Three Recommendations for Improving the Use of p Values

By Daniel J. Benjamin and James O. Berger

Presentation by Satenik Rafayelyan, Merve Tuncer, and Arne Kramer-Sunderbrink

Outline

- Motivation
  - Replication crisis
  - Towards "post $p < 0.05$ era"
- Theoretical Background
- Three Recommendations
- Examples
  - Toy example
  - Real world example

Motivation

Motivation

The 6 principles included in the ASA statement on statistical significance and p-values (Wasserstein and Lazar 2016)

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Replication crisis

The scientific community might not directly move to the "post $p < 0.05$ era" that many would like to see.

- Many (most) investigators will simply find giving up on "statistical significance" to difficult and will end up staying with the current system.
- ... still many misinterpretations of the p-value.

Towards "post $p < 0.05$ era"

? If $p < 0.05$ no longer provides a license for treating a conclusion as "true," then we should start thinking about other elements of the problem in the statistical analysis, e.g.

- transparency of the statistical analysis
- effect size

✓ 3 recommendations as initial, temporary steps that make significant progress immediately.

- Only address the "significance" part of the problem.

Theoretical background

Theoretical background

Definition of p-value

The p-value is the probability, under the null hypothesis, of observing a test statistic as extreme as or more extreme than its observed value.

The Strength of the Evidence Question

How strongly does the evidence from the data favor the alternative hypothesis relative to the null hypothesis?

P-values should not be used to answer the SOEQ!

- The p-value is evaluated only under the null hypothesis.
- The SOEQ doesn't ask about the likelihood of more extreme data than the one actually obtained.

Theoretical Background

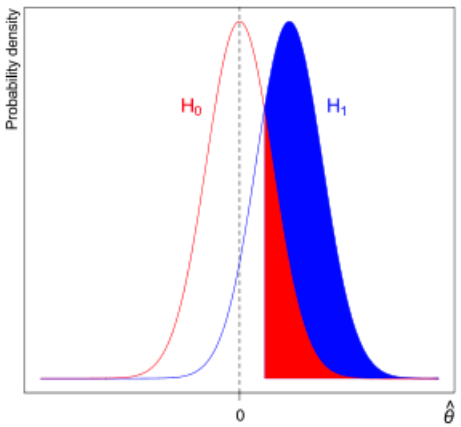
- The Bayesian answer to the SOEQ:

The Bayes Factor

$$BF = \frac{\text{average likelihood of the observed data under the alternative hypothesis}}{\text{likelihood of the observed data under the null hypothesis}}$$

✓ BF has a fully frequentist justification for many common situations [Bayarri et al. (2016)]

- $E_x[R_{post}(x)|H_0, R] = R_{pre}$



Still, it is unrealistic to demand to immediately switch to BFs instead of p-values.

- The computation of the numerator of BF may not be as straightforward as running a prepackaged command.
- The specification of the prior distribution might lead to disagreements.

Converting p-values into BFs

- **Aim**:Translate a p-value into a Bayes Factor.
- **Remark:**

$$BF = \frac{\text{avg. likelihood of the observed data under the } H_1}{\text{avg. likelihood of the observed data under the } H_0}$$

- **Problem**: We need to specify an alternative hypothesis (more specifically, a prior distribution for the parameter values under the alternative hypothesis)

Converting p-Values Into BF

- **Solution**: Instead of calculating BF, calculate Bayes Factor Bound (BFB).
- The highest possible BF consistent with the observed p-value.
- Represents the strongest case for the alternative hypothesis relative to the null hypothesis.

$$BF \leq BFB = \frac{1}{-e\ p\ \log p}$$

- Calculating BFB does not require specifying the distribution of the parameter under the alternative hypothesis because it is **an upper bound** across a large class of reasonable distributions of the parameter.

Converting p-Values Into BF

✓ The following table shows the value of the BFB for a wide range of p-values:

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
BFB	1.60	2.44	8.13	13.9	52.9	400	3226
$\Pr^U(H_1 \mid p)$	0.62	0.71	0.89	0.933	0.981	0.998	0.9997

- P-values often point to much weaker evidence against the null hypothesis than researchers typically assume.
- Conventional levels of significance do not actually provide very strong evidence against the null hypothesis.

Converting p-values Into BFs

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
BFB	1.60	2.44	8.13	13.9	52.9	400	3226
$\Pr^U(H_1 \mid p)$	0.62	0.71	0.89	0.933	0.981	0.998	0.9997

✓ **Consider p=0.05:**

- BFB = 2.44 (BF is at most 2.44)
- The data imply odds in favor of the H_1 relative to the H_0 of at most 2.44 to 1.
- So if the H_0 and H_1 were originally equally likely, there is still **at least** a 29% chance that the H_0 is true (found by 1/3.44).

✓ **Consider p=0.01 ("highly significant"):**

- BFB = 8.13 (8.13 to 1 odds)
- There is still at least a 11% chance that the H_0 is true. (found by 1/9.13)

Three recommendations

Recommendation 2

Idea: Converting a p-value into interpretable odds. Instead of reporting "p = 0.05," report "p = 0.05, BFB = 2.44."

Pros:

- Reporting BFB would alert researchers when seemingly strong evidence is actually not very compelling.
- Prevent researchers from being misled into concluding too much from the p-value of a finding.

Recommendation 2

Cons:

- The BFB is only an upper bound on the Bayes factor, BFB may be far from the BF.

However, note that when BFB is calculated from real data from a range of scientific fields, BFB is often not that far from the BF implied by a scientifically reasonable alternative hypothesis. (Bayarri et al., 2016)

Recommendation 1

Idea: Replace the 0.05 threshold with 0.005. - despite its inappropriateness.

Pros:

- A p-value of 0.05 may have **a high false positive rate** even the original study has no other problems (such as poor study design or multiple hypothesis testing).

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
BFB	1.60	2.44	8.13	13.9	52.9	400	3226
$\Pr^U(H_1 \mid p)$	0.62	0.71	0.89	0.933	0.981	0.998	0.9997

- a p-value of 0.05 corresponds to odds of at most 2.44 to 1 against the null hypothesis, which is fairly weak evidence.
- a p-value of 0.005 can correspond to **much stronger evidence,** with odds of up to 13.9 to 1 against the null hypothesis.

Recommendation 1

Cons:

- Changing the significance threshold from 0.05 to 0.005 would cause an unacceptable increase in the rate of false negatives.
  - But: Failing to reject the null hypothesis does not mean accepting it.
  - Recommendation: Relabeling findings with p-value between 0.05 and 0.005 as "suggestive" instead of "significant".
  - False negative rate will not increase if sample size increases (by roughly 70%) so that statistical power is held constant.

Recommendation 1

Cons:

- The corresponding values of BFB are upper bounds, and the true strength of evidence will sometimes be much weaker. (E.g. if observed evidence is inconsistent with the null hypothesis and/or with the reasonable alternative hypotheses.)

Recommendation 3

Remark:

The Strength of Evidence Question (SOEQ)

- How strongly does the evidence from the data favor the alternative hypothesis relative to the null hypothesis?

More Likely Hypothesis Question (MLHQ)

- How likely is the alternative hypothesis relative to the null hypothesis?

Example:

How likely is it that there is truly an effect of the treatment, as opposed to no effect? $\Rightarrow$ *a crucial question for understanding results of a research study*.

Recommendation 3

- The answer to the MLHQ depends on **the prior odds** of the alternative hypothesis relative to the null hypothesis.

How do we choose prior odds?
This question does not have an exact answer.
It is up to the type of the research study.
For instance,

- For medical treatment, prior odds of 1 to 1 may be assigned.
- For genetic studies, it is often chosen to be 1 to 100,000.
- When testing extrasensory perception the prior odds might be extremely small for most scientists. (Wellcome Trust Case Control Consortium, 2007)

Recommendation 3

Answering MLHQ Question by Using Bayesian Approach:

- Post-experimental odds = Bayes Factor $\times$ Prior Odds
- We do not need a p-value.

