Three Recommendations for Improving the Use of p-Values

Jim Berger

Duke University

with Daniel J. Benjamin (University of Southern California)

NISS Webinar on p-values

November 19, 2019

Premise: Misuse of p-values may gradually disappear, but it will be a slow process and we need immediate "fixes" at various levels of sophistication.

Three Recommendations:

Recommendation 1: *If using the current language of 'statistical significance' for a novel discovery, replace the 0.05 threshold with 0.005. Refer to discoveries with a p-value between 0.05 and 0.005 as 'suggestive,' rather than 'significant.'*

Recommendation 2: *When reporting a p-value, p, in a test of the null hypothesis H_0 versus an alternative H_1, also report that the* data-based odds of H_1 being true to H_0 being true *are* <u>*at most*</u> $1/[-e\ p\ \log p]$, *where* log *is the natural logarithm and e is its constant base.*

Recommendation 3: *Determine and report your prior odds of H_1 to H_0 (i.e., the odds of the hypotheses being true prior to seeing the data), and derive and report the final (posterior) odds of H_1 to H_0, which are the prior odds multiplied by the data-based odds. Alternatively, report that the final (posterior) odds are* <u>*at most*</u> *the prior odds multiplied by* $1/[-e\ p\ \log p]$.

Misinterpretation of p-values

To test: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ based on data $\boldsymbol{X} \sim f(\boldsymbol{x} \mid \theta)$.

- p-value, for test statistic $T(\cdot)$ and actual observation $\boldsymbol{x}$, is $\Pr(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid H_0)$.
- It is common to misinterpret p as the probability that H_0 is true given the data, or to interpret $1/p$ as the odds that H_1 is true compared to H_0, given the data
  - e.g., $p = 0.05$ implies that H_1 is 20 times more likely to be true than H_0.
- This is wrong. The real *data-based odds* (or Bayes factor) of H_1 to H_0, for prior distribution $\pi(\theta)$ under H_1, is

$$B_{10}(\boldsymbol{x}) = \frac{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)d\theta}{f(\boldsymbol{x} \mid 0)} .$$

  This is almost always much smaller than $1/p$, for *any* prior
  - e.g., $B_{10} = 2.3$ when $1/p = 1/[0.05] = 20$.

Recommendation 2: *When reporting a p-value, p, in a test of the null hypothesis H_0 versus an alternative H_1, also report that the* data-based odds of H_1 being true to H_0 being true *are* <u>*at most*</u> $1/[-e\ p\ \log p]$, *where* log *is the natural logarithm and e is its constant base.*

Justification: *Robust Bayesian theory* suggests a way to relate p-values to B_{10}, the data-based odds of H_1 to H_0 (Vovk, 1993, Sellke, Bayarri and Berger, 2001).

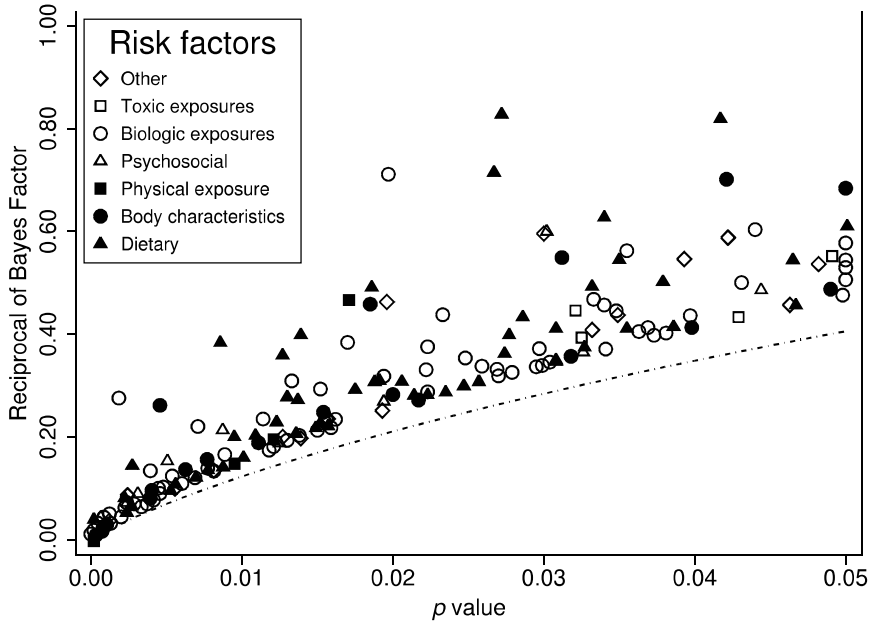
Theorem 1 *A proper p-value satisfies $H_0 : p(X) \sim Uniform(0,1)$, so consider testing this versus $H_1 : p \sim g(p)$, where $Y = -\log(p)$ has a non-increasing failure rate (a natural non-parametric condition on g). Then*

