Tour II Error Probing Tools versus Logics of Evidence

1.4 The Law of Likelihood and Error Statistics

If you want to understand what's true about statistical inference, you should begin with what has long been a holy grail – to use probability to arrive at a type of logic of evidential support – and in the first instance you should look not at full-blown Bayesian probabilism, but at comparative accounts that sidestep prior probabilities in hypotheses. An intuitively plausible logic of comparative support was given by the philosopher Ian Hacking (1965) – the Law of Likelihood. Fortunately, the Museum of Statistics is organized by theme, and the Law of Likelihood and the related Likelihood Principle is a big one.

> *Law of Likelihood (LL):* Data $\boldsymbol{x}$ are better evidence for hypothesis H_1 than for H_0 if $\boldsymbol{x}$ is more probable under H_1 than under H_0: $\Pr(\boldsymbol{x}; H_1) > \Pr(\boldsymbol{x}; H_0)$, that is, the *likelihood ratio (LR)* of H_1 over H_0 exceeds 1.

H_0 and H_1 are statistical hypotheses that assign probabilities to values of the random variable $\boldsymbol{X}$. A fixed value of $\boldsymbol{X}$ is written $\boldsymbol{x}_0$, but we often want to generalize about this value, in which case, following others, I use $\boldsymbol{x}$. The *likelihood of the hypothesis H*, given data $\boldsymbol{x}$, is the probability of observing $\boldsymbol{x}$, under the assumption that H is true or adequate in some sense. Typically, the ratio of the likelihood of H_1 over H_0 also supplies the quantitative measure of comparative support. Note, when $\boldsymbol{X}$ is continuous, the probability is assigned over a small interval around $\boldsymbol{X}$, to avoid probability 0.

Does the Law of Likelihood Obey the Minimal Requirement for Severity?

Likelihoods are vital to all statistical accounts, but they are often misunderstood because the data are fixed and the hypothesis varies. Likelihoods of hypotheses should not be confused with their probabilities. Two ways to see this. First, suppose you discover all of the stocks in Pickrite's promotional letter went up in value ($\boldsymbol{x}$) – all winners. A hypothesis H to explain this is that their method always succeeds in picking winners. H *entails* $\boldsymbol{x}$, so the likelihood of H given $\boldsymbol{x}$ is 1. Yet we wouldn't say H is therefore highly probable, especially without reason to put to rest that they culled the winners post hoc. For a second

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

way, at any time, the same phenomenon may be perfectly predicted or explained by two rival theories; so both theories are equally likely on the data, even though they cannot both be true.

Suppose Bristol-Roach, in our Bernoulli tea tasting example, got two correct guesses followed by one failure. The observed data can be represented as $\boldsymbol{x}_0 = \langle 1,1,0 \rangle$. Let the hypotheses be different values for θ, the probability of success on each independent trial. The likelihood of the hypothesis $H_0 : \theta = 0.5$, given $\boldsymbol{x}_0$, which we may write as Lik(0.5), equals (1/2)(1/2)(1/2) = 1/8. Strictly speaking, we should write Lik(θ;$\boldsymbol{x}_0$), because it's always computed given data $\boldsymbol{x}_0$; I will do so later on. The likelihood of the hypothesis $\theta = 0.2$ is Lik(0.2) = (0.2)(0.2)(0.8) = 0.032. In general, the likelihood in the case of Bernoulli independent and identically distributed trials takes the form: Lik(θ)= $\theta^s(1-\theta)^f$, $0 < \theta < 1$, where s is the number of successes and f the number of failures. Infinitely many values for θ between 0 and 1 yield positive likelihoods; clearly then, likelihoods do not sum to 1, or any number in particular. Likelihoods do not obey the probability calculus.

The Law of Likelihood (LL) will immediately be seen to fail our minimal severity requirement – at least if it is taken as an account of inference. Why? There is no onus on the Likelihoodist to predesignate the rival hypotheses – you are free to search, hunt, and post-designate a more likely, or even maximally likely, rival to a test hypothesis H_0.

Consider the hypothesis that $\theta = 1$ on trials one and two and 0 on trial three. That makes the probability of $\boldsymbol{x}$ maximal. For another example, hypothesize that the observed pattern would always recur in three-trials of the experiment (I. J. Good said in his cryptoanalysis work these were called "kinkera"). Hunting for an impressive fit, or trying and trying again, one is sure to find a rival hypothesis H_1 much better "supported" than H_0 even when H_0 is true. As George Barnard puts it, "there *always* is such a rival hypothesis, viz. that things just had to turn out the way they actually did" (1972, p. 129).

Note that for any outcome of n Bernoulli trials, the likelihood of $H_0 : \theta = 0.5$ is $(0.5)^n$, so is quite small. The likelihood ratio (LR) of a best-supported alternative compared to H_0 would be quite high. Since one could always erect such an alternative,

(*) Pr(LR in favor of H_1 over H_0; H_0) = maximal.

Thus the LL permits BENT evidence. The severity for H_1 is minimal, though the particular H_1 is not formulated until the data are in hand. I call such maximally fitting, but minimally severely tested, hypotheses *Gellerized*, since Uri Geller was apt to erect a way to explain his results in ESP trials. Our Texas sharpshooter is analogous because he can always draw a circle around a cluster of bullet holes, or around each single hole. One needn't go to such an extreme

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

rival, but it suffices to show that the LL does not control the probability of erroneous interpretations.

What do we do to compute (*)? We look beyond the specific observed data to the behavior of the general rule or method, here the LL. The output is always a comparison of likelihoods. We observe one outcome, but we can consider that for any outcome, unless it makes H_0 maximally likely, we can find an H_1 that is more likely. This lets us compute the relevant properties of the method: its inability to block erroneous interpretations of data. As always, a severity assessment is one level removed: you give me the rule, and I consider its latitude for erroneous outputs. We're actually looking at the probability distribution of the rule, over outcomes in the sample space. This distribution is called a *sampling distribution*. It's not a very apt term, but nothing has arisen to replace it. For those who embrace the LL, once the data are given, it's irrelevant what other outcomes could have been observed but were not. Likelihoodists say that such considerations make sense only if the concern is the performance of a rule over repetitions, but not for inference from the data. Likelihoodists hold to "the irrelevance of the sample space" (once the data are given). This is the key contrast between accounts based on error probabilities (error statistical accounts) and logics of statistical inference.

Hacking "There is No Such Thing as a Logic of Statistical Inference"

Hacking's (1965) book was so ahead of its time that by the time philosophers of science started to get serious about philosophy of statistics, he had already broken the law he had earlier advanced. Hacking (1972, 1980) admits to having been caught up in the "logicist" mindset wherein we assume a logical relationship exists between any data and hypothesis; and even denies (1980, p. 145) there is any such thing.

In his review of A. F. Edwards' (1972) book *Likelihood*, Hacking (1972) gives his main reasons for rejecting the LL:

> We capture enemy tanks at random and note the serial numbers on their engines. We know the serial numbers start at 0001. We capture a tank number 2176. How many did the enemy make? On the likelihood analysis, the best-supported guess is: 2176. Now one can defend this remarkable result by saying that it does not follow that we should estimate the actual number as 2176 only that comparing individual numbers, 2176 is better supported than any larger figure. My worry is deeper. Let us compare the relative likelihood of the two hypotheses, 2176 and 3000. Now pass to a situation where we are measuring, say, widths of a grating in which error has a normal distribution with known variance; we can devise data and a pair of hypotheses about the mean which will have the same log-likelihood ratio. I have no inclination to say that the relative support in the

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

tank case is 'exactly the same as' that in the normal distribution case, even though the likelihood ratios are the same. (pp. 136–7)

Likelihoodists will insist that the law may be upheld by appropriately invoking background information, and by drawing distinctions between evidence, belief, and action.

