Tour II Rejection Fallacies: Who's Exaggerating What?

Comedian Jackie Mason will be doing his shtick this evening in the ship's theater: a one-man show consisting of a repertoire of his "Greatest Hits" without a new or updated joke in the mix. A sample:

> If you want to eat nothing, eat nouvelle cuisine. Do you know what it means? No food. The smaller the portion the more impressed people are, so long as the food's got a fancy French name, haute cuisine. An empty plate with sauce!

You'll get the humor only once you see and hear him (Mayo 2012b). As one critic (Logan 2012) wrote, Mason's jokes "offer a window to a different era," one whose caricatures and biases one can only hope we've moved beyond. It's one thing for Jackie Mason to reprise his greatest hits, another to reprise statistical foibles and howlers which could leave us with radical changes to science. Among the tribes we'll be engaging: Large *n*, Jeffreys–Lindley, and Spike and Smear.

How Could a Group of Psychologists Be so Wrong? I'll carry a single tome in our tour: Morrison and Henkel's 1970 classic, *The Significance Test Controversy*. Some abuses of the proper interpretation of significance tests were deemed so surprising even back then that researchers in psychology conducted studies to try to understand how this could be. Notably, Rosenthal and Gaito (1963) discovered that statistical significance at a given level was often fallaciously taken as evidence of a greater discrepancy from the null hypothesis the larger the sample size *n*. In fact, it is indicative of *less* of a discrepancy from the null than if it resulted from a smaller sample size.

> What is shocking is that these psychologists indicated substantially greater confidence or belief in results associated with the larger sample size for the same *p* values. According to the theory, especially as this has been amplified by Neyman and Pearson (1933), the probability of rejecting the null hypothesis for any given deviation from null and *p* values increases as a function of the number of observations. The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect in the population . . . The question is, how could a group of psychologists be so wrong? (Bakan 1970, p. 241)

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

(Our convention is for "discrepancy" to refer to the parametric, not the observed, difference. Their use of "deviation" from the null alludes to our "discrepancy.")

As statistician John Pratt notes, "the more powerful the test, the more a just significant result favors the null hypothesis" (1961, p. 166). Yet we still often hear: "The thesis implicit in the [N-P] approach, [is] that a hypothesis may be rejected with increasing confidence or reasonableness as the power of the test increases" (Howson and Urbach 1993, p. 209). In fact, the thesis implicit in the N-P approach, as Bakan remarks, is the opposite! The fallacy is akin to making mountains out of molehills according to severity (Section 3.2):

> *Mountains out of Molehills* (MM) *Fallacy* (large n problem): The fallacy of taking a rejection of H_0, just at level P, with larger sample size (*higher power*) as indicative of a greater discrepancy from H_0 than with a smaller sample size.

Consider an analogy with two fire alarms: The first goes off with a sensor liable to pick up on burnt toast; the second is so insensitive it doesn't kick in until your house is fully ablaze. You're in another state, but you get a signal when the alarm goes off. Which fire alarm indicates the greater extent of fire? Answer: the second, less sensitive one. When the sample size increases it alters what counts as a *single sample*. It is like increasing the sensitivity of your fire alarm. It is true that a large enough sample size triggers the alarm with an observed mean that is quite "close" to the null hypothesis. But, if the test rings the alarm (i.e., rejects H_0) even for tiny discrepancies from the null value, then the alarm is poor grounds for inferring larger discrepancies. Now this is an analogy, you may poke holes in it. For instance, a test must have a large enough sample to satisfy model assumptions. True, but our interpretive question can't even get started without taking the P-values as legitimate and not spurious.

4.3 Significant Results with Overly Sensitive Tests: Large *n* Problem

> "[W]ith a large sample size virtually every null hypothesis is rejected, while with a small sample size, virtually no null hypothesis is rejected. And we generally have very accurate estimates of the sample size available without having to use significance testing at all!" (Kadane 2011, p. 438).

P-values are sensitive to sample size, but to see this as a problem is to forget what significance tests are for. We want consistent tests, so that as n increases the probability of discerning any discrepancy from the null (i.e., the power) increases. The fact that the test would eventually uncover any discrepancy

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

there may be, regardless of how small, doesn't mean there always is such a discrepancy, by the way. (Another little confusion repeated in the form of "all null hypotheses are false.") Let's focus on the example of Normal testing, T+ with H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$ letting $\sigma = 1$. It's precisely to bring out the effect of sample size that many prefer to write the statistic as

$$\mathrm{d}(\boldsymbol{X}) = \surd n(\overline{X} - 0)/\sigma$$

rather than

$$\mathrm{d}(\boldsymbol{X}) = (\overline{X} - 0)/\sigma_{\overline{X}},$$

where $\sigma_{\overline{X}}$ abbreviates $(\sigma/\surd n)$.

T+ rejects H_0 (at the 0.025 level) iff the sample mean $\overline{X} \geq 0 + 1.96(\sigma/\surd n)$. As n increases, a single $(\sigma/\surd n)$ unit decreases. Thus the value of $\overline{X}$ required to reach significance decreases as n increases.

The test's goal is to distinguish observed effects due to ordinary expected variability under H_0 with those that cannot be readily explained by mere noise. If the inter-ocular test will do, you don't need statistics. As the sample size increases, the ordinary expected variability decreases. The severe tester takes account of the sample size in interpreting the discrepancy indicated. The test is like a thermostat, a fire alarm, or the mesh size in a fishing net. You choose the sensitivity, and it does what you told it to do.

Keep in mind that the hypotheses entertained are not point values, but discrepancies. Informally, for a severe tester, each corresponds to an assertion of form: there's evidence of a discrepancy at least this large, but there's poor evidence it's as large as thus and so. Let's compare statistically significant results at the same level but with different sample sizes.

Consider the 2-standard deviation cut-off for $n = 25, 100, 400$ in test T+, $\sigma = 1$ (Figure 4.1).

Let $\overline{x}_{0.025}$ abbreviate the sample mean that is just statistically significant at the 0.025 level in each test. With $n = 25$, $\overline{x}_{.025} = 2(1/5)$; with $n = 100$, $\overline{x}_{0.025} = 2(1/10)$; with $n = 400$, $\overline{x}_{0.025} = 2(1/20)$. So the cut-offs for rejection are 0.4, 0.2, and 0.1, respectively.

Again, alterations of the sample size change what counts as one unit. If you treat identical values of $(\overline{X} - \mu_0)/\sigma$ the same, ignoring $\surd n$, you will misinterpret your results. With large enough n, the cut-off for rejection can be so close to the null value as to lead some accounts to regard it as evidence *for* the null. This is the Jeffreys–Lindley paradox that we'll be visiting this afternoon (Section 4.4).

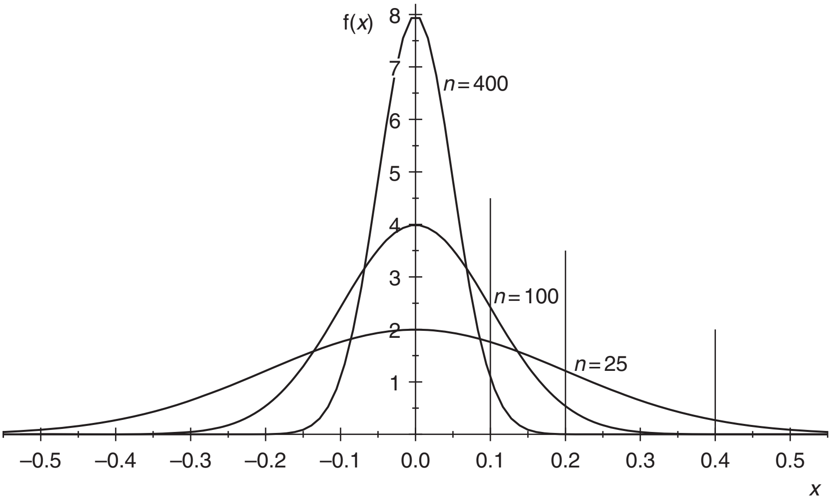

Figure 4.1 $\overline{X} \sim N(\mu, \sigma^2/n)$ for n = 25, 100, 400.

Exhibit (v): Responding to a Familiar Chestnut

> Did you hear the one about the significance tester who rejected H_0 in favor of H_1 even though the result makes H_0 more likely than H_1?