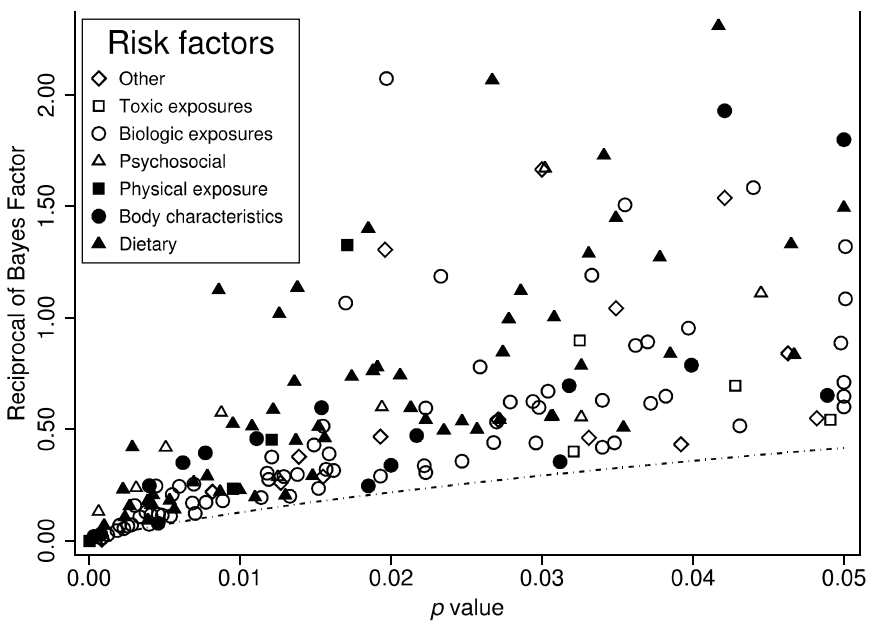
$$B_{10} = \frac{g(p)}{1} \leq \frac{1}{-e\,p\,\log(p)} \quad \textit{for} \quad p < e^{-1}\,.$$

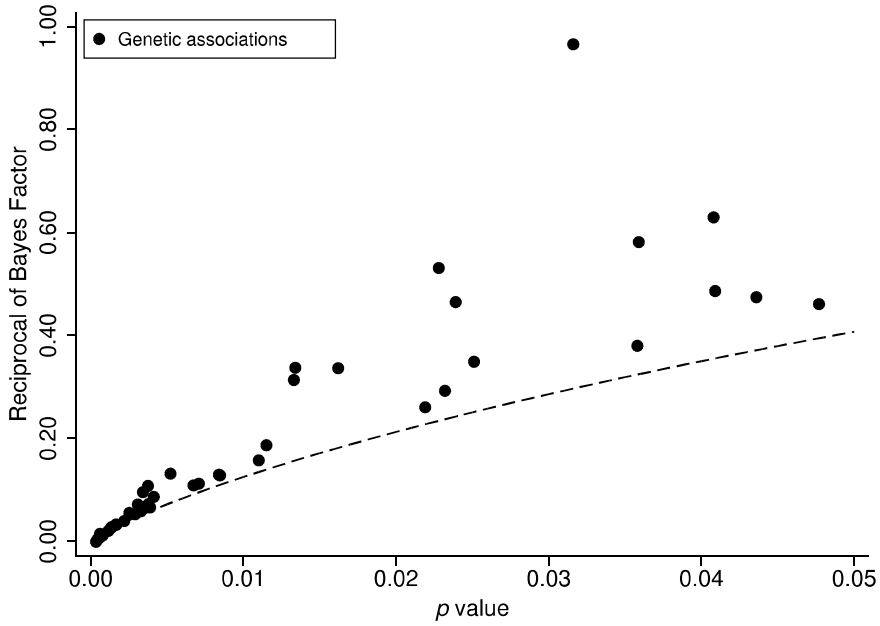
p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001	5×10^{-7}
$\frac{1}{-ep\log(p)}$	1.60	2.44	8.13	13.9	52.9	400	3226	2.0×10^5