Royall's Road to Statistical Evidence

Statistician Richard Royall, a longtime leader of Likelihoodist tribes, has had a deep impact on current statistical foundations. His views are directly tied to recent statistical reforms – even if those reformers go Bayesian rather than stopping, like Royall, with comparative likelihoods. He provides what many consider a neat proposal for settling disagreements about statistical philosophy. He distinguishes three questions: belief, action, and evidence:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. How should I interpret this observation as evidence regarding $[H_0]$ versus $[H_1]$? (Royall 1997, p. 4)

Can we line up these three goals to my probabilism, performance, and probativeness (Section 1.2)? No. Probativeness gets no pigeonhole. According to Royall, what to believe is captured by Bayesian posteriors, how to act is captured by a frequentist performance (in some cases he will add costs). What's his answer to the evidence question? The Law of Likelihood.

Let's use one of Royall's first examples, appealing to Bernoulli distributions again – independent, dichotomous trials, "success" or "failure":

Medical researchers are interested in the success probability, θ, associated with a new treatment. They are particularly interested in how θ relates to the old treatment's success probability, believed to be about 0.2. They have reason to hope that θ is considerably greater, perhaps 0.8 or even greater. (Royall 1997, p. 19)

There is a set of possible outcomes, a sample space, S, and a set of possible parameter values, a parameter space Ω. He considers two hypotheses:

$$\theta = 0.2 \text{ and } \theta = 0.8.$$

These are *simple* or *point* hypotheses. To illustrate take a miniature example with only $n = 4$ trials where each can be a "success" $\{X = 1\}$ or a "failure" $\{X = 0\}$. A possible result might be $\boldsymbol{x}_0 = \langle 1,1,0,1 \rangle$. Since $\Pr(X = 1) = \theta$ and $\Pr(X = 0) = (1 - \theta)$, the probability of $\boldsymbol{x}_0$ is $(\theta)(\theta)(1 - \theta)(\theta)$. Given independent trials, they multiply. Under the two hypotheses, given $\langle 1,1,0,1 \rangle$, the likelihoods are

$$\text{Lik}(H_0) = (0.2)(0.2)(0.8)(0.2) = 0.0064,$$

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

$$\text{Lik}(H_1) = (0.8)(0.8)(0.2)(0.8) = 0.1024.$$

A hypothesis that would make the data most probable would be that $\theta = 1$, on the three trials that yield successes, and 0 where it yields failure.

We typically denigrate "just so" stories, purposely erected to fit the data, as "unlikely." Yet they are *most* likely in the technical sense! So in hearing likelihood used formally, you must continually keep this swap of meanings in mind. (We call them Gellerized only if they pass with minimal severity.) If θ is to be constant on each trial, as in the Bernoulli model, the maximum likely hypothesis equates θ with the relative frequency of success, 0.75. [Exercise for reader: find Lik (0.75)]

Exhibit (i): Law of Likelihood Compared to a Significance Test. Here Royall contrasts his handling of the medical example to the standard significance test:

> A standard statistical analysis of their observations would use a *Bernoulli*(θ) statistical model and test the composite hypotheses H_0: $\theta \leq 0.2$ versus H_1: $\theta > 0.2$. That analysis would show that H_0 can be rejected in favor of H_1 at any significance level greater than 0.003, a result that is conventionally taken to mean that the observations are very strong evidence supporting H_1 over H_0. (Royall 1997, p. 19; substituting H_0 and H_1 for H_1 and H_2.)

So the significance tester looks at the composite hypotheses H_0: $\theta \leq 0.2$ vs. H_1: $\theta > 0.2$, rather than his point hypotheses $\theta = 0.2$ and $\theta = 0.8$. Here, she would look at how much larger the mean success rate is in the sample $(X_1 + X_2 + \ldots X_{17})/17$, which we abbreviate as $\overline{x} = 9/17 = 0.53$, compared to what is expected under H_0, put in standard deviation units. Using Royall's numbers, the observed success rate is

$$\overline{x} = 9/17 = .53;$$

$$\sigma = \sqrt{[\theta\,(1-\theta)]}, \text{ which, under the null, is } \sqrt{[0.2\,(0.8)]} = 0.4.$$

The *test statistic* $\text{d}(\boldsymbol{X})$ is $\sqrt{17}(\overline{X} - 0.2)/\sigma$; it gets larger and larger the more the data deviate from what is expected under H_0 – as is sensible for a good test statistic. Its value is

$$\text{d}(\boldsymbol{x}_0) = \sqrt{17}\,(0.53 - 0.2)/\,0.4 \simeq 3.3.$$

The significance level associated with $\text{d}(\boldsymbol{x}_0)$ is

$$\Pr(\text{d}(\boldsymbol{X}) \geq \text{d}(\boldsymbol{x}); H_0) \simeq 0.003.$$

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

This is read, "the probability d($\boldsymbol{X}$) would be at least as large as the particular value d($\boldsymbol{x}_0$), under the supposition that H_0 adequately describes the data generation procedure" (see Souvenir C). It's not strictly a conditional probability – a subtle point that won't detain us here. We continue to follow Royall's treatment, though we'd want to distinguish the mere *indication* of an isolated significant result from strong *evidence*. We'd also have to audit for model assumptions and selection effects, but we assume these check out; after all, Royall's likelihood account also depends on the model holding.

We'd argue along the following lines: were H_0 a reasonable description of the process, then with very high probability you would not be able to regularly produce d($\boldsymbol{x}$) values as large as this:

$$\Pr(\mathrm{d}(\boldsymbol{X}) < \mathrm{d}(\boldsymbol{x}); H_0) \simeq 0.997.$$

So if you manage to get such a large difference, I may infer that $\boldsymbol{x}$ *indicates* a genuine effect. Let's go back to Royall's contrast, because he's very unhappy with this.

Why Does the LL Reject Composite Hypotheses?

Royall tells us that his account is unable to handle composite hypotheses, even this one (for which there is a uniformly most powerful [UMP] test over all points in H_0). He does not conclude that his test comes up short. He and other Likelihoodists maintain that any genuine test or "rule of rejection" should be restricted to comparing the likelihood of H versus some point alternative H' *relative to* fixed data $\boldsymbol{x}$ (Royall 1997, pp. 19–20). It is a virtue. No wonder the Likelihoodist disagrees with the significance tester. In their view, a simple significance test is not a "real" testing account because it is not a comparative appraisal. Elliott Sober, a well-known philosopher of science, echoes Royall: "The fact that significance tests don't contrast the null with alternatives suffices to show that they do not provide a good rule for rejection" (Sober 2008, p. 56). Now, Royall's significance test *has* an alternative H_1: $\theta > 0.2$! It's just not a point alternative but is compound or composite (including all values greater than 0.2). The form of inference, admittedly, is not of the comparative ("evidence favoring") variety. In this discussion, H_0 and H_1 replace his H_1 and H_2.

What untoward consequences occur if we consider composite hypotheses (according to the Likelihoodist)? The problem is that even though the likelihood of $\theta = 0.2$ is small, there are values within alternative H_1: $\theta > 0.2$ that are even less likely on the data $\overline{x} = 0.53$. For instance consider $\theta = 0.9$.