I follow the treatment in Elliott Sober (2008, p. 56), who is echoing Howson and Urbach (1993, pp. 208–9), who are echoing Lindley (1957). The only difference is that I will allude to a variant on test T+: H_0: $\mu = 0$ vs. H_1: $\mu > 0$ with $\sigma = 1$. "Another odd property of significance tests," says Sober, "concerns the way in which they are sensitive to sample size." Suppose you are applying test T+ with null H_0: $\mu = 0$. If your sample size is $n = 25$, and you choose $\alpha = 0.025$, you will reject H_0 whenever $\overline{x} \geq 0.4$. If you examine $n = 100$, and choose the same value for α, you will reject H_0 whenever $\overline{x} \geq 0.2$. And if you examine $n = 400$, again with $\alpha = 0.025$, you will reject H_0 whenever $\overline{x} \geq 0.1$. "As sample size increases" the sample mean $\overline{x}$ must be closer and closer to 0 for you *not* to reject H_0. "This may not seem strange until you add the following detail. Suppose the alternative to H_0 is the hypothesis" H_1: $\mu = 0.24$. "The Law of Likelihood now entails that observing" $\overline{x} < 0.12$ favors H_0 over H_1, so in particular $\overline{x} = 0.1$ favors H_0 over H_1 (Section 1.4).

Your reply: Hold it right at "add the following detail." You're observing that the significance test disagrees with a Law of Likelihood appraisal to a point vs. point test: H_0: $\mu = 0$ vs. H_1: $\mu = 0.24$. We require the null and alternative

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

hypotheses to exhaust the space of parameters, and these don't. Nor are our inferences to points, but rather to inequalities about discrepancies. That said, we're prepared to consider your example, and make short work of it. We're testing H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$ (one could equally view it as testing H_0: $\mu = 0$ vs. H_1: $\mu > 0$). The outcome $\overline{x} = 0.1$, while indicating *some* positive discrepancy from 0, offers bad evidence and an insevere test for inferring μ as great as 0.24. Since $\overline{x} = 0.1$ rejects H_0, we say the result *accords* with H_1. The severity associated with inference $\mu \geq 0.24$ asks: what's the probability of observing $\overline{X} \leq 0.1$ – i.e., a result *more discordant* with H_1 assuming $\mu = 0.24$.

SEV($\mu \geq 0.24$) with $\overline{x} = 0.1$ and $n = 400$ is computed as Pr($\overline{X} < 0.1$; $\mu = 0.24$). Standardizing $\overline{X}$ yields $Z = \sqrt{400}\ (0.1 - 0.24)/1 = 20(-0.14) = -2.8$. So SEV($\mu \geq 0.24$) = 0.003! Were μ as large as 0.24, we'd have observed a larger observed mean than we did with 0.997 probability! It's terrible evidence for H_1: $\mu = 0.24$.

This is redolent of the Binomial example in discussing Royall (Section 1.4). To underscore the difference between the Likelihoodist's comparative appraisal and the significance tester, you might go further. Consider an alternative that the Likelihoodist takes as favored over H_0: $\mu = 0$ with $\overline{x} = 0.1$, namely, the maximum likely alternative H_1: $\mu = 0.1$. This is one of our key benchmarks for a discrepancy that's poorly indicated. To the Likelihoodist, inferring that H_1: $\mu = 0.1$ "is favored" over H_0: $\mu = 0$ makes sense, whereas to infer a discrepancy of 0.1 from H_0 is highly *un*warranted for a significance tester.[1] Our aims are very different.

We can grant this: starting with any value for $\overline{x}$, however close to 0, there's an n such that $\overline{x}$ is statistically significantly greater than 0 at a chosen level. If one understands the test's intended task, this is precisely what is wanted. How large would n need to be so that 0.02 is statistically significant at the 0.025 level (still retaining σ = l)?

Answer: Setting $0.02 = 2(1/\sqrt{n})$ and solving for n yields $n = 10{,}000$.[2]

Statistics won't tell you what magnitudes are of relevance to you. No matter, we can critique results and purported inferences.

[1] SEV($\mu \geq 0.1$) with $\overline{x} = 0.1$ and $n = 400$ is computed by considering Pr($\overline{X} < 0.1$; $\mu = 0.1$). Standardizing $\overline{X}$ yields $z = \sqrt{400}\ (0.1 - 0.1)/1 = 0$. So SEV($\mu \geq 0.1$) = 0.5!

[2] Let's use this to illustrate the MM fallacy: Compare (i) $n = 100$ and (ii) $n = 10{,}000$ in the same test T+. With $n = 100$, 1SE = 0.1, with $n = 10{,}000$, 1SE = 0.01. The just 0.025 significant outcomes in the two tests are (i) $\overline{x} = 0.2$ and (ii) $\overline{x} = 0.02$. Consider the 0.93 lower confidence bound for each. Subtracting 1.5 SE from the outcome yields $\mu > 0.5(1/\sqrt{n})$: (i) for $n = 100$, the inferred 0.93 lower estimate is. $\mu > 0.5(1/5) = 0.05$, (ii) for $n = 10{,}000$, the inferred 0.93 lower estimate is $\mu > 0.5(1/100) = 0.005$. So a difference that is just statistically significant at the same level, 0.025, permits inferring $\mu > 0.05$ when $n = 25$, but only $\mu > 0.005$ when $n = 10{,}000$ Section 3.7.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

Exhibit (vi): Reforming the Reformers on Confidence Intervals. You will be right to wonder why some of the same tribes who raise a ruckus over *P*-values – to the extent, in some cases, of calling for a "test ban" – are cheerleading for confidence intervals (CIs), given there is a clear duality between the two (Section 3.7). What they're really objecting to is a dichotomous use of significance tests where the report is "significant" or not at a predesignated significance level. I completely agree with this objection, and reject the dichotomous use of tests (which isn't to say there are no contexts where an "up/down" indication is apt). We should reject Unitarianism where a single method with a single interpretation must be chosen. Ironically, some of the most outspoken CI leaders use them in the dichotomous fashion (rightly) deplored when it comes to testing.

Geoffrey Cumming, an acknowledged tribal leader on CIs, tells us that "One-sided CIs are analogous to one-tailed tests but, as usual, the estimation approach is better" (2012, p. 109). Well, it might be better, but like hypothesis testing, it calls for supplements and reinterpretations as begun in Section 3.7.

Our one-sided test T+ (H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, and $\sigma = 1$) at $\alpha = 0.025$ has as its dual the one-sided (lower) 97.5% general confidence interval: $\mu > \overline{X} - 2(1/\surd n)$ – rounding to 2 from 1.96. So you won't have to flip back pages, here's a quick review of the notation we developed to avoid the common slipperiness with confidence intervals. We abbreviate the generic lower limit of a $(1 - \alpha)$ confidence interval as $\hat{\mu}_{1-\alpha}(\overline{X})$ and the particular limit as $\hat{\mu}_{1-\alpha}(\overline{x})$. The general estimating procedure is: Infer $\mu > \hat{\mu}_{1-\alpha}(\overline{X})$. The particular estimate is $\mu > \hat{\mu}_{1-\alpha}(\overline{x})$. Letting $\alpha = 0.025$ we have: $\mu > \overline{x} - 2(1/\surd n)$. With $\alpha = 0.05$, we have $\mu > \overline{x} - 1.65(1/\surd n)$.

Cumming's interpretation of CIs and confidence levels points to their performance-oriented construal: "In the long run 95% of one-sided CIs will include the population mean . . . We can say we're 95% confident our one-sided interval includes the true value . . . meaning that for 5% of replications the [lower limit] will exceed the true value" (Cumming 2012, p. 112). What does it mean to be 95% confident in the particular interval estimate for Cumming? "It means that the values in the interval are plausible as true values for μ, and that values outside the interval are relatively implausible – though not impossible" (ibid., p. 79). The performance properties of the method rub off in a plausibility assessment of some sort.

The test that's dual to the CI would "accept" those parameter values within the corresponding interval, and reject those outside, all at a single predesignated confidence level 1 – α. Our main objection to this is it gives the misleading idea that there's evidence for each value in the interval, whereas, in fact, the interval simply consists of values that aren't rejectable, were one

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

testing at the α level. Not being a rejectable value isn't the same as having evidence for that value. Some values are close to being rejectable, and we should convey this. Standard CIs do not.

To focus on how CIs deal with distinguishing sample sizes, consider again the three instances of test T+ with (i) $n = 25$, (ii) $n = 100$, and (iii) $n = 400$. Imagine the observed mean $\overline{x}$ from each test just hits the significance level 0.025. That is, (i) $\overline{x} = 0.4$, (ii) $\overline{x} = 0.2$, and (iii) $\overline{x} = 0.1$. Form 0.975 confidence interval estimates for each:

(i) for $n = 25$, the inferred estimate is $\mu > \hat{\mu}_{0.975}$, that is, $\mu > \overline{x} - 2(1/5)$;
(ii) for $n = 100$, the inferred estimate is $\mu > \hat{\mu}_{0.975}$, that is, $\mu > \overline{x} - 2(1/10)$;
(iii) for $n = 400$, the inferred estimate is $\mu > \hat{\mu}_{0.975}$, that is, $\mu > \overline{x} - 2(1/20)$.

Substituting $\overline{x}$ in all cases, we get the same one-sided confidence interval:

$$\mu > 0.$$

Cumming writes them as [0, infinity). How are the CIs distinguishing them?