Example:

- BF = 4:1
- Prior Odds = 1:2
- Post-experimental-odds = $\frac{4}{1} \times \frac{1}{2} = \frac{2}{1}$
- The post-experimental-odds are (2:1) in favor of the alternative hypothesis.

Usage areas of the recommendations

- Recommendation 1 should be implemented only temporarily and only in research fields where it is not possible to implement Recommendations 2 or 3.
- For "novel discoveries", the recommendation 1, the threshold of 0.005 might be applied.
- Novelty $\Rightarrow$ Prior odds of a discovery are not higher than 1 to 1.
- Note that, if Recommendations 2 and 3 are adopted, there is no need to worry about whether the discovery is "novel" or not, the issue of its prior probability will be addressed directly.

A toy example

Toy example

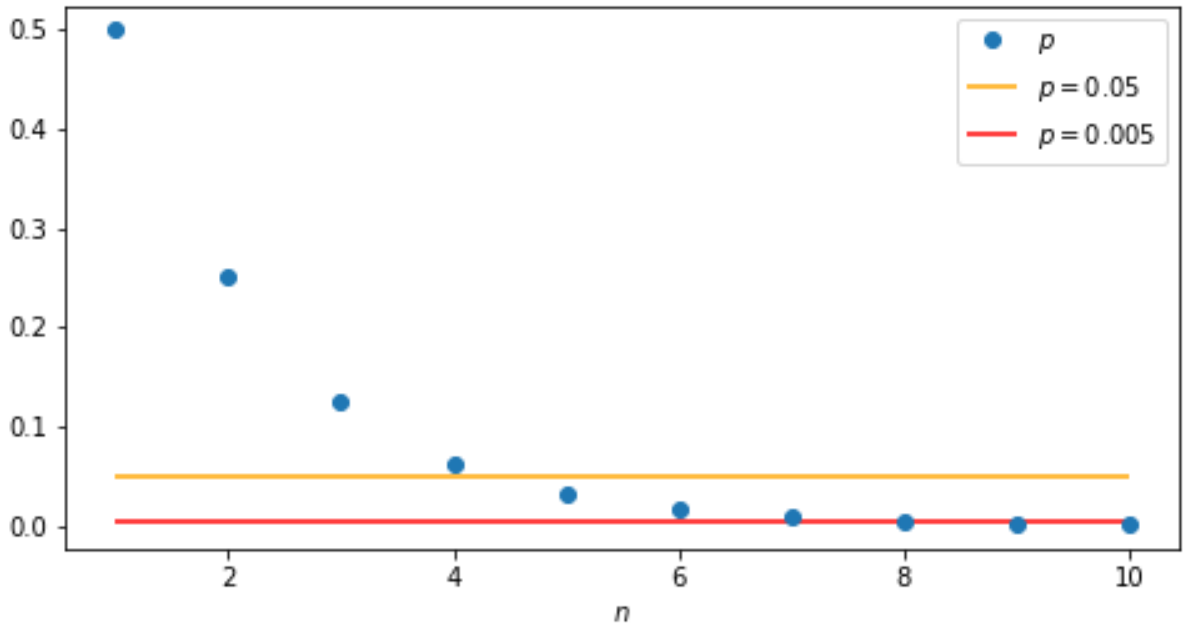
- Imagine we throw a coin n times to see if it is fair or weighted in favor of heads.
- Model: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} bernoulli(\theta)$
- H_0: $\theta = \theta_0 = 0.5$
- Test statistic: $T = \sum_{i=1}^{n} X_i \sim binomial(\theta, n)$

Observation

- Imagine we observe $x_1 = \ldots = x_n = 1$, i.e. we observe n heads in a row out of n throws, not a single tail.
- Test statistic: $t = n$
- Remember: In all of the following toy example discussion we will assume this extreme observation, i.e. when we write n, we mean not only the number of observations but also the number of heads which is also the value of the test statistic!

p-value (one-sided)

- p-value: $p = P(T \geq t|H_0) = P(T = n|H_0) = \frac{1}{2^n}$



- For $n \geq 5$ we get *suggestive* evidence against H_0.
- For $n \geq 8$ we get *significant* evidence against H_0.

Surprise value

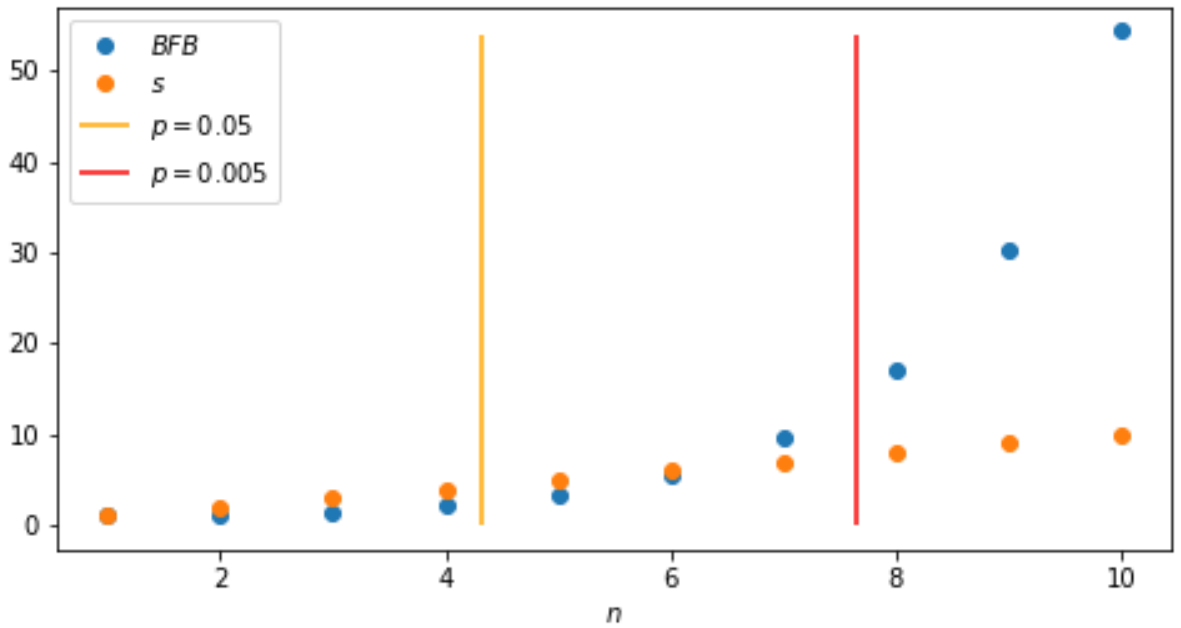
- The surprise value $s = -\log_2(p)$, recommended as an alternative to p by Greenland (2019), would be simply $s = n$.
- For $n = 5$ (suggestive evidence) we get $s = 5$
- For $n = 8$ (significant evidence) we get $s = 8$

Bayes factor bound

- The Bayes factor bound, recommended as an alternative to p by Benjamin and Berger computes to

$$BFB = \frac{1}{-e\ p \log p} = \frac{1}{-e\ 0.5^n \log 0.5^n} = \frac{2^n}{e\ n \log 2}$$

- For $n = 5$ (suggestive evidence) we get $BFB \approx 3.39$
- For $n = 8$ (significant evidence) we get $BFB \approx 16.98$

