P-values:
Preaching to the Choir or Prancing with the Devil?

https://commons.wikimedia.org/wiki/File:DSC_9351_(10142693713).jpg

https://commons.wikimedia.org/wiki/File:MILTON_(1695)_p044_PL_2.jpg

Prof. Naomi S. Altman
Dept. of Statistics and Huck Institutes of Life Sciences
The Pennsylvania State University

Agenda

1. Why talk about p-values?
2. The ASA Statement on P-values
3. P-values and Reproducibility
4. P-values are random variables
5. We can do better

Why talk about p-values

It seems to have started with John Ioannidis (2005 *PLoS Medicine*) provocatively titled "Why Most Published Research Findings are False"

Why talk about p-values

And then it was taken up by the popular press. (@2013)

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

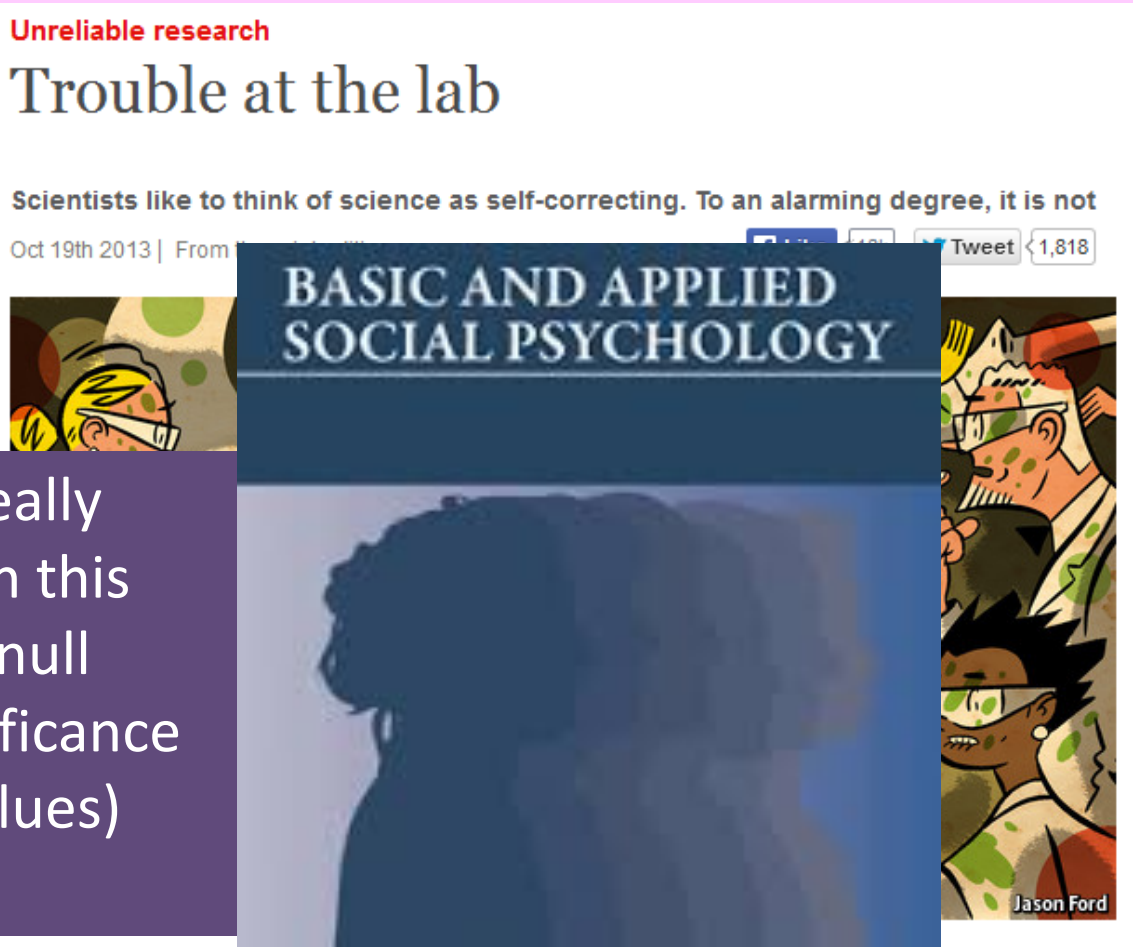
Oct 19th 2013 | From the print edition

http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

It seems to have started with John Ioannidis (2005 *PLoS Medicine*) provocatively titled "Why Most Published Research Findings are False"

Why talk about p-values

And then it was taken up by the popular press. (@2013)



It seems to have started with John Ioannidis (2005 *PLoS* provocative "Why Most Research F False"

But things got really interesting when this journal banned null hypothesis significance testing (i.e. P-values) (2015)

The ASA Statement on P-values

Who?

- A panel of 32 statisticians from across the philosophical divides
- included theoretical and applied statisticians
- some educators and communicators
- Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Ron Wasserstein, Steve Ziliak.

The ASA Statement on P-values

Little p-value

What are you trying to say

Of significance?

Haiku by Steve Ziliak (econometrician, Roosevelt U.)

The ASA Statement on P-values

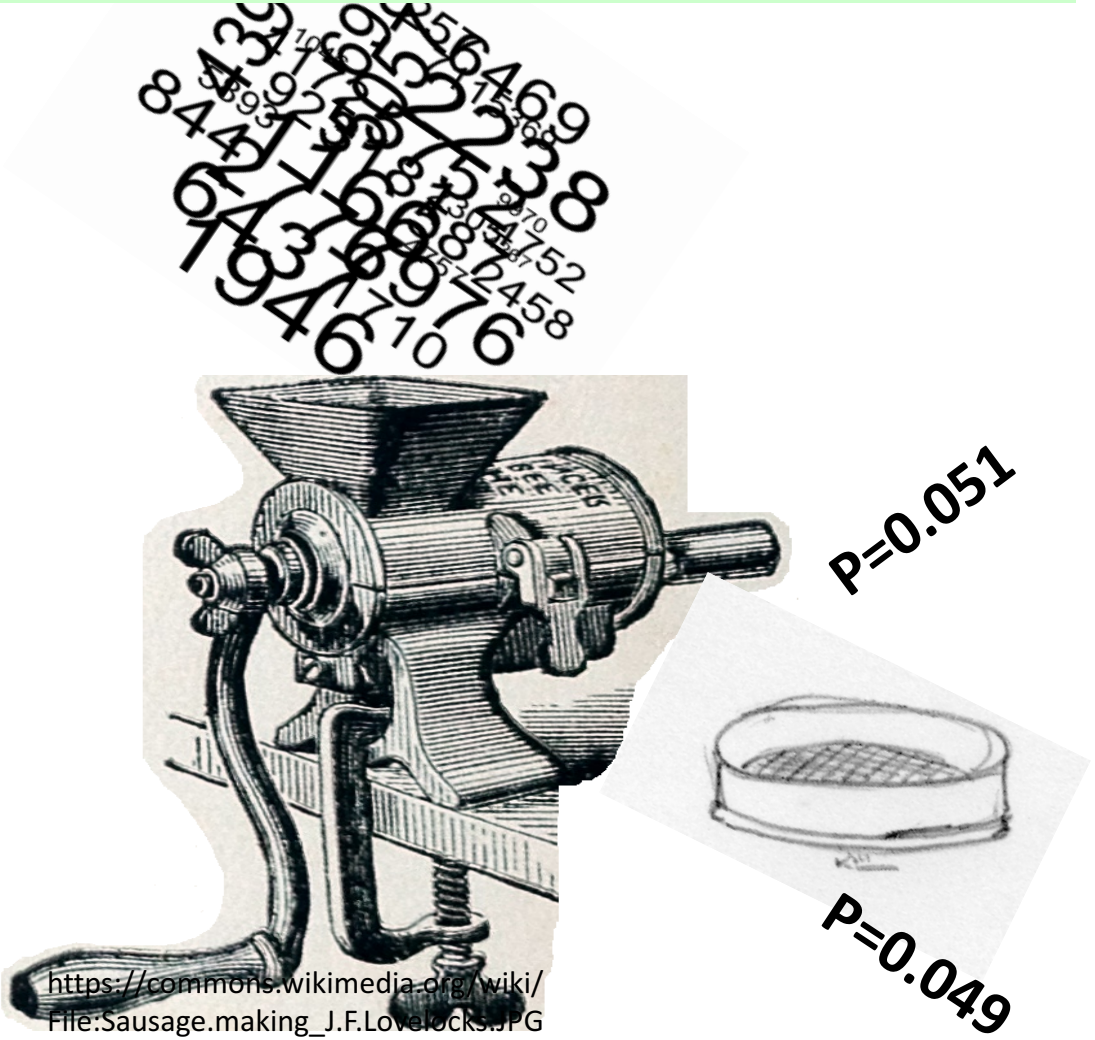
How?

- Oddly, there was substantial agreement that properly used, p-values are an important part of the applied statistical toolkit

The ASA Statement on P-values

How?

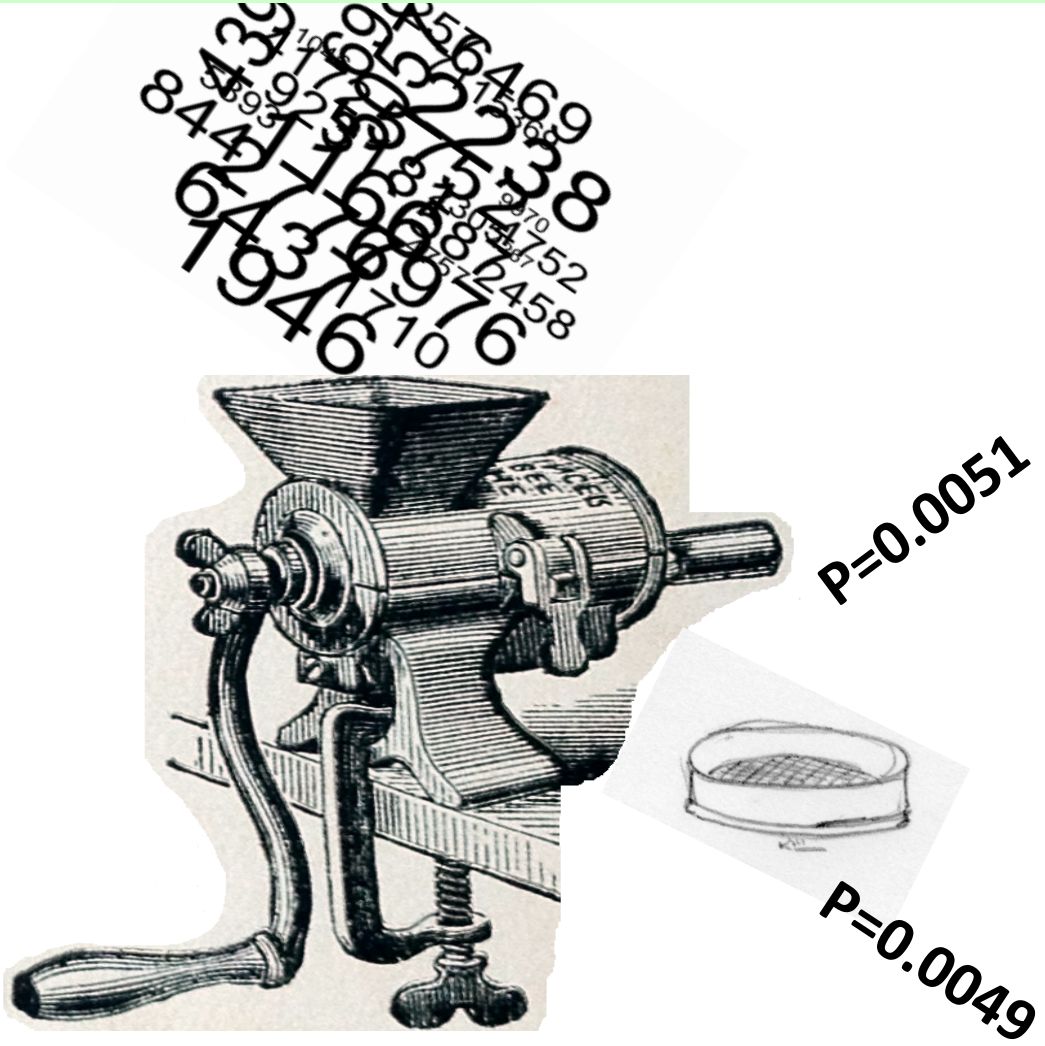
- Oddly, there was substantial agreement that properly used, p-values are an important part of the applied statistical toolkit



The ASA Statement on P-values

How?

