Tour II It's The Methods, Stupid

> There is perhaps in current literature a tendency to speak of the Neyman–Pearson contributions as some static system, rather than as part of the historical process of development of thought on statistical theory which is and will always go on. (Pearson 1962, p. 276)

This goes for Fisherian contributions as well. Unlike museums, we won't remain static.

The lesson from Tour I of this Excursion is that Fisherian and Neyman–Pearsonian tests may be seen as offering clusters of methods appropriate for different contexts within the large taxonomy of statistical inquiries. There is an overarching pattern:

> Just as with the use of measuring instruments, applied to the specific case, we employ the performance features to make inferences about aspects of the particular thing that is measured, aspects that the measuring tool is appropriately capable of revealing. (Mayo and Cox 2006, p. 84)

This information is used to ascertain what claims have, and have not, passed severely, post-data. Any such proposed inferential use of error probabilities gives considerable fodder for criticism from various tribes of Fisherians, Neyman–Pearsonians, and Bayesians. We can hear them now:

- N-P theorists can only report the preset error probabilities, and can't use *P*-values post-data.
- A Fisherian wouldn't dream of using something that skirts so close to power as does the "sensitivity function" $\Pi(\gamma)$.
- Your account cannot be evidential because it doesn't supply posterior probabilities to hypotheses.
- N-P and Fisherian methods preclude any kind of inference since they use "the sample space" (violating the LP).

How can we reply? To begin, we need to uncover how the charges originate in traditional philosophies long associated with error statistical tools. That's the focus of Tour II.

Only then do we have a shot at decoupling traditional philosophies from those tools in order to use them appropriately today. This is especially so when

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

the traditional foundations stand on such wobbly grounds, grounds largely rejected by founders of the tools. There is a philosophical disagreement between Fisher and Neyman, but it differs importantly from the ones that you're presented with and which are widely accepted and repeated in scholarly and popular treatises on significance tests. Neo-Fisherians and N-P theorists, keeping to their tribes, forfeit notions that would improve their methods (e.g., for Fisherians: explicit alternatives, with corresponding notions of sensitivity, and distinguishing statistical and substantive hypotheses; for N-P theorists, making error probabilities relevant for inference in the case at hand).

The spadework on this tour will be almost entirely conceptual: we won't be arguing for or against any one view. We begin in Section 3.4 by unearthing the basis for some classic counterintuitive inferences thought to be licensed by either Fisherian or N-P tests. That many are humorous doesn't mean disentangling their puzzles is straightforward; a medium to heavy shovel is recommended. We can switch to a light to medium shovel in Section 3.5: excavations of the evidential versus behavioral divide between Fisher and N-P turn out to be mostly built on sand. As David Cox observes, Fisher is often more performance-oriented in practice, but not in theory, while the reverse is true for Neyman and Pearson. At times, Neyman exaggerates the behavioristic conception just to accentuate how much Fisher's tests need reining in. Likewise, Fisher can be spotted running away from his earlier behavioristic positions just to derogate the new N-P movement, whose popularity threatened to eclipse the statistics program that was, after all, his baby. Taking the polemics of Fisher and Neyman at face value, many are unaware how much they are based on personality and professional disputes. Hearing the actual voices of Fisher, Neyman, and Pearson (F and N-P), you don't have to accept the gospel of "what the founders really thought." Still, there's an entrenched history and philosophy of F and N-P: A thick-skinned jacket is recommended. On our third stop (Section 3.6) we witness a bit of magic. The very concept of an error probability gets redefined and, hey presto!, a reconciliation between Jeffreys, Fisher, and Neyman is forged. Wear easily removed shoes and take a stiff walking stick. The Unificationist tribes tend to live near underground springs and lakeshore bounds; in the heady magic, visitors have been known to accidentally fall into a pool of quicksand.

3.4 Some Howlers and Chestnuts of Statistical Tests

> The well-known definition of a statistician as someone whose aim in life is to be wrong in exactly 5 per cent of everything they do misses its target. (Sir David Cox 2006a, p. 197)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

Showing that a method's stipulations could countenance absurd or counterintuitive results is a perfectly legitimate mode of criticism. I reserve the term "howler" for common criticisms based on logical fallacies or conceptual misunderstandings. Other cases are better seen as chestnuts – puzzles that the founders of statistical tests never cleared up explicitly. Whether you choose to see my "howler" as a "chestnut" is up to you. Under each exhibit is the purported basis for the joke.

Exhibit (iii): Armchair Science. *Did you hear the one about the statistical hypothesis tester* . . . who claimed that observing "heads" on a biased coin that lands heads with probability 0.05 is evidence of a statistically significant improvement over the standard treatment of diabetes, on the grounds that such an event occurs with low probability (0.05)?

The "armchair" enters because diabetes research is being conducted solely by flipping a coin. The joke is a spin-off from Kadane (2011):

> Flip a biased coin that comes up heads with probability 0.95, and tails with probability 0.05. If the coin comes up tails reject the null hypothesis. Since the probability of rejecting the null hypothesis if it is true is 0.05, this is a valid 5 percent level test. It is also very robust against data errors; indeed it does not depend on the data at all. It is also nonsense, of course, but nonsense allowed by the rules of significance testing. (p. 439)

Basis for the joke: Fisherian test requirements are (allegedly) satisfied by any method that rarely rejects the null hypothesis.

But are they satisfied? I say no. The null hypothesis in Kadane's example can be in any field, diabetes, or the mean deflection of light. (Yes, Kadane affirms this.) He knows the test entirely ignores the data, but avers that "it has the property that Fisher proposes" (Kadane 2016, p. 1). Here's my take: in significance tests and in scientific hypotheses testing more generally, data can disagree with *H* only by being counter to what would be expected under the assumption that *H* is correct. An improbable series of coin tosses or plane crashes does not count as a disagreement from hypotheses about diabetes or light deflection. In Kadane's example, there is accordance so long as a head occurs – but this is a nonsensical distance measure. Were someone to tell you that any old improbable event (three plane crashes in one week) tests a hypothesis about light deflection, you would say that person didn't understand the meaning of testing in science or in ordinary life. You'd be right (for some great examples, see David Hand 2014).

Kadane knows it's nonsense, but thinks the only complaint a significance tester can have is its low power. What's the power of this "test" against any alternative? It's just the same as the probability it rejects, period, namely, 0.05. So an N-P tester could at least complain. Now I agree that bad tests may still be

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

tests; but I'm saying Kadane's is no test at all. If you want to insist Fisher permits this test, fine, but I don't think that's a very generous interpretation. As egregious as is this howler, it is instructive because it shows like nothing else the absurdity of a crass performance view that claims: reject the null and infer evidence of a genuine effect, so long as it is done rarely. Crass performance is bad enough, but this howler commits a further misdemeanor: It overlooks the fact that a test statistic d($\boldsymbol{x}$) must track discrepancies from H_0, becoming bigger (or smaller) as discrepancies increase (I list it as (ii) in Section 3.2). With any sensible distance measure, a misfit with H_0 must be *because* of the falsity of H_0. The probability of "heads" under a hypothesis about light deflection isn't even defined, because deflection hypotheses do not assign probabilities to coin-tossing trials. Fisher wanted test statistics to reduce the data from the generating mechanism, and here it's not even from the mechanism.