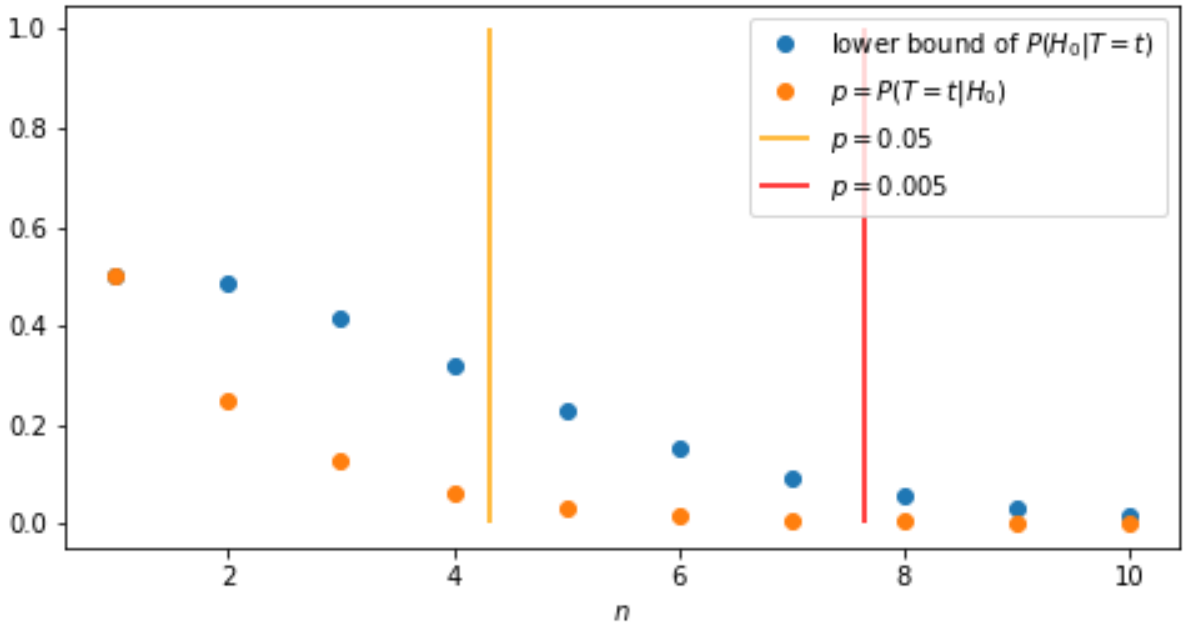


What does it mean??

- Assume our prior odds are 1 : 1, i.e. we are maximally unsure whether our coin is fair.
- This corresponds to a prior $P(H_0) = P(H_1) = 0.5$.
- Hence the posterior odds (in favor of H_1) (= the prior odds $\times$ the Bayes factor) are at most BFB.
- We can translate this posterior odds into the thing the p-value is often confused with: The posterior probability of H_0:

$$P(H_0|T = t) \geq \frac{1}{BFB + 1}$$

p vs posterior



In case we chose different prior odds we simply have to scale the *BFB* by that factor.

Note that the lowest possible posterior is much higher than the p-value, especially for non-significant evidence.

Computing the actual Bayes factor

- Now, lets extend our toy example by a distribution of θ under H_1.
- This will allow us to compute the actual Bayes factor and see how close it is to our Bayes factor bound.

Sharp case

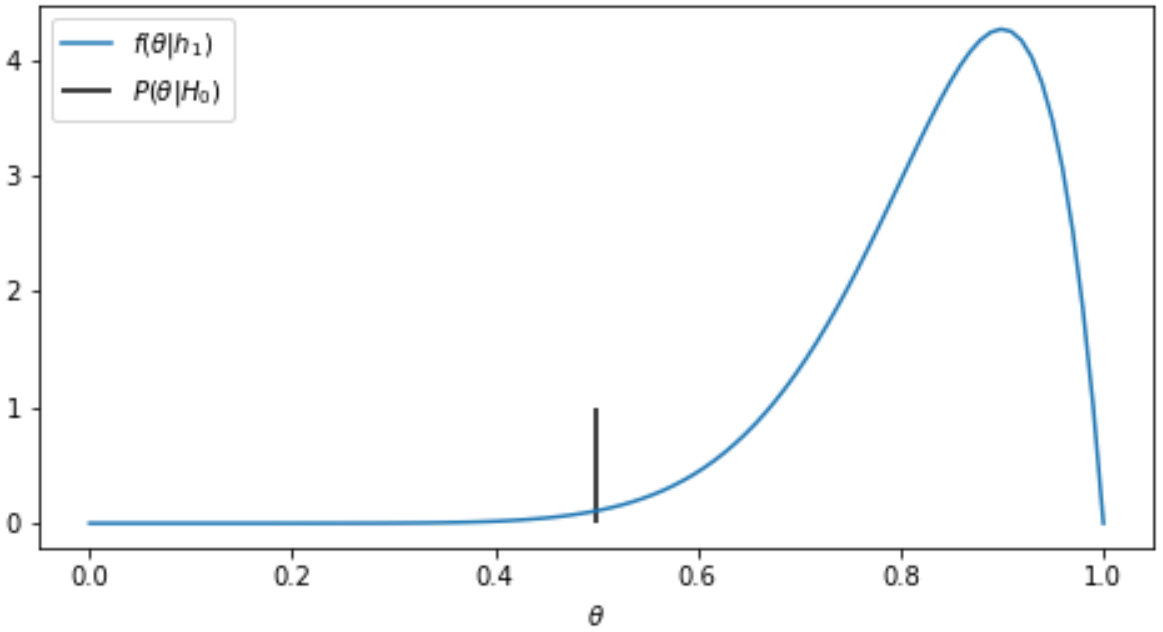
- Lets assume we know that there are only two kinds of coins: fair coins ($\theta_0 = 0.5$) and coins with head on both sides ($\theta_1 = 1$).
- Hence the Bayes factor is $BF = \frac{P(T=t|H_1)}{P(T=t|H_0)} = \frac{\theta_1^n}{\theta_0^n} = 2^n$
- Note that this is the highest possible BF in this situation over all possible distributions of θ under H_1.
- In fact it is so high that it violates our BFB:

$$2^n > \frac{2^n}{e\ n \log 2}$$

- Remember: The BFB is computed under certain assumptions that can be violated in extreme cases like this! (see Sellke, Bayarri & Berger 2001)

Beta case

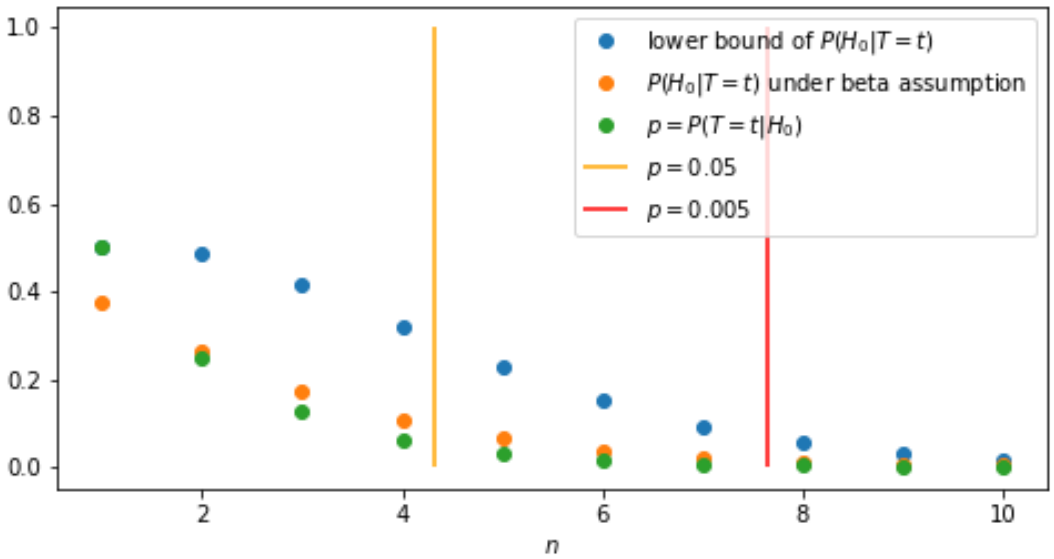
- Lets try again with less extreme assumptions about the distribution of θ under H_1.
- Lets assume coin manipulators on average manage to produce unfair coins with $\theta = 5/6$ but never precisely that value and never values as extreme as in the sharp case.
- This could be modelled as $\theta|H_1 \sim Beta(\alpha = 10, \beta = 2)$.