[B]ecause H_0 contains some simple hypotheses that are better supported than some hypotheses in H_1 (e.g., $\theta = 0.2$ is better supported than $\theta = 0.9$ by a likelihood ratio of

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

LR = $(0.2/0.9)^9(0.8/0.1)^8 = 22.2$), the law of likelihood does not allow the characterization of these observations as strong evidence for H_1 over H_0. (Royall 1997, p. 20)

For Royall, rejecting H_0: $\theta \leq 0.2$ and inferring H_1: $\theta > 0.2$ is to assert *every* parameter point within H_1 is more likely than every point in H_0. That seems an idiosyncratic meaning to attach to "infer evidence of $\theta > 0.2$"; but it explains this particular battle. It still doesn't explain the alleged problem for the significance tester who just takes it to mean what it says:

To reject H_0: $\theta \leq 0.2$ is to infer *some* positive discrepancy from 0.2.

We readily agree with Royall that there's a problem with taking a rejection of H_0: $\theta \leq 0.2$, with $\overline{x} = 0.53$, as evidence of a discrepancy as large as $\theta = 0.9$. It's terrible evidence even that θ is as large as 0.7 or 0.8. Here's how a tester articulates this terrible evidence.

Consider the test rule: infer evidence of a discrepancy from 0.2 as large as 0.9, based on observing $\overline{x} = 0.53$. The data differ from 0.2 in the direction of H_1, but to take that difference as indicating an underlying $\theta > 0.9$ would be wrong with probability ~1. Since the standard error of the mean, $\sigma_{\overline{x}}$, is 0.1, alternative 0.9 is more than $3\sigma_{\overline{x}}$ greater than 0.53. ($\sigma_{\overline{x}} = \sigma/\sqrt{n}$) The inference gets low severity.

We'll be touring significance tests and confidence bounds in detail later. We're trying now to extract some core contrasts between error statistical methods and logics of evidence such as the LL. According to the LL, so long as there is a point within H_1 that is less likely given $\boldsymbol{x}$ than is H_0, the data are "evidence *in favor* of the null hypothesis, not evidence *against* it" (Sober 2008, pp. 55–6). He should add "as compared to" some less likely alternative. We never infer a statistical hypothesis according to the LL, but rather a likelihood ratio of two hypotheses, neither of which might be likely. The significance tester and the comparativist hold very different images of statistical inference.

Can an account restricted to comparisons answer the questions: is $\boldsymbol{x}$ good evidence for H? Or is it a case of bad evidence, no test? Royall says no. He declares that all attempts to say whether $\boldsymbol{x}$ is good evidence for H, or even if $\boldsymbol{x}$ is better evidence for H than is $\boldsymbol{y}$, are utterly futile. Similarly, "What *does* the [LL] say when one hypothesis attaches the same probability to two different observations? It says absolutely nothing . . . [it] applies when two different hypotheses attach probabilities to the same observation" (Royall 2004, p. 148). That cuts short important tasks of inferential scrutiny. Since model checking concerns the adequacy of a single model, the Likelihoodist either forgoes such checks or must go beyond the paradigm.

Still, if the model can be taken as adequate, and the Likelihoodist gives a sufficiently long list of comparisons, the differences between us don't seem so marked. Take Royall:

> One statement that we can make is that the observations are only weak evidence in favor of $\theta = 0.8$ versus $\theta = 0.2$ (LR = 4) . . . and at least moderately strong evidence for $\theta = 0.5$ over any value $\theta > 0.8$ (LR) > 22). (1977, p. 20)

Nonetheless, we'd want to ask: what do these numbers mean? Is 22 a lot? Is 4 small? We're back to Hacking's attempt to compare tank cars with widths of a grating. How do we calibrate them? Neyman and Pearson's answer, we'll see, is to look at the probability of so large a likelihood ratio, under various hypotheses, as in (*).

LRs and Posteriors. Royall is loath to add prior probabilities to the assessment of the import of the evidence. This, he says, allows the LR to be "a precise and objective numerical measure of the strength of evidence" in comparing hypotheses (2004, p. 123). At the same time, Royall argues, the LL "constitutes the essential core of the Bayesian account of evidence . . . the Bayesian who rejects the [LL] undermines his own position" (ibid., p. 146). The LR, after all, is the factor by which the ratio of posterior probabilities is changed by the data. Consider just two hypotheses, switching from the ";" in the significance test to conditional probability "|":[1]

$$\Pr(H_0|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)}{\Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)\ +\ \Pr(\boldsymbol{x}|H_1)\ \Pr(H_1)}.$$

Likewise:

$$\Pr(H_1|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|H_1)\ \Pr(H_1)}{\Pr(\boldsymbol{x}|H_1)\ \Pr(H_1)\ +\ \Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)}.$$

The denominators equal $\Pr(\boldsymbol{x})$, so they cancel in the LR:

$$\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} = \frac{\Pr(\boldsymbol{x}|\ H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|\ H_0)\Pr(H_0)}.$$

All of this assumes the likelihoods and the model are deemed adequate.

[1] Divide the numerator and the denominator by $\Pr(\boldsymbol{x}|H_0)\Pr(H_0)$. Then

$$\Pr(H_0|\boldsymbol{x}) = \frac{1}{1\ +\ \frac{\Pr(\boldsymbol{x}|H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|H_0)\Pr(H_0)}}$$

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Data Dredging: Royall Bites the Bullet

Return now to our most serious problem: The Law of Likelihood permits finding evidence in favor of a hypothesis deliberately arrived at using the data, even in the extreme case that it is Gellerized. Allan Birnbaum, who had started out as a Likelihoodist, concludes, "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). But Royall has a clever response. Royall thinks control of error probabilities arises only in answering his second question about action, not evidence. He is prepared to bite the bullet. He himself gives the example of a "trick deck." You've shuffled a deck of ordinary-looking playing cards; you turn over the top card and find an ace of diamonds:

> According to the law of likelihood, the hypothesis that the deck consists of 52 aces of diamonds (H_1) is better supported than the hypothesis that the deck is normal (H_N) [by the factor 52] . . . Some find this disturbing. (Royall 1997, pp. 13–14)

Royall does not. He admits:

> . . . it seems unfair; no matter what card is drawn, the law implies that the corresponding trick-deck hypothesis (52 cards just like the one drawn) is better supported than the normal-deck hypothesis. Thus even if the deck is normal we will always claim to have found strong evidence that it is not. (ibid.)

What he is admitting then is, given any card:

$$\Pr(\text{LR favors trick deck hypothesis; normal deck}) = 1.$$

Even though different trick deck hypotheses would be formed for different outcomes, we may compute the sampling distribution (*). The severity for "trick deck" would be 0. It need not be this extreme to have BENT results, but you get the idea.

What's Royall's way out? At the level of a report on comparative likelihoods, Royall argues, there's no need for a way out. To Royall, it only shows a confusion between evidence and belief.[2] If you're not convinced the deck has 52 aces of diamonds rather than being a normal deck "it does not mean that the observation is not strong evidence in favor of H_1 versus H_N" where H_N is a normal deck (ibid., p. 14). It just wasn't strong enough to overcome your prior beliefs. If you regard the maximally likely alternative as unpalatable, you should have given it a suitably low prior degree of probability. The more likely hypothesis is still favored on grounds of evidence, but your posterior belief

[2] He notes that the comparative evidence for a trick versus a normal deck is not evidence against a normal deck alone (pp. 14–15).

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

may be low. Don't confuse evidence with belief! For the question of evidence, your beliefs have nothing to do with it, according to Royall's Likelihoodist.

