to speak for itself [Berger and Wolpert, 1988, p. 78].

We say the shift is to the error statistical camp because it reflects agreement with the error statistician's position that one cannot properly 'hear' what the data are saying without knowing how they were generated — whenever that information alters the capabilities of the test to probe errors of interest, as in the case of stopping rules. It is precisely in order to have a place to record such information that Neyman and Pearson were led to go beyond the likelihood ratio (LR):

> If we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis, because a knowledge of [the LR] alone is not adequate to insure control of this error [Pearson and Neyman, 1930, p. 106].

When test T-2 stops, it is true that the LR (in favor of H_0) is small. However, to the error statistician, we cannot thereby infer we should be *justified* in rejecting the hypothesis H_0, because:

> In order to fix a limit between 'small' and 'large' values of [LR] we must know how often such values appear when we deal with a true hypothesis. That is to say we must have knowledge of ... the chance of obtaining [LR as small or smaller than the one observed] in the case where [H_0] is true (*ibid*, p. 106).

Accordingly, without the error probability assessment, Pearson and Neyman are saying we cannot determine if there really is any warranted evidence against H_0.[14] Stopping rules give crucial information for such an error statistical calculation.

It is no surprise, then, that the error statistician regards examples like Armitage's as grounds for rejecting the LP. To those who share the error statistical intuitions, our question is: on what grounds can they then defend the LP?

6.2 *Can Stopping Rules Alter the Prior?*

In order to avoid assigning the high posterior to a false non-null hypothesis, as Berger and Wolpert (1988) point out, "the Bayesian might ... assign some positive prior probability, λ, to μ being equal to zero" (p. 81) perhaps to reflect a suspicion that the agent is using stopping rule T-2 because he thinks the null hypothesis is true.[15] Assume, for example, that one assigns a prior probability mass of 0.50 to the null hypothesis and distributes the rest Normally over the remaining values of

[14] It should be emphasized that it is not that the N-P inference consists merely of a report of the significance level (or other error probabilities), at least not if the tests are being used for inference or evidence. It is rather that determining the warranted inference depends on the actual significance level and other error probabilities of tests. Granted, the onus is on the error-statistician to defend a philosophy of inference that uses and depends on controlling error probabilities (though this is not our concern here). See Note 22.

[15] A positive prior probability, λ, can be assigned to $\mu = 0$ and the rest, $1 - \lambda$, distributed over the remaining values of μ. (This amounts to giving $\mu = 0$ a non-zero mass, and every other hypothesis zero mass.) When $1 - \lambda$ is distributed Normally over the remaining hypotheses with mean 0 and variance ρ^2, the posterior probability distribution will be:

$$P\left(\mu = 0/\bar{x}_n = \frac{K}{\sqrt{n}}\right) = \left[1 + \left(\frac{1}{\lambda} - 1\right) \frac{1}{\sqrt{(1 + n\rho^2)}} e^{\frac{K^2 n\rho^2}{2(1+n\rho^2)}}\right]^{-1}$$

where K is the number of standard deviations stipulated in the stopping rule and n is the number of observations needed to stop [Berger and Wolpert, 1988, p. 81].

See also [Berger and Berry, 1987], [Smith, 1961, p. 36–37].

μ. If it takes $n = 1000$ trials to stop, the posterior probability assignment to $\mu = 0$ is no longer low, but rather, around 0.37. A virtue of such a prior, as Berger and Wolpert note, is that it results in an increasing posterior probability assignment to $\mu = 0$ as the number of trials before stopping increases. For example, with this prior and $n = 10,000$, the posterior for the null is about 0.65.

Granted, in a case where one *had* this prior, the low posterior assignment to the null hypothesis is avoided, but this does nothing to mitigate the problem as it arises with the *uniform* prior — a prior the Bayesian often advocates. Perhaps the Bayesian would wish to suggest that *whenever* one is confronted with an experiment with stopping rule T-2, one should reject the uniform prior in favor of one that appears to avoid the problem. But why should a Bayesian alter the prior upon learning of the stopping rule?

There is the motivation suggested by Berger and Wolpert [1988], that if you *suspected* that the person generating the observations was using stopping-rule T-2 *for the purpose of misleading you*, you would raise your prior probability assignment to $\mu = 0$. Does this not violate the LP? Perhaps one could retain the LP on the grounds that one is only allowing the stopping-rule to affect the *prior* rather than the likelihoods (and hence not "*what the data say*").[16]

Nevertheless, a Bayesian should have serious objections to this response to the stopping rule problem. Why, after all, should we think that the experimenter is using T-2 to *deceive* you? Why not regard his determination to demonstrate evidence against the null hypothesis as a sign that the null is *false*? Perhaps he is using T-2 only because he *knows* that $\mu \neq 0$ and he is trying to *convince* you of the truth!