Hence the Bayes factor can be computed as

$$\begin{aligned} BF &= \frac{P(T = t|H1)}{P(T = t|H0)} \\ &= \frac{\int_0^1 P(T = t|\theta) f(\theta|H_1) d\theta}{\theta_0^n} \\ &= 2^n \frac{\Gamma(12)}{\Gamma(10)\Gamma(2)} \int_0^1 \theta^{n+9} - \theta^{n+10} d\theta \\ &= \frac{110 * 2^n}{n^2 + 21n + 110} \end{aligned}$$

Lets again transform our BF to a posterior probability of H_0 (see slide "What does it mean??") and compare it to the bound we calculated using the BFB and the p-value.



- Again, the lower bound does not hold. To be honest, it is not completely clear to us why that is. Maybe it is because our data is discrete and hence p is only approximately uniformly distributed under the null or because our data (heads only) is unusually unambiguous in favoring the alternative?

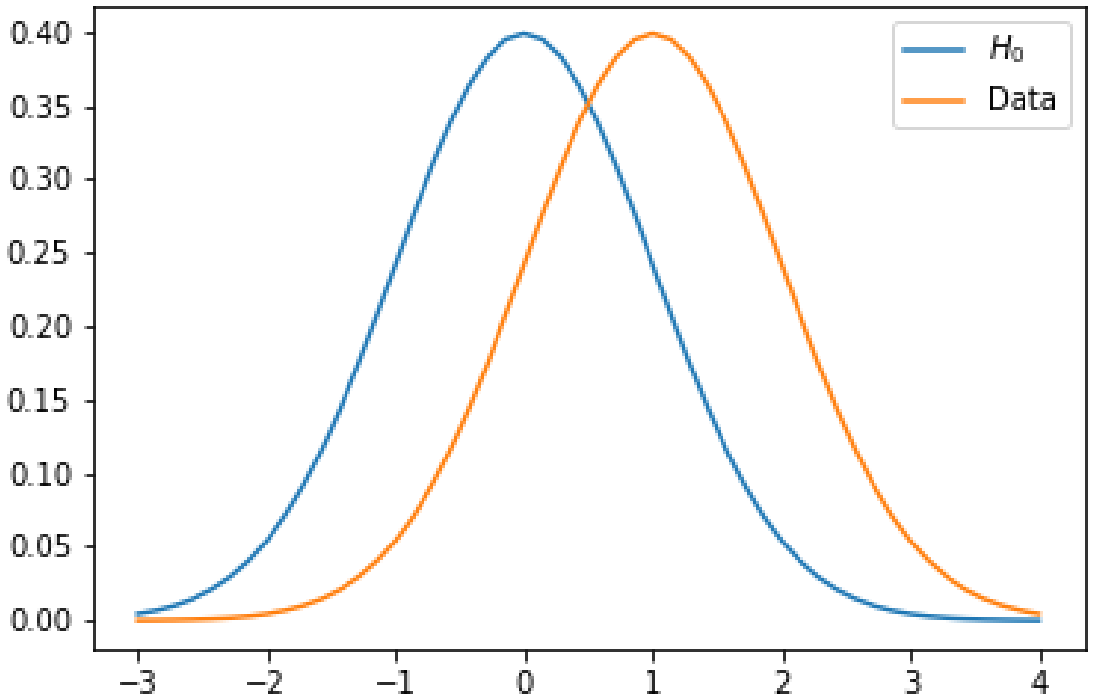
Another toy example

Toy example

- Since the previous example did not yield the results the paper promised, we decided to double check with another example, this time with a more common Gauß test:
- Model: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} norm(\mu, \sigma = 1)$
- H_0: $\mu \leq \mu_0 = 0$
- Test statistic: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \sqrt{n}\bar{X} \sim norm(0, 1)$ under H_0

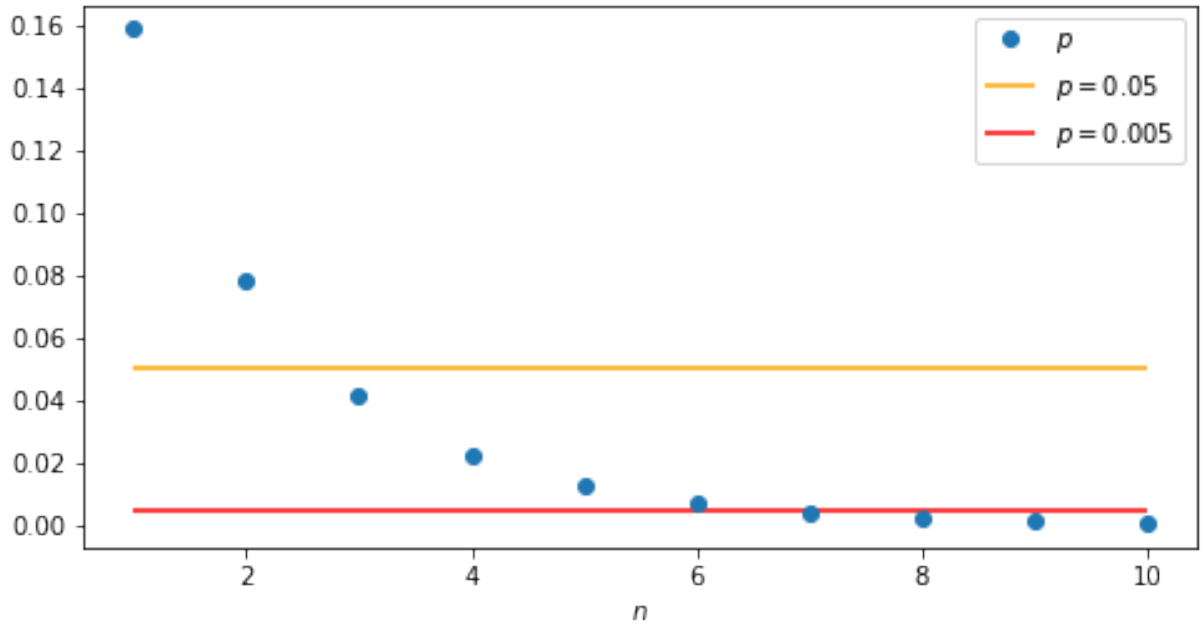
Observation

- Imagine we observe data with $\bar{x} = 1$, i.e. the average observation is one standard deviation away from μ_0.
- Test statistic: $z = \sqrt{n}$
- Lets observe how our different measures of evidence change as we increase n, i.e. collect more data.



p-value (one-sided)

- p-value: $p = P(Z \geq \sqrt{n}|H_0) = 1 - \Phi(\sqrt{n})$, where Φ is the cumulative distribution function of the standard normal distribution



- For $n \geq 3$ we get *suggestive* evidence against H_0.
- For $n \geq 7$ we get *significant* evidence against H_0.