- Oddly, there was substantial agreement that properly used, p-values are an important part of the applied statistical toolkit



Summary of the ASA Statement on P-values

1. P-values measure the compatibility between an observed sample and a given model (the null hypothesis).
2. P-values do not measure the probability that the null or alternative hypothesis is true or that the sample result could be obtained by chance.
3. Decisions should not be based on "bright line" cut-offs between $p<=\alpha$, $p> \alpha$
4. Proper inference requires full disclosure and transparency.
5. P-values do not measure effect size or practical importance.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. P-values do not measure the probability that the null or alternative hypothesis is true or that the sample result could be obtained by chance.
3. Decisions should not be based on “bright line” cut-offs between $p<=\alpha$, $p>\alpha$
4. Proper inference requires full disclosure and transparency.
5. P-values do not measure effect size or practical importance.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. We have to do a better job of teaching and explaining p-values. Bunnies explain p-values
3. Decisions should not be based on "bright line" cut-offs between $p<=\alpha$, $p>\alpha$
4. Proper inference requires full disclosure and transparency.
5. P-values do not measure effect size or practical importance.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. We have to do a better job of teaching and explaining p-values. Bunnies explain p-values
3. P-values are random variables. (So are BF if you think in terms of sampling.)
4. Proper inference requires full disclosure and transparency.
5. P-values do not measure effect size or practical importance.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. We have to do a better job of teaching and explaining p-values. Bunnies explain p-values
3. P-values are random variables. (So are BF if you think in terms of sampling.)
4. Not only shouldn't we p-hack, we should consider how we chose our analysis method (Gelman).
5. P-values do not measure effect size or practical importance.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. We have to do a better job of teaching and explaining p-values. Bunnies explain p-values
3. P-values are random variables. (So are BF if you think in terms of sampling.)
4. Not only shouldn't we p-hack, we should consider how we chose our analysis method (Gelman).
5. We should use interval estimates that emphasize effect size and variability when available.
6. Other supporting evidence is needed to give a fuller picture than that provided by the p-value.

Summary of the ASA Statement on P-values

1. Assumptions such as independence affect p-values but are not often explicitly part of the null hypothesis.
2. We have to do a better job of teaching and explaining p-values. Bunnies explain p-values
3. P-values are random variables. (So are BF if you think in terms of sampling.)
4. Not only shouldn't we p-hack, we should consider how we chose our analysis method (Gelman).
5. We should use interval estimates that emphasize effect size and variability when available.
6. We need to supplement p-values in various ways including evidential and expert opinion.

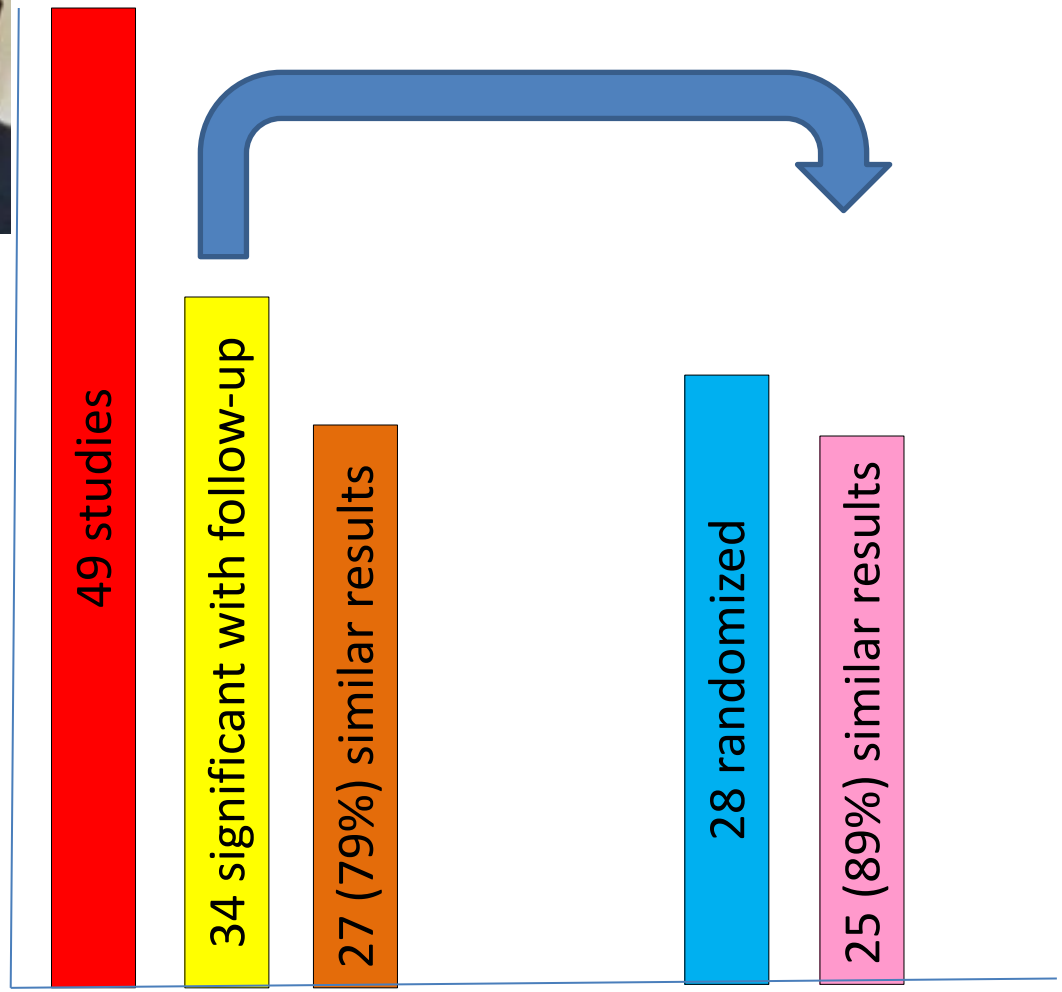
P-values and Reproducibility

- Reproducibility is a misunderstood goal.
- Usually people only look at studies with statistically significant results.

- Suppose all the tests were done honestly at size α and power 1-β.
- We expect α or (1-β) to reproduce depending on whether or not H_0 is true and assuming the second study used the same sampling population, methods, etc.

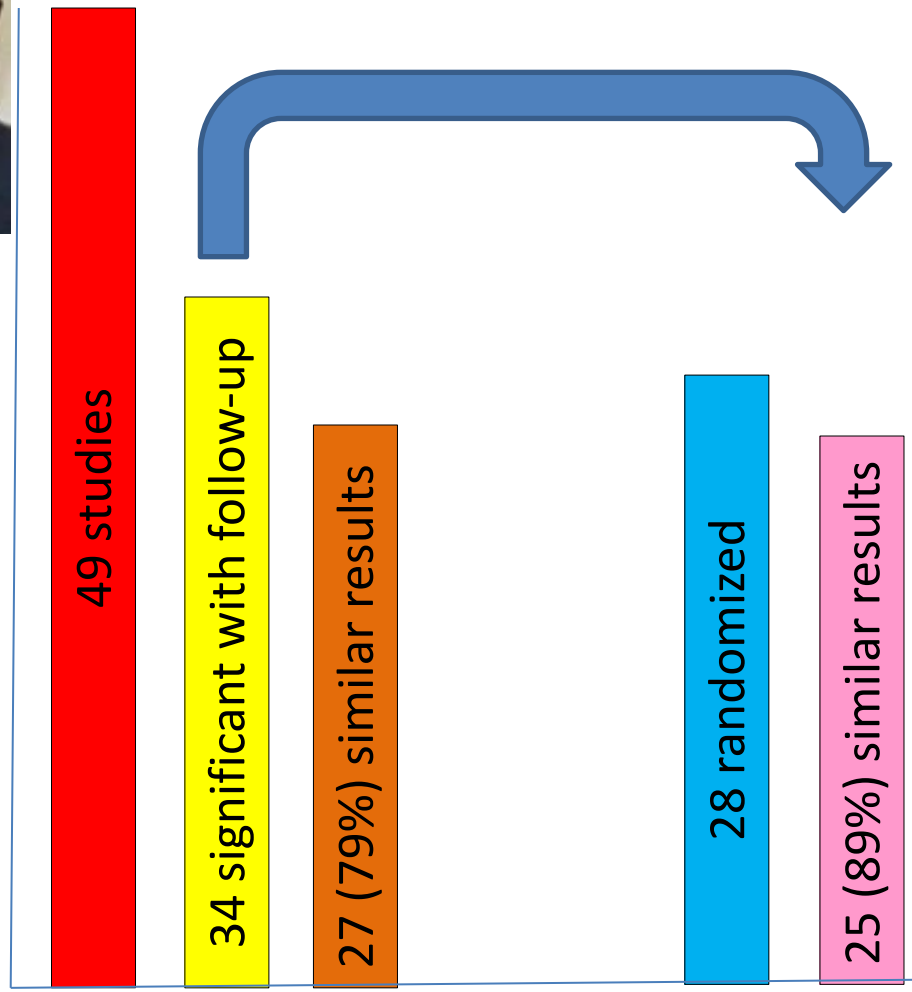
P-values and Reproducibility

- Ioannidis 2005 *JAMA*: "Contradicted and initially stronger effects in highly cited clinical research."



P-values and Reproducibility

- Ioannidis 2005 *JAMA*: "Contradicted and initially stronger effects in highly cited clinical research."