Kadane regards this example as "perhaps the most damaging critique" of significance tests (2016, p. 1). Well, Fisher can get around this easily enough.

Exhibit (iv): Limb-sawing Logic. *Did you hear the one about significance testers sawing off their own limbs?*

> As soon as they reject the null hypothesis H_0 based on a small P-value, they no longer can justify the rejection because the P-value was computed under the assumption that H_0 holds, and now it doesn't.

Basis for the joke: If a test assumes H, then as soon as H is rejected, the grounds for its rejection disappear!

This joke, and I swear it is widely repeated but I won't name names, reflects a serious misunderstanding about ordinary conditional claims. The assumption we use in testing a hypothesis H, statistical or other, is an *implicationary* or *i-assumption*. We have a conditional, say: If H then expect $\boldsymbol{x}$, with H the antecedent. The entailment from H to $\boldsymbol{x}$, whether it is statistical or deductive, does not get sawed off after the hypothesis or model H is rejected when the prediction is not borne out. A related criticism is that statistical tests assume the truth of their test or null hypotheses. No, once again, they may serve only as i-assumptions for drawing out implications. The howler occurs when a test hypothesis that serves merely as an i-assumption is purported to be an actual assumption, needed for the inference to go through. A little logic goes a long way toward exposing many of these howlers. As the point is general, we use H.

This next challenge is by Harold Jeffreys. I won't call it a howler because it hasn't, to my knowledge, been excised by testers: it's an old chestnut, and a very revealing one.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

Exhibit (v): Jeffreys' Tail Area Criticism. *Did you hear the one about statistical hypothesis testers rejecting H_0 because of outcomes it failed to predict?*

> What's unusual about that?
>
> What's unusual is that they do so even when these unpredicted outcomes haven't occurred!

Actually, one can't improve upon the clever statement given by Jeffreys himself. Using *P*-values, he avers, implies that "*a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*" (1939/1961 p. 385).

Basis for the joke: The *P*-value, $\Pr(d \geq d_0; H_0)$, uses the "tail area" of the curve under H_0. d_0 is the observed difference, but $\{d \geq d_0\}$ includes differences even further from H_0 than d_0.

This has become the number one joke in comical statistical repertoires. Before debunking it, let me say that Jeffreys shows a lot of admiration for Fisher: "I have in fact been struck repeatedly in my own work . . . to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would either be identical with mine or would differ only in cases where we should both be very doubtful" (ibid., p. 393). The famous quip is funny because it seems true, yet paradoxical. Why consider more extreme outcomes that didn't occur? The non-occurrence of more deviant results, Jeffreys goes on to say, "might more reasonably be taken as evidence for the law [in this case, H_0], not against it" (ibid., p. 385). The implication is that considering outcomes beyond d_0 is to unfairly discredit H_0, in the sense of finding more evidence against it than if only the actual outcome d_0 is considered.

The opposite is true.

Considering the tail area makes it harder, not easier, to find an outcome statistically significant (although this isn't the only function of the tail area). Why? Because it requires not merely that $\Pr(d = d_0; H_0)$ be small, but that $\Pr(d \geq d_0; H_0)$ be small. This alone squashes the only sense in which this could be taken as a serious criticism of tests. Still, there's a legitimate question about why the tail area probability is relevant. Jeffreys himself goes on to give it a rationale: "If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected. Everybody rejects the conclusion" (ibid., p. 385), so some other criterion is needed. Looking at the tail area supplies one, another would be a prior, which is Jeffreys' preference.

It's worth reiterating Jeffreys' correctly pointing out that "everybody rejects" the idea that the improbability of data under *H* suffices for evidence against *H*.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

Shall we choose priors or tail areas? Jeffreys chooses default priors. Interestingly, as Jeffreys recognizes, for Normal distributions "the tail area represents the probability, given the data" that the actual discrepancy is in the direction opposite to that observed – d_0 is the wrong "sign" (ibid., p. 387). (This relies on a uniform prior probability for the parameter.) This connection between P-values and posterior probabilities is often taken as a way to "reconcile" them, at least for one-sided tests (Sections 4.4, 4.5). This was not one of Fisher's given rationales.

Note that the joke talks about outcomes the null does not predict – just what we wouldn't know without an assumed test statistic or alternative. One reason to evoke the tail area in Fisherian tests is to determine what H_0 "has not predicted," that is, to identify a sensible test statistic d($\boldsymbol{x}$). Fisher, strictly speaking, has only the null distribution, with an implicit interest in tests with sensitivity of a given type. Fisher discusses this point in relation to the lady tasting tea (1935a, pp. 14–15). Suppose I take an observed difference d_0 as grounds to reject H_0 on account of it's being improbable under H_0, when in fact larger differences (larger d values) are even more probable under H_0. Then, as Fisher rightly notes, the improbability of the observed difference would be a poor indication of underlying discrepancy. (In N-P terms, it would be a biased test.) Looking at the tail area would reveal this fallacy; whereas it would be readily committed, Fisher notes, in accounts that only look at the improbability of the observed outcome d_0 under H_0.

When E. Pearson (1970) takes up Jeffreys' question: "Why did we use tail-area probabilities . . .?", his reply is that "this interpretation was not part of our approach" (p. 464). Tail areas simply fall out of the N-P desiderata of good tests. Given the lambda criterion one needed to decide at what point H_0

> should be regarded as no longer tenable, that is where should one choose to bound the rejection region? To help in reaching this decision it appeared that the probability of falling into the region chosen, if H_0 were true, was one necessary piece of information. (ibid.)

So looking at the tail area could be seen as the result of formulating a sensible distance measure (for Fisher), or erecting a good critical region (for Neyman and Pearson).

Pearson's reply doesn't go far enough; it does not by itself explain why reporting the probability of falling into the rejection region is relevant for *inference*. It points to a purely performance-oriented justification that I know Pearson shied away from: It ensures data fall in a critical region rarely under H_0 and sufficiently often under alternatives in H_1 – but this tends to be left as

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

a pre-data, performance goal (recall Birnbaum's Conf, Souvenir D). It is often alleged the N-P tester only reports whether or not $\boldsymbol{x}$ falls in the rejection region. Why are N-P collapsing all outcomes in this region?