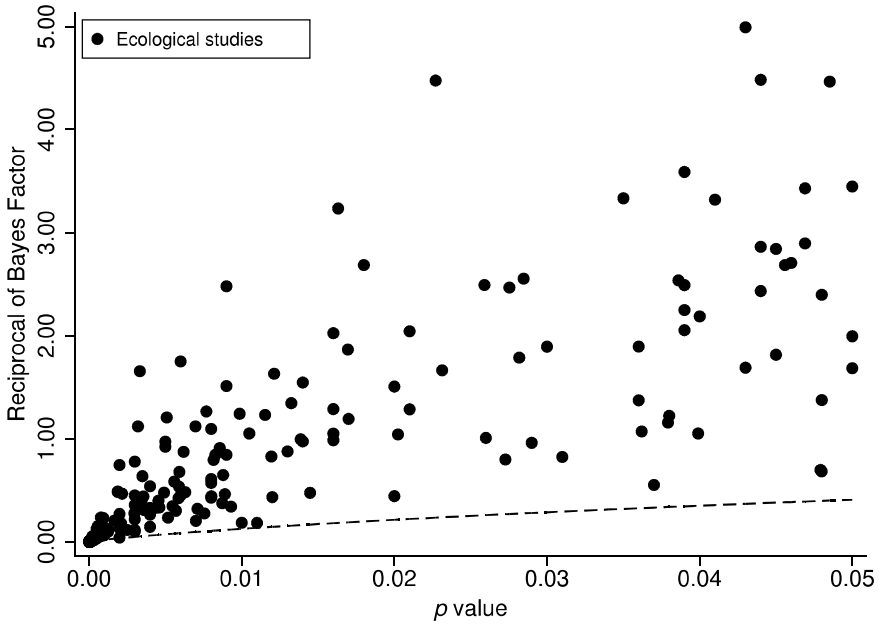
- Although very simple, there was initially concern that the $\frac{1}{-ep\log(p)}$ bound is too large, since it is known that Bayes factors can depend strongly on the sample size n, and the bounds do not.
- But the following studies indicate that this might not typically be a problem. These studies
  - look at large collections of published studies where $0 < p < 0.05$;
  - compute a Bayes factor, $B_{01} = 1/B_{10}$, for each study;
  - graph the Bayes factors versus the corresponding p-values.