They are not. The construal is dichotomous: in or out, plausible or not. Would we really want to say "the values in the interval are plausible as true values for μ"? Clearly not, since that includes values to infinity. I don't want to step too hard on the CI champion's toes, since CIs are in the frequentist, error statistical tribe. Yet, to avoid fallacies, this standard use of CIs won't suffice. Severity directs you to avoid taking your result as indicating a discrepancy beyond what's warranted. For an example, we can show the same inference is poorly indicated with $n = 400$, while fairly well indicated when $n = 100$. For a poorly indicated claim, take our benchmark for severity of 0.5; for fairly well, 0.84:

$$\text{For } n = 400,\ \overline{x}_{0.025} = 0.1,\ \text{so } \mu > 0.1 \text{ is poorly indicated;}$$

$$\text{For } n = 100,\ \overline{x}_{0.025} = 0.2,\ \text{and } \mu > 0.1 \text{ is fairly well indicated.}$$

The reasoning based on severity is counterfactual: were μ less than or equal to 0.1, it is fairly probable, 0.84, that a smaller $\overline{X}$ would have occurred. This is not part of the standard CI account, but enables the distinction we want. Another move would be for a CI advocate to require we always compute a two-sided interval. The upper 0.975 bound would reflect the greater sensitivity with increasing sample sizes:

(i) $n = 25$: (0, 0.8], (ii) $n = 100$: (0, 0.4], (iii) $n = 400$: (0, 0.2].

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

But we cannot just deny one-sided tests, nor does Cumming. In fact, he encourages their use: “it’s unfortunate they are usually ignored” (2012, p. 113). (He also says he is happy for people to decide afterwards whether to report it as a one- or two-sided interval (ibid., p. 112), only doubling α, which I do not mind.) Still needed is a justification for bringing in the upper limit when applying a one-sided estimator, and severity supplies it. You should always be interested in at least *two benchmarks*: discrepancies well warranted and those terribly warranted. In test T+, our handy benchmark for the terrible is to set the lower limit to $\overline{x}$. The severity for $(\mu > \overline{x})$ is 0.5. Two side notes:

First I grant it would be wrong to charge Cumming with treating all parameter values within the confidence interval *on par*, because he does suggest distinguishing them by their likelihoods (by how probable each renders the outcome). Take just the single 0.975 lower CI bound with $n = 100$ and $\overline{x} = 0.2$. A μ value closer to the observed 0.2 has higher likelihood (in the technical sense) than ones close to the 0.975 lower limit 0. For example, $\mu = 0.15$ is more likely than $\mu = 0.05$. However, this moves away from CI reasoning (toward likelihood comparisons). The claim $\mu > 0.05$ has a *higher* confidence level (0.93) than does $\mu > 0.15$ (0.7)[3] even though the point hypothesis $\mu = 0.05$ is less likely than $\mu = 0.15$ (the latter is closer to $\overline{\mathrm{x}} = 0.2$ than is the former). Each point in the lower CI corresponds to a different lower bound, each associated with a different confidence level, and corresponding severity assessment. That’s how to distinguish them.

Second there’s an equivocation, or at least a potential equivocation, in Cumming’s assertion “that for [2.5%] of replications the [lower limit] will exceed the true value” (Cumming 2012, p. 112 replacing 5% with 2.5%). This is not a true claim if “lower limit” is replaced by a *particular* lower limit: $\hat{\mu}_{0.025}(\overline{x})$, it holds only for the *generic* lower limit $\hat{\mu}_{0.025}(\overline{X})$. That is, we can’t say μ exceeds zero 2.5% of the time, which would be to assign a probability of 0.975 to $\mu > 0$. Yet this misinterpretation of CIs is legion, as we’ll see in a historical battle about fiducial intervals (Section 5.8).

4.4 Do *P*-Values Exaggerate the Evidence?

“Significance levels overstate the evidence against the null hypothesis,” is a line you may often hear. Your first question is:

> What do you mean by overstating the evidence against a hypothesis?

Several (honest) answers are possible. Here is one possibility:

[3] Subtract 1.5 SE and 0.5 SE from $\overline{x} = 0.2$, respectively.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

> What I mean is that when I put a lump of prior weight π_0 of 1/2 on a point null H_0 (or a very small interval around it), the P-value is smaller than my Bayesian posterior probability on H_0.

More generally, the "P-values exaggerate" criticism typically boils down to showing that if inference is appraised via one of the probabilisms – Bayesian posteriors, Bayes factors, or likelihood ratios – the evidence against the null (or against the null and in favor of some alternative) isn't as big as $1 - P$.

You might react by observing that: (a) P-values are not intended as posteriors in H_0 (or Bayes ratios, likelihood ratios) but rather are used to determine if there's an indication of discrepancy from, or inconsistency with, H_0. This might only mean it's worth getting more data to probe for a real effect. It's not a degree of belief or comparative strength of support to walk away with. (b) Thus there's no reason to suppose a P-value should match numbers computed in very different accounts, that differ among themselves, and are measuring entirely different things. Stephen Senn gives an analogy with "height and stones":

> . . . [S]ome Bayesians in criticizing P-values seem to think that it is appropriate to use a threshold for significance of 0.95 of the probability of the alternative hypothesis being true. This makes no more sense than, in moving from a minimum height standard (say) for recruiting police officers to a minimum weight standard, declaring that since it was previously 6 foot it must now be 6 stone. (Senn 2001b, p. 202)

To top off your rejoinder, you might ask: (c) Why assume that "the" or even "a" correct measure of evidence (relevant for scrutinizing the P-value) is one of the probabilist ones?

All such retorts are valid, and we'll want to explore how they play out here. Yet, I want to push beyond them. Let's be open to the possibility that evidential measures from very different accounts can be used to scrutinize each other.

Getting Beyond "I'm Rubber and You're Glue". The danger in critiquing statistical method X from the standpoint of the goals and measures of a distinct school Y, is that of falling into begging the question. If the P-value is exaggerating evidence against a null, meaning it seems too small from the perspective of school Y, then Y's numbers are too big, or just irrelevant, from the perspective of school X. Whatever you say about me bounces off and sticks to you. This is a genuine worry, but it's not fatal. The goal of this journey is to identify minimal theses about "bad evidence, no test (BENT)" that enable some degree of scrutiny of any statistical inference account – at least on the meta-level. Why assume all schools of statistical inference embrace the minimum severity principle? I don't, and they don't. But by identifying when methods violate

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

severity, we can pull back the veil on at least one source of disagreement behind the battles.

Thus, in tackling this latest canard, let's resist depicting the critics as committing a gross blunder of confusing a P-value with a posterior probability in a null. We resist, as well, merely denying we care about their measure of support. I say we should look at exactly what the critics are on about. When we do, we will have gleaned some short-cuts for grasping a plethora of critical debates. We may even wind up with new respect for what a P-value, the least popular girl in the class, really *does*.

To visit the core arguments, we travel to 1987 to papers by J. Berger and Sellke, and Casella and R. Berger. These, in turn, are based on a handful of older ones (Cox 1977, E, L, & S 1963, Pratt 1965), and current discussions invariably revert back to them. Our struggles through quicksand of Excursion 3, Tour II, are about to pay large dividends.

J. Berger and Sellke, and Casella and R. Berger. Berger and Sellke (1987a) make out the conflict between P-values and Bayesian posteriors by considering the two-sided test of the Normal mean, H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$. "Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$, where the X_i are IID N(μ, σ^2), σ^2 known" (p. 112). Then the test statistic $d(\boldsymbol{X}) = \sqrt{n}\,|\overline{X} - \mu_0|/\sigma$, and the P-value will be twice the P-value of the corresponding one-sided test.

Starting with a lump of prior, generally 0.5, on the point hypothesis H_0, they find the posterior probability in H_0 is larger than the P-value for a variety of different priors on the alternative. However, the result depends entirely on how the remaining 0.5 is allocated or smeared over the alternative (a move dubbed spike and smear). Using what they call a Jeffreys-type prior, the 0.5 is spread out over the alternative parameter values as if the parameter is itself distributed N(μ_0, σ). Now Harold Jeffreys recommends the lump prior only to capture cases where a special value of a parameter is deemed plausible, for instance, the GTR deflection effect $\lambda = 1.75''$, after about 1960. The rationale is to avoid a 0 prior on H_0 and enable it to receive a reasonable posterior probability .