In my reading, the error statistician does not collapse the result beyond what the minimal sufficient statistic requires for the question at hand. From our Translation Guide, Souvenir C, considering $(\mathrm{d}(\boldsymbol{X}) \geq \mathrm{d}(\boldsymbol{x}_0))$ signals that we're interested in the method, and we insert "the test procedure would have yielded" before $\mathrm{d}(\boldsymbol{X})$. We report what was observed $\boldsymbol{x}_0$ and the corresponding $\mathrm{d}(\boldsymbol{x}_0)$ – or d_0 – but we require the methodological probability, via the sampling distribution of $\mathrm{d}(\boldsymbol{X})$ – abbreviated as d. This could mean looking at other stopping points, other endpoints, and other variables. We require that with high probability our test would have warned us if the result could easily have come about in a universe where the test hypothesis is true, that is $\Pr(\mathrm{d}(\boldsymbol{X}) < \mathrm{d}(\boldsymbol{x}_0); H_0)$ is high. Besides, we couldn't throw away the detailed data, since they're needed to audit model assumptions.

To conclude this exhibit, considering the tail area does not make it easier to reject H_0 but harder. Harder because it's not enough that the outcome be improbable under the null, outcomes even greater must be improbable under the null. $\Pr(d(\boldsymbol{X}) = d(\boldsymbol{x}_0); H_0)$ could be small while $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0)$ not small. This leads to blocking a rejection when it should be because it means the test could readily produce even larger differences under H_0. Considering other possible outcomes that could have arisen is essential for assessing the test's capabilities. To understand the properties of our inferential tool is to understand what it would do under different outcomes, under different conjectures about what's producing the data. (Yes, the sample space matters post-data.) I admit that neither Fisher nor N-P adequately pinned down an inferential justification for tail areas, but now we have.

A bit of foreshadowing of a later shore excursion: some argue that looking at $d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0)$ actually *does* make it easier to find evidence against H_0. How can that be? Treating $(1 - \beta)/\alpha$ as a kind of likelihood ratio in favor of an alternative over the null, then fed into a Likelihoodist or Bayesian algorithm, it can appear that way. Stay tuned.

Exhibit (vi): Two Measuring Instruments of Different Precisions. *Did you hear about the frequentist who, knowing she used a scale that's right only half the time, claimed her method of weighing is right 75% of the time*?

> She says, "I flipped a coin to decide whether to use a scale that's right 100% of the time, or one that's right only half the time, so, overall, I'm right 75% of the time." (She wants credit because she could have used a better scale, even knowing she used a lousy one.)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

Basis for the joke: An N-P test bases error probabilities on all possible outcomes or measurements that could have occurred in repetitions, but did not.

As with many infamous pathological examples, often presented as knock-down criticisms of all of frequentist statistics, this was invented by a frequentist, Cox (1958). It was a way to highlight what could go wrong in the case at hand, if one embraced an unthinking behavioral-performance view. Yes, error probabilities are taken over hypothetical repetitions of a process, but not just any repetitions will do. Here's the statistical formulation.

We flip a fair coin to decide which of two instruments, E_1 or E_2, to use in observing a Normally distributed random sample $\boldsymbol{Z}$ to make inferences about mean θ. E_1 has variance of 1, while that of E_2 is 10^6. Any randomizing device used to choose which instrument to use will do, so long as it is irrelevant to θ. This is called a *mixture* experiment. The full data would report both the result of the coin flip and the measurement made with that instrument. We can write the report as having two parts: First, which experiment was run and second the measurement: $(E_i, \boldsymbol{z})$, $i = 1$ or 2.

In testing a null hypothesis such as $\theta = 0$, the same $\boldsymbol{z}$ measurement would correspond to a much smaller *P*-value were it to have come from E_1 rather than from E_2: denote them as $p_1(\boldsymbol{z})$ and $p_2(\boldsymbol{z})$, respectively. The overall significance level of the mixture: $[p_1(\boldsymbol{z}) + p_2(\boldsymbol{z})]/2$, would give a misleading report of the precision of the actual experimental measurement. The claim is that N-P statistics would report the average *P*-value rather than the one corresponding to the scale you actually used! These are often called the unconditional and the conditional test, respectively. The claim is that the frequentist statistician must use the unconditional test.

Suppose that we know we have observed a measurement from E_2 with its much larger variance:

> The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution [with the larger variance]. (Cox 1958, p. 361)

Once it is known which E_i has produced $\boldsymbol{z}$, the *P*-value or other inferential assessment should be made with reference to the experiment actually run. As we say in Cox and Mayo (2010):

> The point essentially is that the marginal distribution of a *P*-value averaged over the two possible configurations is misleading for a particular set of data. It would mean that an individual fortunate in obtaining the use of a precise instrument in effect sacrifices some of that information in order to rescue an investigator who has been unfortunate enough to have the randomizer choose a far less precise tool. From the perspective of interpreting the specific data that are actually available, this makes no sense. (p. 296)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

To scotch his famous example, Cox (1958) introduces a principle: weak conditionality.

Weak Conditionality Principle (WCP)**:** If a mixture experiment (of the aforementioned type) is performed, then, if it is known which experiment produced the data, inferences about θ *are appropriately drawn in terms of the sampling behavior* in the experiment known to have been performed (Cox and Mayo 2010, p. 296).

It is called weak conditionality because there are more general principles of conditioning that go beyond the special case of mixtures of measuring instruments.

While conditioning on the instrument actually used seems obviously correct, nothing precludes the N-P theory from choosing the procedure "which is best on the average over both experiments" (Lehmann and Romano 2005, p. 394), and it's even possible that the average or unconditional power is better than the conditional. In the case of such a conflict, Lehmann says relevant conditioning takes precedence over average power (1993b). He allows that in some cases of acceptance sampling, the average behavior may be relevant, but in scientific contexts the conditional result would be the appropriate one (see Lehmann 1993b, p. 1246). Context matters. Did Neyman and Pearson ever weigh in on this? Not to my knowledge, but I'm sure they'd concur with N-P tribe leader Lehmann. Admittedly, if your goal in life is to attain a precise α level, then when discrete distributions preclude this, a solution would be to flip a coin to decide the borderline cases! (See also Example 4.6, Cox and Hinkley 1974, pp. 95–6; Birnbaum 1962 p. 491.)

Is There a Catch?

The "two measuring instruments" example occupies a famous spot in the pantheon of statistical foundations, regarded by some as causing "a subtle earthquake" in statistical foundations. Analogous examples are made out in terms of confidence interval estimation methods (Tour III, Exhibit (viii)). It is a warning to the most behavioristic accounts of testing from which we have already distinguished the present approach. Yet justification for the conditioning (WCP) is fully within the frequentist error statistical philosophy, for contexts of scientific inference. There is no suggestion, for example, that only the particular data set be considered. That would entail abandoning the sampling distribution as the basis for inference, and with it the severity goal. Yet we are told that "there is a catch" and that WCP leads to the Likelihood Principle (LP)!