What if we grant the Likelihoodist this position? What do we do to tackle the essential challenge to the credibility of statistical inference today, when it's all about Texas Marksmen, hunters, snoopers, and cherry pickers? These moves, which play havoc with a test's ability to control erroneous interpretations, do not alter the evidence at all, say Likelihoodists. The fairest reading of Royall's position might be this: the data indicate only the various LRs. If they are the same, it matters not whether hypotheses arose through data dredging – at least, so long as you are in the category of "what the data say." As soon as you're troubled, you slip into the category of belief. What if we're troubled by the ease of exaggerating findings when you're allowed to rummage around? What if we wish to clobber the Texas sharpshooter method, never mind my beliefs in the particular claims they infer. You might aver, we should never be considering trick deck hypotheses, but this is the example Royall gives, and he is a, if not the, leading Likelihoodist.

To him, appealing to error probabilities is relevant only pre-data, which wouldn't trouble the severe tester so much if Likelihoodists didn't regard them as relevant only for a performance goal, not inference. Given that frequentists have silently assented to the performance use of error probabilities, it's perhaps not surprising that others accept this. The problem with cherry picking is not about long runs, it's that a poor job has been done in the case at hand. The severity requirement reflects this intuition. By contrast, Likelihoodists hold that likelihood ratios, and unadjusted *P*-values, still convey what the data say, even with claims arrived at through data dredging. It's true you can explore, arrive at H, then test H on other data; but isn't the reason there's a need to test on new data that your assessment will otherwise fail to convey how well tested H is?

Downsides to the "Appeal to Beliefs" Solution to Inseverity

What's wrong with Royall's appeal to prior beliefs to withhold support to a "just so" hypothesis? It may get you out of a jam in some cases. Here's why the severe tester objects. First, she insists on distinguishing the *evidential* warrant for one and the same hypothesis H in two cases: one where it was constructed post hoc, cherry picked, and so on, a second where it was predesignated. A cherry-picked hypothesis H could well be believable, but we'd still want to distinguish the evidential credit H deserves in the two cases. Appealing to priors can't help, since here there's one and the same H.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Perhaps someone wants to argue that the mode of testing alters the degree of belief in *H*, but this would be non-standard (violating the Likelihood Principle to be discussed shortly). Philosopher Roger Rosenkrantz puts it thus: The LL entails the irrelevance "of whether the theory was formulated in advance or suggested by the observations themselves" (Rosenkrantz 1977, p. 121). For Rosenkrantz, a default Bayesian last I checked, this irrelevance of predesignation is altogether proper. By contrast, he admits, "Orthodox (non-Bayesian) statisticians have found this to be strong medicine indeed!" (ibid.). Many might say instead that it is bad medicine. Take, for instance, something called the CONSORT, the Consolidated Standards of Reporting Trials from RCTs in medicine:

> Selective reporting of outcomes is widely regarded as misleading. It undermines the validity of findings, particularly when driven by statistical significance or the direction of the effect [4], and has memorably been described in the New England Journal of Medicine as "Data Torturing" [5]. (COMpare Team 2015)

This gets to a second problem with relying on beliefs to block data-dredged hypotheses. Post-data explanations, even if it took a bit of data torture, are often incredibly convincing, and you don't have to be a sleaze to really believe them. Goldacre (2016) expresses shock that medical journals continue to report outcomes that were altered post-data – he calls this *outcome-switching*. Worse, he finds, some journals defend the practice because they are convinced that their very good judgment entitles them to determine when to treat post-designated hypotheses as if they were predesignated. Unlike the LL, the CONSORT and many other best practice guides view these concerns as an essential part of reporting what the data say. Now you might say this is just semantics, as long as, in the end, they report that outcome-switching occurred. Maybe so, provided the report mentions why it would be misleading to hide the information. At least people have stopped referring to frequentist statistics as "Orthodox."

There is a third reason to be unhappy with supposing the only way to block evidence for "just so" stories is by the *deus ex machina* of a low prior degree of belief: it misidentifies what the problem really *is*. The influence of the biased selection is not on the believability of *H* but rather on the capability of the test to have unearthed errors. The error probing capability of the testing procedure is being diminished. If you engage in cherry picking, you are not "sincerely trying," as Popper puts it, to find flaws with claims, but instead you are finding evidence in favor of a well-fitting hypothesis that you deliberately construct – barred only if your intuitions say it's unbelievable. The job that was supposed to be accomplished by an account of statistics now has to be performed by *you*. Yet you are the one most likely to follow your preconceived opinions, biases, and pet

theories. If an account of statistical inference or evidence doesn't supply self-critical tools, it comes up short in an *essential* way. So says the severe tester.

Souvenir B: Likelihood versus Error Statistical

Like pamphlets from competing political parties, the gift shop from this tour proffers pamphlets from these two perspectives.

To the Likelihoodist, points in favor of the LL are:

- The LR offers "a precise and objective numerical measure of the strength of statistical evidence" for one hypotheses over another; it is a frequentist account and does not use prior probabilities (Royall 2004, p. 123).
- The LR is fundamentally related to Bayesian inference: the LR is the factor by which the ratio of posterior probabilities is changed by the data.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike *P*-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

To the error statistician, problems with the LL include:

- LRs do not convey the same evidential appraisal in different contexts.
- The LL denies it makes sense to speak of how well or poorly tested a single hypothesis is on evidence, essential for model checking; it is inapplicable to composite hypothesis tests.
- A Likelihoodist account does not consider outcomes other than the one observed, unlike *P*-values, and Type I and II errors. (Irrelevance of the sample space.)
- Fishing for maximally fitting hypotheses and other gambits that alter error probabilities do not affect the assessment of evidence; they may be blocked by moving to the "belief" category.

Notice, the last two points are identical for both. What's a selling point for a Likelihoodist is a problem for an error statistician.

1.5 Trying and Trying Again: The Likelihood Principle

> The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. (Edwards, Lindman, and Savage 1963, p. 193)

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Several well-known gambits make it altogether easy to find evidence in support of favored claims, even when they are unwarranted. A responsible statistical inference report requires information about whether the method used is capable of controlling such erroneous interpretations of data or not. Now we see that adopting a statistical inference account is also to buy into principles for processing data, hence criteria for "what the data say," hence grounds for charging an inference as illegitimate, questionable, or even outright cheating. The best way to survey the landscape of statistical debates is to hone in on some pivotal points of controversy – saving caveats and nuances for later on.

Consider for example the gambit of "trying and trying again" to achieve statistical significance, stopping the experiment only when reaching a nominally significant result. Kosher, or not? Suppose somebody reports data showing a statistically significant effect, say at the 0.05 level. Would it matter to your appraisal of the evidence if you found out that each time they failed to find significance, they went on to collect more data, until finally they did? A rule for when to stop sampling is called a *stopping rule*.

The question is generally put by considering a random sample $\boldsymbol{X}$ that is Normally distributed with mean μ and standard deviation $\sigma = 1$, and we are testing the hypotheses:

$$H_0\text{: } \mu = 0 \text{ against } H_1\text{: } \mu \neq 0.$$

This is a two-sided test: a discrepancy in either direction is sought. (The details of testing are in Excursions 3 and thereafter.) To ensure a significance level of 0.05, H_0 is rejected whenever the sample mean differs from 0 by more than $1.96\sigma/\sqrt{n}$, and, since $\sigma = 1$, the rule is: Declare $\boldsymbol{x}$ is statistically significant at the 0.05 level whenever $|\overline{X}| > 1.96/\sqrt{n}$. However, instead of fixing the sample size in advance, n is determined by the optional stopping rule:

$$\textit{Optional stopping rule:} \text{ keep sampling until } |\overline{X}| \geq (1.96/\sqrt{n}).$$

Equivalently, since the test statistic $\mathrm{d}(\boldsymbol{X}) = (\overline{X} - 0)/\sqrt{n}$:

$$\text{Keep sampling until } |\mathrm{d}(\boldsymbol{X})| \geq 1.96.$$

Our question was: would it be relevant to your evaluation of the evidence if you learned she'd planned to keep running trials until reaching 1.96? Having failed to rack up a 1.96 difference after, say, 10 trials, she goes on to 20, and failing yet again, she goes to 30 and on and on until finally, say, on trial 169 she gets a 1.96 difference. Then she stops and declares the statistical significance is ~0.05.