By subtitling their paper "The irreconcilability of P-values and evidence," Berger and Sellke imply that if P-values disagree with posterior assessments, they can't be measures of evidence at all. Casella and R. Berger (1987) retort that "reconciling" is at hand, if you move away from the lump prior. So let's see how this unfolds. I assume throughout, as do the critics, that the P-values are "audited," so that neither selection effects nor violated model assumptions are in question at this stage. I see no other way to engage their arguments.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

Table 4.1 Pr($H_0|\boldsymbol{x}$) for Jeffreys-type prior

		***n* (sample size)**				
P one-sided	z_α	**10**	**20**	**50**	**100**	**1000**
0.05	1.645	0.47	0.56	0.65	0.72	0.89
0.025	1.960	0.37	0.42	0.52	0.60	0.82
0.005	2.576	0.14	0.16	0.22	0.27	0.53
0.0005	3.291	0.024	0.026	0.034	0.045	0.124

(From Table 1, J. Berger and T. Sellke (1987) p. 113 using the one-sided *P*-value)

Table 4.1 gives the values of Pr($H_0|\boldsymbol{x}$). We see that we would declare no evidence against the null, and even evidence for it (to the degree indicated by the posterior) whenever d($\boldsymbol{x}$) fails to reach a 2.5 or 3 standard error difference. With n = 50, "one can classically 'reject H_0 at significance level p = 0.05,' although Pr($H_0|\boldsymbol{x}$) = 0.52 (which would actually indicate that the evidence *favors* H_0)" (J. Berger and Sellke 1987, p. 113).

If n = 1000, a result statistically significant at the 0.05 level results in the posterior probability to $\mu = 0$ going up from 0.5 (the lump prior) to 0.82! From their Bayesian perspective, this suggests *P*-values are exaggerating evidence against H_0. Error statistical testers, on the other hand, balk at the fact that using the recommended priors allows statistically significant results to be interpreted as no evidence against H_0 – or even evidence for it. They point out that 0 is excluded from the two-sided confidence interval at level 0.95. Although a posterior probability doesn't have an error probability attached, a tester can evaluate the error probability credentials of these inferences. Here we'd be concerned with a Type II error: failing to find evidence against the null, and providing a fairly high posterior for it, when it's false (Souvenir I).

Let's use a less extreme example where we have some numbers handy: our water-plant accident. We had $\sigma = 10$, $n = 100$ leading to the nice ($\sigma/\sqrt{n}$) value of 1. Here it would be two-sided, to match their example: H_0: $\mu = 150$ vs. H_1: $\mu \neq 150$. Look at the second entry of the 100 column, the posterior when $z_\alpha = 1.96$. With the Jeffreys prior, perhaps championed by the water coolant company, J. Berger and Sellke assign a posterior of 0.6 to H_0: $\mu = 150$ degrees when a mean temperature of 152 (151.96) degrees is observed – reporting decent evidence the cooling mechanism is working just fine. How often would this occur even if the actual underlying mean temperature is, say, 151 degrees? With a two-sided test, cutting off 2 standard errors on either side, we'd reject whenever either $\overline{X} \geq 152$ or $\overline{X} \leq 148$. The probability of the second is negligible under μ = 151, so the probability we want is

$\Pr(\overline{X} < 152;\ \mu = 151) = 0.84\ (Z = (152 - 151) = 1)$. The probability of declaring evidence for 150 degrees (with posterior of 0.6 to H_0) even if the true increase is actually 151 degrees is around 0.84; 84% of the time they erroneously fail to ring the alarm, and would boost their probability of $\mu = 150$ from 0.5 to 0.6. Thus, from our minimal severity principle, the statistically significant result can't even be taken as evidence for compliance with 151 degrees, let alone as evidence *for* the null of 150 (Table 3.1).

Is this a problem for them? It depends what you think of that prior. The N-P test, of course, does not use a prior, although, as noted earlier, one needn't rule out a frequentist prior on mean water temperature after an accident (Section 3.2). For now our goal is making out the criticism.

Jeffreys–Lindley "Paradox" or Bayes/Fisher Disagreement

But how, intuitively, does it happen that a statistically significant result corresponds to a Bayes boost for H_0? Go back to J. Berger and Sellke's example of Normal testing of H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$. Some sample mean $\bar{x}$ will be close enough to 0 to increase the posterior for H_0. By choosing a sufficiently large n, even a statistically significant result can correspond to large posteriors on H_0. This is the Jeffreys–Lindley "paradox," which some more aptly call the Bayes/Fisher disagreement. Lindley's famous result dealt with just this example, two-sided Normal testing with known variance. With a lump given to the point null, and the rest appropriately spread over the alternative, an n can be found such that an α significant result corresponds to $\Pr(H_0|\boldsymbol{x}) = (1 - \alpha)$! We can see by extending Table 4.1 to arbitrarily large n, we can get a posterior for the null of 0.95, when the (two-sided) P-value is 0.05. Many say you should decrease the required P-value for significance as n increases; and Cox and Hinkley (1974, p. 397) provide formulas to achieve this and avoid the mismatch. There's nothing in N-P or Fisherian theory to oppose this. I won't do that here, as I want to make out the criticism. We need only ensure that the interpretation takes account of the (obvious) fact that, with a fixed P-value and increasing n, the test is more and more sensitive to smaller and smaller discrepancies. Using a smaller plate at the French restaurant may make the portion appear bigger, but, Jackie Mason notwithstanding, knowing the size of the plate, I can see there's not much there.

Why assign the lump of ½ as prior to the point null? "The choice of $\pi_0 = 1/2$ has obvious intuitive appeal in scientific investigations as being 'objective'" say J. Berger and Sellke (1987, p. 115). But is it? One starts by making H_0 and H_1 equally probable, then the 0.5 accorded to H_1 is spread out over all the values in H_1: "The net result is that all values of $[\mu]$ are far from being equally likely"

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

(Senn 2015a). Any small group of μ values in H_1 gets a tiny prior. David Cox describes how it happens:

> . . . if a sample say at about the 5% level of significance is achieved, then either H_0 is true or some alternative in a band of order $1/\sqrt{n}$; the latter possibility has, as $n \to \infty$, prior probability of order $1/\sqrt{n}$ and hence at a fixed level of significance the posterior probabilities shift in favour of H_0 as n increases. (Cox 1977, p. 59)

What justifies the lump prior of 0.5?

A Dialogue at the Water Plant Accident

EPA Rep: The mean temperature of the water was found statistically significantly higher than 150 degrees at the 0.025 level.

Spiked Prior Rep: This even strengthens my belief the water temperature's no different from 150. If I update the prior of 0.5 that I give to the null hypothesis, my posterior for H_0 is still 0.6; it's not 0.025 or 0.05, that's for sure.

EPA Rep: Why do you assign such a high prior probability to H_0?

Spiked Prior Rep: If I gave H_0 a value lower than 0.5, then, if there's evidence to reject H_0, at most I would be claiming an improbable hypothesis has become more improbable.

[W]ho, after all, would be convinced by the statement 'I conducted a Bayesian test of H_0, assigning prior probability 0.1 to H_0, and my conclusion is that H_0 has posterior probability 0.05 and should be rejected?' (J. Berger and Sellke 1987, p. 115).

This quote from J. Berger and Sellke is peculiar. They go on to add: "We emphasize this obvious point because some react to the Bayesian–classical conflict by attempting to argue that [prior] π_0 should be made small in the Bayesian analysis so as to force agreement" (ibid.). We should not force agreement. But it's scarcely an obvious justification for a lump of prior on the null H_0 – one which results in a low capability to detect discrepancies – that it ensures, if they *do* reject H_0, there will be a meaningful drop in its probability. Let's listen to the pushback from Casella and R. Berger (1987a), the Berger being Roger now (I use initials to distinguish them).

The Cult of the Holy Spike and Slab. Casella and R. Berger (1987a) charge that the problem is not P-values but the high prior, and that "concentrating mass on the point null hypothesis is biasing the prior in favor of H_0 as much as possible" (p. 111) whether in one- or two-sided tests. According to them:

> The testing of a point null hypothesis is one of the most misused statistical procedures. In particular, in the location parameter problem, the point null

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

> hypothesis is more the mathematical convenience than the statistical method of choice. (ibid., p. 106)

Most of the time "there is a direction of interest in many experiments, and saddling an experimenter with a two-sided test would not be appropriate" (ibid.). The "cult of the holy spike" is an expression I owe to Sander Greenland (personal communication).

By contrast, we can reconcile *P*-values and posteriors in one-sided tests if we use more diffuse priors. (e.g., Cox and Hinkley 1974, Jeffreys 1939/1961, Pratt 1965). In fact, Casella and Berger show that for sensible priors in that case, the *P*-value is at least as big as the minimum value of the posterior probability on the null, again contradicting claims that *P*-values exaggerate the evidence.[4]

J. Berger and Sellke (1987) adhere to the spikey priors, but following E, L, & S (1963), they're keen to show that P-values exaggerate evidence even in cases less extreme than the Jeffreys posteriors in Table 4.1. Consider the likelihood ratio of the null hypothesis over the hypothesis most generous to the alternative, they say. This is the point alternative with maximum likelihood, $H_{\max}$ – arrived at by setting $\mu = \overline{x}$. Through their tunnel, it's disturbing that even using this likelihood ratio, the posterior for H_0 is still larger than 0.05 – when they give a 0.5 spike to both H_0 and $H_{\max}$. Some recent authors see this as the key to explain today's lack of replication of significant results. Through the testing tunnel, things look different (Section 4.5).