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

> It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma. Conditioning is warranted to achieve objective frequentist goals, and the [weak] conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The 'dilemma' argument is therefore an illusion. (Cox and Mayo 2010, p. 298)

There is a large literature surrounding the argument for the Likelihood Principle, made famous by Birnbaum (1962). Birnbaum hankered for something in between radical behaviorism and throwing error probabilities out the window. Yet he himself had apparently proved there is no middle ground (if you accept WCP)! Even people who thought there was something fishy about Birnbaum's "proof" were discomfited by the lack of resolution to the paradox. It is time for post-LP philosophies of inference. So long as the Birnbaum argument, which Savage and many others deemed important enough to dub a "breakthrough in statistics," went unanswered, the frequentist was thought to be boxed into the pathological examples. She is not.

In fact, I show there is a flaw in his venerable argument (Mayo 2010b, 2013a, 2014b). That's a relief. Now some of you will howl, "Mayo, not everyone agrees with your disproof! Some say the issue is not settled." Fine, please explain where my refutation breaks down. It's an ideal brainbuster to work on along the promenade after a long day's tour. Don't be dismayed by the fact that it has been accepted for so long. But I won't revisit it here.

3.5 P-values Aren't Error Probabilities Because Fisher Rejected Neyman's Performance Philosophy

> Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. (Lehmann 1993b, p. 1248)
>
> Thus, Fisher rather incongruously appears to be attacking his own past position rather than that of Neyman and Pearson. (Lehmann 2011, p. 55)

By and large, when critics allege that Fisherian *P*-values are not error probabilities, what they mean is that Fisher wanted to interpret them in an evidential manner, not along the lines of Neyman's long-run behavior. I'm not denying there is an important difference between using error probabilities inferentially and behavioristically. The truth is that N-P and Fisher used

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

P-values and other error probabilities in both ways.[1] What they didn't give us is a clear account of the former. A big problem with figuring out the "he said/they said" between Fisher and Neyman–Pearson is that "after 1935 so much of it was polemics" (Kempthorne 1976) reflecting a blow-up which had to do with professional rivalry rather than underlying philosophy. Juicy details later on.

We need to be clear on the meaning of an error probability. A method of statistical inference moves from data to some inference about the source of the data as modeled. Associated error probabilities refer to the probability the method outputs an erroneous interpretation of the data. Choice of test rule pins down the particular error; for example, it licenses inferring there's a genuine discrepancy when there isn't (perhaps of a given magnitude). The test method is given in terms of a test statistic $\mathrm{d}(\boldsymbol{X})$, so the error probabilities refer to the probability distribution of $\mathrm{d}(\boldsymbol{X})$, the sampling distribution, computed under an appropriate hypothesis. Since we need to highlight subtle changes in meaning, call these ordinary "frequentist" error probabilities. (I can't very well call them error statistical error probabilities, but that's what I mean.)[2] We'll shortly require subscripts, so let this be error probability$_1$. Formal error probabilities have almost universally been associated with N-P statistics, and those with long-run performance goals. I have been disabusing you of such a straightjacketed view; they are vital in assessing how well probed the claim in front of me is. Yet my reinterpretation of error probabilities does not change their mathematical nature.

We can attach a frequentist performance assessment to any inference method. Post-data, these same error probabilities can, though they need not, serve to quantify the severity associated with an inference. Looking at the mathematics, it's easy to see the P-value as an error probability. Take Cox and Hinkley (1974):

> For given observations $\mathbf{y}$ we calculate $t = t_{\mathrm{obs}} = t(\mathbf{y})$, say, and the *level of significance* p_{obs} by $p_{\mathrm{obs}} = \Pr(T \geq t_{\mathrm{obs}}; H_0)$.
>
> . . . Hence p_{obs} is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0. (p. 66)

Thus p_{obs} would be the Type I error probability associated with the test procedure consisting of finding evidence against H_0 when reaching p_{obs}.[3]

[1] Neyman (1976) said he was "not aware of a conceptual difference between a 'test of a statistical hypothesis' and a 'test of significance' and uses these terms interchangeably" (p. 737). We will too, with qualifications as needed.

[2] Thanks to the interpretation being fairly intimately related to the test, we get the error probabilities (formal or informal) attached to the interpretation.

[3] Note that p_{obs} and t_{obs} are the same as our p_0 and d_0. (or $\mathrm{d}(\boldsymbol{x}_0)$)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

Thus the *P*-value equals the corresponding Type I error probability. [I've been using upper case *P*, but it's impossible to unify the literature.] Listen to Lehmann, speaking for the N-P camp:

> [I]t is good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level . . . at which the hypothesis would be rejected for the given observation. This number, the so-called *P-value* gives an idea of how strongly the data contradict the hypothesis. It also enables others to reach a verdict based on the significance level of their choice. (Lehmann and Romano 2005, pp. 63–4)

N-P theorists have no compunctions in talking about N-P tests using attained significance levels or *P*-values. Bayesians Gibbons and Pratt (1975) echo this view:

> The *P*-value can then be interpreted as the smallest level of significance, that is, the 'borderline level', since the outcome observed would . . . not [be] significant at any smaller levels. Thus it is sometimes called the 'level attained' by the sample . . . Reporting a *P*-value . . . permits each individual to choose his own . . . maximum tolerable probability for a Type I error. (p. 21)

Is all this just a sign of texts embodying an inconsistent hybrid? I say no, and you should too.

A certain tribe of statisticians professes to be horrified by the remarks of Cox and Hinkley, Lehmann and Romano, Gibbons and Pratt and many others. That these remarks come from leading statisticians, members of this tribe aver, just shows the depth of a dangerous "confusion over the evidential content (and mixing) of *p*'s and *α*'s" (Hubbard and Bayarri 2003, p. 175). On their view, we mustn't mix what they call "evidence and error": F and N-P are incompatible. For the rest of this tour, we'll alternate between the museum and engaging the Incompatibilist tribes themselves. When viewed through the tunnel of the Incompatibilist statistical philosophy, these statistical founders appear confused.

> The distinction between evidence (*p*'s) and error (*α*'s) is not trivial . . . it reflects the fundamental differences between Fisher's ideas on significance testing and inductive inference, and [N-P's] views on hypothesis testing and inductive behavior. (Hubbard and Bayarri 2003, p. 171)

What's fascinating is that the Incompatibilists admit it's the philosophical difference they're on about, not a mathematical one. The paper that has become *the* centerpiece for the position in this subsection is Berger and Sellke (1987). They ask:

> Can *P* values be justified on the basis of how they perform in repeated use? We doubt it. For one thing, how would one measure the performance of *P* values? With

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

> significance tests and confidence intervals, they are either right or wrong, so it is possible to talk about error rates. If one introduces a decision rule into the situation by saying that H_0 is rejected when the P value < 0.05, then of course the classical error rate is 0.05. (p. 136)

Good. Then we can agree a P-value is, mathematically, an error probability. Berger and Sellke are merely opining that Fisher wouldn't have *justified* their use on grounds of error rate performance. That's different. Besides, are we so sure Fisher wouldn't sully himself with crass error probabilities, and dichotomous tests? Early on at least, Fisher appears as a behaviorist par excellence. That he is later found "attacking his own position," as Lehmann puts it, is something else.

Mirror Mirror on the Wall, Who's the More Behavioral of Them All?