I feel Ioannidis overstated the case.

The problem is not reproducibility but the assumption that significant means "real" or failure to replicate means "wrong".

Quantifying Uncertainty of p

- p-values are statistics: numerical summaries computed from a sample
- They are not Fisher consistent – i.e. they are not functionals of the empirical CDF that can be related to a functional of the actual CDF
- For this reason we cannot compute e.g. a CI for a p-value
- However, they have a sampling distribution so we can consider interval estimates based on the sampling distribution

Quantifying Uncertainty of p

Continuous test statistic

Under H_0:

- p~U(0,1) regardless of the sample size.
- A 95% PI is any subinterval of [0,1] of length 0.95.
- An “optimistic” interval is [0,0.95].

Quantifying Uncertainty of p

Continuous test statistic

Under H_A:

The interval should get smaller with the sample size.

e.g. Suppose our data are N(μ,1) and we are testing H_0: μ=0.

To obtain a (1-α) CI for μ, from data $x_1 \ldots x_n$ we would use

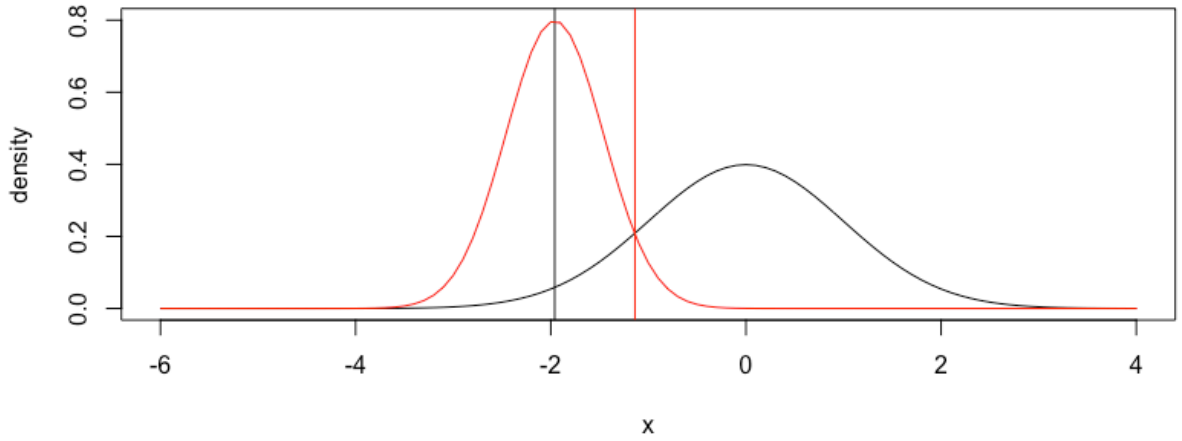
$$\bar{x} \pm z_{\alpha/2}/\sqrt{n}$$

In this spirit, we could obtain a τ interval estimate for p

Quantifying Uncertainty of p

Continuous test statistic

Under H_A:



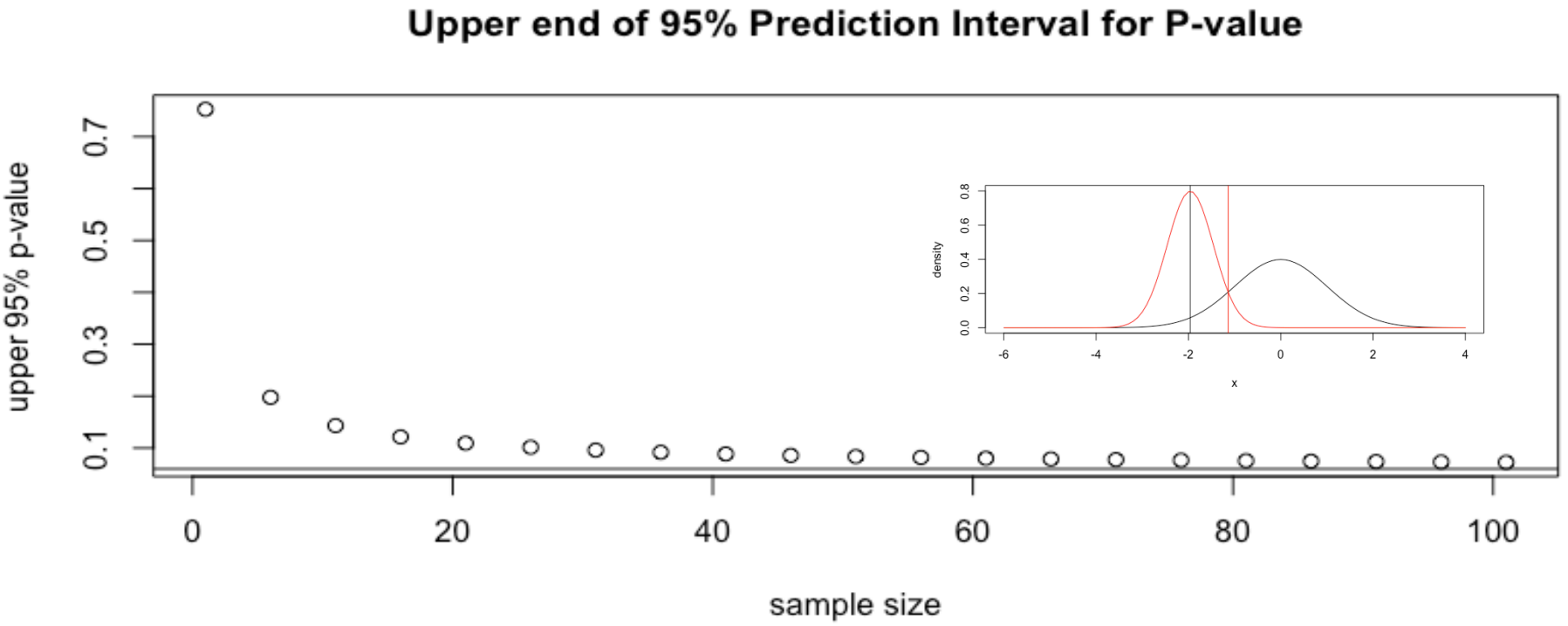
e.g. Suppose $\bar{x} < 0$.

The τ interval with the smallest upper bound corresponds to the p-value corresponding to the τ quantile of N($\bar{x}$,1/$\sqrt{n}$).

Quantifying Uncertainty of p

Continuous test statistic

Under H_A: e.g. Suppose we observe $\bar{x}$=-1.96

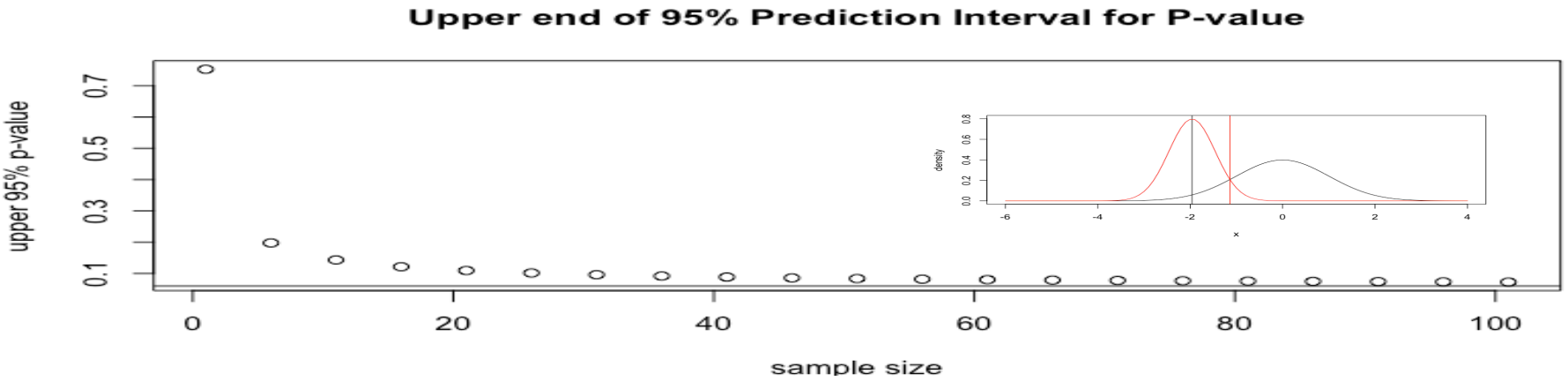


Quantifying Uncertainty of p

Continuous test statistic

However, we cannot actually present a figure like this as an "interval estimate" for the p-value, because it has correct coverage only if the alternative is true (and has effect size at least as large as the observed effect size).

If the null is true, then the smallest upper bound for the interval estimate is 0.95 no matter what the sample size.

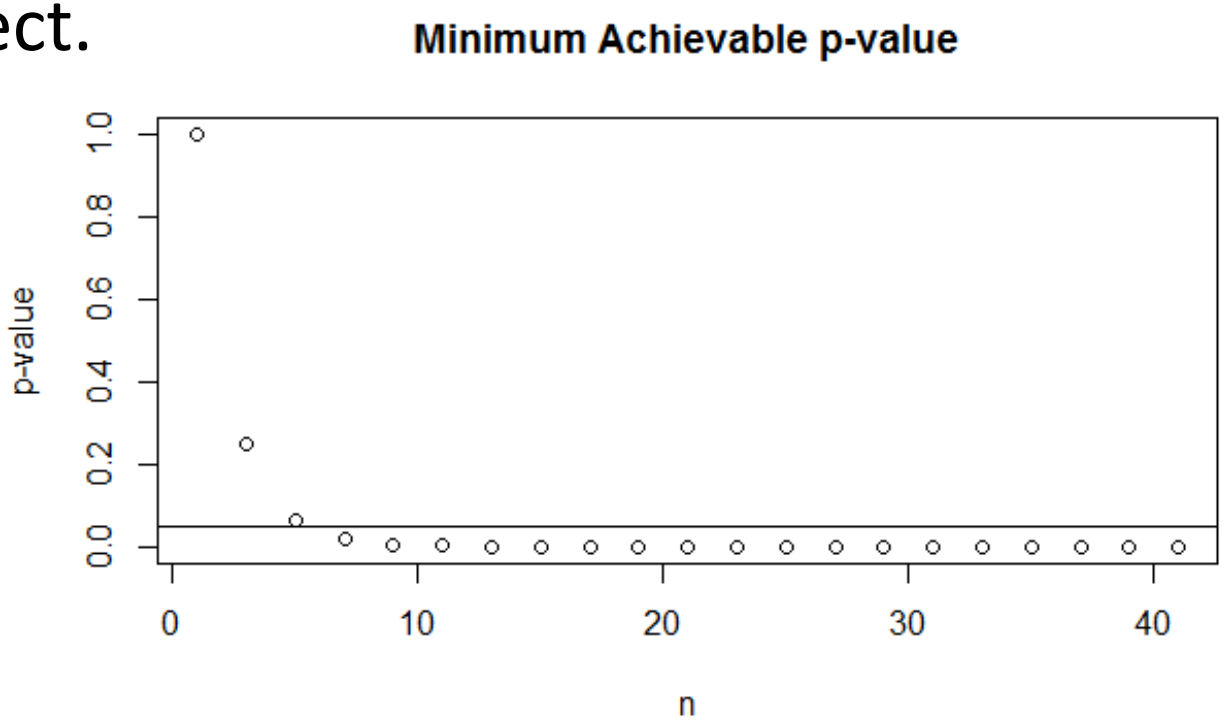


Quantifying Uncertainty of p

Discrete test statistic

Things get weirder for discrete tests. Lets start with the Binomial test and test π=0.5.