This is an example of what's called a *proper stopping rule*: the probability it will stop in a finite number of trials is 1, regardless of the true value of μ. Thus, in one of the most seminal papers in statistical foundations, by Ward Edwards,

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Harold Lindman, and Leonard (Jimmie) Savage (E, L, & S) tell us, "if an experimenter uses this procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true" (1963, p. 239). Understandably, they observe, the significance tester frowns on optional stopping, or at least requires the auditing of the *P*-value to require an adjustment. Had *n* been fixed, the significance level would be 0.05, but with optional stopping it increases.

Imagine instead if an account advertised itself as ignoring stopping rules. What if an account declared:

> 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. (ibid., pp. 238–9)

I've been teasing you, because these same authors who warn that to ignore stopping rules is to guarantee rejecting the null hypothesis even if it's true are the individuals who tout the irrelevance of stopping rules in the above citation – E, L, & S. They call it the *Stopping Rule Principle*. Are they contradicting themselves?

No. It is just that what looks to be, and indeed is, cheating from the significance testing perspective is not cheating from these authors' Bayesian perspective. "[F]requentist test results actually depend not only on what $\boldsymbol{x}$ was observed, but on how the experiment was stopped" (Carlin and Louis 2008, p. 8). Yes, but shouldn't they? Take a look at Table 1.1: by the time one reaches 50 trials, the probability of attaining a nominally significant 0.05 result is not 0.05 but 0.32. The actual or overall significance level is the probability of finding a 0.05 nominally significant result at some stopping point *or other*, up to the point it stops. The actual significance level accumulates.

Well-known statistical critics from psychology, Joseph Simmons, Leif Nelson, and Uri Simonsohn, place at the top of their list of requirements the need to block flexible stopping: "Researchers often decide when to stop data collection on the basis of interim data analysis . . . many believe this practice exerts no more than a trivial influence on false-positive rates" (Simmons et al. 2011, p. 1361). "Contradicting this intuition" they show the probability of erroneous rejections balloons. "A researcher who starts with 10 observations per condition and then tests for significance after every new . . . observation finds a significant effect 22% of the time" erroneously (ibid., p. 1362). Yet the followers of the Stopping Rule Principle deny it makes a difference to evidence. On their account, it *doesn't*. It's easy to see why there's disagreement.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Table 1.1 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

The Likelihood Principle

By what magic can such considerations disappear? One way to see the vanishing act is to hold, with Royall, that "what the data have to say" is encompassed in likelihood ratios. This is the gist of a very important principle of evidence, the *Likelihood Principle* (LP). Bayesian inference requires likelihoods plus prior probabilities in hypotheses; but the LP has long been regarded as a crucial part of their foundation: to violate it is to be *incoherent* Bayesianly. Disagreement about the LP is a pivot point around which much philosophical debate between frequentists and Bayesians has turned. Here is a statement of the LP:

> According to Bayes's Theorem, $\Pr(\boldsymbol{x}|\mu)$ ... constitutes the entire evidence of the experiment, that is it tells all that the experiment has to tell. More fully and more precisely, if $\boldsymbol{y}$ is the datum of some other experiment, and if it happens that $\Pr(\boldsymbol{x}|\mu)$ and $\Pr(\boldsymbol{y}|\mu)$ are proportional functions of μ (that is constant multiples of each other), then each of the two data $\boldsymbol{x}$ and $\boldsymbol{y}$ have exactly the same thing to say about the value of μ ... (Savage 1962, p. 17; replace λ with μ)

Some go further and claim that if $\boldsymbol{x}$ and $\boldsymbol{y}$ give the same likelihood, "they should give the same inference, analysis, conclusion, decision, action or anything else" (Pratt et al. 1995, p. 542). Does the LP entail the LL? No. Bayesians, for

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

example, generally hold to the LP, but would insist on priors that go beyond the LL. Even the converse may be denied (according to Hacking) but this is not of concern to us.

Weak Repeated Sampling Principle. For sampling theorists (my error statisticians), by contrast, this example "taken in the context of examining consistency with $\theta = 0$, is enough to refute the strong likelihood principle" (Cox 1978, p. 54), since, with probability 1, it will stop with a "nominally" significant result even though $\theta = 0$. It contradicts what Cox and Hinkley call "the weak repeated sampling principle" (Cox and Hinkley 1974, p. 51). "[W]e should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time" (ibid., pp. 45–6).

For Cox and Hinkley, to report a 1.96 standard deviation difference from optional stopping just the same as if the sample size had been fixed, is to discard relevant information for inferring inconsistency with the null, while "according to any approach that is in accord with the strong likelihood principle, the fact that this particular stopping rule has been used is irrelevant" (ibid., p. 51). What they call the "strong" likelihood principle will just be called the LP here. (A weaker form boils down to sufficiency, see Excursion 3.)

Exhibit (ii): How Stopping Rules Drop Out. Our question remains: by what magic can such considerations disappear? Formally, the answer is straightforward. Consider two versions of the above experiment: In the first, 1.96 is reached via fixed sample size (n = 169); in the second, by means of optional stopping that ended at 169. While $\mathrm{d}(\boldsymbol{x}) = \mathrm{d}(\boldsymbol{y})$, because of the stopping rule, the likelihood of $\boldsymbol{y}$ differs from that of $\boldsymbol{x}$ by a constant k, that is,

$$\Pr(\boldsymbol{x}|H_i) = k\Pr(\boldsymbol{y}|H_i) \text{ for constant } k.$$

Given that likelihoods enter as ratios, such proportional likelihoods are often said to be the "same." Now suppose inference is by Bayes' Theorem. Since likelihoods enter as ratios, the constant k drops out. This is easily shown. I follow E, L, & S; p. 237.

For simplicity, suppose the possible hypotheses are exhausted by two, H_0 and H_1, neither with probability of 0.

To show $\Pr(H_0|\boldsymbol{y}) = \Pr(H_0|\boldsymbol{x})$:

(1) We are given the proportionality of likelihoods, for an arbitrary value of k:

$$\Pr(\boldsymbol{y}|H_0) = k\Pr(\boldsymbol{x}|H_0),$$

$$\Pr(\boldsymbol{y}|H_1) = k\Pr(\boldsymbol{x}|H_1).$$

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

(2) By definition:

$$\Pr(H_0|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{y}|H_0)\Pr(H_0)}{\Pr(\boldsymbol{y})}.$$

The denominator $\Pr(\boldsymbol{y}) = \Pr(\boldsymbol{y}|H_0)\ \Pr(H_0) + \Pr(\boldsymbol{y}|H_1)\ \Pr(H_1)$.

Now substitute for each term in (2) the proportionality claims in (1). That is, replace $\Pr(\boldsymbol{y}|H_0)$ with $k\Pr(\boldsymbol{x}|H_0)$ and $\Pr(\boldsymbol{y}|H_1)$ with $k\Pr(\boldsymbol{x}|H_1)$.