N-P were striving to emulate the dichotomous interpretation they found in Fisher:

> It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. . . . It is usual and convenient for the experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. (Fisher 1935a, pp. 13–14)

Fisher's remark can be taken to justify the tendency to ignore negative results or stuff them in file drawers, somewhat at odds with his next lines, the ones that I specifically championed in Excursion 1: "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result..." (1935a, p. 14).[4] This would require us to keep the negative results around for a while. How else could we see if we are rarely failing, or often succeeding?

What I mainly want to call your attention to now are the key phrases "willing to admit," "satisfy him," "deciding to ignore." What are these, Neyman asks, but actions or behaviors? He'd learned from R. A. Fisher! So, while many take

[4] Fisher, in a 1926 paper, gives another nice rendering: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high odds sometimes claimed for experimental results should usually be discounted, for inaccurate methods of estimating error have far more influence than has the particular standard of significance chosen" (pp. 504–5).

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

the dichotomous "up-down" spirit of tests as foreign to Fisher, it is not foreign at all. Again from Fisher (1935a):

> Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations … those which show a significant discrepancy from a certain hypothesis; … and on the other hand, results which show no significant discrepancy from this hypothesis. (pp. 15–16)

No wonder Neyman could counter Fisher's accusations that he'd turned his tests into tools for inductive behavior by saying, in effect, look in the mirror (for instance, in the acrimonious exchange of 1955–6, 20 years after the blow-up): Pearson and I were only systematizing your practices for how to interpret data, taking explicit care to prevent untoward results that you only managed to avoid on intuitive grounds!

Fixing Significance Levels. What about the claim that N-P tests fix the Type I error probability in advance, whereas *P*-values are post-data? Doesn't *that* prevent a *P*-value from being an error probability? First, we must distinguish between fixing the significance level for a test prior to data collection, and fixing a threshold to be used across one's testing career. Fixing α and power is part of specifying a test with reasonable capabilities of answering the question of interest. Having done so, there's nothing illicit about reporting the *achieved* or *attained* significance level, and it is even recommended by Lehmann. As for setting a threshold for habitual practice, that's actually more Fisher than N-P.

Lehmann is flummoxed by the association of fixed levels of significance with N-P since "[U]nlike Fisher, Neyman and Pearson (1933, p. 296) did not recommend a standard level but suggested that 'how the balance [between the two kinds of error] should be struck must be left to the investigator'" (Lehmann 1993b, p. 1244). From their earliest papers, Neyman and Pearson stressed that the tests were to be "used with discretion and understanding" depending on the context (Neyman and Pearson 1928, p. 58). In a famous passage, Fisher (1956) raises the criticism – but without naming names:

> A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection … However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (pp. 44–5)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

It is assumed Fisher is speaking of N-P, or at least Neyman. But N-P do not recommend such habitual practice.

Long Runs Are Hypothetical. What about the allegation that N-P error probabilities allude to actual long-run repetitions, while the *P*-value is a *hypothetical* distribution? N-P error probabilities are also about hypothetical would-be's. Each sample of size *n* gives a single value of the test statistic $d(\boldsymbol{X})$. Our inference is based on this one sample. The third requirement (Pearson's "Step 3") for tests is that we be able to compute the distribution of $d(\boldsymbol{X})$, under the assumption that the world is approximately like H_0, and under discrepancies from H_0. Different outcomes would yield different $d(\boldsymbol{X})$ values, and we consider the frequency distribution of $d(\boldsymbol{X})$ over hypothetical repetitions.

At the risk of overkill, the sampling distribution is all about hypotheticals: the relative frequency of outcomes under one or another hypothesis. These also equal the relative frequencies assuming you really did keep taking samples in a long run, tiring yourself out in the process. It doesn't follow that the value of the hypothetical frequencies depends on referring to, much less actually carrying out, that long run. A statistical hypothesis has implications for some hypothetical long run in terms of how frequently this or that would occur. A statistical test uses the data to check how well the predictions are met. The sampling distribution is the testable meeting-ground between the two.

The same pattern of reasoning is behind resampling from the one and only sample in order to generate a sampling distribution. (We meet with resampling in Section 4.10.) The only gap is to say why such a hypothetical (or counterfactual) is relevant for inference in the case at hand. Merely proposing that error probabilities give a vague "strength of evidence" to an inference won't do. Our answer is that they capture the capacities of tests, which in turn tell us how severely tested various claims may be said to be.

It's Time to Get Beyond the "Inconsistent Hybrid" Charge

Gerd Gigerenzer is a wonderful source of how Fisherian and N-P methods led to a statistical revolution in psychology. He is famous for, among much else, arguing that the neat and tidy accounts of statistical testing in social science texts are really an inconsistent hybrid of elements from N-P's behavioristic philosophy and Fisher's more evidential approach (Gigerenzer 2002, p. 279). His tribe is an offshoot of the Incompatibilists, but with a Freudian analogy to illuminate the resulting tension and anxiety that a researcher is seen to face.

N-P testing, he says, "functions as the Superego of the hybrid logic" (ibid., p. 280). It requires alternatives, significance levels, and power to be prespecified, while strictly outlawing evidential or inferential interpretations about the

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

truth of a particular hypothesis. The Fisherian "Ego gets things done . . . and gets papers published" (ibid.). Power is ignored, and the level of significance is found after the experiment, cleverly hidden by rounding up to the nearest standard level. "The Ego avoids . . . exact predictions of the alternative hypothesis, but claims support for it by rejecting a null hypothesis" and in the end is "left with feelings of guilt and shame for having violated the rules" (ibid.). Somewhere in the background lurks his Bayesian Id, driven by wishful thinking into misinterpreting error probabilities as degrees of belief.

As with most good caricatures, there is a large grain of truth in Gigerenzer's Freudian metaphor – at least as the received view of these methods. I say it's time to retire the "inconsistent hybrid" allegation. Reporting the attained significance level is entirely legitimate and is recommended in N-P tests, so long as one is not guilty of other post-data selections causing *actual* P-values to differ from *reported* or nominal ones. By failing to explore the inferential basis for the stipulations, there's enormous unclarity as to what's being disallowed and why, and what's mere ritual or compulsive hand washing (as he might put it (ibid., p. 283)). Gigerenzer's Ego might well *deserve* to feel guilty if he has chosen the hypothesis, or characteristic to be tested, based on the data, or if he claims support for a research hypothesis by merely rejecting a null hypothesis – the illicit NHST animal. A post-data choice of test statistic may be problematic, but not an attained significance level.

Gigerenzer recommends that statistics texts teach the conflict and stop trying "to solve the conflict between its parents by denying its parents" (2002, p. 281). I, on the other hand, think we should take responsibility for interpreting the tools according to their capabilities. Polemics between Neyman and Fisher, however lively, taken at face value, are a highly unreliable source; we should avoid chiseling into even deeper stone the hackneyed assignments of statistical philosophy – "he's inferential, he's an acceptance sampler." The consequences of the "inconsistent hybrid" allegation are dire: both schools are caricatures, robbed of features that belong in an adequate account.