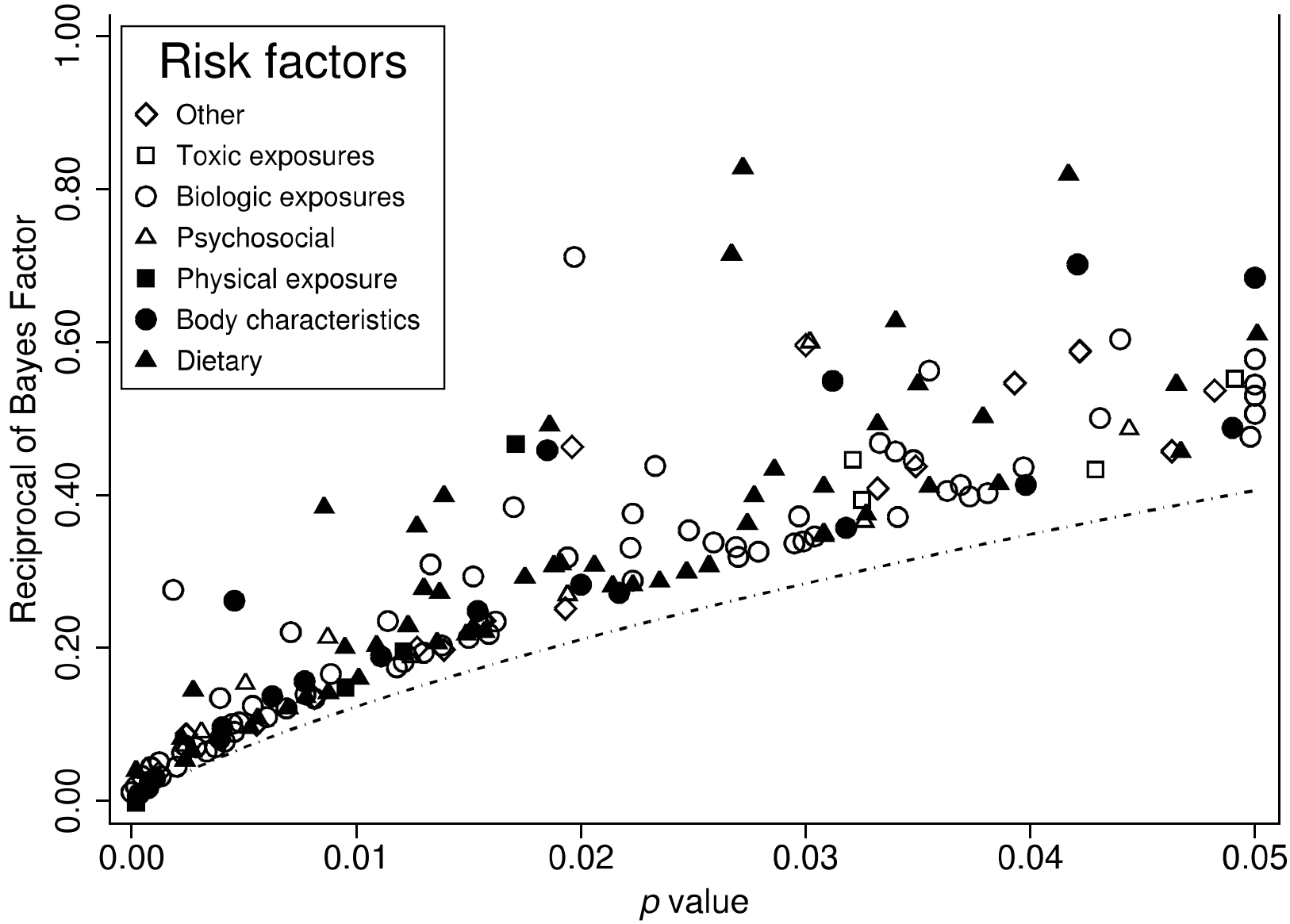
The first two graphs are for 272 'significant' epidemiological studies with two different choices of the prior; the third for 50 'significant' meta-analyses (these three from J.P. Ioannides, Am J Epidemiology, 2008); and the last is for 314 ecological studies (reported in Elgersma and Green, 2011).


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.50
1.00
1.50
2.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Genetic associations
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Ecological studies
Reciprocal of Bayes Factor
0.00
1.00
2.00
3.00
4.00
5.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value

- The data-based odds bound for B_{01} is $B_{01} \geq -e\ p\ \log p\,.$
- The lower boundary in all figures is close to the lower bound $-e\,p\log(p)$ (the corresponding bound for $B_{01} = 1/B_{10}$, given by the dashed lines in the figures), indicating that it is often an accurate bound.

Recommendation 1: *If using the current language of 'statistical significance' for a novel discovery, replace the 0.05 threshold with 0.005. Refer to discoveries with a p-value between 0.05 and 0.005 as 'suggestive,' rather than 'significant.'*

Justification: From the previous table we have

$$B_{10} \leq 2.44 \text{ when } p = 0.05; \qquad B_{10} \leq 13.9 \text{ when } p = 0.005\,.$$

Having 2.44 to 1 odds in favor of H_1 is hardly compelling evidence.