There is a smallest achievable p-value so we may never reject.

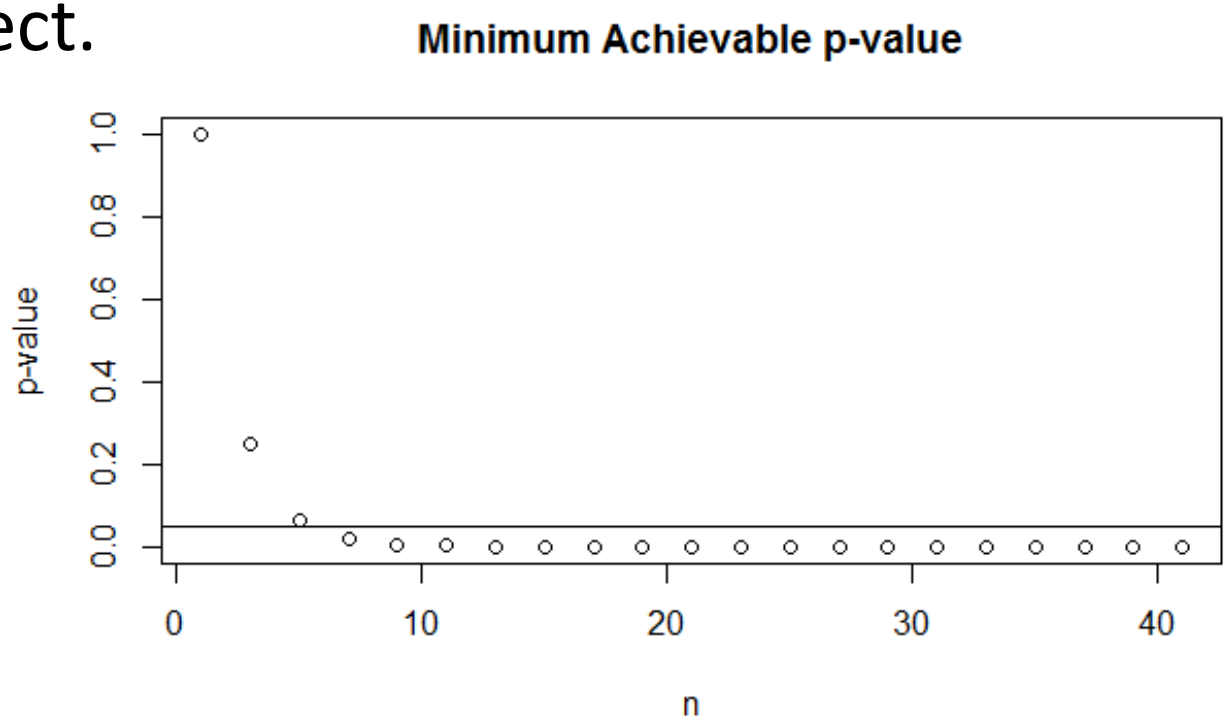


Quantifying Uncertainty of p

Discrete test statistic

Things get weirder for discrete tests. Lets start with the Binomial test and test π=0.5.

There is a smallest achievable p-value so we may never reject.



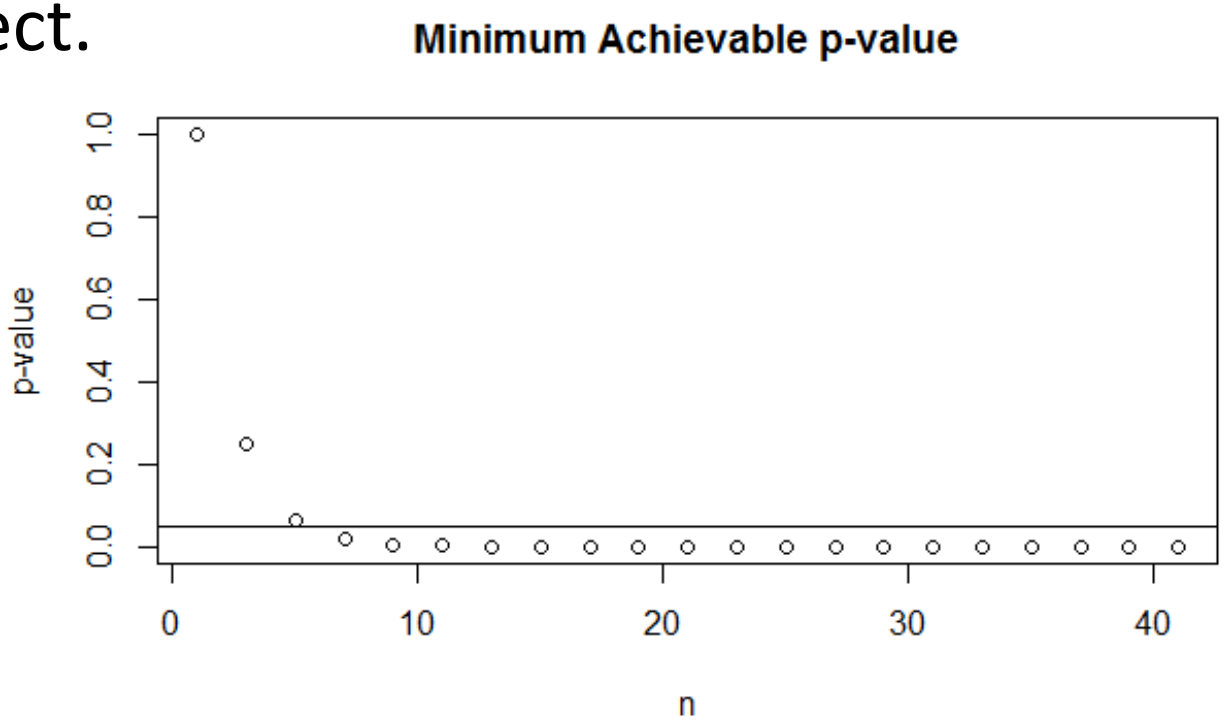
Now consider multiple testing adjustments

Quantifying Uncertainty of p

Discrete test statistic

Things get weirder for discrete tests. Lets start with the Binomial test and test π=0.5.

There is a smallest achievable p-value so we may never reject.



Now consider multiple testing adjustments.

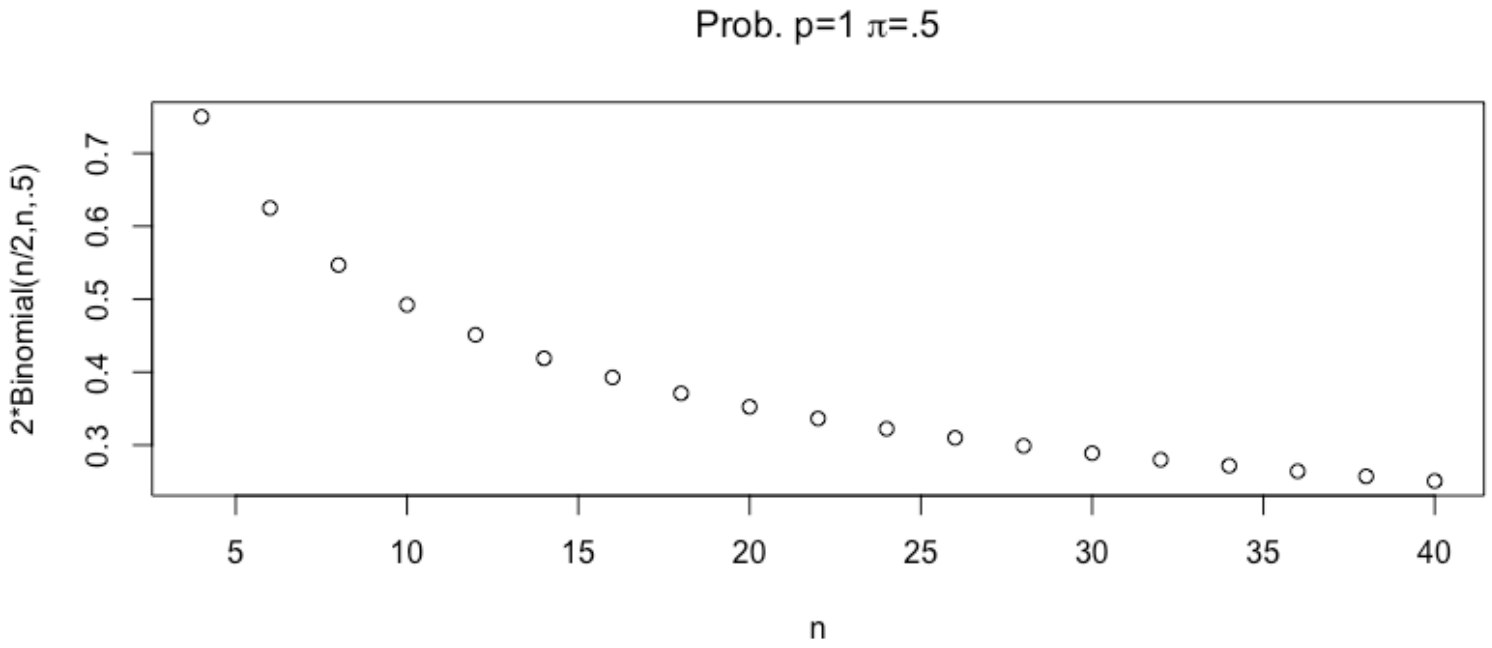
I have seen even experienced statisticians opt to use a discrete test and then accept all the Nulls.

Quantifying Uncertainty of p

Discrete test statistic

Things get weirder for discrete tests. Lets start with the Binomial test and test π=0.5.

There is a large probability under the null that the p-value is 1.

