Day 10 (A) Excursion 4 Tour II: Rejection Fallacies: Whose Exaggerating What?

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.

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.

The “P-values exaggerate” criticism typically boils down to: if inference is appraised via one of the probabilisms–Bayesian posteriors, Bayes factors, or likelihood ratios–the evidence against the null isn’t as big as 1 – P.

Really it’s comparative: evidence against the null and in favor of some alternative)

You might react by observing that:

- P-values are not intended as posteriors in H_0 (or Bayes ratios, likelihood ratios)
- Thus there’s no reason to suppose a P-value should match numbers computed in very different accounts.

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 2001, p. 202).

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, yet, I want to push beyond them.

Getting Beyond "I'm Rubber and You're Glue". P. 247

- The danger in critiquing statistical method X from the standpoint of a distinct school Y, is that of falling into begging the question.
- Whatever you say about me bounces off and sticks to you. This is a genuine worry, but it's not fatal. T
- The rationale of this journey is that we may 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 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.

J. Berger and Sellke, and Casella and R. Berger. Berger and Sellke (1987) make out the conflict between P-values and Bayesian posteriors using the two-sided test of the Normal mean, H_0: $\mu = 0$ versus H_1: $\mu \neq 0$. "Suppose that $\boldsymbol{X} = (X_1,\ldots,X_n)$, where the X_i are IID $N(\mu, \sigma^2)$, σ^2 known" (p. 112).

Then the test statistic $d(\boldsymbol{X}) = \sqrt{n}\,|\bar{X} - \mu_0|/\sigma$, and the P-value will be twice the P-value of the corresponding one-sided test.

- Starting with a lump of prior, generally .5, on H_0, they find the posterior probability in H_0 is larger than the P-value for a variety of different priors to the alternative.
- However, the result depends entirely on how the remaining .5 is allocated or smeared over the alternative (a move dubbed spike and smear)!

- Using a Jeffreys-type prior, the .5 is spread out over the alternative parameter values as if the parameter is itself distributed $N(\mu_0,\sigma)$.
- Harold Jeffreys recommends the lump prior only for cases where a special value of a parameter is deemed plausible, eg, the GTR deflection effect $\lambda = 1.75$" , after ~1960.
- The rationale is to enable it to receive a reasonable posterior probability, and avoid a 0 prior to H_0

By titling 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.

Table 4.1 SIST p. 249 (From J. Berger and T. Sellke (1987))

- Declare no evidence against the null, and even evidence for it whenever d(***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 = .05,' although $Pr(H_0|\boldsymbol{x}) = .52$ (which would actually indicate that the evidence *favors* H_0)" (Berger and Sellke, p. 113).

If $n = 1000$, a result statistically significant at the .05 level has the posterior probability to μ = 0 go up from .5 (the lump prior) to .82!

From their Bayesian perspective, this appears to show P-values are exaggerating evidence against H_0.

Error statistical testers balk that using the recommended priors allows statistically significant results to be interpreted as no evidence against H_0–or even evidence *for* it!

- After all, 0 is excluded from the 2-sided confidence interval at level .95.
- Although a posterior probability doesn't have an error probability attached, a tester can evaluate the error probability credentials of these inferences.

The probability of declaring evidence for the null even if false is high.

Jeffreys-Lindley "Paradox" or Bayes/Fisher Disagreement (p. 250)

Lindley's result dealt with just this example, two-sided Normal testing with known variance: H_0: μ = 0 versus H_1: μ ≠ 0.

With a lump given to the point null, and the rest appropriately spread over the alternative, an *n* can be found such an α significant result corresponds to

$\Pr(H_0|\boldsymbol{x}) = (1 - \alpha)$!

Many would 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.

- 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’” Berger and Sellke (1987, p. 115).

- But is it? One starts by making H_0 and H_1 equally probable, then the .5 accorded to H_1 is spread out over all the values in H_1:

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 $n \rightarrow \infty$, a 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. 59/2005, p. 41).

What justifies the lump prior of .5?

A Dialogue on the Water Plant Accident (p.251).