Why Blame Us Because You Can't Agree on Your Posterior? Stephen Senn argues that the reason for the wide range of variation of the posterior is the fact that it depends radically on the choice of alternative to the null and its prior.[5] According to Senn, "... the reason that Bayesians can regard *P*-values as overstating the evidence against the null is simply a reflection of the fact that Bayesians can disagree *sharply* with each other" (Senn 2002, p. 2442). Senn

[4] Casella and R. Berger (1987b) argue, "We would be surprised if most researchers would place even a 10% prior probability of H_0. We hope that the casual reader of Berger and Delampady realizes that the big discrepancies between P-values $P(H_0|\boldsymbol{x})$ ... are due to a large extent to the large value of [the prior of 0.5 to H_0] that was used." The most common uses of a point null, asserting the difference between means is 0, or the coefficient of a regression coefficient is 0, merely describe a potentially interesting feature of the population, with no special prior believability. "Berger and Delampady admit ..., P-values are reasonable measures of evidence when there is no a priori concentration of belief about H_0" (ibid., p. 345). Thus, "the very argument that Berger and Delampady use to dismiss P-values can be turned around to argue *for* P-values" (ibid., p. 346).

[5] In defending spiked priors, Berger and Sellke move away from the importance of effect size. "Precise hypotheses ... ideally relate to, say, some precise theory being tested. Of primary interest is whether the theory is right or wrong; the amount by which it is wrong may be of interest in developing alternative theories, but the initial question of interest is that modeled by the precise hypothesis test" (1987, p. 136).

illustrates how "two Bayesians having the same prior probability that a hypothesis is true and having seen the same data can come to radically different conclusions because they differ regarding the alternative hypothesis" (Senn 2001b, p. 195). One of them views the problem as a one-sided test and gets a posterior on the null that matches the *P*-value; a second chooses a Jeffreys-type prior in a two-sided test, and winds up with a posterior to the null of $1 - p$!

Here's a recap of Senn's example (ibid., p. 200): Two scientists A and B are testing a new drug to establish its treatment effect, δ, where positive values of δ are good. Scientist A has a vague prior whereas B, while sharing the same distribution about the probability of positive values of δ, is less pessimistic than A regarding the effect of the drug. If it's not useful, B believes it will have no effect. They "share the same belief that the drug has a positive effect. Given that it has a positive effect, they share the same belief regarding its effect. . . . They differ only in belief as to how harmful it might be." A clinical trial yields a difference of 1.65 standard units, a one-sided *P*-value of 0.05. The result is that A gives 1/20 posterior probability to H_0: the drug does *not* have a positive effect, while B gives a probability of 19/20 to H_0. B is using the two-sided test with a lump of prior on the null (H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$), while A is using a one-sided test T+ (H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$). The contrast, Senn observes, is that of Cox's distinction between "precise and dividing hypothesis" (Section 3.3). "[F]rom a common belief in the drug's efficacy they have moved in opposite directions" (ibid., pp. 200–201). Senn riffs on Jeffreys' well-known joke that we heard in Section 3.4:

> It would require that a procedure is dismissed [by significance testers] because, when combined with information which it doesn't require and which may not exist, it disagrees with a [Bayesian] procedure that disagrees with itself. (ibid., p. 195)

In other words, if Bayesians disagree with each other even when they're measuring the same thing – posterior probabilities – why be surprised that disagreement is found between posteriors and *P*-values? The most common argument behind the "*P*-values exaggerate evidence" appears not to hold water. Yet it won't be zapped quite so easily, and will reappear in different forms.

Exhibit (vii): Contrasting Bayes Factors and Jeffreys–Lindley Paradox. We've uncovered some interesting bones in our dig. Some lead to seductive arguments purporting to absolve the latitude in assigning priors in Bayesian tests. Take Wagenmakers and Grünwald (2006, p. 642): "Bayesian hypothesis tests are often criticized because of their dependence on prior distributions

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

... [yet] no matter what prior is used, the Bayesian test provides substantially less evidence against H_0 than" P-values, in the examples we've considered. Be careful in translating this. We've seen that what counts as "less" evidence runs from seriously underestimating to overestimating the discrepancy we are entitled to infer with severity. Begin with three types of priors appealed to in some prominent criticisms revolving around the Fisher – Jeffreys disagreement.

1. *Jeffreys-type prior with the "spike and slab" in a two-sided test.* Here, with large enough n, a statistically significant result becomes evidence *for* the null; the posterior to H_0 exceeds the lump prior.
2. *Likelihood ratio most generous to the alternative.* Here, there's a spike to a point null, H_0: $\theta = \theta_0$ to be compared to the point alternative that's maximally likely $\theta_{\max}$. Often, both H_0 and $H_{\max}$ are given 0.5 priors.
3. *Matching.* Instead of a spike prior on the null, it uses a smooth diffuse prior, as in the "dividing" case. Here, the P-value "is an approximation to the posterior probability that $\theta < 0$" (Pratt 1965, p. 182).

In sync with our attention to high-energy particle physics (HEP) in Section 3.6, consider an example that Aris Spanos (2013b) explores in relation to the Jeffreys–Lindley paradox. The example is briefly noted in Stone (1997).

A large number (n = 527,135) of independent collisions that can be of either type A or type B are used to test if the proportion of type A collisions is exactly 0.2, as opposed to any other value. It's modeled as n Bernoulli trials testing H_0: $\theta = 0.2$ vs. H_1: $\theta \neq 0.2$. The observed proportion of type A collisions is scarcely greater than the point null of 0.2:

$$\overline{x} = k/n = 0.20165233, \text{ where } n = 527{,}135,\ k = 106{,}298.$$

The significance level against H_0 is small *(so there's evidence against H_0)*

The *test statistic* $\mathrm{d}(\boldsymbol{X}) = [\sqrt{n}(\overline{X} - 0.2)/\sigma] = 3$, $\sigma = \sqrt{[\theta(1-\theta)]}$, which under the null is $\sqrt{[0.2(0.8)]} = 0.4$. The significance level associated with $\mathrm{d}(\boldsymbol{x}_0)$ in this two-sided test is

$$\Pr(|\mathrm{d}(\boldsymbol{X})| > |\mathrm{d}(\boldsymbol{x}_0)|; H_0) = 0.0027.$$

So the result $\overline{x}$ is highly significant, even though it's scarcely different from the point null.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

The Bayes factor in favor of H_0 is high

H_0 is given the spiked prior of 0.5, and the remaining 0.5 is spread equally among the values in H_1. I follow Spanos' computations:[6]

$$\Pr(k|H_0) = \binom{n}{k} 0.2^k (0.8)^{n-k},$$

$$\Pr(k|H_1) = \int_1^0 \binom{n}{k} \theta^k (1-\theta)^{n-k} \mathrm{d}\theta = 1/(n+\ 1),$$

where n = 527,135 and k = 106,298.

$$\text{The Bayes factor } B_{01} = \Pr(k|H_0)/\ \Pr(k|H_1) = 0.000015394/0.000001897$$
$$= 8.115.$$

While the likelihood of H_0 in the numerator is tiny, the likelihood of H_1 is even tinier. Since B_{01} in favor of H_0 is 8, which is greater than 1, the posterior for H_0 goes up, even though the outcome is statistically significantly greater than the null.

There's no surprise once you consider the Bayesian question here: compare the likelihood of a result scarcely different from 0.2 being produced by a universe where θ = 0.2 – where this has been given a spiked prior of 0.5 under H_0 – with the likelihood of that result being produced by any θ in a small band of θ values, which have been given a very low prior under H_1. Clearly, $\theta = 0.2$ is more likely, and we have an example of the Jeffreys–Fisher disagreement.

Who should be afraid of this disagreement (to echo the title of Spanos' paper)? Many tribes, including some Bayesians, think it only goes to cast doubt on this particular Bayes factor. Compare it with proposal 2 in Exhibit (vii): the *Likelihood ratio most generous to the alternative*: Lik(0.2)/Lik($\theta_{\max}$). We know the maximally likely value for θ, $\theta_{\max} = \overline{x}$:

$$\overline{x} = k/n = 0.20165233 = \theta_{\max},$$

$$\Pr(k|H_{\max}) = \binom{n}{k} 0.20165233^k (1-0.20165233)^{n-k} = 0.0013694656,$$

$$\mathrm{Lik}(0.2) = 0.000015394, \text{ and } \mathrm{Lik}(\theta_{\max}) = 0.0013694656.$$

Now B_{01} is 0.01 and B_{10}, Lik($\theta_{\max}$)/Lik(0.2) = 89.

Why should a result 89 times more likely under alternative $\theta_{\max}$ than under $\theta = 0.2$ be taken as strong evidence *for* $\theta = 0.2$? It shouldn't, according to some, including Lindley's own student, default Bayesian José Bernardo (2010).