Hubbard and Bayarri (2003) are a good example of this; they proclaim an N-P tester is forbidden – forbidden! – from reporting the observed P-value. They eventually concede that an N-P test "could be defined equivalently in terms of the p value . . . the null hypothesis should be rejected if the observed $p < \alpha$, and accepted otherwise" (p. 175). But they aver "no matter how small the p value is, the appropriate report is that the procedure guarantees a $100\alpha\%$ false rejection of the null on repeated use" (ibid.). An N-P tester must robotically obey the reading that has grown out of the Incompatibilist tribe to which they belong. A user must round up to the predesignated α. This type of prohibition

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

gives a valid guilt trip to Gigerenzer's Ego; yet the hang-up stems from the Freudian metaphor, not from Neyman and Pearson, who say:

> it is doubtful whether the knowledge that P_z [the *P*-value associated with test statistic *z*] was really 0.03 (or 0.06) rather than 0.05, ... would in fact ever modify our judgment ... regarding the origin of a single sample. (Neyman and Pearson 1928, p. 27)

But isn't it true that rejection frequencies needn't be indicative of the evidence against a null? Yes. Kadane's example, if allowed, shows how to get a small rejection frequency with no evidence. But this was to be a problem for Fisher, solved by N-P (even if Kadane is not fond of them either). Granted, even in tests not so easily dismissed, crude rejection frequencies differ from an evidential assessment, especially when some of the outcomes leading to rejection vary considerably in their evidential force. This is the lesson of Cox's famous "two machines with different precisions." Some put this in terms of selecting the relevant reference set which "need not correspond to all possible repetitions of the experiment" (Kalbfleisch and Sprott 1976, p. 272). We've already seen that relevant conditioning is open to a N-P tester. Others prefer to see it as a matter of adequate model specification. So once again it's not a matter of Fisher vs. N-P.

I'm prepared to admit Neyman's behavioristic talk. Mayo (1996, Chapter 11) discusses: "Why Pearson rejected the (behavioristic) N-P theory" (p. 361). Pearson does famously declare that "the behavioristic conception is Neyman's not mine" (1955, p. 207). Furthermore, Pearson explicitly addresses "the situation where statistical tools are applied to an isolated investigation of considerable importance ..." (1947, p. 170).

> In other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules ... Why do we do this? ... Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment?
>
> Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgment which we control at a low figure? (ibid., p. 172)

While tantalizingly leaving the answer dangling, it's clear that for Pearson: "the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment" (ibid.) in learning about the particular case at hand. He gives an example from his statistical work in World War II:

> Two types of heavy armour-piercing naval shell of the same caliber are under consideration; they may be of different design or made by different firms ... Twelve

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

> shells of one kind and eight of the other have been fired; two of the former and five of the latter failed to perforate the plate . . . (Pearson 1947, 171)

> Starting from the basis that individual shells will never be identical in armour-piercing qualities, . . . he has to consider how much of the difference between (i) two failures out of twelve and (ii) five failures out of eight is likely to be due to this inevitable variability. (ibid.)

He considers what other outcomes could have occurred, and how readily, in order to learn what variability alone is capable of producing.[5] Pearson opened the door to the evidential interpretation, as I note in 1996, and now I go further.

Having looked more carefully at the history before the famous diatribes, and especially at Neyman's applied work, I now hold that Neyman largely rejected it as well! Most of the time, anyhow. But that's not the main thing. Even if we couldn't point to quotes and applications that break out of the strict "evidential versus behavioral" split: *we* should be the ones to interpret the methods for inference, and supply the statistical philosophy that directs their right use.

Souvenir L: Beyond Incompatibilist Tunnels

What people take away from the historical debates is Fisher (1955) accusing N-P, or mostly Neyman, of converting his tests into acceptance sampling rules more appropriate for five-year plans in Russia, or making money in the USA, than for science. Still, it couldn't have been too obvious that N-P distorted his tests, since Fisher tells us only in 1955 that it was Barnard who explained that, despite agreeing mathematically in very large part, there is this distinct philosophical position. Neyman suggests that his terminology was to distinguish what he (and Fisher!) were doing from the attempts to define a unified rational measure of belief on hypotheses. N-P both denied there was such a thing. Given Fisher's vehement disavowal of subjective Bayesian probability, N-P thought nothing of crediting Fisherian tests as a step in the development of "inductive behavior" (in their 1933 paper).

The myth of the radical difference in either methods or philosophy is a myth. Yet, as we'll see, the hold it has over people continues to influence the use and discussion of tests. It's based almost entirely on sniping between Fisher and Neyman from 1935 until Neyman leaves for the USA in 1938. Fisher didn't engage much with statistical developments during World War II. Barnard describes Fisher as cut off "by some mysterious personal or political agency. Fisher's isolation occurred, I think, at a particularly critical

[5] Pearson said that a statistician has an α and a β side, the former alludes to what they say in theory, the latter to what they do in practice. In practice, even Neyman, so often portrayed as performance-oriented, was as inferential as Pearson.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

time, when opportunities existed for a fruitful fusion of ideas stemming from Neyman and Pearson and from Fisher" (Barnard 1985, p. 2). Lehmann observes that Fisher kept to his resolve not to engage in controversy with Neyman until the highly polemical exchange of 1955 at age 65. Fisher alters some of the lines of earlier editions of his books. For instance, Fisher's disinterest in the attained *P*-value was made clear in *Statistical Methods for Research Workers* (SMRW) (1934a, p. 80):

> . . . in practice we do not want to know the exact value of P for any observed value of [the test statistic], but, in the first place, whether or not the observed value is open to suspicion.
>
> If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05.

Lehmann explains that it was only "fairly late in life, Fisher's attitude had changed" (Lehmann 2011, p. 52). In the 13th edition of SMRW, Fisher changed his last sentence to:

> The actual value of P obtainable . . . indicates the strength of the evidence against the hypothesis. [Such a value] is seldom to be disregarded. (p. 80)

Even so, this at most suggests how the methodological (error) probability is thought to provide a measure of evidential strength – it doesn't abandon error probabilities. There's a deeper reason for this backtracking by Fisher; I'll save it for Excursion 5. One other thing to note: F and N-P were creatures of their time. Their verbiage reflects the concern with "operationalism" and "behaviorism," growing out of positivistic and verificationist philosophy. I don't deny the value of tracing out the thrust and parry between Fisher and Neyman in these excursions. None of the founders solved the problem of an inferential interpretation of error probabilities – though they each offered tidbits. Their name-calling: "you're too mechanical," "no *you* are," at most shows, as Gigerenzer and Marewski observe, that they all rejected mechanical statistics (2015, p. 422).