(3) The result is

$$\Pr(H_0|\boldsymbol{y}) = \frac{k\Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)}{k\Pr(\boldsymbol{x})} = \Pr(H_0|\boldsymbol{x}).$$

The posterior probabilities are the same whether the 1.96 result emerged from optional stopping, $\boldsymbol{Y}$, or fixed sample size, $\boldsymbol{X}$.

This essentially derives the LP from inference by Bayes' Theorem, and shows the equivalence for the particular case of interest, optional stopping. As always, when showing a Bayesian computation I use the conditional probability "|" rather than the ";" of the frequentist.[3]

The 1959 Savage Forum: What Counts as Cheating?

My colleague, well-known Bayesian I. J. Good, would state it as a "paradox":

> [I]f a Fisherian is prepared to use optional stopping (which usually he is not) he can be sure of rejecting a true null hypothesis provided that he is prepared to go on sampling for a long time. The way I usually express this 'paradox' is that a Fisherian [but not a Bayesian] can cheat by pretending he has a plane to catch like a gambler who leaves the table when he is ahead. (Good 1983, p. 135)

The lesson about who is allowed to cheat depends on your statistical philosophy. Error statisticians require that the overall and not the "computed" significance level be reported. To them, cheating would be to report the significance level you got after trying and trying again in just *the same way* as if the test had a fixed sample size (Mayo 1996, p. 351). Viewing statistical methods as tools for severe tests, rather than as probabilistic logics of evidence, makes a deep difference to the tools we seek. Already we find ourselves thrust into some of the knottiest and most intriguing foundational issues.

This is Jimmie Savage's message at a 1959 forum deemed sufficiently important to occupy a large gallery of the Museum of Statistics (hereafter "The Savage Forum" (Savage 1962)). Attendees include Armitage, Barnard,

[3] $\Pr(\boldsymbol{x}) = \Pr(\boldsymbol{x}\ \&\ H_0) + \Pr(\boldsymbol{x}\ \&\ H_1)$, where H_0 and H_1 are exhaustive.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Bartlett, Cox, Good, Jenkins, Lindley, Pearson, Rubin, and Smith. Savage announces to this eminent group of statisticians that if adjustments in significance levels are required for optional stopping, which they are, then the fault must be with significance levels. Not all agreed. Needling Savage on this issue, was Peter Armitage:

> 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)

Armitage, an expert in sequential trials in medicine, is fully in favor of them, but he thinks stopping rules should be reflected in overall inferences. He goes further:

> [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? (ibid., p. 72)

He has in mind using a type of uniform prior probability for μ, wherein the posterior for the null hypothesis matches the significance level. (We return to this in Excursion 6. For $\sigma = 1$, its distribution is Normal($\overline{x}$, $1/n$).)

Not all cases of trying and trying again injure error probabilities. Think of trying and trying again until you find a key that fits a lock. When you stop, there's no probability of being wrong. (We return to this in Excursion 4.)

Savage's Sleight of Hand

Responding to Armitage, Savage engages in a bit of sleight of hand. Moving from the problematic example to one of two predesignated point hypotheses, H_0: $\mu = \mu_0$, and H_1: $\mu = \mu_1$, he shows that the error probabilities are controlled in that case. In particular, the probability of obtaining a result that makes H_1 r times more likely than H_0 is less than $1/r$: $\Pr(\text{LR} > r; H_0) < 1/r$. But, that wasn't Armitage's example; nor does Savage return to it. Now, it is open to Likelihoodists to resist being saddled "with ideas that are alien to them" (Sober 2008, p. 77). Since the Likelihoodist keeps to this type of comparative appraisal, they can set bounds to the probabilities of error. However, the bounds are no longer impressively small as we add hypotheses, even if they are predesignated[4] (Mayo and Kruse 2001).

[4] A general result, stated in Kerridge (1963, p. 1109), is that with k simple hypotheses, where H_0 is true and $H_1, \ldots, H_{k-1}$ are false, and equal priors, "*the frequency with which, at the termination of sampling the posterior probability of the true hypothesis is* p *or less cannot exceed* $(k-1)p/(1-p)$." Such bounds depend on having countably additive probability, while the uniform prior in Armitage's example imposes finite additivity.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Something more revealing is going on when the Likelihoodist sets pre-data bounds. Why the sudden concern with showing the rule for comparative evidence would very improbably find evidence in favor of the wrong hypothesis? This is an error probability. So it appears they also care about error probabilities – at least before-trial – or they are noting, for those of us who do, that they also have error control in the simple case of predesignated point hypotheses. The severe tester asks: If you want to retain these pre-data safeguards, why allow them to be spoiled by data-dependent hypotheses and stopping rules?

Some have said: the evidence is the same, but you take into account things like stopping rules and data-dependent selections *afterwards*. When making an inference, this *is* afterwards, and we need an epistemological rationale to pick up on their influences *now*. Perhaps knowing someone uses optional stopping warrants a high belief he's trying to deceive you, leading to a high enough prior belief in the null. Maybe so, but this is to let priors reflect methods in a non-standard way. Besides, Savage (1961, p. 583) claimed optional stopping "is no sin," so why should it impute deception? So far as I know, subjective Bayesians have resisted the idea that rules for stopping alter the prior. Couldn't you pack the concern in some background *B*? You could, but you would need another account to justify doing so, thereby only pushing back the issue. I've discussed an assortment of attempts elsewhere: Mayo (1996), Mayo and Kruse (2001), Mayo (2014b). Others have too, discussed here and elsewhere; please see our online sources (preface).

Arguments from Intentions: All in Your Head?

A funny thing happened at the Savage Forum: George Barnard announces he no longer holds the LP for the two-sided test under discussion, only for the predesignated point alternatives. Savage is shocked to hear it:

> I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right. (Savage 1962, p. 76)

The argument Barnard gave him was that the plan for when to stop was a matter of the researchers' intentions, all wrapped up in their heads. While Savage denies he was ever sold on the argument from intentions, it's a main complaint you will hear about taking account, not just of stopping rules, but of error probabilities in general. Take the subjective Bayesian philosophers Howson and Urbach (1993):

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 experimenters, embodying their stopping rule. It seems to us that this fact precludes a significance test delivering any kind of judgment about empirical support. (p. 212)

The truth is, whether they're hidden or not turns on your methodology being able to pick up on them. So the deeper question is: *ought* your account pick up on them?

The answer isn't a matter of mathematics, it depends on your goals and perspective – yes on your philosophy of statistics. Ask yourself: What features lead you to worry about cherry picking, and selective reporting? Why do the CONSORT and myriad other best practice manuals care? Looking just at the data and hypotheses – as a "logic" of evidence would – you will not see the machinations. Nevertheless, these machinations influence the capabilities of the tools. Much of the handwringing about irreproducibility is the result of wearing blinders as to the construction and selection of both hypotheses and data. In one sense, all test specifications are determined by a researcher's intentions; that doesn't make them private or invisible to us. They're visible to accounts with antennae to pick up on them!

You might try to deflect the criticism of stopping rules by pointing out that some stopping rules do alter priors. Armitage wasn't ignoring that, nor are we. These are called informative stopping rules, and examples are rather contrived. For instance, "a man who wanted to know how frequently lions watered at a certain pool was chased away by lions" (E, L, & S 1963, p. 239). They add, "we would not give a facetious example had we been able to think of a serious one." In any event, this is irrelevant for the Armitage example, which is non-informative.

