Excursion 4 Tour II: Rejection
Fallacies: Whose Exaggerating What?

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!
SIR: The Severity Interpretation of a Rejection in test $T^+$: (small $P$-value)

(i): [Some discrepancy is indicated]: $d(x_0)$ 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 $d(x_0)$ if $\mu = \mu_0 + \gamma$.

(ii): [I’m not that impressed]: $d(x_0)$ 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 $d(x_0)$ even if $\mu = \mu_0 + \gamma$. 
Tiny illustration of Power & sample size

\( H_0: \mu \leq 150 \) vs. \( H_1: \mu > 150 \)

(Let \( \sigma = 10 \), \( n = 100 \))

let \( \alpha = .025 \)

\[ \text{POW(T+, } \mu_1 \text{ )} = \Pr(\text{Test T+ rejects } H_0; \mu_1), \]

Consider \( \mu_1 = 153 \)

\[ \text{POW(T+, 153) } \Pr(\bar{X} > 152; \mu = 153) \]

\[ Z = \frac{(152 - 153)}{\sigma_{\bar{X}}} = -1 \]

\[ \Pr (Z > -1) = .84 \]

(however, it’s poor evidence \( \mu > 153 \))
$H_0$: $\mu \leq 150$ vs. $H_1$: $\mu > 150$

(Let $\sigma = 10$, $n = 25$) Now $\sigma_{\bar{X}} = 2$ (i.e., 10/5)
let $\alpha = .025$

$POW(T^+, \mu_0) = Pr(\text{Test } T^+ \text{ rejects } H_0; \mu_0)$,

Again consider $\mu_1 = 153$

$POW(T^+, 153) Pr(\bar{X} \geq 154; \mu = 153)$
$Z = (154 - 153)/2 = .5$
$Pr (Z > .5) = .3$
Do P-Values Exaggerate the Evidence?
I. J. Berger and Sellke and Casella and R. Berger

II. Jeffreys-Lindley Paradox; Bayes/Fisher Disagreement

III. Redefine Statistical Significance
Common criticism: “Significance levels (or P-values) exaggerate the evidence against the null hypothesis”

**What do you mean by exaggerating the evidence against \( H_0 \)?**

**Answer:** The P-value is too small, for ex.: What I mean is that when I put a lump of prior weight \( \pi_0 \) 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 \( H_0 \).

(p.246)
“P-values exaggerate”: 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 $1 - P$.

• On the other hand, the probability $H_0$ would have survived is $1 - P$

• Difference in the role for probabilities
You might react by observing that:

- P-values are not intended as posteriors in $H_0$ (or Bayes factors, likelihood ratios)

- Why suppose a P-value should match numbers computed in very different accounts.
When the criticism is in the form of a posterior:

...[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).
• 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.
• The minimal theses about “bad evidence no test (BENT)” enables 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.
This is a “how to” book

• We do not depict critics as committing a gross blunder (confusing a P-value with a posterior probability in a null).

• Nor just deny we care about their measure of support: I say we should look at exactly what the critics are on about.
Bayes Factor (bold part)

\[
\frac{\Pr(H_0|x)}{\Pr(H_1|x)} = \frac{\Pr(x|H_0) \Pr(H_0)}{\Pr(x|H_1) \Pr(H_1)}
\]

• Likelihood ratio but not limited to point hypothesis

• The parameter is viewed as a random variable with a distribution
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 $X = (X_1, \ldots, X_n)$, where the $X_i$ are IID $N(\mu, \sigma^2)$, $\sigma^2$ known” (p. 112).

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

Starting with a lump of prior, 0.5, on $H_0$, they find the posterior probability in $H_0$ is larger than the P-value for a variety of different priors assigned to the alternative.

The result depends entirely on how the remaining .5 is smeared over the alternative
• Using a Jeffreys-type prior, the .5 is spread out over the alternative parameter values as if the parameter is itself distributed $N(\mu_0, \sigma)$.

• Actually Jeffreys recommends the lump prior only when a special value of a parameter is deemed plausible.*

• The rationale is to enable it to receive a reasonable posterior probability, and avoid a 0 prior to $H_0$.