[6] The spiked prior drops out, so the result is the same as a uniform prior on the null and alternative.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

Presumably, the Likelihoodist concurs. There are family feuds within and between the diverse tribes of probabilisms.[7]

Greenland and Poole Given how often spiked priors arise in foundational arguments, it's worth noting that even Bayesians Edwards, Lindman, and Savage (1963, p. 235), despite raising the "*P*-values exaggerate" argument, aver that for Bayesian statisticians, "no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence." Epidemiologists Sander Greenland and Charles Poole, who claim not to identify with any one statistical tribe, but who often lead critics of significance tests, say:

> Our stand against spikes directly contradicts a good portion of the Bayesian literature, where null spikes are used too freely to represent the belief that a parameter 'differs negligibly' from the null. In many settings . . . even a tightly concentrated probability near the null has no basis in genuine evidence. Many scientists and statisticians exhibit quite a bit of irrational prejudice in favor of the null . . . (2013, p. 77).

They angle to reconcile *P*-values and posteriors, and to this end they invoke the matching result in # 3, Exhibit (vii). An uninformative prior, assigning equal probability to all values of the parameter, allows the *P*-value to approximate the posterior probability that $\theta < 0$ in one-sided testing ($\theta \leq 0$ vs. $\theta > 0$). In two-sided testing, the posterior probability that θ is on the opposite side of 0 than the observed is *P*/2. They proffer this as a way "to live with" *P*-values. Commenting on them, Andrew Gelman (2013, p. 72) raises this objection:

> [C]onsider what would happen if we routinely interpreted one-sided *P* values as posterior probabilities. In that case, an experimental result that is 1 standard error from zero – that is, exactly what one might expect from chance alone – would imply an 83% posterior probability that the true effect in the population has the same direction as the observed pattern in the data at hand. It does not make sense to me to claim 83% certainty – 5 to 1 odds [to H_1] . . .

(The *P*-value is 0.16.) Rather than relying on non-informative priors, Gelman prefers to use prior information that leans towards the null. This avoids as high a posterior to H_1 as when using the matching result.

Greenland and Poole respond that Gelman is overlooking the hazard of "strong priors that are not well founded. . . . Whatever our prior opinion and

[7] Bernardo shocked his mentor in announcing that the Lindley paradox is really an indictment of the Bayesian computations: "Whether you call this a paradox or a disagreement, the fact that the Bayes factor for the null may be arbitrarily large for sufficiently large *n, however relatively unlikely the data may be under H_0* is, … deeply disturbing" (Bernardo 2010, p. 59).

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

its foundation, we still need reference analyses with weakly informative priors to alert us to how much our prior probabilities are driving our posterior probabilities" (2013, p. 76). They rightly point out that, in some circles, giving weight to the null can be the outgrowth of some ill-grounded metaphysics about "simplicity." Or it may be seen as an assumption akin to a presumption of innocence in law. So the question turns on the appropriate prior on the null.

Look what has happened! The problem was simply to express "I'm not impressed" with a result reaching a *P*-value of 0.16: Differences even larger than 1 standard error are not so very infrequent – they occur 16% of the time – even if there's zero effect. So I'm not convinced of the reality of the effect, based on this result. *P*-values did their job, reflecting as they do the severity requirement. H_1 has passed a lousy test. That's that. No prior probability assignment to H_0 is needed. Problem solved.

But there's a predilection for changing the problem (if you're a probabilist). Greenland and Poole feel they're helping us to live with *P*-values without misinterpretation. By choosing the prior so that the *P*-value matches the posterior on H_0, they supply us "with correct interpretations" (ibid., p. 77) where "correct interpretations" are those where the misinterpretation (of a *P*-value as a posterior in the null) is not a misinterpretation. To a severe tester, this results in completely changing the problem from an assessment of how well tested the reality of the effect is, with the given data, to what odds I would give in betting, or the like. We land in the same uncharted waters as with other attempts to fix *P*-values, when we could have stayed on the cruise ship, interpreting *P*-values as intended.

Souvenir Q: Have We Drifted From Testing Country? (Notes From an Intermission)

Before continuing, let's pull back for a moment, and take a coffee break at a place called Spike and Smear. Souvenir Q records our notes. We've been exploring the research program that appears to show, quite convincingly, that significance levels exaggerate the evidence against a null hypothesis, based on evidential assessments endorsed by various Bayesian and Likelihoodist accounts. We suspended the impulse to deny it can make sense to use a rival inference school to critique significance tests. We sought to explore if there's something to the cases they bring as ammunition to this conflict. The Bayesians say the disagreement between their numbers and *P*-values is relevant for impugning *P*-values, so we try to go along with them.

Reflect just on the first argument, pertaining to the case of two-sided Normal testing H_0: $\mu = 0$ vs. H_0: $\mu \neq 0$, which was the most impressive, particularly with $n \geq 50$. It showed that a statistically significant difference from a test hypothesis

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

at familiar levels, 0.05 or 0.025, can correspond to a result that a Bayesian takes as evidence *for* H_0. The prior for this case is the spike and smear, where the smear will be of the sort leading to J. Berger and Sellke's results, or similar. The test procedure is to move from a statistically significant result at the 0.025 level, say, and infer the posterior for H_0.

Now our minimal requirement for data $\boldsymbol{x}$ to provide evidence for a claim H is that

> (S-1) H accords with (agrees with) $\boldsymbol{x}$, and
> (S-2) there's a reasonable, preferably a high, probability that the procedure would have produced disagreement with H, if in fact H were false.

So let's apply these severity requirements to the data taken as evidence for H_0 here.

Consider (S-1). Is a result that is 1.96 or 2 standard errors away from 0 in good accord with 0? Well, 0 is excluded from the corresponding 95% confidence interval. That does not seem to be in accord with 0 at all. Still, they have provided measures whereby $\boldsymbol{x}$ does accord with H_0, the likelihood ratio or posterior probability on H_0. So, in keeping with the most useful and most generous way to use severity, let's grant (S-1) holds.

What about (S-2)? Has anything been done to probe the falsity of H_0? Let's allow that H_0 is not a precise point, but some very small set of values around 0. This is their example, and we're trying to give it as much credibility as possible. Did the falsity of H_0 have a good chance of showing itself? The falsity of H_0 here is H_1: $\mu \neq 0$. What's troubling is that we found the probability of failing to pick up on population discrepancies as much as 1 standard error in excess of 0 is rather high (0.84) with $n = 100$. Larger sample sizes yield even less capability. Nor are they merely announcing "no discrepancy from 0" in this case. They're finding evidence for 0!

So how did the Bayesian get the bump in posterior probability on the null? It was based on a spiked prior of 0.5 to H_0. All the other points get minuscule priors having to share the remaining 0.5 probability. What was the warrant for the 0.5 prior to H_0? J. Berger and Sellke are quite upfront about it: if they allowed the prior spike to be low, then a rejection of the null would merely be showing an improbable hypothesis got more improbable. "[W]ho, after all, would be convinced," recall their asking: if "my conclusion is that H_0 has posterior probability 0.05 and should be rejected" since it previously had probability, say 0.1 (1987, p. 115). A slight lowering of probability won't cut it. Moving from a low prior to a slightly higher one also lacks punch.

This explains their high prior (at least 0.5) on H_0, but is it evidence for it? Clearly not, nor does it purport to be. We needn't deny there are cases where a theoretical parameter value has passed severely (we saw this in the case of GTR in Excursion 3). But that's not what's happening here. Here they intend for the 0.5 prior to show, *in general*, that statistically significant results problematically exaggerate evidence.[8]

A tester would be worried when the rationale for a spike is to avoid looking foolish when rejecting with a small drop; she'd be worried too by a report: "I don't take observing a mean temperature of 152 in your 100 water samples as indicating it's hotter than 150, because I give a whopping spike to our coolants being in compliance." That is why Casella and R. Berger describe J. Berger and Sellke's spike and smear as maximally biased toward the null (1987a, p. 111). Don't forget the powerful role played by the choice of how to smear the 0.5 over the alternative! Bayesians might reassure us that the high Bayes factor for a point null doesn't depend on the priors given to H_0 and H_1, when what they mean is that it depends only on the priors given to discrepancies under H_1. It was the diffuse prior to the effect size that gave rise to the Jeffreys–Lindley Paradox. It affords huge latitude in what gets supported.

We thought we were traveling in testing territory; now it seems we've drifted off to a different place. It shouldn't be easy to take data as evidence for a claim when that claim is false; but here it is easy (the claim here being H_0). How can this be one of a handful of main ways to criticize significance tests as exaggerating evidence? Bring in a navigator from a Popperian testing tribe before we all feel ourselves at sea:

> Mere supporting instances are as a rule too cheap to be worth having . . . any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, if it can be refuted. (Popper 1983, p. 130)

The high spike and smear tactic can't be take as a basis from which to launch a critique of significance tests because it fails rather glaringly a minimum requirement for evidence, let alone a test. We met Bayesians who don't approve of these tests either, and I've heard it said that Bayesian testing is still a work in progress (Bernardo). Yet a related strategy is at the heart of some recommended statistical reforms.