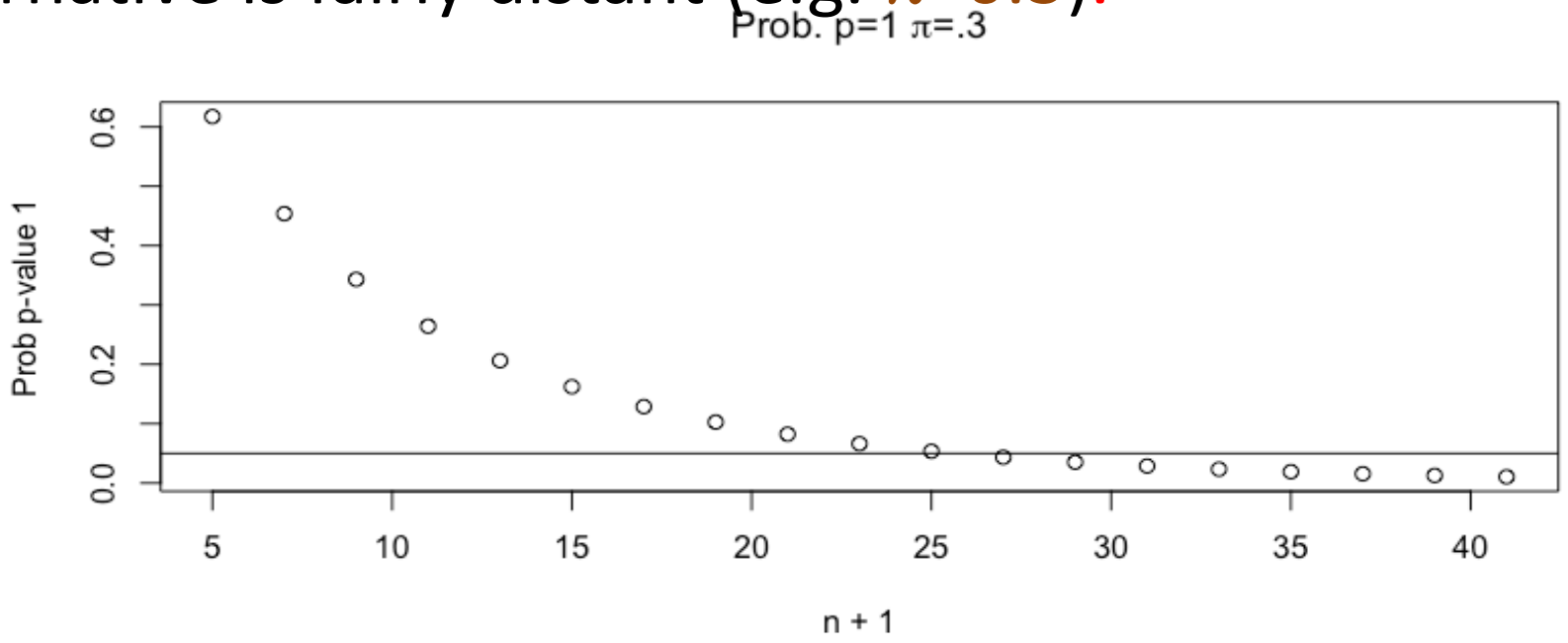


Quantifying Uncertainty of p

Discrete test statistic

Things get weirder for discrete tests. Lets start with the Binomial test and test π=0.5.

But the probability p=1 remains large even when the alternative is fairly distant (e.g. π=0.3).



Quantifying Uncertainty of p

Conclusions:

- P-values are random variables.
- We should emphasize the randomness when discussing p-values, possibly by simulated examples, etc.
- There is no easy way to summarize the variability of p-values; stick to effect size variability

What is Better than p<0.05?

Some ideas:

- Confidence Intervals for Effect (Just about everyone)
- False Positive Proportion (Altman 2016)
- Bayes Factor (Bayarri et al 2016)

(Note:

I am a pragmatist with frequentist leanings)

What is Better than p<0.05?

Confidence Intervals

Pro:

1) Incorporates effect size and uncertainty

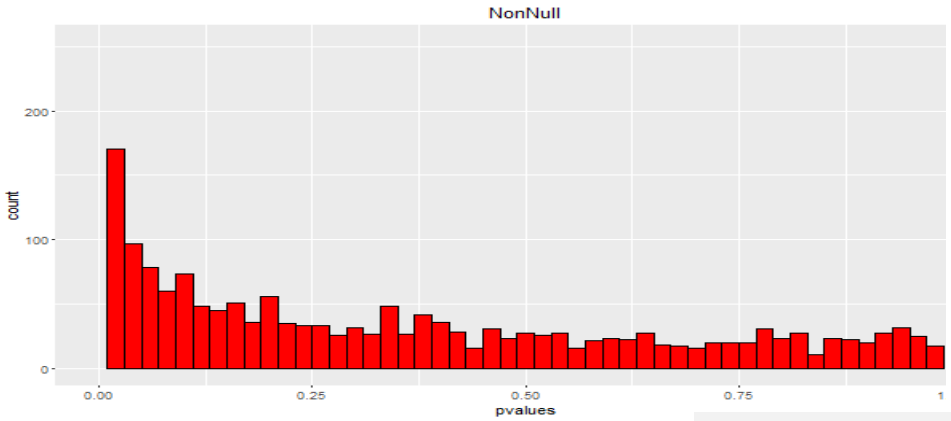
Con:

1)P-hacking leads to intervals too far from the null and too narrow

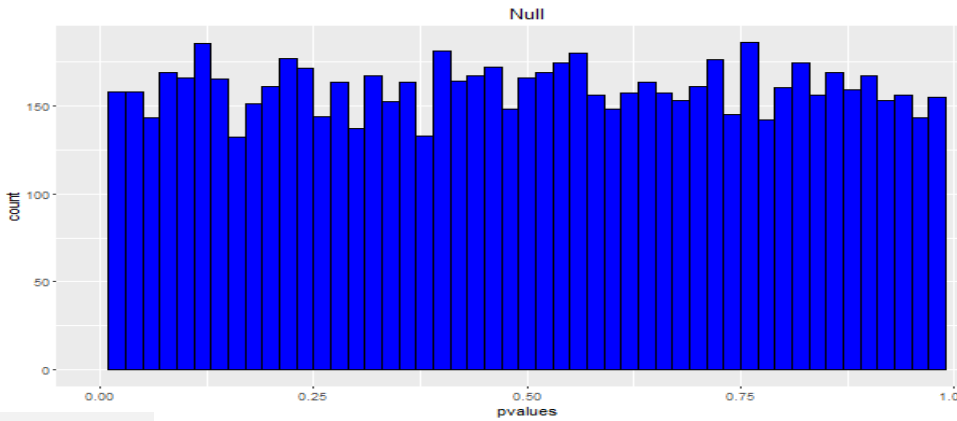
2) Frequentist – width decreases even if bias does not