Error Probabilities Violate the LP

> [I]t seems very strange that a frequentist could not analyze a given set of data, such as $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ if the stopping rule is not given . . . [D]ata should be able to speak for itself. (Berger and Wolpert 1988, p. 78)

Inference by Bayes' Theorem satisfies this intuition, which sounds appealing; but for our severe tester, data no more speak for themselves in the case of stopping rules than with cherry picking, hunting for significance, and the like. We may grant to the Bayesian that

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

[The] 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). (E, L, & S 1963, p. 239)

The question is whether this latitude is desirable. If you are keen to use statistical methods critically, as our severe tester, you'll be suspicious of a simplicity and freedom to mislead.

Admittedly, this should have been more clearly spelled out by Neyman and Pearson. They rightly note:

In order to fix a limit between 'small' and 'large' values of [the likelihood ratio] we must know how often such values appear when we deal with a true hypothesis. (Pearson and Neyman 1930, p. 106)

That's true, but putting it in terms of the desire "to control the error involved in rejecting a true hypothesis" it is easy to dismiss it as an affliction of a frequentist concerned only with long-run performance. Bayesians and Likelihoodists are free of this affliction. Pearson and Neyman should have said: ignoring the information as to how readily true hypotheses are rejected, we cannot determine if there really is evidence of inconsistency with them.

Our minimal requirement for evidence insists that data only provide genuine or reliable evidence for H if H survives a severe test – a test H would probably have failed if false. Here the hypothesis H of interest is the non-null of Armitage's example: the existence of a genuine effect. A warranted inference to H depends on the test's ability to find H false when it is, i.e., when the null hypothesis is true. The severity conception of tests provides the link between a test's error probabilities and what's required for a warranted inference.

The error probability computations in significance levels, confidence levels, power, all depend on violating the LP! Aside from a concern with "intentions," you will find two other terms used in describing the use of error probabilities: a concern with (i) outcomes other than the one observed, or (ii) the sample space. Recall Souvenir B, where Royall, who obeys the LP, speaks of "the irrelevance of the sample space" once the data are in hand. It's not so obvious what's meant. To explain, consider Jay Kadane: "Significance testing violates the Likelihood Principle, which states that, having observed the data, inference must rely only on what happened, and not on what might have happened but did not" (Kadane 2011, p. 439). According to Kadane, the probability statement: $\Pr(|d(\boldsymbol{X})| > 1.96) = 0.05$ "is a statement about $d(\boldsymbol{X})$ before it is observed. After it is observed, the event $\{d(\boldsymbol{X}) > 1.96\}$ either

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

happened or did not happen and hence has probability either one or zero" (ibid.).

Knowing $d(\boldsymbol{x}) = 1.96$, Kadane is saying there's no more uncertainty about it. But would he really give it probability 1? That's generally thought to invite the problem of "known (or old) evidence" made famous by Clark Glymour (1980). If the probability of the data $\boldsymbol{x}$ is 1, Glymour argues, then $\Pr(\boldsymbol{x}|H)$ also is 1, but then $\Pr(H|\boldsymbol{x}) = \Pr(H)\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}) = \Pr(H)$, so there is no boost in probability given $\boldsymbol{x}$. So does that mean known data don't supply evidence? Surely not. Subjective Bayesians try different solutions: either they abstract to a context prior to knowing $\boldsymbol{x}$, or view the known data as an instance of a general type, in relation to a sample space of outcomes. Put this to one side for now in order to continue the discussion.[5]

Kadane is emphasizing that Bayesian inference is *conditional* on the particular outcome. So once $\boldsymbol{x}$ is known and fixed, other possible outcomes that could have occurred but didn't are irrelevant. Recall finding that Pickrite's procedure was to build k different portfolios and report just the one that did best. It's as if Kadane is asking: "Why are you considering other portfolios that you might have been sent but were not, to reason from the one that you got?" Your answer is: "Because that's how I figure out whether your boast about Pickrite is warranted." With the "search through k portfolios" procedure, the possible outcomes are the success rates of the k different attempted portfolios, each with its own null hypothesis. The actual or "audited" P-value is rather high, so the severity for H: Pickrite has a reliable strategy, is low $(1 - p)$. For the holder of the LP to say that, once $\boldsymbol{x}$ is known, we're not allowed to consider the other chances they gave themselves to find an impressive portfolio, is to put the kibosh on a crucial way to scrutinize the testing process.

Interestingly, nowadays, non-subjective or default Bayesians concede they "have to live with some violations of the likelihood and stopping rule principles" (Ghosh, Delampady, and Samanta 2010, p. 148) since their prior probability distributions are influenced by the sampling distribution. Is it because ignoring stopping rules can wreak havoc with the well-testedness of inferences? If that is their aim, too, then that is very welcome. Stay tuned.

[5] Colin Howson, a long-time subjective Bayesian, has recently switched to being a non-subjective Bayesian at least in part because of the known evidence problem (Howson 2017, p. 670).

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

Souvenir C: A Severe Tester's Translation Guide

Just as in ordinary museum shops, our souvenir literature often probes treasures that you didn't get to visit at all. Here's an example of that, and you'll need it going forward. There's a confusion about what's being done when the significance tester considers the set of all of the outcomes leading to a $d(\boldsymbol{x})$ greater than or equal to 1.96, i.e., $\{\boldsymbol{x}: d(\boldsymbol{x}) \geq 1.96\}$, or just $d(\boldsymbol{x}) \geq 1.96$. This is generally viewed as throwing away the particular $\boldsymbol{x}$, and lumping all these outcomes together. What's really happening, according to the severe tester, is quite different. What's actually being signified is that we are interested in the method, not just the particular outcome. Those who embrace the LP make it very plain that data-dependent selections and stopping rules drop out. To get them to drop in, we signal an interest in what the test procedure *would have* yielded. This is a counterfactual and is altogether essential in expressing the properties of the method, in particular, the probability it would have yielded some nominally significant outcome *or other*.

When you see $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x_0}); H_0)$, or $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x_0}); H_1)$, for any particular alternative of interest, insert:

> "the test procedure would have yielded"

just before the $d(\boldsymbol{X})$. In other words, this expression, with its inequality, is a signal of interest in, and an abbreviation for, the error probabilities associated with a test.

Applying the Severity Translation. In Exhibit (i), Royall described a significance test with a Bernoulli(θ) model, testing H_0: $\theta \leq 0.2$ vs. H_1: $\theta > 0.2$. We blocked an inference from observed difference $d(\boldsymbol{x}) = 3.3$ to $\theta = 0.8$ as follows. (Recall that $\overline{x} = 0.53$ and $d(\boldsymbol{x}_0) \simeq 3.3$.)

> *We computed* $\Pr(d(\boldsymbol{X}) > 3.3; \theta = 0.8) \simeq 1$.
>
> *We translate it as* $\Pr(\text{The test would yield } d(\boldsymbol{X}) > 3.3; \theta = 0.8) \simeq 1$.

We then reason as follows:

> *Statistical inference*: If $\theta = 0.8$, then the method would virtually always give a difference larger than what we observed. Therefore, the data indicate $\theta < 0.8$.

(This follows for rejecting H_0 in general.) When we ask: "How often would your test have found such a significant effect even if H_0 is approximately true?" we are asking about the properties of the experiment that *did* happen.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

The counterfactual "would have" refers to how the procedure would behave in general, not just with these data, but with other possible data sets in the sample space.