Surprise value

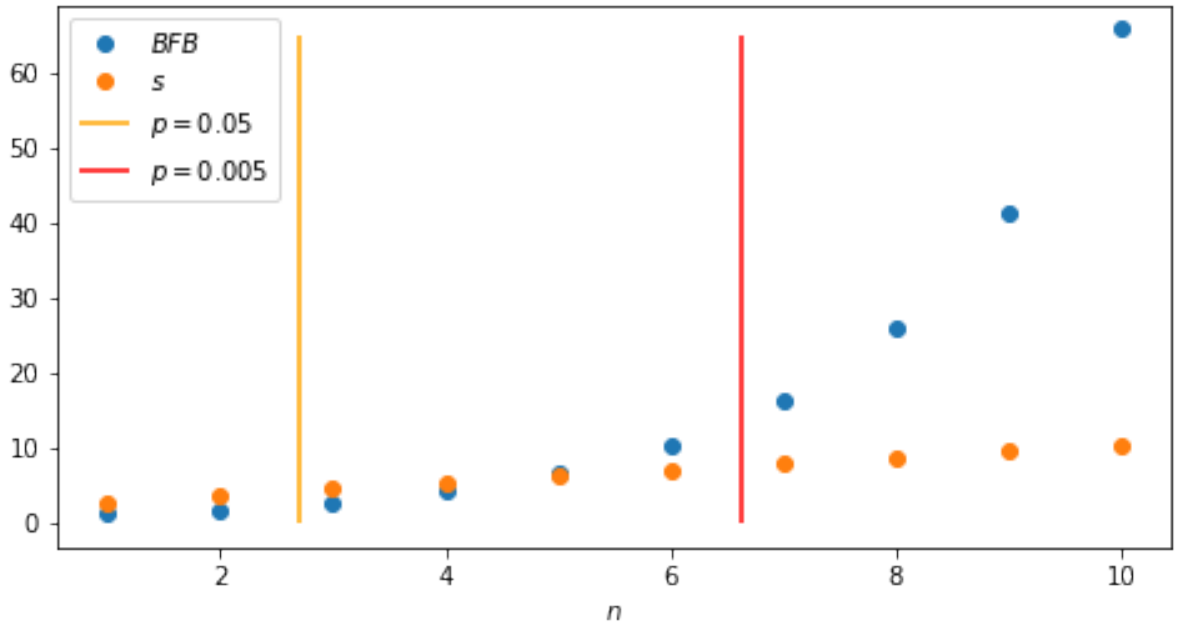
- $s = -\log_2(p) = -\log_2(1 - \Phi(\sqrt{n}))$
- For $n = 3$ (suggestive evidence) we get $s \approx 4.6$
- For $n = 7$ (significant evidence) we get $s \approx 7.9$

Bayes factor bound

- The Bayes factor bound computes to

$$BFB = \frac{1}{-e\ p \log p} = \frac{1}{-e\ (1-\Phi(\sqrt{n}))\log(1-\Phi(\sqrt{n}))}$$

- For $n = 3$ (suggestive evidence) we get $BFB \approx 2.78$
- For $n = 7$ (significant evidence) we get $BFB \approx 16.4$

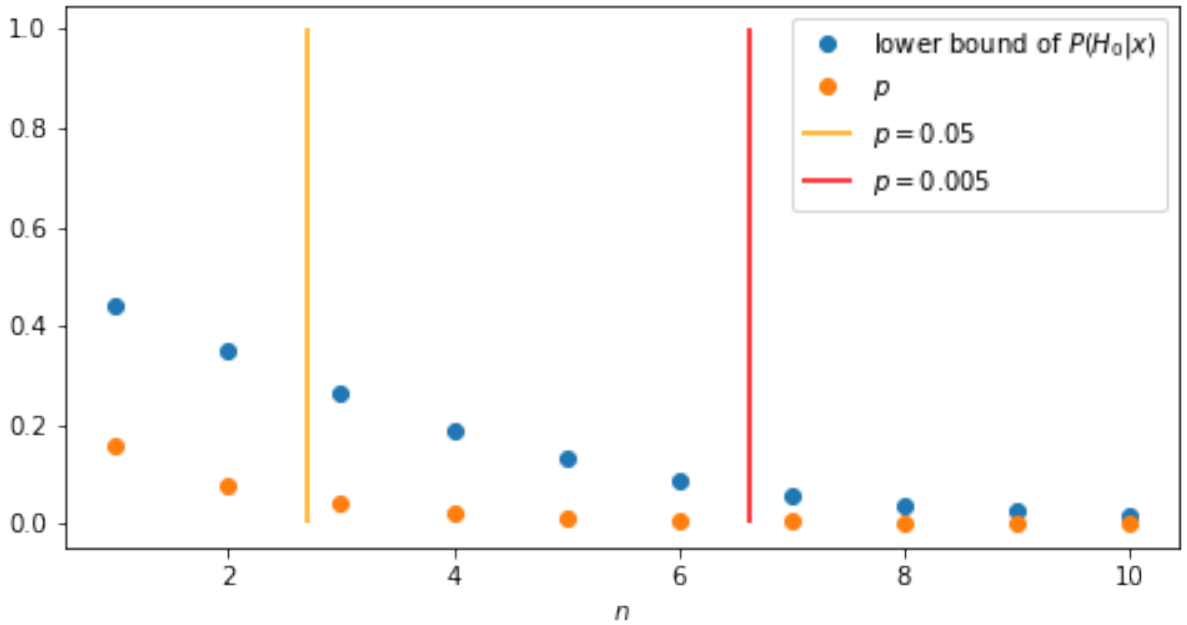


What does it mean??

- Again, assume our prior odds are 1 : 1, i.e. we are maximally unsure whether our coin is fair.
- Hence the posterior odds (in favor of H_1) (= the prior odds $\times$ the Bayes factor) are at most BFB.
- Again, we can translate this posterior odds into the posterior probability of H_0:

$$P(H_0|Z = z) \geq \frac{1}{BFB + 1}$$

p vs posterior



Note, again, that the lowest possible posterior is much higher than the p-value, especially for non-significant evidence.

Computing the actual Bayes factor

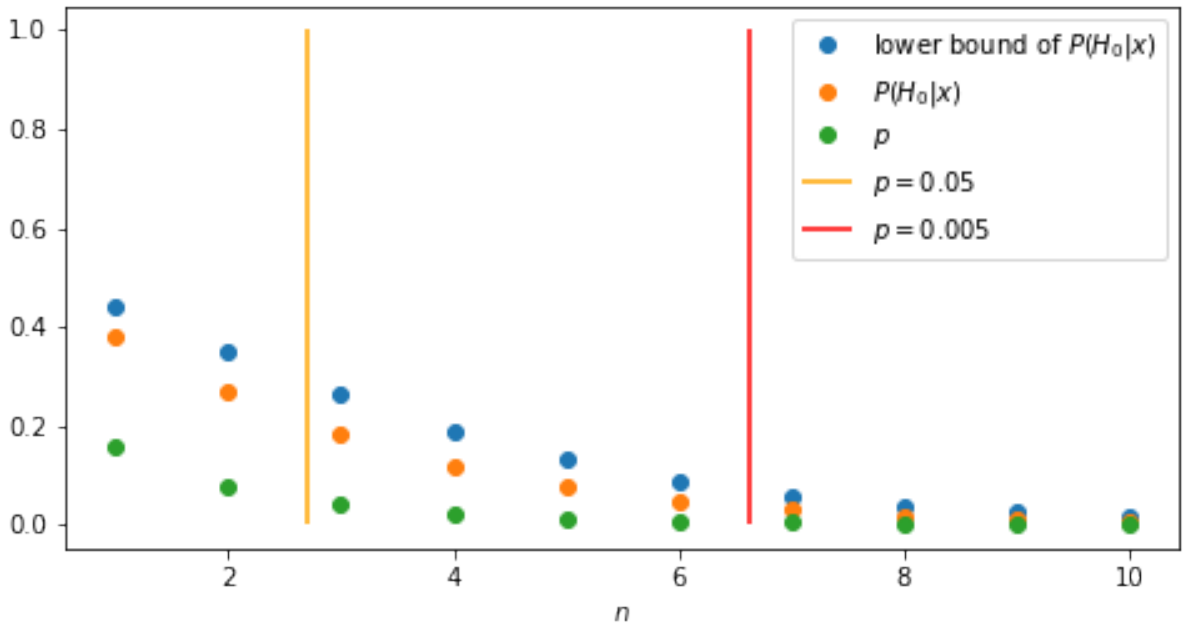
- Now, lets extend our toy example by a distribution of μ under H_1.
- This will allow us to compute the actual Bayes factor and see how close it is to our Bayes factor bound.