The danger is when one group's interpretation is the basis for a historically and philosophically "sanctioned" reinterpretation of one or another method. Suddenly, rigid rules that the founders never endorsed are imposed. Through the Incompatibilist philosophical tunnel, as we are about to see, these reconstruals may serve as an effective way to dismiss the entire methodology – both F and N-P. After completing this journey, you shouldn't have to retrace this "he said/they said" dispute again. It's the methods, stupid.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

3.6 Hocus-Pocus: *P*-values Are Not Error Probabilities, Are Not Even Frequentist!

> Fisher saw the *p* value as a measure of evidence, not as a frequentist evaluation. Unfortunately, as a measure of evidence it is very misleading. (Hubbard and Bayarri 2003, p. 181)

This entire tour, as you know, is to disentangle a jungle of conceptual issues, not to defend or criticize any given statistical school. In sailing forward to scrutinize Incompatibilist tribes who protest against mixing *p*'s and *α*'s, we need to navigate around a pool of quicksand. They begin by saying *P*-values are for evidence and inference, unlike error probabilities. N-P error probabilities are too performance oriented to be measures of evidence. In the next breath we're told *P*-values aren't good measures of evidence either. A good measure of evidence, it's assumed, should be probabilist, in some way, and *P*-values disagree with probabilist measures, be they likelihood ratios, Bayes factors, or posteriors. If you reinterpret error probabilities, they promise, you can make peace with all tribes. Whether we get on firmer ground or sink in a marshy swamp will have to be explored.

Berger's Unification of Jeffreys, Neyman, and Fisher

With "reconciliation" and "unification" in the air, Jim Berger, a statistician deeply influential in statistical foundations, sets out to see if he can get Fisher, Neyman, and (non-subjective) Bayesian Jeffreys to agree on testing (2003). A compromise awaits, if we nip and tuck the meaning of "error probability" (Section 3.5). If you're an N-P theorist and like your error probability$_1$, you can keep it he promises, but he thinks you will want to reinterpret it. It then becomes possible to say that a *P*-value is not an error probability (full stop), meaning it's not the newly defined error probability$_{2.}$ What's error probability$_2$? It's a type of posterior probability in a null hypothesis, conditional on the outcome, given a prior. It may still be frequentist in some sense. On this reinterpretation, *P*-values are not error probabilities. Neither are N-P Type I and II, α and β. Following the philosopher's clarifying move via subscripts, there is error probability$_1$ – the usual frequentist notion – and error probability$_2$ – notions from probabilism that had never been called error probabilities before.

In commenting on Berger (2003), I noted my surprise at his redefinition (Mayo 2003b). His reply: "Why should the frequentist school have exclusive right to the term 'error probability?' It is not difficult to simply add the designation 'frequentist' (or Type I or Type II) or 'Bayesian' to the term to differentiate between the schools" (Berger 2003, p. 30). That would work splendidly. So let error probability$_2$ = Bayesian error probability. Frankly, I

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

didn't think Bayeslans would want the term. In a minute, however, Berger will claim they alone are the true frequentist error probabilities! If you feel yourself sinking in a swamp of sliding meanings, remove your shoes, flip onto your back atop your walking stick and you'll stop sinking. Then, you need only to pull yourself to firm land. (See Souvenir M.)

The Bayes Factor. In 1987, Berger and Sellke said that in order to consider *P*-values as error probabilities we need to introduce a decision or test rule. Berger (2003) proposes such a rule and error probability$_2$ is born. In trying to merge different methodologies, there's always a danger of being biased in favor of one, begging the question against the others. From the severe tester's perspective, this is what happens here, but so deftly that you might miss it if you blink.[6]

His example involves $X_1, \ldots, X_n$ IID data from N(θ, σ^2), with σ^2 known, and the test is of two simple hypotheses H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$. Consider now their two *P*-values: "for $i = 0, 1$, let p_i be the *p*-value in testing H_i against the other hypothesis" (ibid., p. 6). Then reject H_0 when $p_0 \leq p_1$, and accept H_0 otherwise. If you reject H_0 you next compute the posterior probability of H_0 using one of Jeffreys' default priors giving 0.5 to each hypothesis. The computation rests on the *Bayes factor* or likelihood ratio $\mathrm{B}(\boldsymbol{x}) = \Pr(\boldsymbol{x}|H_0)/\Pr(\boldsymbol{x}|H_1)$:

$$\Pr(H_0|\boldsymbol{x}) = \mathrm{B}(\boldsymbol{x})/[1 + \mathrm{B}(\boldsymbol{x})].$$

The priors drop out, being 0.5. As before, $\boldsymbol{x}$ refers to a generic value for $\boldsymbol{X}$.

This was supposed to be something Fisher would like, so what happened to *P*-values? They have a slight walk-on part: the rejected hypothesis is the one that has the lower *P*-value. Its value is irrelevant, but it directs you to which posterior to compute. We might understand his Bayesian error probabilities this way: If I've rejected H_0, I'd be wrong if H_0 were true, so $\Pr(H_0|\boldsymbol{x})$ is a probability of being wrong about H_0. It's the *Bayesian Type I error probability*$_2$. If instead you reject H_1, then you'd be wrong if H_1 were true. So in that case you report the Bayesian Type II error probability$_2$, which would be $\Pr(H_1|\boldsymbol{x}) = 1/[1 + \mathrm{B}(\boldsymbol{x})]$. Whatever you think of these, they're quite different from error probability$_1$, which does not use priors in H_i.

Sleight of Hand? Surprisingly, Berger claims to give a "dramatic illustration of the nonfrequentist nature of *P*-values" (ibid., p. 3). Wait a second, how did they become *non-frequentist*? What he means is that the *P*-value can be shown to disagree with the special posterior probability for H_0, defined as error

[6] We are forced to spend more time on *P*-values than one would wish simply because so many of the criticisms and proposed reforms are in terms of them.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

probability$_2$. They're not called Bayesian error probabilities any more but frequentist conditional error probabilities (CEPs). Presto! A brilliant sleight of hand.

This 0.5 prior is not supposed to represent degree of belief, but it is Berger's "objective" default Bayesian prior. Why does he call it frequentist? He directs us to an applet showing if we imagine randomly selecting our test hypothesis from a population of null hypotheses, 50% of which are true, the rest false, and then compute the relative frequency of true nulls conditional on its having been rejected at significance level p, we get a number that is larger than p. This violates what he calls the frequentist principle (not to be confused with FEV):

> *Berger's frequentist principle*: $\Pr(H_0 \text{ true} \mid H_0 \text{ rejected at level } p)$ should equal p.

This is very different from what a P-value gives us, namely, $\Pr(P \leq p; H_0) = p$ (or $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0) = p$).