“P-values are reasonable measures of evidence when there is no a priori concentration of belief about $H_0$ (Berger and Delampady)
Table 4.1 Pr($H_0|x$) for Jeffreys-type prior

<table>
<thead>
<tr>
<th>P one-sided</th>
<th>$z_\alpha$</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.645</td>
<td>0.47</td>
<td>0.56</td>
<td>0.65</td>
<td>0.72</td>
<td>0.89</td>
</tr>
<tr>
<td>0.025</td>
<td>1.960</td>
<td>0.37</td>
<td>0.42</td>
<td>0.52</td>
<td>0.60</td>
<td>0.82</td>
</tr>
<tr>
<td>0.005</td>
<td>2.576</td>
<td>0.14</td>
<td>0.16</td>
<td>0.22</td>
<td>0.27</td>
<td>0.53</td>
</tr>
<tr>
<td>0.0005</td>
<td>3.291</td>
<td>0.024</td>
<td>0.026</td>
<td>0.034</td>
<td>0.045</td>
<td>0.124</td>
</tr>
</tbody>
</table>

(From Table 1, J. Berger and T. Sellke (1987) p. 113 using the one-sided $P$-value)
• With $n = 50$, “one can classically ‘reject $H_0$ at significance level $p = .05,’$ although $\Pr(H_0|x) = .52$ (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 $\mu = 0$ go up from .5 (the lump prior) to .82!
From their Bayesian perspective, this seems to show P-values are exaggerating evidence against $H_0$.

From an error statistical perspective, this allows statistically significant results to be interpreted as no evidence against $H_0$—or even evidence for it!
(posterior $H_0$. is higher than the prior-B-boost)

- After all, 0 is excluded from the 2-sided confidence interval at level .95.
- The probability of declaring evidence for the null even if false is high.
• Why assign the lump of \( \frac{1}{2} \) 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 \( H_1 \):
A Dialogue at the Water Plant Accident (p.251)

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).
But it’s scarcely an obvious justification for a lump of prior on the null $H_0$ 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.).
Jeffreys-Lindley “Paradox” or Bayes/Fisher Disagreement (p. 250)

The disagreement (between the P-value and the posterior can be dramatic

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

$$\Pr(H_0 | x) = (1 - \alpha)!$$
Contrasting Bayes Factors p. 254

They arise in 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 $\theta_{\text{max}}$.

3. **Matching.** Instead of a spike prior on the null, it uses a smooth diffuse prior. Here, the P-value “is an approximation to the posterior probability that $\theta < 0$” (Pratt 1965, p. 182).
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 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)
Exhibit (vii). Jeffrey-Lindley ‘paradox’

A large number \( (n = 527,135) \) of independent collisions either of type A or type B will test if the proportion of type A collisions is exactly .2, as opposed to any other value.

\[ n \text{ Bernoulli trials, testing } H_0: \theta = .2 \text{ vs. } H_1: \theta \neq .2. \]

The observed proportion of type A collisions is scarcely greater than the point null of .2:

\[ \bar{x} = k/n = .20165233 \text{ where } n=527,135; \ k = 106,298. \]

Example from Aris Spanos (2013) (from Stone 1997.)
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 0.2 being produced by a universe where $\theta = 0.2$ – where this has been given a spiked prior of 0.5 under $H_0$ – with the likelihood of that result being produced by any $\theta$ in a small band of $\theta$ 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.

Clearly, $\theta = .2$ is more likely, and we have an example of the Jeffreys-Fisher disagreement. SIST p. 255
Compare it with the second kind of prior:

Here Bayes factor $B_{01} = 0.01; B_{10} = \frac{\text{Lik}(\theta_{\text{max}})}{\text{Lik}(0.2)} = 89$

Why should a result 89 times more likely under alternative $\theta_{\text{max}}$ than under $\theta = 0.2$ be taken as strong evidence for $\theta = 0.2$?
Contrasting Bayes Factors p. 254

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 $\theta_{\text{max}}$.

3. **Matching.** Instead of a spike prior on the null, it uses a smooth diffuse prior. Here, the P-value “is an approximation to the posterior probability that $\theta < 0$” (Pratt 1965, p. 182).
Bayesian Family feuds

It shouldn’t, according to some, including Lindley’s own student, default Bayesian José Bernardo (2010). (SIST p. 256, Note 7)

Yet it’s at the heart of recommended reforms

First, look at p. 256 on matching priors
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 $\theta$ is on the opposite side of 0 than the observed is $P/2$. They proffer this as a way “to live with” $P$-values.
4.5 Reforms (Redefine Significance) Based on Bayes Factor Standards

“Redefine Significance” is recent, but, like other reforms, is based on old results:
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 (Edwards, Lindman, and Savage 1963, p. 228).

Normal testing case of Berger and Sellke, but as a one-tailed test of \( H_0: \mu = 0 \) vs. \( H_1: \mu = \mu_1 = \theta_{\text{max}} \).