Lessons from Highly Multiple Testing

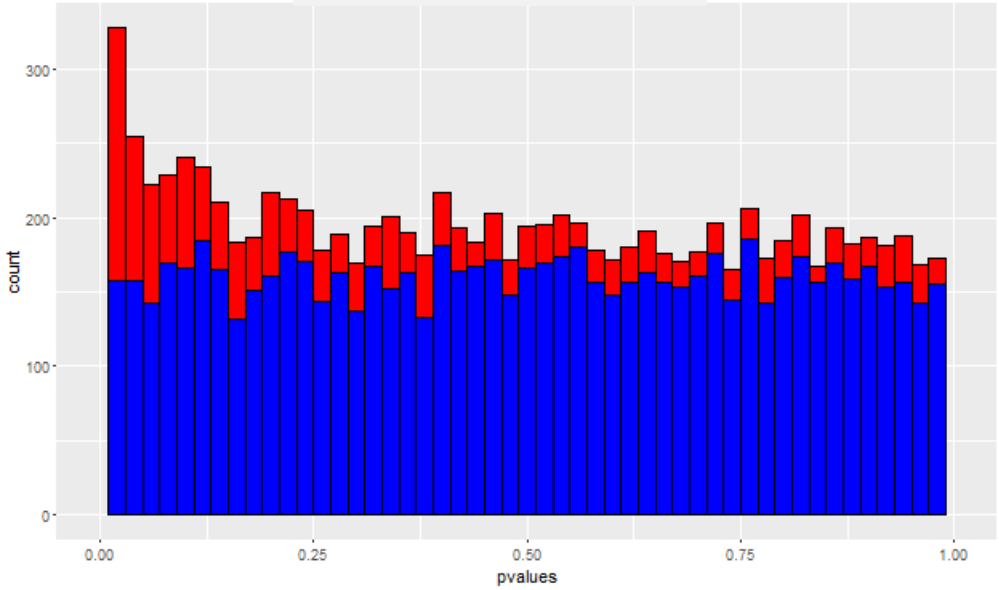
P-values from non-Null



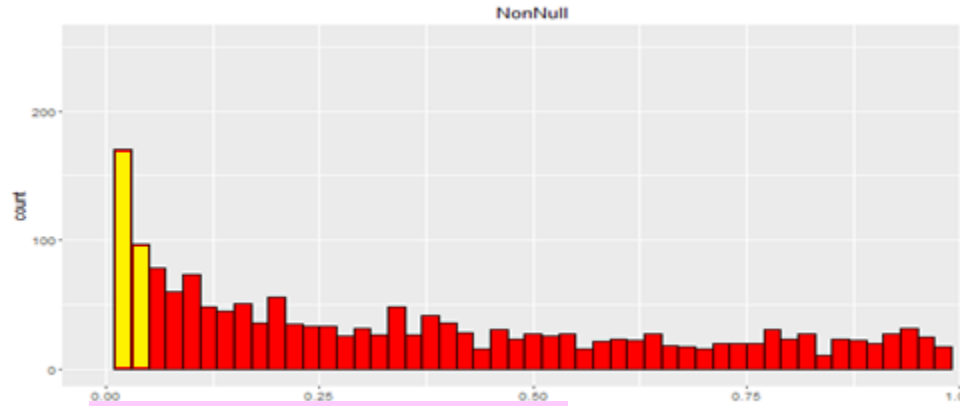
P-values from Null



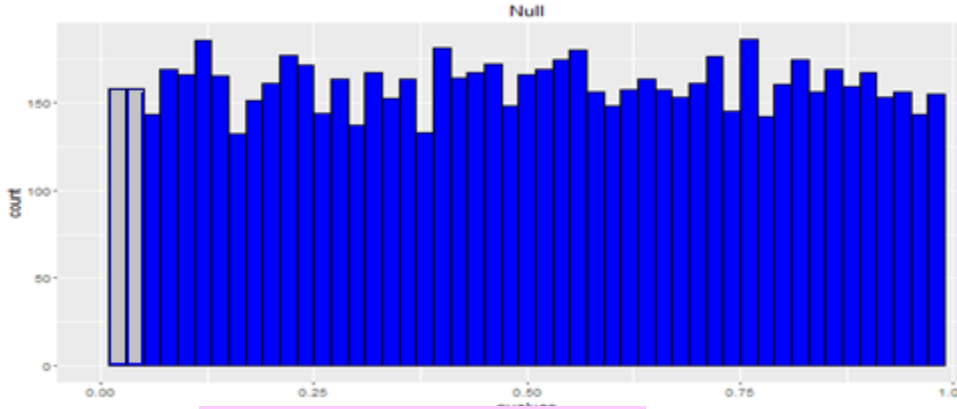
Observed P-values



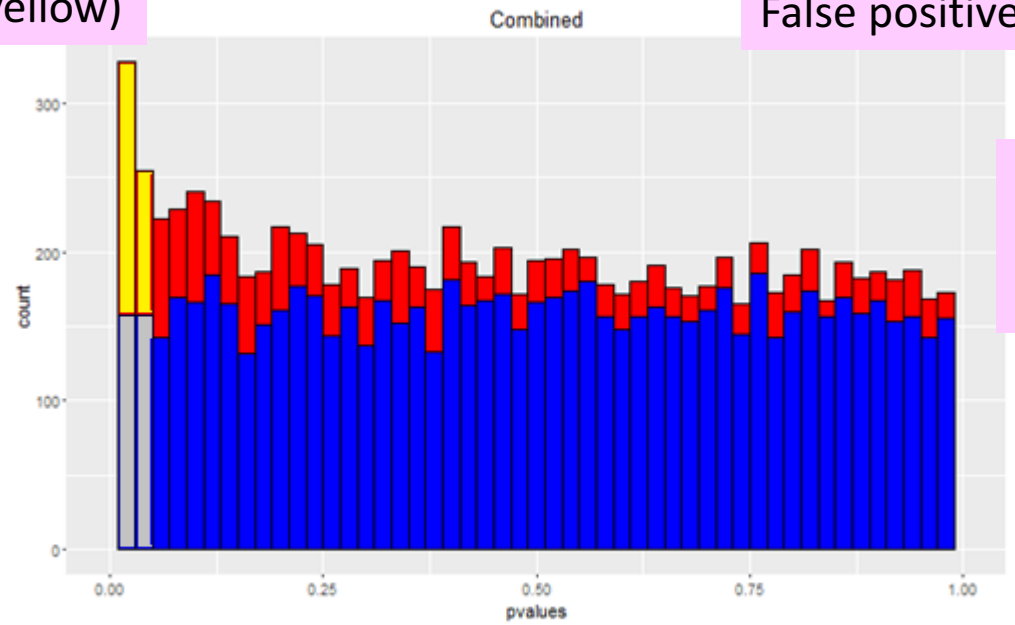
Lessons from Highly Multiple Testing



True positives (yellow)



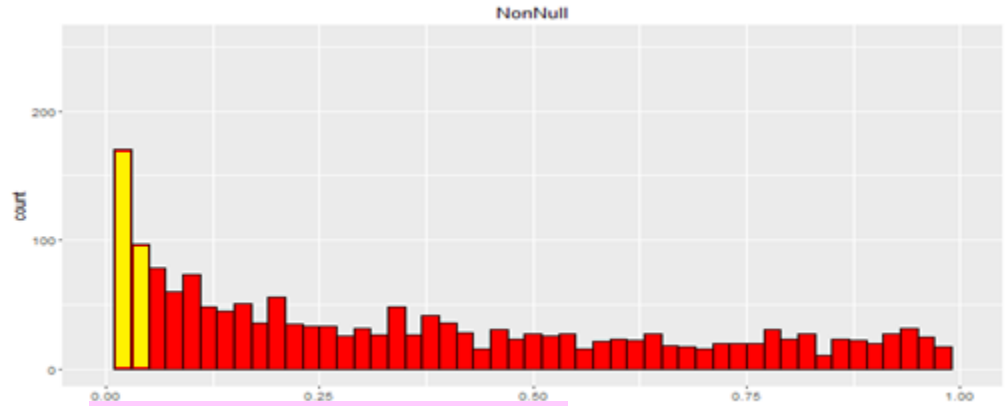
False positives (grey)



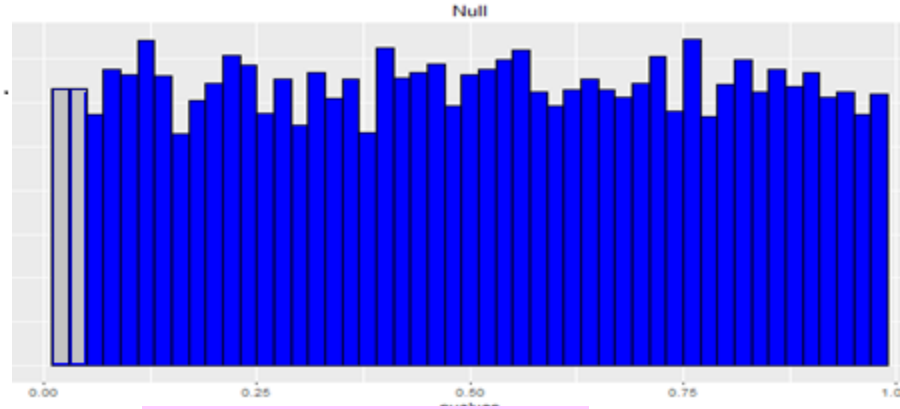
False positive proportion

grey/(grey+yellow)

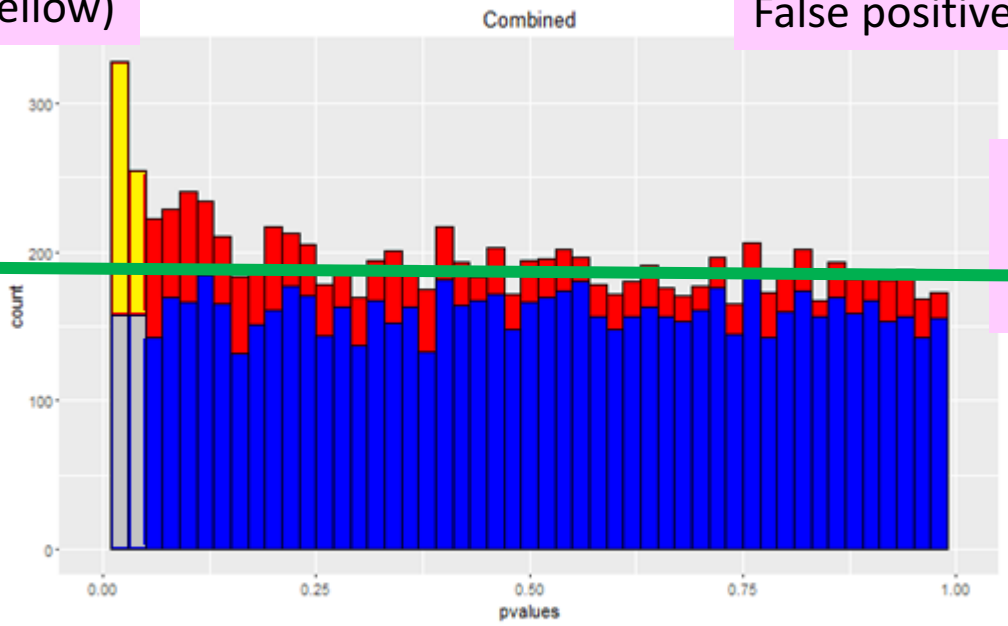
Lessons from Highly Multiple Testing



True positives (yellow)



False positives (grey)



False positive proportion

grey/(grey+yellow)

Estimated π_0

(True is 80%)

Lessons from Highly Multiple Testing

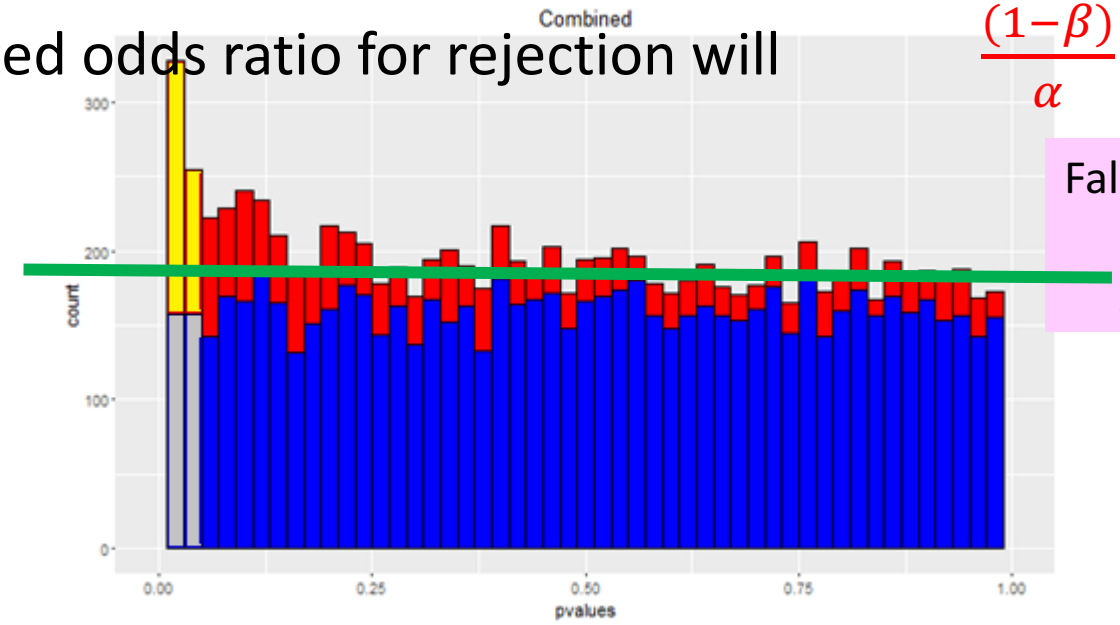
Suppose we know π_0 and reject at some level α with power $(1-\beta)$

The expected false positives will be $\alpha\pi_0$

The expected true positives will be $(1-\beta)(1-\pi_0)$

The estimated false positive proportion will be $\frac{\alpha\pi_0}{\alpha\pi_0+(1-\beta)(1-\pi_0)}$

The estimated odds ratio for rejection will $\frac{(1-\beta)}{\alpha}$



False positive proportion

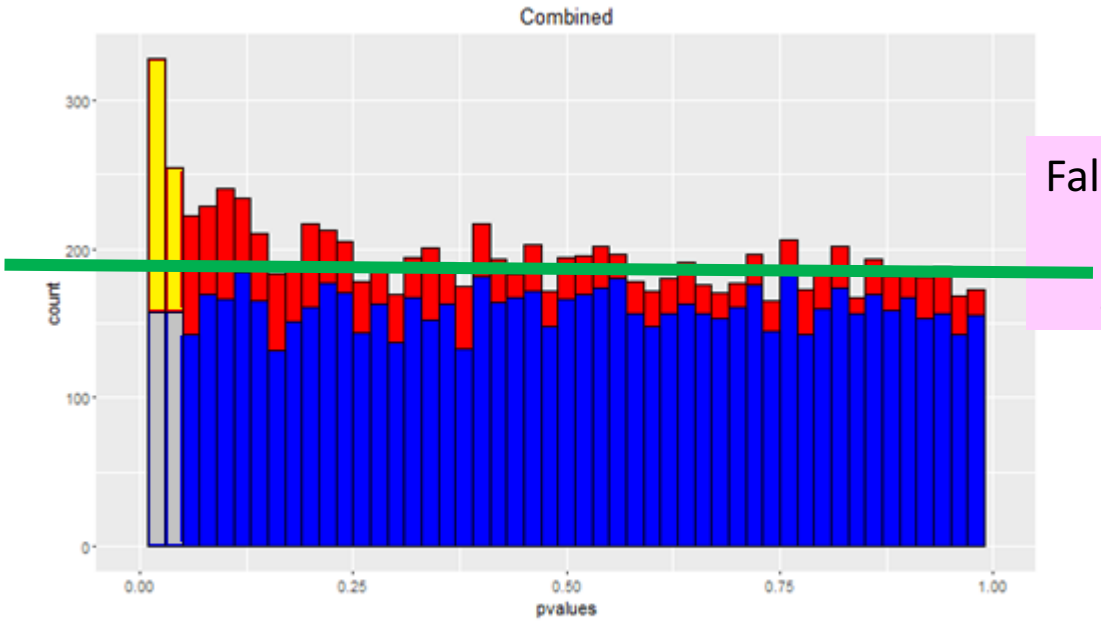
grey/(grey+yellow)