Sharp case

- Lets assume we know that the expectation of our data is either 0 (H_0) or 1 (H_1). I.e. $P(\mu = \mu_1 | H_1) = 1$ for $\mu_1 = 1$ and else 0.
- Hence the probability density function of $\bar{X}$ computes to $f(\bar{x}|H_1) = f(\bar{x}|\mu = \mu_1) = \varphi(\frac{\bar{x}-\mu_1}{\sigma/\sqrt{n}}) = \varphi(\sqrt{n}(1-\mu_1)) = \varphi(0)$, where φ is the probability density function of the standard normal distribution.
- Hence the Bayes factor is $BF = \frac{f(\bar{x}|H_1)}{f(\bar{x}|H_1)} = \frac{\varphi(0)}{\varphi(1)} = \frac{\exp(0)}{\exp(-n/2)} = \exp(n/2)$

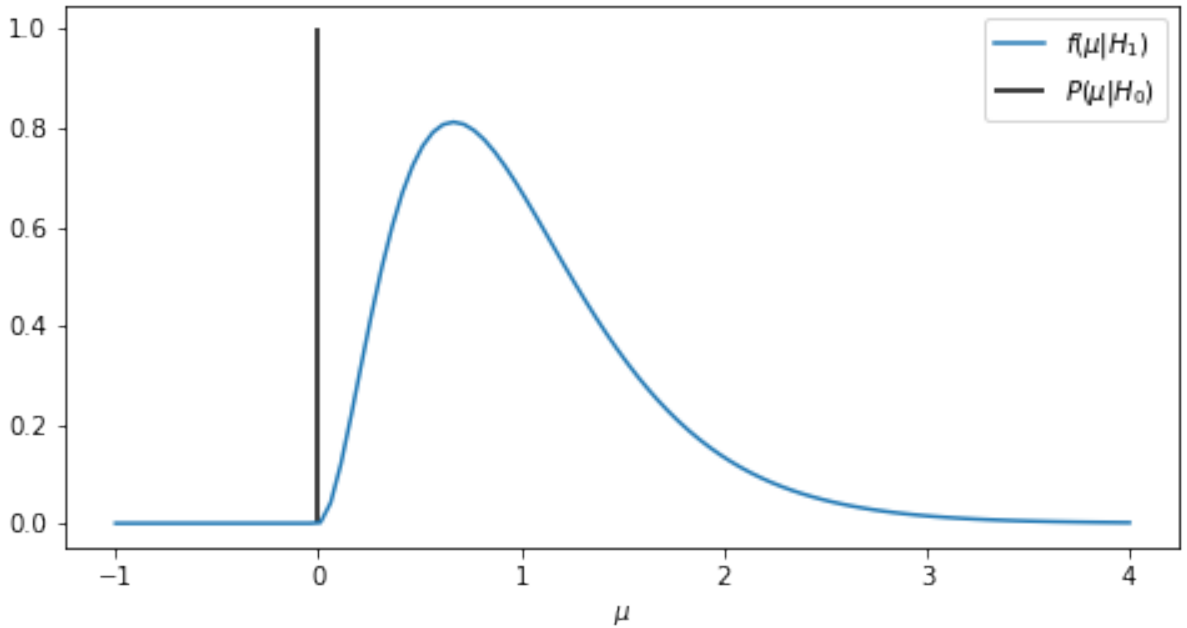
Sharp case plot

- Note that this is the highest possible BF in this situation over all possible distributions of μ under H_1.
- In fact it is, again, so high that it violates our BFB:



Gamma case

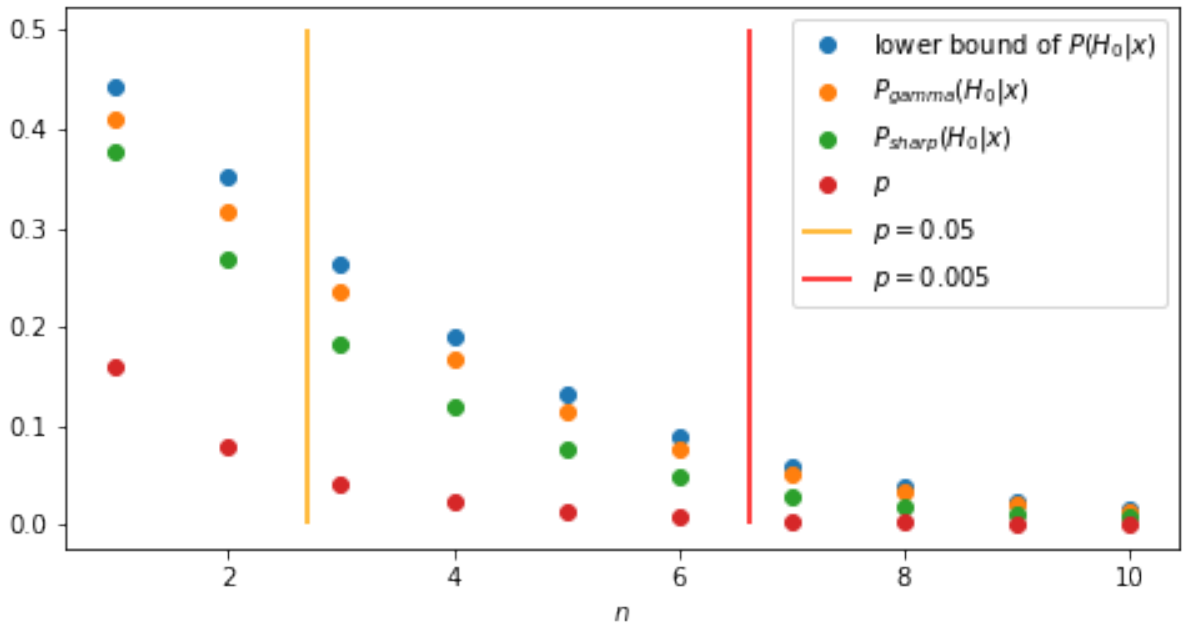
- Lets try again with less extreme assumptions about the distribution of μ under H_1.
- Lets say $\mu|H_1 \sim Gamma(\alpha = 3, \beta = 3)$.



Hence the Bayes factor can be computed as

$$
\begin{aligned}
BF &= \frac{f(\bar{x}|H1)}{f(\bar{X}|H0)} \\
&= \frac{\int_0^\infty f(\bar{X} = \bar{x}|\mu = \mu_1) f(\mu = \mu_1|H_1) d\mu_1}{f(\bar{X} = \bar{x}|H0)} \\
&= \frac{\int_0^\infty \varphi(\sqrt{n}(1-\mu_1)) g(\mu_1) d\mu_1}{\varphi(\sqrt{n})} \\
&= \exp(n/2) \int_0^\infty \exp(-n(1-\mu_1)^2/2) g(\mu_1) d\mu_1
\end{aligned}
$$

For the plot, the integral is approximated numerically.
Lets again transform our BF to a posterior probability of H_0 (see slide "What does it mean??") and compare it to the BFB.



Again, even though the posterior is much closer to the bound as in the the first example, it still violates the bound, even though this time there is really nothing unusual with the example, so we have no idea what the problem is ...