Surely it would be unfair to suppose that those who, like Savage, touted the irrelevance of the stopping rule were sanctioning *deception* when they asserted: "Many experimenters would like to feel free to collect data until they have either conclusively proved their point, [or] conclusively disproved it" [Edwards *et al.*, 1963, p. 239]. Plainly, what they *meant* to be saying is that there is no reason to interpret the data differently because they arose from optional stopping. Equating optional stopping (with rule T-2) with deception runs counter to Savage's insistence that, because "optional stopping is no sin," any measure that *is* altered by the stopping rule, such as the significance level, is thereby inappropriate for assessing evidence [Savage, 1964, p. 185].[17]

Those who advocate the above move, then, should ask why the sensitivity of significance levels to stopping rules violates the LP — and thus is a *bad* thing — but the *same* kind of sensitivity of priors is acceptable. The LP, after all, asserts that *all* the information contained in the data that is relevant to comparisons between different parameter values is given in the likelihood function. But what else could

[16]This 'solution' demands that the agent know not only the *stopping*-rule used, but *why* the experimenter chose that particular stopping-rule, since knowing he wanted to *deceive* rather than to *help* you could make all the difference in the prior you use. Yet Bayesians have delighted in the fact that the LP renders irrelevant the *intentions* of experimenters to the import of the experiment.

[17]There is nothing in the LP to prevent Bayesians from deciding in advance to prohibit certain kinds of stopping rules, but again, one would like to know why.

it mean to say that one's choice of *priors* depends on the stopping-rule other than that the stopping-rule contains information relevant to comparisons between values of μ? It is little wonder that many Bayesians have balked at allowing the stopping rule to alter one's prior probability: "Why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it" [Lindley, 1972, p. 71]. Why indeed?[18]

Finally, even if we put aside the question of stopping rules leading to problematic final posterior probabilities, as long as the Bayesian conceives of *likelihood* as determining "what the data have to say", it is *still* the case that the data from T-2 are regarded as much stronger support for the non-null than the null, according to the Bayesian criterion of support.[19]

6.3 *Does the LP Provide Bounds on Being Misled?*

A third kind of response *grants* that the stopping rule makes no difference at all to either the likelihood function or the priors, and instead attempts to argue that, nonetheless, one who holds the LP can avoid having a high probability of being misled by the data. This argument is sound only for tests that differ in essential ways from the type leading to the Armitage result. Nevertheless, this response is important, if only because it is the one first put forward by Savage in responding to Armitage (Savage 1962).[20]

"Let us examine first a simple case" Savage proposes, where we are testing a simple or point null hypothesis H_0 against a point alternative H_1: that is H_0 asserts $\mu = \mu_0$, and the alternative H_1 asserts $\mu = \mu_1$. Call this a *point against point* test. In that case, if one is intent on sampling until the likelihood ratio (LR) in favor of H_1 exceeds r (for any value of $r > 1$), it can be shown that if H_0 is true, the probability is only $1/r$ that one will succeed in stopping the trials. This response turns on the fact that when we have a (true) simple null H_0 against a simple alternative H_1, then there is an upper bound to the probability of obtaining a result that makes $H_1 r$ times more likely than H_0, namely, $1/r$, i.e. $\mathrm{P}(LR > r; H_0) \leq 1/r$.

[18]Lindley is referring to Bayesians like Jeffreys [1961] and Rosenkrantz [1977] who determine 'objective' or 'non-subjective' priors by appealing to formal information-theoretic criteria. They would, for example, recommend different priors in the Binomial vs. the Negative Binomial case [Box and Tiao, 1973]. Doing so apparently violates the LP, and has led many Bayesians to be suspicious of such priors [Hill, 1987; Seidenfeld, 1979], or even to declare that "no theory which incorporates non-subjective priors can truly be called Bayesian, and no amount of wishful thinking can alter this reality" (Dawid, in [Bernardo, 1997, p. 179]). For related discussions contrasting subjective and objective priors, see also [Akaike, 1982; Barnett, 1982; Bernardo, 1979; Bernardo, 1997].