Having 13.9 to 1 odds in favor of H_1 would be reasonably strong evidence.

The article *Redefining Statistical Significance* appearing in 2018 in *Nature Human Behavior*

- Argued for Recommendation 1 above.
- It had 72 authors, leading scientists from a wide variety of disciplines.

Daniel Benjamin, James Berger, Magnus Johannesson, Brian Nosek, E.J. Wagenmakers, Richard Berk, Kenneth Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher Chambers, Merlise Clyde, Thomas Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy Field, Malcolm Forster, Edward George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald Green, Anthony Greenwald, Jarrod Hadfield, Larry Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, James Holland Jones, Daniel Hruschka, Kosuke Imai, Guido Imbens, John Ioannidis, Minjeong Jeon, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafo, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix Schonbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan Watts, Christopher Winship, Robert Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, **Valen Johnson**

Recommendation 3: *Determine and report your prior odds of H_1 to H_0 (i.e., the odds of the hypotheses being true prior to seeing the data), and derive and report the final (posterior) odds of H_1 to H_0, which are the prior odds multiplied by the data-based odds. Alternatively, report that the final (posterior) odds are <u>at most</u> the prior odds multiplied by $1/[-e\ p\ \log p]$.*

Justification: Letting $\Pr(H_1)$ and $\Pr(H_0)$ denote the prior probabilities of H_1 and H_0, a form of Bayes theorem gives

$$\underbrace{\frac{\Pr(H_1 \mid \boldsymbol{x})}{\Pr(H_0 \mid \boldsymbol{x})}}_{\text{posterior odds}} = \underbrace{\frac{\Pr(H_1)}{\Pr(H_0)}}_{\text{prior odds}} \times \underbrace{B_{10}}_{\substack{\text{data-based odds} \\ \text{(Bayes factor)}}},$$

where $\Pr(H_1 \mid \boldsymbol{x})$ and $\Pr(H_0 \mid \boldsymbol{x})$ are posterior probabilities of H_1 and H_0.

From the robust Bayesian bound,

$$\frac{\Pr(H_1 \mid \boldsymbol{x})}{\Pr(H_0 \mid \boldsymbol{x})} \leq \frac{\Pr(H_1)}{\Pr(H_0)} \times \frac{1}{[-e\ p\ \log p]}.$$

Example: Genome-wide Association Studies (GWAS)

- Early GWAS studies – testing H_0 : *gene X is not associated with disease D* versus H_1 : *gene X is associated with disease D* – almost universally failed to replicate (estimates of the replication rate are as low as 1%), because they were doing extreme multiple testing at non-extreme p-values.
- A very influential paper in Nature (2007) by the Wellcome Trust Case Control Consortium proposed the cutoff $p < 5 \times 10^{-7}$.
  - Found 21 genome/disease associations; 20 have been replicated.
- In the analysis, they assessed the prior odds of H_1 to H_0 to be

$$\frac{\Pr(H_1)}{\Pr(H_0)} = \frac{1}{100,000}\,.$$

- The data-based odds for the 21 claimed associations ranged from

$$B_{10} = 10^4 \quad \text{to} \quad B_{10} = 10^{73}\,,$$

resulting in final posterior odds ranging from

$$\frac{\Pr(H_1 \mid \boldsymbol{x})}{\Pr(H_0 \mid \boldsymbol{x})} = \frac{1}{100,000} \times 10^4 = \frac{1}{10} \quad \text{to} \quad \frac{\Pr(H_1 \mid \boldsymbol{x})}{\Pr(H_0 \mid \boldsymbol{x})} = \frac{1}{100,000} \times 10^{73} = 10^{68}\,.$$

Some Caveats:

- The bound on the data-based odds only applies if H_0 is a plausible nested precise hypotheses.
  - A nested precise hypothesis is one with at least one of the unknown parameters in H_1 specified.
  - A plausible hypothesis is one that has nonzero prior probability. (In the GWAS example, the prior probability of H_0 was $1 - 10^5$).
- We felt somewhat guilty about Recommendation 1, proposing the 0.005 significance cutoff.
  - But a cutoff can really simplify things, if it is chosen scientifically.
    - In the GWAS example, they chose an odds cutoff of 10 to 1 for a true discovery to a false discovery (odds of H_1 to H_0), presumably based on an analysis comparing the costs of a false positive and a false negative.
  - Note that the Bayes factor bound $\frac{1}{[-e\ p\ \log p]}$ is not a cutoff.
- Simply deciding between H_0 and H_1 is usually not enough; the size of the effect under H_1 must typically also be considered, preferably through decision analysis.

Frequentist justification for use of $B_{10}(\boldsymbol{x})$

Consider Neyman-Pearson testing with a fixed rejection region $\mathcal{R}$, type I error $\alpha = Pr(\mathcal{R} \mid H_0)$ and power $1 - \beta(\theta) = Pr(\mathcal{R} \mid \theta)$.

Lemma: *The frequentist expectation of* $B_{10}(\boldsymbol{x})$*, over the rejection region and under* H_0*, is*

$$E[B_{10}(\boldsymbol{X}) \mid H_0, \mathcal{R}] = \frac{(1-\bar{\beta})}{\alpha},$$

where $(1-\bar{\beta}) = \int (1-\beta(\theta))\pi(\theta)d\theta$ is the average power wrt the prior $\pi(\theta)$.

- $\frac{(1-\bar{\beta})}{\alpha}$, called the 'rejection odds,' is typically interpreted as the odds that the experiment will yield a correct rejection to an incorrect rejection (or a true positive to a false positive), assuming that the prior odds are 1 to 1.
- The lemma guarantees that, under H_0, the "average of the reported Bayes factors when rejecting" equals the actual rejection odds, so $B_{10}(\boldsymbol{x})$ *is as valid a frequentist report as is* $\frac{(1-\bar{\beta})}{\alpha}$.

GWAS example continued:

- They wanted an experiment with pre-experimental odds of a true to false positive equal to $10:1$.
- The pre-experimental odds of a true to false positive are the prior odds times the rejection odds, so they wanted

$$\frac{1}{100,000} \times \frac{1-\bar{\beta}}{\alpha} = \frac{10}{1}\,.$$

- Typical GWAS studies had power $(1-\bar{\beta}) = 0.5$.
- Solving $[\frac{1}{100{,}000} \times \frac{0.5}{\alpha} = \frac{10}{1}]$ gave $\alpha = 5 \times 10^{-7}$.
- Instead of reporting $10:1$ as the odds of a true to false positive, the Lemma says one has the same frequentist justification for reporting, as the odds,

$$\frac{1}{100,000} \times B_{10}(\boldsymbol{x})\,,$$

which varied from $\frac{1}{10}$ to 10^{63}.

Conclusions and impact on scientific culture

- Optimal is to report the posterior odds $\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} \times B_{10}(\boldsymbol{x})$.
  - Incorporation of the prior odds $\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})}$ radically changes scientific culture, but necessarily so; otherwise nonsense rules in today's multiple testing environments. (For single tests, this is less important.)
  - If $\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})}$ is unavailable, report $B_{10}(\boldsymbol{x})$ as the data-based odds.
  - If $B_{10}(\boldsymbol{x})$ is unavailable, report the upper bound $\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})} \times \frac{1}{[-e\ p\ \log p]}$.
- If $\frac{\Pr(H_1|\boldsymbol{x})}{\Pr(H_0|\boldsymbol{x})}$ and $B_{10}(\boldsymbol{x})$ are unavailable, report the upper bound on the data-based odds $\frac{1}{[-e\ p\ \log p]}$.
- If only a p-value is allowed, set the cutoff for significance to be 0.005.

These recommendations need not change scientific culture, i.e. $p \leq 0.05$ could still be the basis for publication, perhaps with the argument that even 'suggestive' findings are worth reporting.

The recommendations are to better communicate the strength of findings.

Thanks!