Lessons from Highly Multiple Testing

The estimated false positive proportion will be $\frac{\alpha\pi_0}{\alpha\pi_0+(1-\beta)(1-\pi_0)}$

e.g. π_0=50% and α=.05 and (1-β)=.8

The estimated FPP is 4.25%



False positive proportion

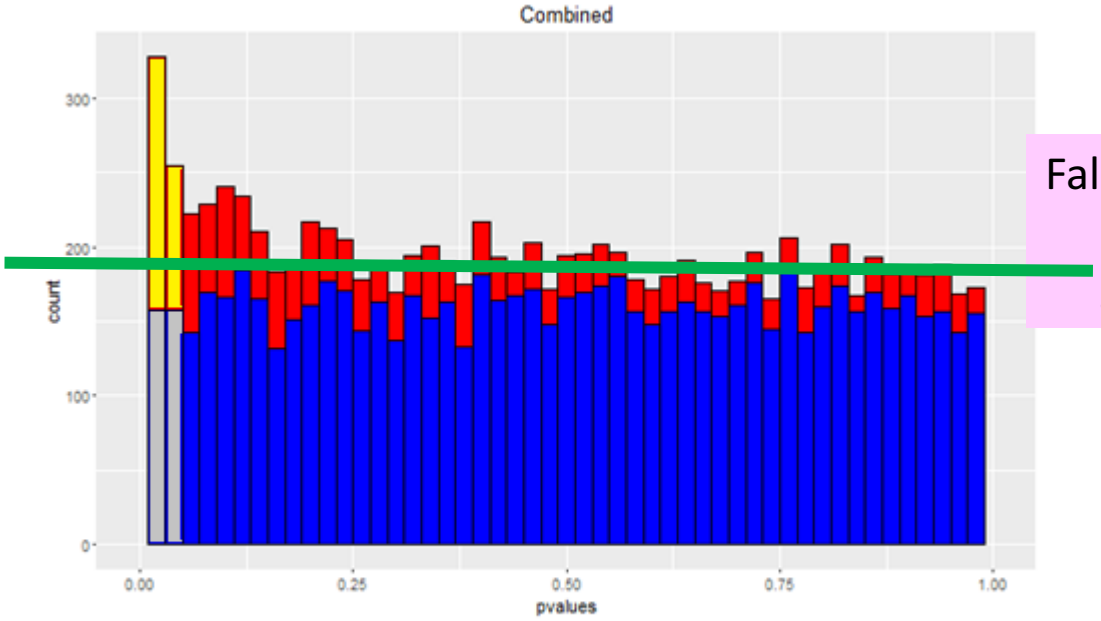
grey/(grey+yellow)

Lessons from Highly Multiple Testing

The estimated false positive proportion will be $\frac{\alpha\pi_0}{\alpha\pi_0+(1-\beta)(1-\pi_0)}$

e.g. π_0=95% and α=.05 and (1-β)=.8

The estimated FPP is 54%



False positive proportion

grey/(grey+yellow)

Lessons from Highly Multiple Testing

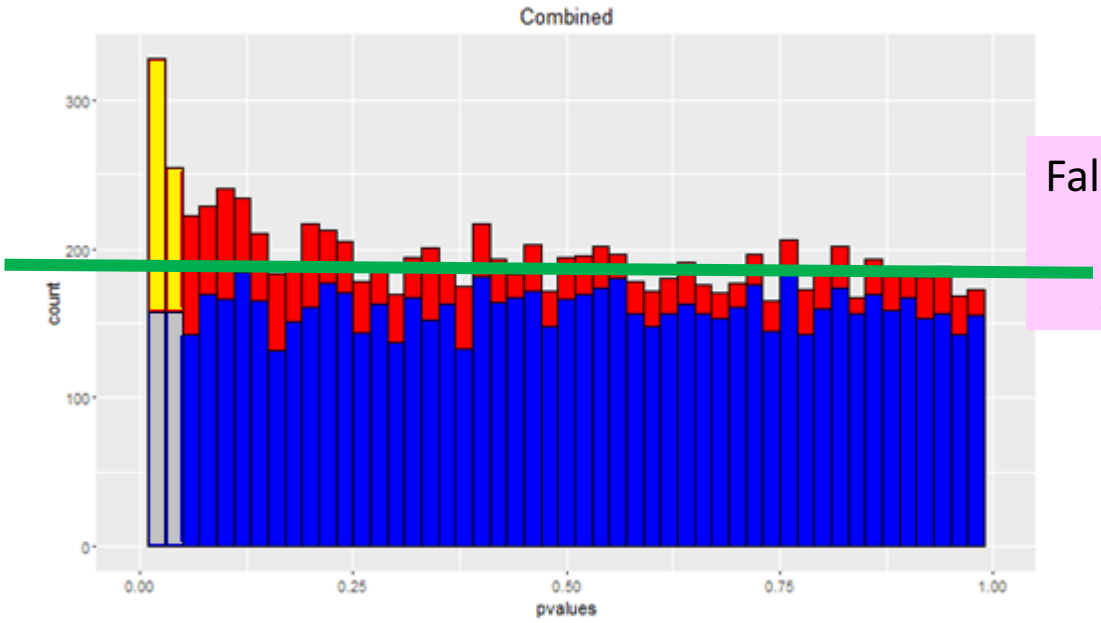
The estimated false positive proportion will be $\frac{\alpha\pi_0}{\alpha\pi_0+(1-\beta)(1-\pi_0)}$

e.g. π_0=95% and α=.05 and (1-β)=.6

The estimated FPP is 61%

Estimated π_0

(True is 80%)



False positive proportion

grey/(grey+yellow)

Using the FPP

The estimated false positive proportion will be $\frac{\alpha\pi_0}{\alpha\pi_0+\beta(1-\pi_0)}$

How can we apply this to studies that do not have high multiplicity?

We should be able to estimate everything but π_0

So lets apply some rules of thumb.

Using the FPP

Some rules of thumb for π_0

Well supported hypotheses with preliminary data and literature	50%
Fortuitous findings	95%
Findings after model selection	99%

Using the FPP

Some rules of thumb for π_0

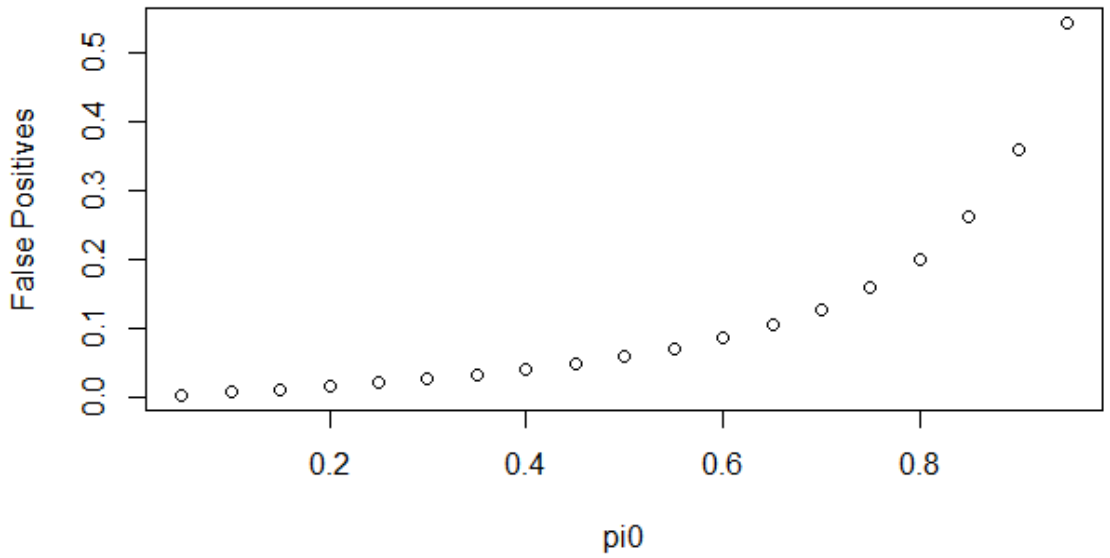
Well supported hypotheses with preliminary data and literature	50%
Fortuitous findings	95%
Findings after model selection	99%

These should be modified by field and expert analysis of the literature –

e.g. in parapsychology I might prefer 95%, 99% and 99.9% .

Using the FPP

FPP as a Function of π_0, $\alpha=0.05$, $(1-\beta)=.8$

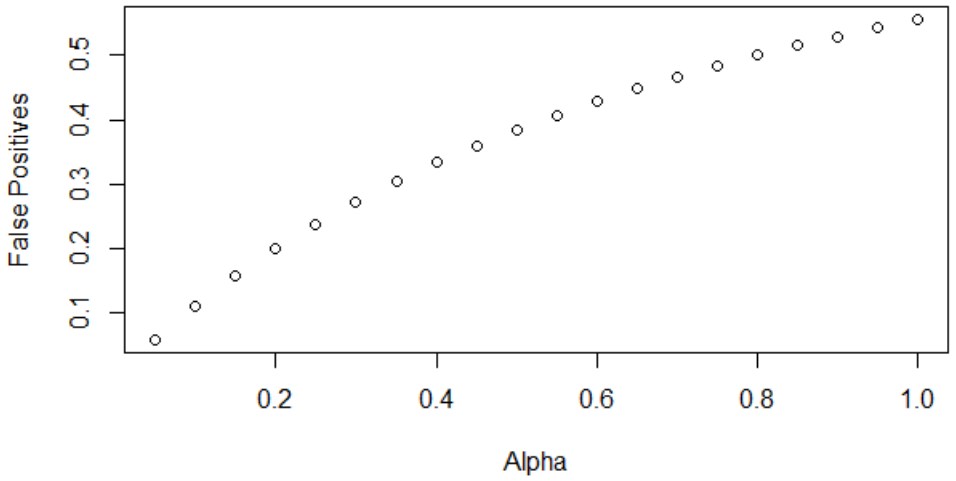


Lesson:

Findings based on testing hypotheses with solid prior support are more likely to be true