EPA Rep: The mean temperature of the water was found statistically significantly higher than 150 degrees at the .025 (or .05 level).

Spiked Prior Bayesian: This even strengthens my belief the water temperature's no different from 150. If I update the prior of .5 that I give to the null hypothesis, my posterior for H_0 is still .6; it's not .025 or .05, that's for sure.

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

Spiked Prior Bayesian: If I gave H_0 a value lower than .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 .1 to H_0, and my conclusion is that H_0 has

> posterior probability .05 and should be rejected'? (Berger and Sellke 1987, p. 115).

But it's scarcely an obvious justification for a lump of prior on the null H_0–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.

Casella and R. Berger (1987) charge 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 1 or 2-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 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.).

That P-value and posteriors match well in one-sided tests with "uninformative" or "diffuse" priors has been long known (e.g., Cox and Hinkley 1974, Jeffreys 1961, Pratt 1965).

Why Blame Us Because You Can't Agree on Your Posterior?

Stephen Senn argues, "…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 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 2001, 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! (SIST P. 253)

Senn riffs on the well-known joke of Jeffreys that we heard in 3.4 (1961, p. 385):

> 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. Senn (ibid. p. 195)

Contrasting Bayes Factors p. 254

Let's identify three of the types of priors appealed to in some prominent criticisms and/or reforms of significance tests.

1. *Jeffrey-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*. Second, there's a spike to a point null, to be compared to the point alternative that's maximally likely θ_{max}.
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).

Exhibit (vii). *Jeffrey-Lindley 'paradox'*. Consider an example that Aris Spanos (2013) explores in relation to the Jeffreys-Lindley 'paradox'. It is briefly noted in Stone 1997.

A large number (*n* =527,135) of independent collisions that can be either of type A or type B are used to test if the proportion of type A collisions is exactly .2, as opposed to any other value.

It's modeled as *n* Bernoulli trials testing H_0: θ = .2 vs. H_1: θ ≠ .2. The observed proportion of type A collisions is scarcely greater than the point null of .2:

$\bar{x}$ = k/*n* = *.20165233* where *n*=527,135; k = 106,298.

The significance level against H_0 is small

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

The Bayes Factor in favor of H_0 is high

- H_0 is given the spiked prior of .5, and the remaining .5 is spread equally among the values in H_1.

The Bayes factor $B_{01} = Pr(k|H_0)/\ Pr(k|H_1)$ =
.000015394/.000001897 = 8.115
While the likelihood of H_0 in the numerator is tiny, the likelihood of H_1 is even tinier.

There's no surprise once you consider the Bayesian question here: compare the likelihood of a result scarcely different from .2 being produced by a universe where θ = .2– where this has been given a spiked prior of .5 under H_0–with 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, θ = .2 is more likely, and we have an example of the Jeffreys-Fisher disagreement.

Compare it with the second kind of prior:

Here Bayes factor B_{01} L= 0.01; Lik(θ_{max})/Lik(.2) = 89.

Why should a result 89 times more likely under alternative θ_{max} than under $\theta = .2$ be taken as strong evidence *for* $\theta = .2$?

It shouldn't, according to some, including Lindley's own student, default Bayesian José Bernardo (2010).

Presumably, the likelihoodist concurs. There are family feuds within and between the diverse tribes of probabilisms.

.

4.5 Who's Exaggerating?

Edwards, Lindman, and Savage–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 Berger and Sellke, but they view it as a one-tailed test of H_0: $\mu = 0$ vs. H_1: $\mu = \mu_1 = \theta_{max}$.

We abbreviate H_1 by H_{max}.

Here the likelihood ratio $Lik(\theta_{max})/Lik(\theta_0) = \exp [z^2/2]$;

the inverse is $Lik(\theta_0)/Lik(\theta_{max})$, is $\exp [-z^2/2]$.

What is θ_{max}? It's the observed mean $\bar{x}$ (whatever it is), and we're to consider $\bar{x}$ = the result that is just statistically significant at the indicated P-value.