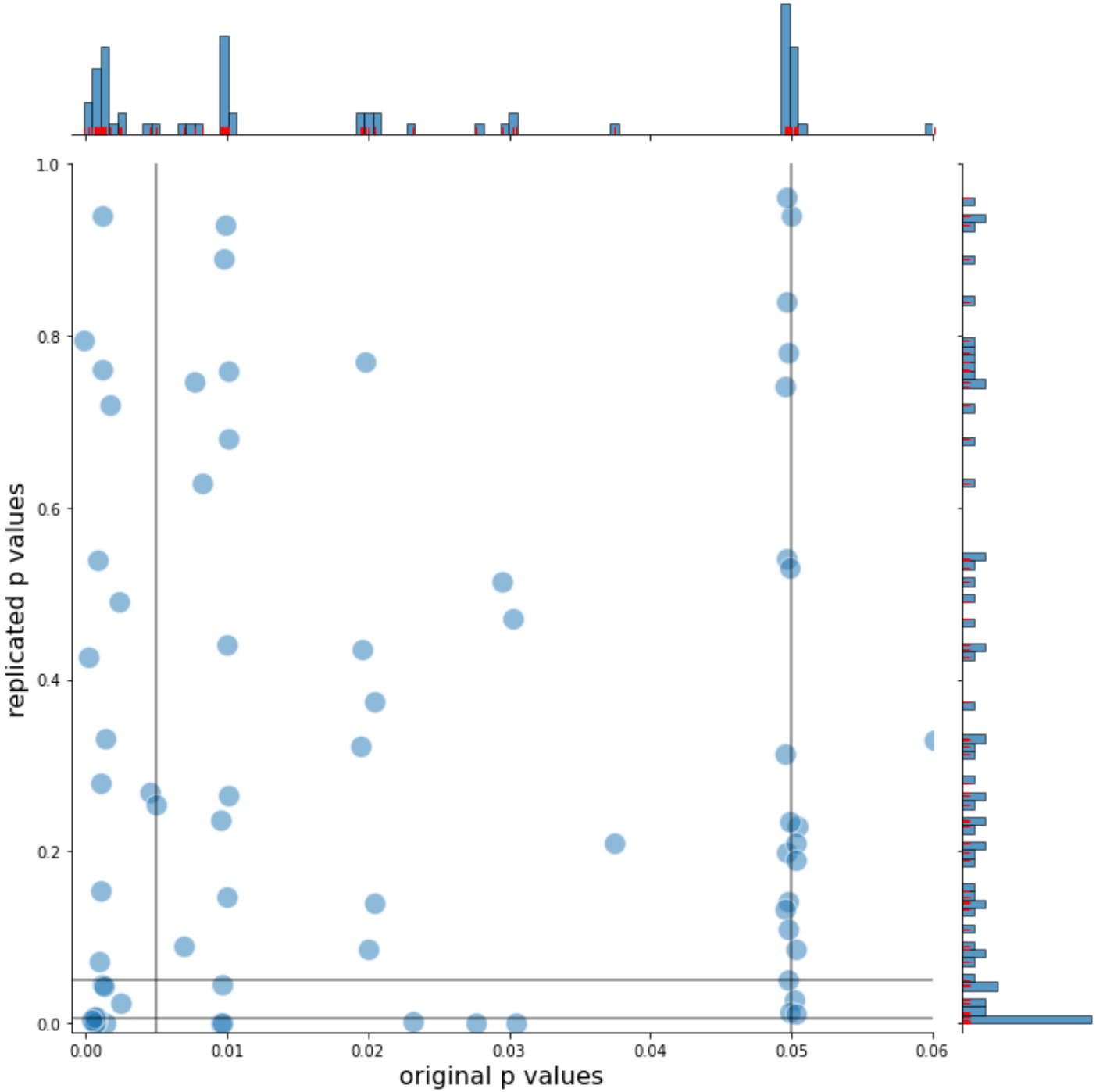
(In search of) A real world example

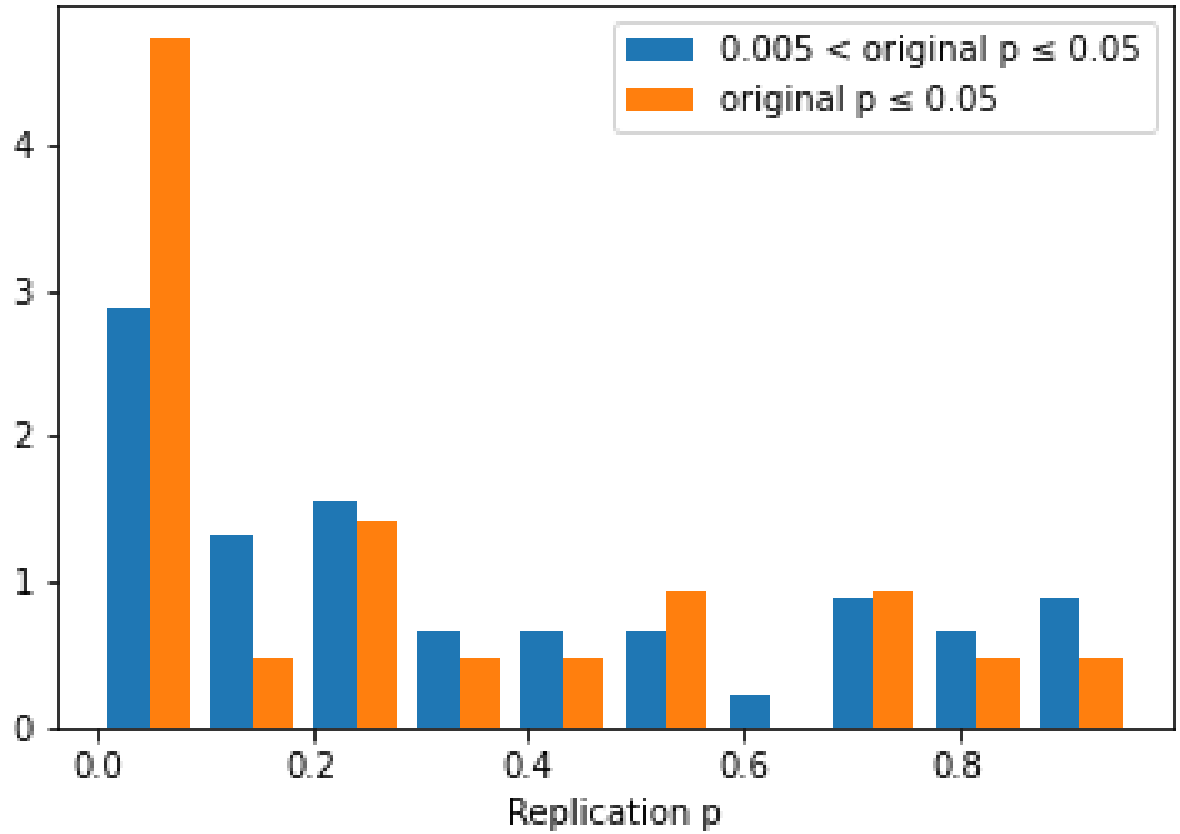
Real world example

The idea for this last section of our presentation was to find a real world example of a study that claimed significant results but turned out not to be reproducible and see if following the proposed recommendations would have helped.
We found a meta study (Aarts et al 2015) that looked at 100 studies in three major psychological journals from 2008 to try to replicate them. (An undertaking of epic proportions with 270 contributing authors!) Surely we would find what we were looking for there...