[8] In the special case, where there's appreciable evidence for a special parameter, Senn argues that Jeffreys only required H_1's posterior probability to be greater than 0.5. One has, so to speak, used up the prior belief by using the spiked prior (Senn 2015a).

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

4.5 Who's Exaggerating? How to Evaluate Reforms Based on Bayes Factor Standards

Edwards, Lindman, and Savage (E, L, & S) – who were perhaps first to raise this criticism – say this:

> Imagine all the density under the alternative hypothesis concentrated at ***x***, the place most favored by the data. . . .
>
> Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest. (1963, p. 228)

The example is the Normal testing case of J. Berger and Sellke, but they compare it to a one-tailed test of H_0: $\mu = 0$ vs. H_1: $\mu = \mu_1 = \mu_{\max}$ (entirely sensibly in my view). We abbreviate H_1 by $H_{\max}$. Here the likelihood ratio Lik($\mu_{\max}$)/Lik(μ_0) = exp[$z^2/2$]); the inverse is Lik(μ_0)/Lik($\mu_{\max}$) = exp[$-z^2/2$]. I think the former makes their case stronger, yet you will usually see the latter. (I record their values in a Note[9]). What is $\mu_{\max}$? It's the observed mean $\overline{x}$, the place most "favored by the data." In each case we consider $\overline{x}$ as the result that is just statistically significant at the indicated P-value, or its standardized z form.

With a P-value of 0.025, $H_{\max}$ is "only" 6.84 times as likely as the null. I put quotes around "only" not because I think 6.84 is big; I'm never clear what's to

Table 4.2 Upper Bounds on the Comparative Likelihood

P-value: one-sided	z_α	Lik($\mu_{\max}$)/Lik(μ_0)
0.05	1.65	3.87
0.025	1.96	6.84
0.01	2.33	15
0.005	2.58	28
0.0005	3.29	227

[9] The entries for the inverse are useful. This is adapted from Berger and Sellke (1987) Table 3.

P-value: one-sided	z_α	Lik(μ_0)/Lik($\mu_{\max}$)
0.05	1.65	0.258
0.025	1.96	0.146
0.01	2.33	0.067
0.005	2.58	0.036
0.0005	3.29	0.0044

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

count as big until I have information about the associated error probabilities. If you seek to ensure $H_{\max}$: $\mu = \mu_{\max}$ is 28 times as likely as is H_0: $\mu = \mu_0$, you need to use a P-value ~0.005, with z value of 2.58, call it 2.6. Compare the corresponding error probabilities. Were there 0 discrepancy from the null, a difference smaller than 1.96 would occur 97.5% of the time; one smaller than 2.6 would occur 99.5% of the time. In both cases, the 95% two-sided and 97.5% confidence intervals entirely exclude 0. The two one-sided lower intervals are $\mu > 0$ and $\mu >$ ~0.64. Both outcomes are good indications of $\mu > 0$: the difference between the likelihood ratios 6.8 and 28 doesn't register as very much when it comes to indicating a positive discrepancy. Surely E, L, & S couldn't expect Bayes factors to match error probabilities when they are the ones who showed how optional stopping can alter the latter and not the former (Section 1.5).

Valen Johnson (2013a,b) offers a way to bring the likelihood ratio more into line with what counts as strong evidence, according to a Bayes factor. He begins with a review of "Bayesian hypotheses tests." "The posterior odds between two hypotheses H_1 and H_0 can be expressed as"

$$\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} = \mathrm{BF}_{10}(\boldsymbol{x}) \times \frac{\Pr(H_1)}{\Pr(H_0)}.$$

> Like classical statistical hypothesis tests, the tangible consequence of a Bayesian hypothesis test is often the rejection of one hypothesis, say H_0, in favor of the second, say H_1. In a Bayesian test, the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold. (Johnson 2013b, pp. 1720–1)

According to Johnson, Bayesians reject hypotheses based on a sufficiently high posterior and "the alternative hypothesis is accepted if $\mathrm{BF}_{10} > k$" (ibid., p. 1726, k for his γ). A weaker stance might stop with the comparative report Lik($\mu_{\max}$)/Lik(μ_0). It's good that he supplies a falsification rule.

Johnson views his method as showing how to specify an alternative hypothesis – he calls it the "implicitly tested" alternative (ibid., p. 1739) – when H_0 is rejected. H_0 and H_1 are each given a 0.5 prior. Unlike N-P, the test does not exhaust the parameter space, it's just two points.

> [D]efining a Bayes factor requires the specification of both a null hypothesis and an alternative hypothesis, and in many circumstances there is no objective mechanism for defining an alternative hypothesis. The definition of the alternative hypothesis therefore involves an element of subjectivity and it is for this reason that scientists generally eschew the Bayesian approach toward hypothesis testing. (Johnson 2013a, p. 19313)

He's right that comparative inference, as with Bayes factors, leaves open a wide latitude of appraisals by dint of the alternative chosen, and any associated priors.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

Table 4.3 V. Johnson's implicit alternative analysis for T+: H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$

P-value one-sided	z_α	Lik(μ_{max})/Lik(μ_0)	μ_{max}	Pr(H_0|x)	Pr(H_{max} |x)
0.05	1.65	3.87	$1.65\sigma/\surd n$	0.2	0.8
0.025	1.96	6.84	$1.96\sigma/\surd n$	0.128	0.87
0.01	2.33	15	$2.33\sigma/\surd n$	0.06	0.94
0.005	2.58	28	$2.58\sigma/\surd n$	0.03	0.97
0.0005	3.29	227	$3.3\sigma/\surd n$	0.004	0.996
	$\surd(2 \log k)$	$exp\left(\frac{z_\alpha^2}{2}\right)$	$z_\alpha\, \sigma/\surd n$	$1/(1+k)$	$k/(1+k)$

In his attempt to rein in that choice, Johnson offers an illuminating way to relate the Bayes factor and the standard cut-offs for rejection, at least in UMP tests such as this. (He even calls it a uniformly most powerful Bayesian test!) We focus on the cases where we just reach statistical significance at various levels. Setting k as the Bayes factor you want, you can obtain the corresponding cut-off for rejection by computing $\surd(2 \log k)$: this matches the z_α corresponding to a N-P, UMP one-sided test. The UMP test T+ is of the form: Reject H_0 iff $\overline{X} \geq \overline{x}_\alpha$, where $\overline{x}_\alpha = \mu_0 + z_\alpha\, \sigma/\surd n$, which is $z_\alpha \sigma/\surd n$ for the case $\mu_0 = 0$. Thus he gets (2013b, p. 1730)

$$H_1\colon \mu_1 = \sigma\sqrt{\frac{2 \log k}{n}}.$$

Since this is the alternative under which the observed data, which we are taking to be $\overline{x}_\alpha$, have maximal probability, write it as H_{max} and μ_1 as μ_{max}. The computations are rather neat, see Note 10. (The last row of Table 4.3 gives an equivalent form.) The reason the LR in favor of the (maximal) alternative gets bigger and bigger is that Pr($\boldsymbol{x}$; H_0) is getting smaller and smaller with increasingly large $\boldsymbol{x}$ values.

Johnson's approach is intended to "provide a new form of default, non subjective Bayesian tests" (2013b, p. 1719), and he extends it to a number of other cases as well. Given it has the same rejection region as a UMP error statistical test, he suggests it "can be used to translate the results of classical significance tests into Bayes factors and posterior model probabilities" (ibid.). To bring them into line with the BF, however, you'll need a smaller α level. Johnson recommends levels more like 0.01 or 0.005. *Is there anything lost in translation?*

There's no doubt that if you reach a smaller significance level in the same test, the discrepancy you are entitled to infer is larger. You've made the hurdle

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

for rejection higher: any observed mean that makes it over must be larger. It also means that more will fail to make it over the hurdle: the Type II error probability increases. Using the 1.96 cut-off, a discrepancy of 2.46, call it 2.5, will be detected 70% of the time – add 0.5 SE to the cut-off – (the Type II error is 0.3) whereas using a 2.6 cut-off has less than 50% (0.46) chance of detecting a 2.5 discrepancy (Type II error of 0.54!). Which is a better cut-off for rejection? The severe tester eschews rigid cut-offs. In setting up a test, she looks at the worst cases she can live with; post-data she reports the discrepancies well or poorly warranted at the attained levels. (Recall, discrepancy always refers to parameter values.) Johnson proposes to make up for the loss of power by increasing the sample size, but it's not that simple. We know that as sample size increases, the discrepancy indicated by results that reach a given level of significance decreases. Still, you get a Bayes factor and a default posterior probability that you didn't have with ordinary significance tests. What's not to like?