[19]This point does not rely on the technical Bayesian definition of "support" as an increase in the posterior, but holds for any conception based on the likelihood, *e.g.*, weight of evidence [Good, 1983]. Bayesians who reject all such notions of Bayesian support need to tell us what notion of support or evidence they condone.

[20]It is also the first one mentioned by many defenders of the LP, e.g., [Berger and Wolpert, 1988; Oakes, 1986; Royall, 1997].

This impressively small upper bound, however, does nothing to ameliorate the consequences of the Armitage optional stopping example because that example is not a case of a point against point test.[21]

6.4 *Extrapolating From Our Intuitions in Simple Cases*

The simple case of testing "point against point" hypotheses has encouraged some to suppose that the LP offers such protection in *all* cases — yet it does not. Perhaps the tendency to turn to the point against point test when confronted with stopping rule problems explains why the Armitage-type consequence has not received more attention by Bayesians. But there seems to be a different kind of strategy often at work in alluding to the point against point test in defending the LP, and we may regard this as a distinct response to the stopping rule problem.

In appraising the LP, say some, we should trust our intuitions about its plausibility when we focus on certain *simple* kinds of situations, such as testing point against point hypotheses, "rather than in extremely complex situations such as [Armitage's example]" [Berger and Wolpert, 1988, p. 83]. Since looking at just the likelihood ratio (and ignoring the stopping rule) seems intuitively plausible in point against point testing, they urge, it stands to reason that the LP must be adhered to in the more 'complex situation' — even if its consequences in the latter case seem unpalatable. Regarded as an argument for deflecting the Armitage example it is clearly unsound. Bracketing a whole class of counterexamples simply on the basis that they are "extremely complicated" is *ad hoc* — preventing the LP from being subject to the relevant kind of test here. Moreover, such sequential tests are hardly exotic, being standard in medicine and elsewhere.

But perhaps it is only intended as a kind of pragmatic appeal to what is imagined to be the lesser of two evils: their reasoning seems to be that even if the LP leads to unintuitive consequences in the complex (optional stopping) case, its rejection would be so unappealing in the simple cases that it is better to uphold the LP and instead discount our intuitions in the complex cases. By contrast, some have gone the route of George Barnard — the statistician credited with first articulating the LP [Barnard, 1949] — who confessed at the 1959 Savage Forum that the Armitage-type example led him to conclude that whereas the LP is fine for the simple cases it must be abandoned in the more complex ones (see [Barnard, 1962]). The LP adherent owes us an argument as to why Barnard's move should not be preferred.

[21] The existence of an upper bound less than 1 can also be shown in more general cases such as when deling with k simple hypotheses, though as k increases, the upper bound is no longer impressively small. The general result, stated in [Kerridge, 1963] is that with $k + 1$ simple hypothses where H_0 is true and $H_1, \ldots, H_k$ are false and $Pr(H_i) = (k+1)^{-1}$ for $i = 0, 1, \ldots, k$:

$$P(P(H_0/X_n) \leq p) \leq \frac{kp}{(1-p)}.$$

Moreover, such bounds depend on having countably addititve probability, while the uniform prior in Armitage's example imposes finite additivity.

7 CONCLUDING REMARKS

Philosophers who appeal to principles and methods from statistical theory in tackling problems in philosophy of science need to recognize the *consequences* of the statistical theory they endorse. Nowhere is this more crucial than in the on-going debate between Bayesian and non-Bayesian approaches to scientific reasoning. Since Bayesianism — which is committed to the LP — has emerged as the dominant view of scientific inference among philosophers of science, it becomes all the more important to be aware of the LP's many implications regarding evidence, inference and methodology.

Some of the most important of these implications concern the LP's effect on our ability to control error and thereby the reliability and severity of our inferences and tests — generally regarded as important goals of science. A consequence of our discussion is that there is no obvious way in which approaches consistent with the LP can deliver these goods. In giving the spotlight to the kind of unreliability that can result from ignoring *stopping rules*, our goal is really to highlight some of the consequences for reliability of accepting the LP, not to argue that examples such as Armitage's are common. At the same time, however, it should be realized that examining the effect of stopping rules is just *one* of the ways that facts about how the data are generated can affect error probabilities. Embracing the LP is at odds with the goal of distinguishing the import of data on grounds of the error statistical characteristics of the procedure that generated them.