We abbreviate \( H_1 \) by \( H_{\text{max}} \).
Here the likelihood ratio $\text{Lik}(\theta_{\text{max}})/\text{Lik}(\theta_0) = \exp [z^2/2]$; 
the inverse is $\text{Lik}(\theta_0)/\text{Lik}(\theta_{\text{max}})$, is $\exp [-z^2/2]$.

What is $\theta_{\text{max}}$?

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.

SIST p. 260 (see note #9)
Normal Distribution

\[ y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \]

- \( \mu = \) Mean
- \( \sigma = \) Standard Deviation
- \( \pi \approx 3.14159 \ldots \)
- \( e \approx 2.71828 \ldots \)
To ensure $H_{max}: \mu = \mu_{max}$ is 28 times as likely as $H_0: \theta = \theta_0$, you’d need to use a P-value $\sim .005$, z value of 2.58.
• Valen Johnson (2013a,b): a way to bring the likelihood ratio more into line with what counts as strong evidence, according to a Bayes factor.

• “The posterior odds between two hypotheses $H_1$ and $H_0$ can be expressed as”

$$\frac{\Pr(H_1|x)}{\Pr(H_0|x)} = BF_{10}(x) \times \frac{\Pr(H_1)}{\Pr(H_0)}.$$  

“In a Bayesian test, the null hypothesis is rejected if the posterior probability of $H_1$ exceeds a certain threshold. ….”(Johnson 2013b, p. 1721)
• and “the alternative hypothesis is accepted if $BF_{10} > k$”

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

• It will be $H_{max}$

• 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 in UMP tests

- (SIST p. 262) Setting $k$ as the Bayes factor you want, you get the corresponding cut-off for rejection by computing $\sqrt{2\log k}$: this matches the $z_\alpha$ corresponding to a N-P, UMP one-sided test.

- The UMP test (with $\mu > \mu_0$) is of the form:

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

Table 4.3 (SIST p. 262), computations note #10 p. 264
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 | $z_\alpha$ | $\text{Lik}(\mu_{\text{max}})/\text{Lik}(\mu_0)$ | $\mu_{\text{max}}$ | $\text{Pr}(H_0|x)$ | $\text{Pr}(H_{\text{max}}|x)$ |
|-------------------|-----------|-----------------------------------------------|---------------------|---------------------|---------------------|
| 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               |

\[
\sqrt{(2 \log k \exp \left(\frac{z_\alpha^2}{2}\right))} \quad \frac{z_\alpha \sigma}{\sqrt{n}} \quad \frac{1}{1 + k} \quad \frac{k}{1 + k}
\]
\[ P_r(H_0 | x) = \frac{P_r(x | H_0) P_r(H_0)}{P_r(x | H_0) P_r(H_0) + P_r(x | H_{max}) P_r(H_{max})} \]

Erase priors - both are \( \frac{1}{2} \)

Divide by \( P_r(x | H_0) \)

\[ 1 \]

\[ \frac{1}{1 + P_r(x | H_{max})} \]

\[ \frac{P_r(x | H_0)}{P_r(x | H_{max})} \]

\[ \frac{\sqrt{k}}{k} \]
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 $\alpha$ level.

Johnson recommends levels more like .01 or .005.

• True, if you reach a smaller significance level, say .01 rather than .025, you may infer a larger discrepancy.

• But more will fail to make it over the hurdle: the Type II error probability increases.
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.

*\( H_{max} : \mu = \bar{x}_{\alpha} \) accords with \( \bar{x}_{\alpha} \) --they’re equal

Next slide: SIST p. 263
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 $H_{\text{max}}$: $\mu = \mu_{\text{max}}$. Giving priors of 0.5 to both $H_0$ and $H_{\text{max}}$ you can report the posteriors. Clearly, (S-1) holds: $H_{\text{max}}$ accords with $\bar{x}$ – it’s equal to it. Our worry is with (S-2). $H_0$ is being rejected in favor of $H_{\text{max}}$, but should we infer it? The severity associated with inferring $\mu$ is as large as $\mu_{\text{max}}$ is

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

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 $\mu \geq \mu_{\text{max}}$ (answer 0.5). Yet Johnson assigns $\Pr(H_{\text{max}}|x) = 0.97$. $H_{\text{max}}$ 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 $H_{\text{max}}$. If we add a column to Table 4.1 for SEV($\mu \geq \mu_{\text{max}}$) it would be 0.5 all the way down!
To conclude....
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)
**Souvenir (R)**
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).