He actually states the frequentist principle more vaguely; namely, that the reported error probability should equal the actual one, but the computation is to error probability$_2$. If I'm not being as clear as possible, it's because Berger isn't, and I don't want to prematurely saddle him with one of at least two interpretations he moves between. For instance, Berger says the urn of nulls applet is just a heuristic, showing how it could happen. So suppose the null was randomly selected from an urn of nulls 50% of which are true. Wouldn't 0.5 be its frequentist prior? One has to be careful. First consider a legitimate frequentist prior. Suppose I selected the hypothesis H_0: that the mean temperature in the water, θ, is 150 degrees (Section 3.2). I can see this value resulting from various features of the lake and cooling apparatus, and identify the relative frequency that θ takes different values. $\{\Theta = \theta\}$ is an event associated with random variable Θ. Call this an *empirical* or *frequentist* prior just to fix the notion. What's imagined in Berger's applet is very different. Here the analogy is with diagnostic screening for disease, so I will call it that (Section 5.6). We select one null from an urn of nulls, which might include all hypotheses from a given journal, a given year, or lots of other things.[7] If 50% of the nulls in this urn are true, the experiment of

[7] It is ironic that it's in the midst of countering a common charge that he requires repeated sampling from the same population that Neyman (1977) talks about a series of distinct scientific inquiries (presumably independent) with Type I and Type II error probabilities (for specified alternatives) $\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots$ and $\beta_1, \beta_2, \ldots, \beta_n, \ldots$

> I frequently hear a particular regrettable remark … that the frequency interpretation of either the level of significance α or of power $(1 - \beta)$ is only possible when one deals many times WITH THE SAME HYPOTHESIS H, TESTED AGAINST THE SAME ALTERNATIVE. (Neyman 1977, 109, his use of capitals)

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

randomly selecting a null from the urn could be seen as a Bernoulli trial with two outcomes: a null that is true or false. The probability of selecting a null that has the property "true" is 0.5. Suppose I happen to select H_0: $\theta = 150$, the hypothesis from the accident at the water plant. It would be incorrect to say 0.5 was the relative frequency that $\theta = 150$ would emerge with the empirical prior. So there's a frequentist computation, but it differs from what Neyman's empirical Bayesian would assign it. I'll come back to this later (Excursion 6).

Suppose instead we keep to the default Bayesian construal that Berger favors. The priors come from one or another conventional assignment. On this reading, his frequentist principle is: the *P*-value should equal the default posterior on H_0. That is, a reported *P*-value should equal error probability$_2$. By dropping the designation "Bayesian" that he himself recommended "to differentiate between the schools" (p. 30), it's easy to see how confusion ensues.

Berger emphasizes that the confusion he is on about "is different from the confusion between a *P*-value and the posterior probability of the null hypothesis" (p. 4). What confusion? That of thinking *P*-values are frequentist error probabilities$_2$ – but he has just introduced the shift of meaning! But the only way error probability$_2$ inherits a frequentist meaning is by reference to the heuristic (where the prior is the proportion of true nulls in a hypothetical urn of nulls), giving a diagnostic screening posterior probability. The subscripts are a lifesaver for telling what's true when definitions shift about throughout an argument. The frequentist had only ever wanted error probabilities$_1$ – the ones based solely on the sampling distribution of $d(\boldsymbol{X})$. Yet now he declares that error probability$_2$ – Bayesian error probability – is the only real or relevant frequentist error probability! If this is the requirement, preset α, β aren't error probabilities either.

It might be retorted, however, that this was to be a compromise position. We can't dismiss it out of hand because it requires Neyman and Fisher to become default Bayesians. To smoke the peace pipe, everyone has to give a little. According to Berger, "Neyman criticized p-values for violating the frequentist principle." (p. 3) With Berger's construal, it is not violated. So it appears Neyman gets something. Does he? We know N-P used *P*-values, and never saw them as non-frequentist; and surely Neyman wouldn't be criticizing a *P*-value for not being equal to a default (or other) posterior probability. Hence Nancy Reid's quip: "the Fisher/Jeffreys agreement is essentially to have Fisher"

From the Central Limit Theorem, Neyman remarks:

> The relative frequency of the first kind of errors will be close to the arithmetic mean of numbers $\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots$ Also the relative frequency of detecting the falsehood of the hypotheses tested, when false . . . will differ but little from the average of [the corresponding powers, for specified alternatives].

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

kowtow to Jeffreys (N. Reid 2003). The surest sign that we've swapped out meanings are the selling points.

Consider the Selling Points

"Teaching statistics suddenly becomes easier . . . it is considerably less important to disabuse students of the notion that a frequentist error probability is the probability that the hypothesis is true, given the data" (Berger 2003, p. 8), since his error probability$_2$ actually has that interpretation. We are also free of having to take into account the stopping rule used in sequential tests (ibid.). As Berger dangles his tests in front of you with the labels "frequentist," "error probabilities," and "objectivity," there's one thing you know: if the methods enjoy the simplicity and freedom of paying no price for optional stopping, you'll want to ask if they're also controlling error probabilities$_1$. When that handwringing disappears, unfortunately, so does our assurance that we block inferences that have passed with poor severity.

Whatever you think of default Bayesian tests, Berger's error probability$_2$ differs from N-P's error probability$_1$. N-P requires controlling the Type I and II error probabilities at low values regardless of prior probability assignments. The scrutiny here is not of Berger's recommended tests – that comes later. The scrutiny here is merely to shine a light on the type of shifting meanings that our journey calls for. Always carry your walking stick – it serves as a metaphorical subscript to keep you afloat.

Souvenir M: Quicksand Takeaway

The howlers and chestnuts of Section 3.4 call attention to: the need for an adequate test statistic, the difference between an i-assumption and an actual assumption, and that tail areas serve to raise, and not lower, the bar for rejecting a null hypothesis. The stop in Section 3.5 pulls back the curtain on one front of typical depictions of the N-P vs. Fisher battle, and Section 3.6 disinters equivocal terms in a popular peace treaty between the N-P, Fisher, and Jeffreys tribes. Of these three stops, I admit that the last may still be murky. One strategy we used to clarify are subscripts to distinguish slippery terms. Probabilities of Type I and Type II errors, as well as *P*-values, are defined exclusively in terms of the sampling distribution of d($\boldsymbol{X}$), under a statistical hypothesis of interest. That's error probability$_1$. Error probability$_2$, in addition to requiring priors, involves conditioning on the particular outcome, with the hypothesis varying. There's no consideration of the sampling distribution of d($\boldsymbol{X}$), if you've conditioned on the actual

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press

outcome. A second strategy is to consider the selling points of the new "compromise" construal, to gauge what it's asking you to buy.

Here's from our guidebook:

> You're going to need to be patient. Depending on how much quicksand is around you, it could take several minutes or even hours to slowly, methodically get yourself out . . .
>
> *Relax*. Quicksand usually isn't more than a couple feet deep . . . If you panic you can sink further, but if you relax, your body's buoyancy will cause you to float.
>
> Breathe deeply . . . It is impossible to "go under" if your lungs are full of air (WikiHow 2017).

In later excursions, I promise, you'll get close enough to the edge of the quicksand to roll easily to hard ground. More specifically, all of the terms and arguments of Section 3.6 will be excavated.

https://doi.org/10.1017/9781107286184.007 Published online by Cambridge University Press