Now Bayesians and likelihoodists may deny that this appeal to error probabilities is what matters in assessing data for inference. They often deny, for example, that the error statistician's concern with the behavior of a test in a series of repetitions is relevant for inference.[22] Strict adherence to this position would lead one to expect that they would be unfazed by the Armitage result. In reality, however, existing Bayesian and Likelihoodist reactions to Armitage-type examples are strikingly and surprisingly equivocal, and the Bayesian attempts to deflect the Armitage result have been unclear e.g. see [Johnstone *et al.*, 1986]. Sometimes they say "It's not a problem, we do not care about error rates", while at other times the claim is "Even though we don't care about error rates, we can still satisfy one who does." The former response is consistent for a holdr of the LP, but it demands renouncing error probabilities, as we understand that notion. The latter attitude demands an argument showing how to resolve the apparent tension with the LP. We have tried to locate the most coherent and consistent arguments, and found that

[22]The long-standing challenge of how to interpret error statistical tests "evidentially" cannot be delved into here, but we can see the directions in which such an interpretation (or reinterpretation) might take us, by extending what we said about why the error statistician regards the stopping rule as relevant. The error statistician regards data as evidence for a hypothesis H to the extent that H has passed a reliable or severe test of H, and this requires not just that H fit $\boldsymbol{x}$ but also that test T would very probably not have resulted in so good a fit with H were H false or specifiably in error. See [Mayo, 2000], [Mayo and Spanos, 2000]. By contrast, the Armitage stopping rule makes it maximally probable that $\mathbf{x}$ fits a false H, so H passes a test with *minimal* severity.

they failed to live up to this demand. We invite anyone who can further clarify the Bayesian and Likelihoodist position on the Armitage example to do so.

ACKNOWLEDGEMENTS

We are indebted to Aris Spanos for numerous, highly important statistical insights regarding the Armitage case. We thank Teddy Seidenfeld, and the participants of D. Mayo's 1999 National Endowment for the Humanities Summer Seminar, for a number of challenging questions, criticisms, and suggestions regarding earlier drafts. D. Mayo gratefully acknowledges support for this research from the National Science Foundation, grant no. SBR-9731505.

Virginia Tech, USA.

BIBLIOGRAPHY

[Akaike, 1982] H. Akaike. On the fallacy of the likelihood principle. *Statistics and Probability Letters* **1**, 75–78, 1982.

[Armitage, 1962] P. Armitage. Contribution to discussion in L. Savage, ed. 1962.

[Armitage, 1975] P. Armitage. *Sequential Medical Trials*. Oxford: Blackwell, 1975.

[Barnard, 1949] G. A. Barnard. Statistical inference. *Journal of the Royal Statistical Society*, Series B (Methodologial), **11**, 115–149, 1949.

[Barnard, 1962] G. A. Barnard. Contribution to discussion in L. Savage, ed. 1962.

[Barnard and Godambe, 1982] G. A. Barnard and V. P. Godambe. Memorial article: Allan Birnbaum 1923–1976. *The Annals of Statistics* **10**, 1033–1039, 1982.

[Barnett, 1982] V. Barnett. *Comparative Statistical Inference*, 2nd edition. John Wiley, New York 1982.

[Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. 2nd edition. Springer-Verlag, New York, 1985.

[Berger and Berry, 1987] J. O. Berger and D. A. Berry. The relevance of stopping rules in statistical inference. In *Statistical Decision Theory and Related Topics IV*, vol. 1, S. S. Gupta and J. Berger, eds. Springer-Verlag, 1987.

[Berger and Wolpert, 1988] J. O. Berger and R. L. Wolpert. *The Likelihood Principle*, 2nd edition. Institute of Mathematical Statistics, Hayward, CA, 1988.

[Bernardo, 1979] J. M. Bernardo. Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society*, series B:41, 113–147, 1979.

[Bernardo, 1997] J. M. Bernardo. Noninformative priors do not exist: A discussion with José M. Bernardo (with discussion). *Journal of Statistical Planning and Inference* **65**, 159–189, 1997.

[Birnbaum, 1961] A. Birnbaum. On the foundations of statistical inference: binary experiments. *Annals of Mathematical Statistics* **32**, 414–435, 1961.

[Birnbaum, 1962] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, **57**, 269–306, 1962.

[Birnbaum, 1969] A. Birnbaum. Concepts of statistical evidence. In *Essays in Honor of Ernest Nagel*, Sidney Morgenbesser, Patrick Suppes and Morton White, eds. St. Martin's Press, 1969.