Exhibit (iii). Analogous situations to the optional stopping example occur even without optional stopping, as with selecting a data-dependent, maximally likely, alternative. Here's an example from Cox and Hinkley (1974, 2.4.1, pp. 51–2), attributed to Allan Birnbaum (1969).

A single observation is made on $\boldsymbol{X}$, which can take values 1,2, . . .,100. "There are 101 possible distributions conveniently indexed by a parameter θ taking values 0, 1,..., 100" (ibid.). We are not told what θ is, but there are 101 possible point hypotheses about the value of θ: from 0 to 100. If $\boldsymbol{X}$ is observed to be $\boldsymbol{r}$, written $\boldsymbol{X} = \boldsymbol{r}$ ($\boldsymbol{r} \neq 0$), then the most likely hypothesis is $\theta = \boldsymbol{r}$: in fact, $\Pr(\boldsymbol{X} = \boldsymbol{r}; \theta = \boldsymbol{r}) = 1$. By contrast, $\Pr(\boldsymbol{X} = \boldsymbol{r}; \theta = 0) = 0.01$. Whatever value $\boldsymbol{r}$ that is observed, hypothesis $\theta = \boldsymbol{r}$ is 100 times as likely as is $\theta = 0$. Say you observe $\boldsymbol{X} = 50$, then H: $\theta = 50$ is 100 times as likely as is $\theta = 0$. So "even if in fact $\theta = 0$, we are certain to find evidence apparently pointing strongly against $\theta = 0$, if we allow comparisons of likelihoods chosen in the light of the data" (Cox and Hinkley 1974, p. 52). This does not happen if the test is restricted to two preselected values. In fact, if $\theta = 0$ the probability of a ratio of 100 in favor of the false hypothesis is 0.01.[6]

Allan Birnbaum gets the prize for inventing chestnuts that deeply challenge both those who do, and those who do not, hold the Likelihood Principle!

Souvenir D: Why We Are So New

What's Old? You will hear critics say that the reason to overturn frequentist, sampling theory methods – all of which fall under our error statistical umbrella – is that, well, they've been around a long, long time. First, they are scarcely stuck in a time warp. They have developed with, and have often been the source of, the latest in modeling, resampling, simulation, Big Data, and machine learning techniques. Second, all the methods have roots in long-ago ideas. Do you know what is really up-to-the-minute in this time of massive, computer algorithmic methods and "trust me" science? A new vigilance about retaining hard-won error control techniques. Some thought that, with enough data, experimental design

[6] From Cox and Hinkley 1974, p. 51. The likelihood function corresponds to the normal distribution of $\overline{X}$ around μ with SE $\sigma/\sqrt{n}$. The likelihood at $\mu = 0$ is $\exp(-0.5k^2)$ times that at $\mu = \bar{x}$. One can choose k to make the ratio small. "That is, even if in fact $\mu = 0$, there always appears to be strong evidence against $\mu = 0$, at least if we allow comparison of the likelihood at $\mu = 0$ against any value of μ and hence in particular against the value of μ giving maximum likelihood". However, if we confine ourselves to comparing the likelihood at $\mu = 0$ with that at some fixed $\mu = \mu'$, this difficulty does not arise.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press

could be ignored, so we have a decade of wasted microarray experiments. To view outcomes other than what you observed as irrelevant to what x_0 says is also at odds with cures for irreproducible results. When it comes to cutting-edge fraud-busting, the ancient techniques (e.g., of Fisher) are called in, refurbished with simulation.

What's really old and past its prime is the idea of a logic of inductive inference. Yet core discussions of statistical foundations today revolve around a small cluster of (very old) arguments based on that vision. Tour II took us to the crux of those arguments. Logics of induction focus on the relationships between given data and hypotheses – so outcomes other than the one observed drop out. This is captured in the Likelihood Principle (LP). According to the LP, trying and trying again makes no difference to the probabilist: it is what someone intended to do, locked up in their heads.

> It is interesting that frequentist analyses often need to be adjusted to account for these 'looks at the data,'. . . That Bayesian analysis claims no need to adjust for this 'look elsewhere' effect – called the *stopping rule principle* – has long been a controversial and difficult issue. . . (J. Berger 2008, p. 15)

The irrelevance of optional stopping is an asset for holders of the LP. For the task of criticizing and debunking, this puts us in a straightjacket. The warring sides talk past each other. We need a new perspective on the role of probability in statistical inference that will illuminate, and let us get beyond, this battle.

New Role of Probability for Assessing What's Learned. A passage to locate our approach within current thinking is from Reid and Cox (2015):

> Statistical theory continues to focus on the interplay between the roles of probability as representing physical haphazard variability . . . and as encapsulating in some way, directly or indirectly, aspects of the uncertainty of knowledge, often referred to as epistemic. (p. 294)

> We may avoid the need for a different version of probability by appeal to a notion of calibration, as measured by the behavior of a procedure under hypothetical repetition. That is, we study assessing uncertainty, as with other measuring devices, by assessing the performance of proposed methods under hypothetical repetition. Within this scheme of repetition, probability is defined as a hypothetical frequency. (p. 295)

This is an ingenious idea. Our meta-level appraisal of methods proceeds this way too, but with one important difference. A key question for us is the proper epistemic role for probability. It is standardly taken as providing a probabilism, as an assignment of degree of actual or rational belief in a claim, absolute or comparative. We reject this. We proffer an alternative theory: a severity assessment. An account of what is warranted and unwarranted to infer – a normative epistemology – is not a matter of using probability to assign rational beliefs, but to control and assess how well probed claims are.

If we keep the presumption that the epistemic role of probability is a degree of belief of some sort, then we can "avoid the need for a different version of probability" by supposing that good/poor performance of a method warrants high/low belief in the method's output. Clearly, poor performance is a problem, but I say a more nuanced construal is called for. The idea that partial or imperfect knowledge is all about degrees of belief is handed down by philosophers. Let's be philosophical enough to challenge it.

New Name? An error statistician assesses inference by means of the error probabilities of the method by which the inference is reached. As these stem from the sampling distribution, the conglomeration of such methods is often called "sampling theory." However, sampling theory, like classical statistics, Fisherian, Neyman–Pearsonian, or frequentism are too much associated with hardline or mish-mashed views. Our job is to clarify them, but in a new way. Where it's apt for taking up discussions, we'll use "frequentist" interchangeably with "error statistician." However, frequentist error statisticians tend to embrace the long-run performance role of probability that I find too restrictive for science. In an attempt to remedy this, Birnbaum put forward the "confidence concept" (Conf), which he called the "one rock in a shifting scene" in statistical thinking and practice. This "one rock," he says, takes from the Neyman–Pearson (N-P) approach "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p.1033). Extending his notion to a composite alternative:

> Conf: An adequate concept of statistical evidence should find strong evidence against H_0 (for $\sim H_0$) with small probability α when H_0 is true, and with much larger probability $(1 - \beta)$ when H_0 is false, increasing as discrepancies from H_0 increase.

This is an entirely right-headed pre-data performance requirement, but I agree with Birnbaum that it requires a reinterpretation for evidence post-data (Birnbaum 1977). Despite hints and examples, no such evidential interpretation has been given. The switch that I'm hinting at as to what's required for an evidential or epistemological assessment is key. Whether one uses a frequentist or a propensity interpretation of error probabilities (as Birnbaum did) is not essential. *What we want is an error statistical approach that controls and assesses a test's stringency or severity*. That's not much of a label. For short, we call someone who embraces such an approach a severe tester. For now I will just venture that a severity scrutiny illuminates all statistical approaches currently on offer.

https://doi.org/10.1017/9781107286184.003 Published online by Cambridge University Press