Using the FPP

FPP as a Function of α, $\pi_0=0.5$, $(1-\beta)=.8$

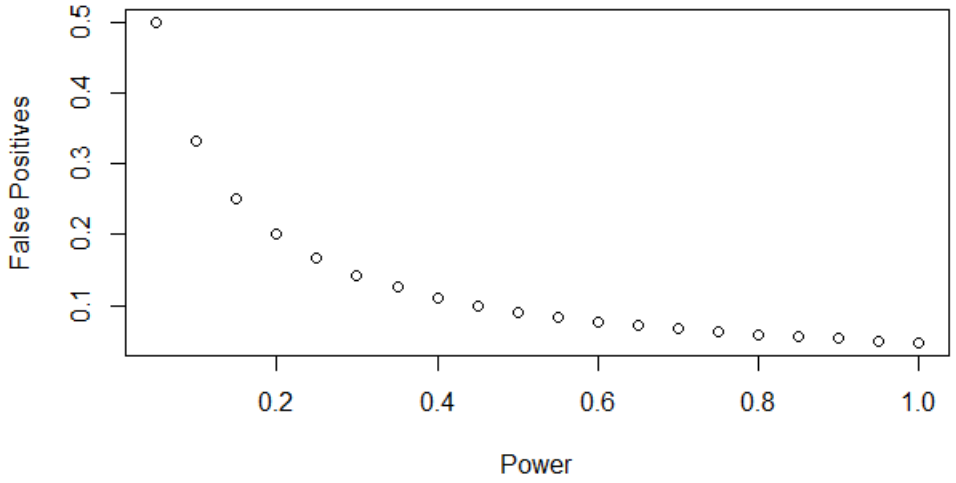


Lesson:

If you can hold power fixed, then findings with smaller p-value are more likely to be true.

Using the FPP

FPP as a Function of $(1-\beta)$, $\alpha=.05$, $\pi_0=0.5$



Lesson:

When there is more power, findings are more likely to be true.

Using the FPP

Applying the rule of thumb is not that easy.

e.g. In a study of the effects of diet and exercise on blood chemistry, several (correlated) blood chemistry variables were measured.

The p-value for diet was about 0.06 for several of the variables, with effect size in the expected direction.

Should we do a multivariate ANOVA and if so, how should we adjust for having already done the univariate analyses?

What is Better than p<0.05?

Pre- and post-experimental odds Bayarri et al (2016)

They replace the type II error rate β by the average (over the prior) $\bar{\beta}$

The pre-experimental rejection odds are $\frac{(1-\bar{\beta})(1-\pi_0)}{\alpha\pi_0}$.

The pre-experimental rejection ratio is $\frac{(1-\bar{\beta})}{\alpha}$.

What is Better than p<0.05?

Pre- and post-experimental odds Bayarri et al (2016)

They replace the type II error rate β by the average (over the prior) $\bar{\beta}$

The pre-experimental rejection odds are $\frac{(1-\bar{\beta})(1-\pi_0)}{\alpha\pi_0}$.

The pre-experimental rejection ratio is $\frac{(1-\bar{\beta})}{\alpha}$.

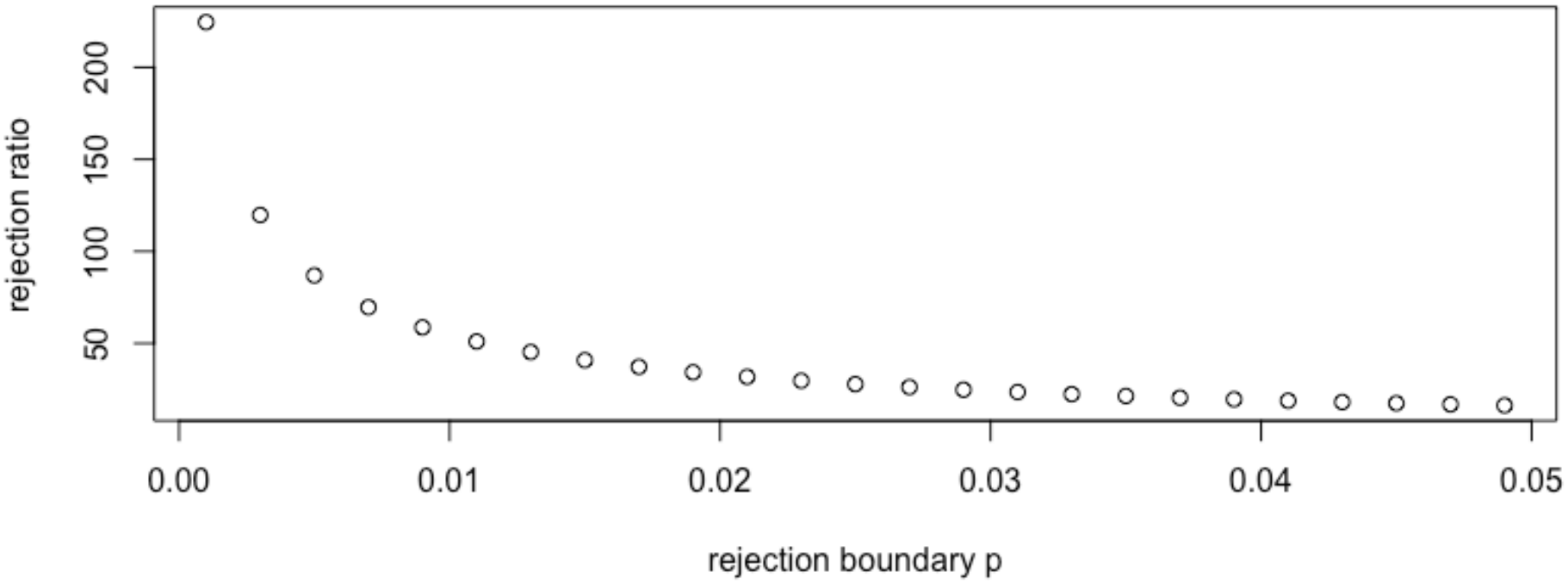
They note that the usual rule of thumb is 80/5=16.

Studies with a lower pre-experimental rejection ratio need a very small value of π_0 otherwise even a p-value less than a very small α is not much evidence against the null.

What is Better than p<0.05?

The pre-experimental rejection ratio:

power/size



What is Better than p<0.05?

Pre- and post-experimental rejection odds (BF) Bayarri et al (2016)

The post-experimental rejection odds is generally referred to as the Bayes Factor

$$\text{BF}=\frac{average\ likelihood\ of\ data\ over\ H_A}{average\ likelihood\ of\ data\ over\ H_0}$$

The post-experimental odds are BF $\frac{1-\pi_0}{\pi_0}$

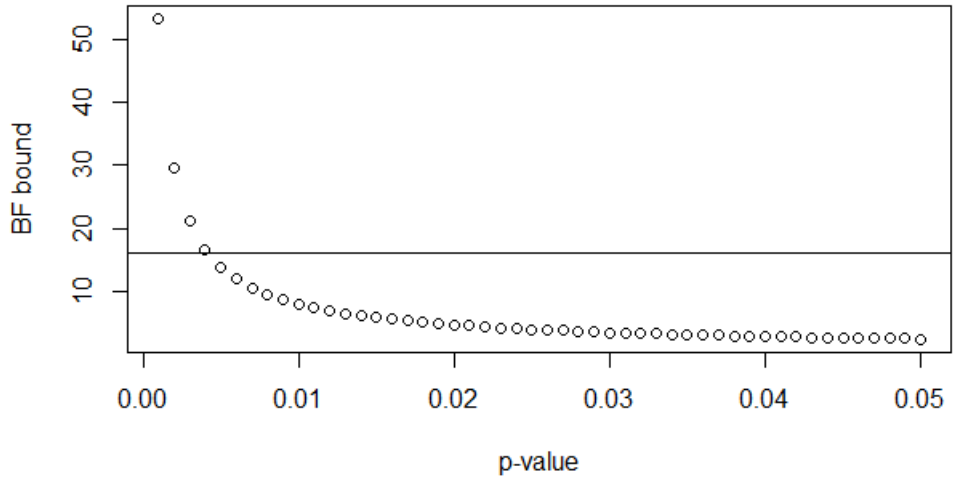
One way to assess the post-experimental odds is to use the BF with the prior most favorable to the alternative.

What is Better than p<0.05?

Bayarri et al (2016) note that for p≤1/e

$$BF \leq 1/[-ep \log(p)]$$

under quite general conditions (for 2-sided tests), so some of the computations can be done using ONLY the p-value

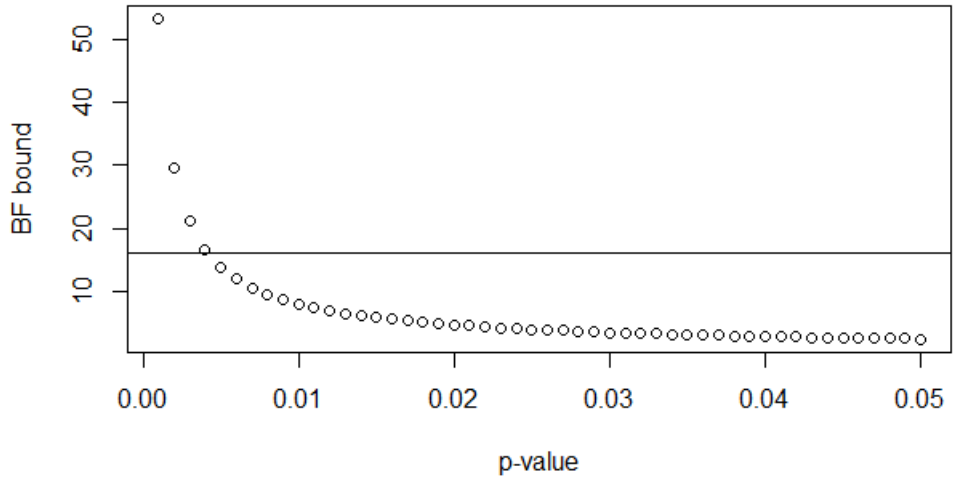


What is Better than p<0.05?

Bayarri et al (2016) note that for p≤1/e

$$BF \leq 1/[-ep \log(p)]$$

under quite general conditions (for 2-sided tests), so some of the computations can be done using ONLY the p-value



This has been cited as an argument for using much smaller α (0.005).

(V. Johnson, 2013).

However, I worry about the loss of power.

Conclusions

P-values are useful, but need to be supplemented by measures that:

- Quantify sampling variability
- Quantify effect size
- Quantify evidence in favor of a hypothesis

There are measures that can be framed to be acceptable to most frequentist and Bayesian statisticians.

Many thanks to:

The ASA Committee on P-values

(The opinions expressed here are entirely my own)

References:

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

M. J. Bayarri et al (2016) Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. of Mathematical Psychology* 72, 90-103, DOI: 10.1016/j.jmp.2015.12.007

Johnson, Valen E. (2013) "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences* 110.48 : 19313-19317.