[Birnbaum, 1972] A. Birnbaum. More on concepts of statistical evidence. *Journal of the American Statistical Association*. **67**, 858–861, 1972.

[Bjornstad, 1992] J. F. Bjornstad. Birnbaum (1962) on the foundations of statistical inference. In *Breakthroughs in Statistics*, vol. 1, 461–477. Samuel Kotz and Norman L. Johnson, eds. Springer-Verlag, New York, 1992.

[Box and Tiao, 1973] G. Box and G. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

[Edwards, 1992] A. W. F. Edwards. *Likelihood* (2nd edition). Cambridge University Press, 1992.

[Edwards *et al.*, 1963] W. Edwards, H. Lindman and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review* **70**, 450–499, 1963.

[Feller, 1940] W. K. Feller. Statistical aspects of ESP. *Journal of Parapsychology* **4**, 271–298, 1940.

[Giere, 1977] R. N. Giere. Alan Birnbaum's conception of statistical evidence. *Synthese*, **36**, 5–13, 1977.

[Gillies, 1990] D. A. Gillies. Bayesianism versus falsificationism. *Ratio*, **3**, 82–98, 1990.

[Good, 1983] I. J. Good. *Good Thinking*. University of Minnesota Press, Minneapolis, MN, 1983.

[Hill, 1987] B. M. Hill. The validity of the likelihood principle. *The American Statistician*, **47**, 95–100, 1987.

[Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*, second edition. Open Court, Chicago, 1993.

[Jeffreys, 1961] H. Jeffreys. *Theory of Probability*, 3rd edition. Clarendon Press, Oxford, 1961.

[Johnstone *et al.*, 1986] D. J. Johnstone, G. A. Barnard and D. V. Lindley. Tests of significance in theory and practice. *The American Statistician*, **35**, 491–504, 1986.

[Kadane *et al.*, 1999] J. B. Kadane, M.J. Schervish and T. Seidenfeld. *Rethinking the Foundations of Statistics*. Cambridge University Press, Cambridge, 1999.

[Kerridge, 1963] D. Kerridge. Bounds for the frequency of misleading Bayes' inferences. *Annals of Mathematical Statistics* **34**, 1109–1110, 1963.

[Lindley, 1972] D. V. Lindley. *Bayesian Statistics — A Review*. J. W. Arrowsmith, Bristol, 1972.

[Mayo, 1996] D. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago, 1996.

[Mayo, 2000] D. Mayo. experimental practice and an error statistical account of evidence. *Philosophy of Science*, **67**, (Proceedings), S193–S207, 2000.

[Mayo and Spanos, 2000] D. Mayo and A. Spanos. A Post-data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing. *Measurements in Physics and Economics Discussion Paper Series*, DP MEAS 8/00. Centre for Philosophy of Natural & Social Science, London School of Economics, 2000.

[Oakes, 1986] M. Oakes. *Statistical Inference*, Wiley, 1986.

[Pearson and Neyman, 1930] E. S. Pearson and J. Neyman. On the problem of two samples. *Bull. Acad. Pol. Sci.*, 73–96, 1930. Reprinted in J. Neyman and E. S. Pearson, *Joint Statistical Papers*. pp. 81–106 University of California Press, Berkeley, 1967.

[Pratt *et al.*, 1995] J. W. Pratt, H. Raffia and R. Schlaifer. *Introduction to Statistical Decision Theory*. The MIT Press, Cambridge, MA, 1995.

[Roberts, 1967] H. V. Roberts. Informative stopping rules and inferences about population size. *Journal of the American Statistical Association*. **62**, 763–775, 1967.

[Rosenkrantz, 1977] R. D. Rosenkrantz. *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*. Boston: Reidel, 1977.

[Royall, 1992] R. Royall. The elusive concept of statistical evidence (with discussion). In Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds. pp. 405–418. Oxford University Press, Oxford, 1992.

[Royall, 1997] R. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London, 1997

[Savage, 1962] L. J. Savage. *The Foundations of Statistical Inference: A Discussion*. Methuen, London, 1962.

[Savage, 1964] L. J. Savage. The foundations of statistics reconsidered. In *Studies in Subjective Probability*, H. Kyberg and H. Smokler, eds. John Wiley, New York, 1964.

[Seidenfeld, 1979] T. Seidenfeld. Why I am not an objective Bayesian. *Theory and Decision* **11**, 413–440, 1979.

[Smith, 1961] C. A. B. Smith. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society*, (B), Vol. 23, No. 1, 1–37, 1961.