We perform our two-part criticism, based on the minimal severity requirement. The procedure under the looking glass is: having obtained a statistically significant result, say at the 0.005 level, reject H_0 in favor of $H_{\max}$: $\mu = \mu_{\max}$. Giving priors of 0.5 to both H_0 and $H_{\max}$ you can report the posteriors. Clearly, (S-1) holds: $H_{\max}$ accords with $\bar{x}$ – it's equal to it. Our worry is with (S-2). H_0 is being rejected in favor of $H_{\max}$, but should we infer it? The severity associated with inferring μ is as large as $\mu_{\max}$ is

$$\Pr(Z < z_{\alpha}; \mu = \mu_{\max}) = 0.5.$$

This is our benchmark for poor evidence. So (S-2) doesn't check out. You don't have to use severity, just ask: what confidence level would permit the inference $\mu \geq \mu_{\max}$ (answer 0.5). Yet Johnson assigns $\Pr(H_{\max}|\boldsymbol{x}) = 0.97$. $H_{\max}$ is comparatively more likely than H_0 as $\bar{x}$ moves further from 0 – but that doesn't mean we'd want to infer there's evidence for $H_{\max}$. If we add a column to Table 4.1 for SEV($\mu \geq \mu_{\max}$) it would be 0.5 all the way down!

To have some numbers, in our example (H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$), $\sigma = 1$, $n = 25$, and the 0.005 cut-off is $2.58\sigma/\sqrt{n} = 0.51$, round to 0.5. When a significance tester says the difference $\bar{x} = 0.5$ is statistically significantly greater than 0 at the 0.005 level, she isn't saying anything as strong as "there is fairly good evidence that $\mu = 0.5$." Here it gets a posterior of 0.97. While the goal of the reform was to tamp down on "overstated evidence," it appears to do just the opposite from a severe tester's perspective.

How can I say it's lousy if it's the maximally likely estimate? Because there is the variability of the estimator, and statistical inference must take this into account. It's true that the error statistician's inference isn't the point alternative

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

these critics want us to consider ($H_{\max}$), but they're the ones raising the criticism of ostensive relevance to us, and we're struggling in good faith to see what there might be in it. Surely to infer $\mu = 0.5$ is to infer $\mu > 0.4$. Our outcome of 0.5 is 0.5 standard error in excess of 0.4, resulting in SEV($\mu > 0.4$) = 0.7. Still rather poor. Equivalently, it is to form the 0.7 lower confidence limit ($\mu > 0.4$).

Johnson (2013a, p. 19314) calls the 0.5 spikes equipoise, but what happened to the parameter values in between H_0 and $H_{\max}$? Do they get a prior of 0? To be clear, it may be desirable or at least innocuous for a significance tester to require smaller *P*-values. What is not desirable or innocuous is basing the altered specification on a BF appraisal, if in fact it is an error statistical justification you're after. Defenders of the argument may say, they're just showing the upper bound of evidence that can accrue, even if we imagine being as biased as possible against the null and for the alternative. But are they? A fair assessment, say Casella and R. Berger, wouldn't have the spike prior on the null – yes, it's still there. If you really want a match, why not use the frequentist matching priors for this case? (Prior 3 in Exhibit vii) The spiked prior still has a mismatch between BF and *P*-value.[10] This is the topic of megateam battles. (Benjamin et al. 2017 and Lakens et al. 2018).

Exhibit (viii): Whether *P*-values Exaggerate Depends on Philosophy. When a group of authors holding rather different perspectives get together to examine a position, the upshot can take them out of their usual comfort zones. We need more of that. (See also the survey in Hurlbert and Lombardi 2009, and Haig 2016.) Here's an exhibit from Greenland et al. (2016). They greet each member of a list of incorrect interpretations of *P*-values with "No!", but then make this exciting remark:

[10] Computations

1. Suppose the outcome is just significant at the α level: $\overline{x} = \mu_0 + z_\alpha \sigma/\surd n$.
2. So the most likely alternative is $H_{\max}$: $\mu_1 = \overline{x} = \mu_0 + z_\alpha \sigma/\surd n$.
3. The ratio of the maximum likely alternative $H_{\max}$ to the likelihood of H_0 is:

$$\frac{\mathrm{Lik}(x|H_{\max})}{\mathrm{Lik}(x|H_0)} = \frac{1}{\exp[-z^2/2]} = \exp[z^2/2].$$

This gives the Bayes factor: BF_{10}. (BF_{01} would be $\exp[-z^2/2]$.)

4. Set $\mathrm{Lik}(\boldsymbol{x}|\ H_{\max})/\mathrm{Lik}(\boldsymbol{x}|H_0) = k$.
5. So $\exp[z^2/2] = k$.
   Since the natural log (ln) and exp are inverses:

$$\log k = \log(\exp[z^2/2]) = [z^2/2];$$
$$2 \log k = z^2, \text{ so } \surd(2 \log k) = z.$$

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press

> There are other interpretations of *P* values that are controversial, in that whether a categorical "No!" is warranted depends on one's philosophy of statistics and the precise meaning given to the terms involved. The disputed claims deserve recognition if one wishes to avoid such controversy. . . .
>
> For example, it has been argued that *P* values overstate evidence against test hypotheses, based on directly comparing *P* values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis . . . Nonetheless, many other statisticians do not accept these quantities as gold standards, and instead point out that *P* values summarize crucial evidence needed to gauge the error rates of decisions based on statistical tests (even though they are far from sufficient for making those decisions). Thus, from this frequentist perspective, *P* values do not overstate evidence and may even be considered as measuring one aspect of evidence . . . with $1 - P$ measuring evidence against the model used to compute the *P* value. (p. 342)

It's erroneous to fault one statistical philosophy from the perspective of a philosophy with a different and incompatible conception of evidence or inference. The severity principle always evaluates a claim as against its denial within the framework set. In N-P tests, the frame is within a model, and the hypotheses exhaust the parameter space. Part of the problem may stem from supposing N-P tests infer a point alternative, and then seeking that point. Whether you agree with the error statistical form of inference, you can use the severity principle to get beyond this particular statistics battle.

Souvenir R: The Severity Interpretation of Rejection (SIR)

In Tour II you have visited the tribes who lament that *P*-values are sensitive to sample size (Section 4.3), and they exaggerate the evidence against a null hypothesis (Sections 4.4, 4.5). We've seen that significance tests take into account sample size in order to critique the discrepancies indicated objectively. A researcher may choose to decrease the *P*-value as *n* increases, but there's no problem in understanding that the same *P*-value reached with a larger sample size indicates fishing with a finer mesh. Surely we should not commit the fallacy exposed over 50 years ago.

Here's a summary of the severe tester's interpretation (of a rejection) putting it in terms that seem most clear:

> ***SIR: The Severity Interpretation of a Rejection in test T+:*** *(small P-value)*
>
> (i): [*Some* discrepancy is indicated]: $d(\boldsymbol{x}_0)$ is a good indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of observing a *less* statistically significant difference than $d(\boldsymbol{x}_0)$ if $\mu = \mu_0 + \gamma$.

N-P and Fisher tests officially give the case with $\gamma = 0$. In that case, what does a small P-value mean? It means the test very probably $(1 - P)$ would have produced a result more in accord with H_0, were H_0 an adequate description of the data-generating process. So it indicates a discrepancy from H_0, especially if I can bring it about fairly reliably. To avoid making mountains out of molehills, it's good to give a second claim about the discrepancies that are *not* indicated:

> (ii): [I'm not *that* impressed]: $d(\boldsymbol{x}_0)$ is a poor indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of an even more statistically significant difference than $d(\boldsymbol{x}_0)$ even if $\mu = \mu_0 + \gamma$.

As for the exaggeration allegation, merely finding a single statistically significant difference, even if audited, is indeed weak: it's an indication of *some* discrepancy from a null, a first step in a task of identifying a genuine effect. But, a legitimate significance tester would never condone rejecting H_0 in favor of alternatives that correspond to a low severity or confidence level such as 0.5. Stephen Senn sums it up: "Certainly there is much more to statistical analysis than P-values but they should be left alone rather than being deformed . . . to become second class Bayesian posterior probabilities" (Senn 2015a). Reformers should not be deformers.

There is an urgency here. Not only do some reforms run afoul of the minimal severity requirement, to suppose things are fixed by lowering P-values ignores or downplays the main causes of non-replicability. According to Johnson:

> [I]t is important to note that this high rate of nonreproducibility is not the result of scientific misconduct, publication bias, file drawer biases, or flawed statistical designs; it is simply the consequence of using evidence thresholds that do not represent sufficiently strong evidence in favor of hypothesized effects. (2013a, p. 19316)

This sanguine perspective sidesteps the worry about the key sources of spurious statistical inferences: biasing selection effects and violated assumptions, at all levels. (Fortunately, recent reforms admit this; Benjamin et al. 2017.) Catching such misdemeanors requires *auditing*, the topic of Tours III and IV of this Excursion.

https://doi.org/10.1017/9781107286184.010 Published online by Cambridge University Press