Upper bounds on the comparative likelihood

P-value:1-sided	z_α	Lik(θ_{max})/Lik(θ_0)
.05	1.65	3.87
.025	1.96	6.84
.01	2.33	15
.005	2.58	28
0.0005	3.29	227

Table 4.2

If you're seeking to ensure H_{max}: $\mu = \mu_{max}$ is 28 times as likely as is H_0: $\theta = \theta_0$, trying to match the .05 value, you'd need to use a P-value ~.005, with z value of 2.58.

- 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 reviews of "Bayesian hypotheses tests". "The posterior odds between two hypotheses H_1 and H_0 can be expressed as"

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

"In a Bayesian test, the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold. …."(Johnson 2013b, p. 1721)

$\mathrm{BF}_{10}(\boldsymbol{x})$ here will be $\mathrm{Lik}(\theta_{max})/\mathrm{Lik}(\theta_0)$.

- For Johnson, Bayesians reject hypotheses based on a sufficiently high posterior and “the alternative hypothesis is accepted if $BF_{10} > k$” (ibid., p. 1726, k for his γ).

- Johnson views his method as showing how to specify an alternative hypothesis–he calls it the “implicit alternative”

- Unlike N-P, the test does not exhaust the parameter space, it’s just two points.

Johnson offers an illuminating way to relate Bayes factors and standard cut-offs for rejection, at least in UMP tests

- (SIST p. 262) Setting k as the Bayes factor you want, you can obtain the corresponding cut-off for rejection by computing $\sqrt{}$(2log k): this matches the z_α corresponding to a N-P, UMP one-sided test.
- The UMP test (with $\mu > \mu_0$) is of the form:

Reject H_0 iff $\bar{X} \geq \bar{x}_\alpha$ where $\bar{x}_\alpha = \mu_0 + z_\alpha \sigma/\sqrt{n}$, which is $\sigma/\sqrt{n}$ for the case $\mu_0 = 0$.

The computations are straightforward for this case and are really rather neat: Table 4.3 (SIST p. 262)

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	$\text{Lik}(\mu_{\max})/\text{Lik}(\mu_0)$	$\mu_{\max}$	$\Pr(H_0\|x)$	$\Pr(H_{\max}\|x)$
0.05	1.65	3.87	$1.65\sigma/\sqrt{n}$	0.2	0.8
0.025	1.96	6.84	$1.96\sigma/\sqrt{n}$	0.128	0.87
0.01	2.33	15	$2.33\sigma/\sqrt{n}$	0.06	0.94
0.005	2.58	28	$2.58\sigma/\sqrt{n}$	0.03	0.97
0.0005	3.29	227	$3.3\sigma/\sqrt{n}$	0.004	0.996
	$\sqrt{(2\log k)}$	$exp\left(\frac{z_\alpha^2}{2}\right)$	$z_\alpha\,\sigma/\sqrt{n}$	$1/(1+k)$	$k/(1+k)$

His approach is intended to "provide a new form of default, non subjective Bayesian tests" (2013b, p. 1719)

- It has the same rejection region as a UMP error statistical test, but to bring them into line with the BF you need a smaller α level. Johnson recommends levels more like .01 or .005.

- Of course if you reach a smaller significance level, say .01 rather than .025, you may infer a larger discrepancy.

- It also means more will fail to make it over the hurdle: the Type II error probability increases (he recommends increasing n).

So, you get a Bayes Factor and a default posterior probability. What's not to like?

We perform our two-part criticism, based on the minimal severity requirement. SIST p. 263

(S-1) holds, but (S-2) fails; the SEV is .5.

(Lakens et.al., 2018)

Exhibit (viii). *Whether P-values exaggerate depends on philosophy.*

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 (4.3), and they exaggerate the evidence against a null hypothesis (4.4, 4.5).

Stephen Senn says "reformers" should stop deforming P-values to turn them into second class Bayesian posterior probabilities (Senn 2015a). I agree.

There is an urgency here. Not only do the replacements run afoul of the minimal severity requirement, to suppose all is fixed by lowering P-values ignores the biasing selection effects at the bottom of nonreplicability.

[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.” (Johnson 2013a, p. 19316).