Aarts et al

- And indeed: 97% of the original papers claimed significant results while only 36% of the replication studies could find significant results.
- BUT: Would the recommendations have helped?
- So we looked through the original studies but struggled to find clear examples of p-values that were high enough that calling them significant would be an misinterpretation.
- Than we got the raw data and it turns out that:

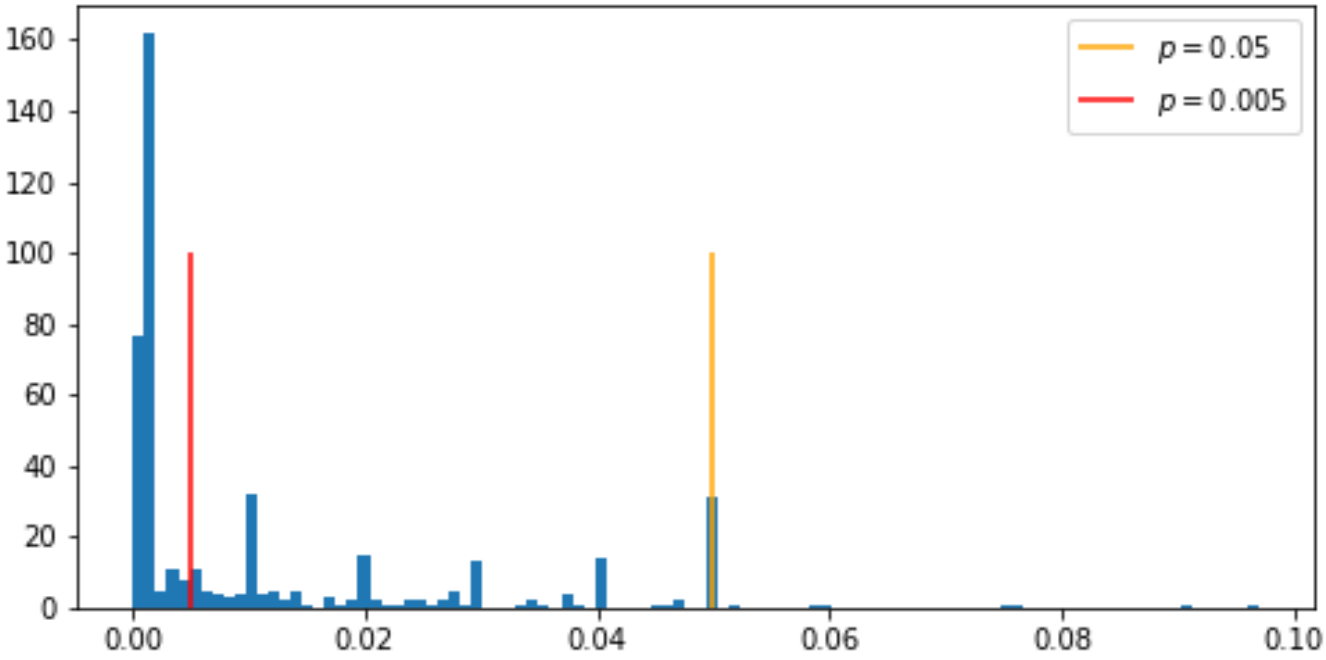




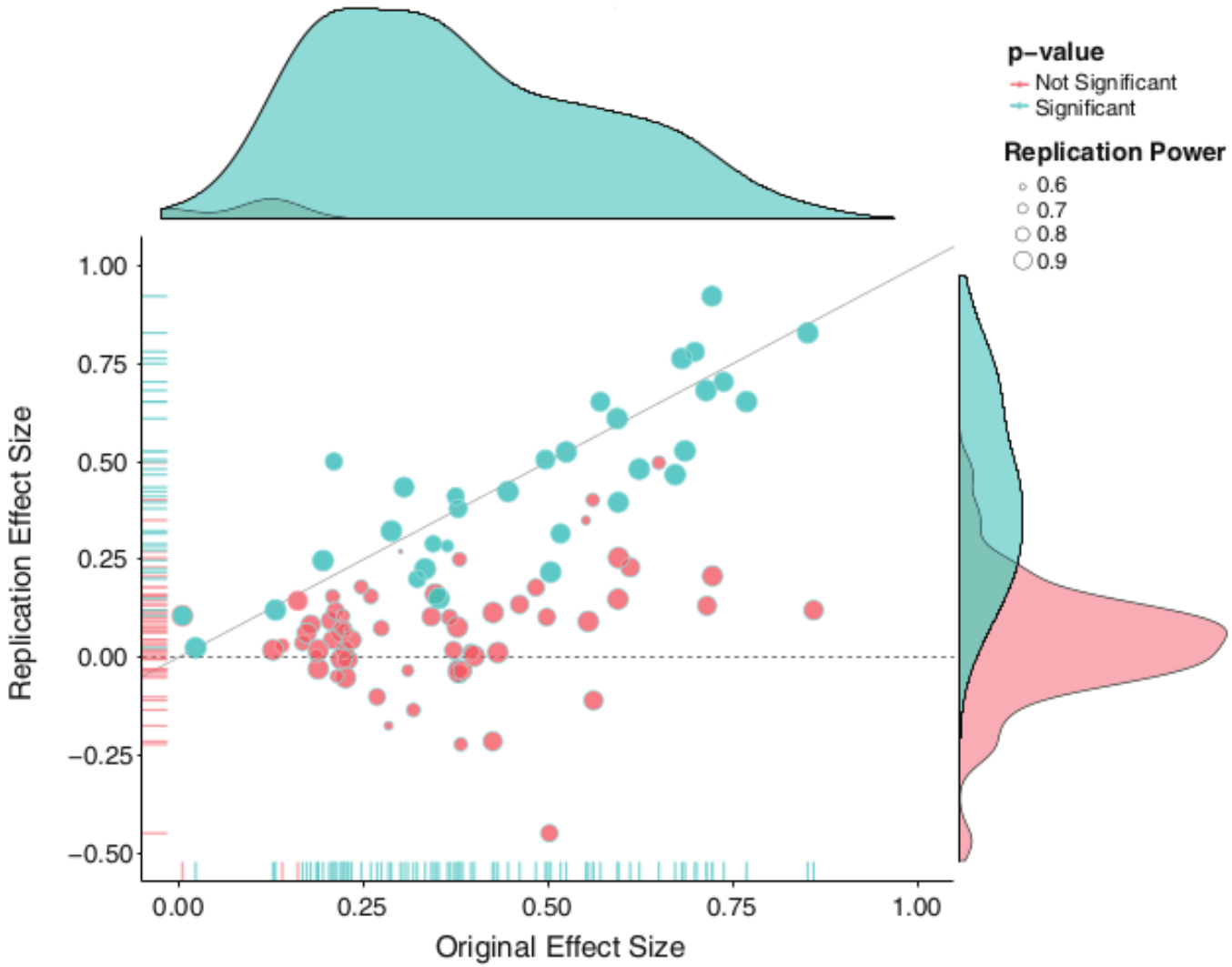
- 27% of the p-values bellow 0.05 are already below 0.005.
- Whats more: This significant third does not look much better than the merely suggestive p-values when it comes to replicability.
- $\Rightarrow$ Misinterpretation of merely suggestive p-values as significant does not seem to be the main issue here...

More data

- We checked another data set with 195 medical trial abstracts from the field of diabetes and glaucoma from 2000 to 2018.
- 524 p-values were reported, 79% bellow 0.05, 50% bellow 0.005.
- If we only look at the smallest value reported per paper, 90% were bellow 0.05 and 69% bellow 0.005.



Effect size to the rescue?



- The Graph (from Aarts et al 2015) shows effect sizes measured in Pearsons correlation coefficient r.
- $r \in [-1, 1]$, $|r| \approx 0.3$ is considered a medium size effect, $|r| \approx 0.5$ is considered big.
- So most of the originally measured effect sizes are already medium to big, that does not seem to help either...

We're done, thanks for your attention!

Literature

- Alexander A. Aarts; et al; & Stephanie C. Lin (2015) *Estimating the reproducibility of psychological science*, Science, 349:6251, 943-950
- Bayarri, M. J., Benjamin, D., Berger, J., and Sellke, T. (2016), *Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses*, Journal of Mathematical Psychology, 72, 90–103
- Daniel J. Benjamin & James O. Berger (2019) *Three Recommendations for Improving the Use of p-Values*, The American Statistician, 73:sup1, 186-191
- Sander Greenland (2019) *Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values*, The American Statistician, 73:sup1, 106-114

Literature

- Thomas Sellke, M. J. Bayarri & James O. Berger (2001) *Calibration of p Values for Testing Precise Null Hypotheses*, The American Statistician, 55:1, 62-71
- Wellcome Trust Case Control Consortium (2007), *Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls*, Nature, 447, 661–678