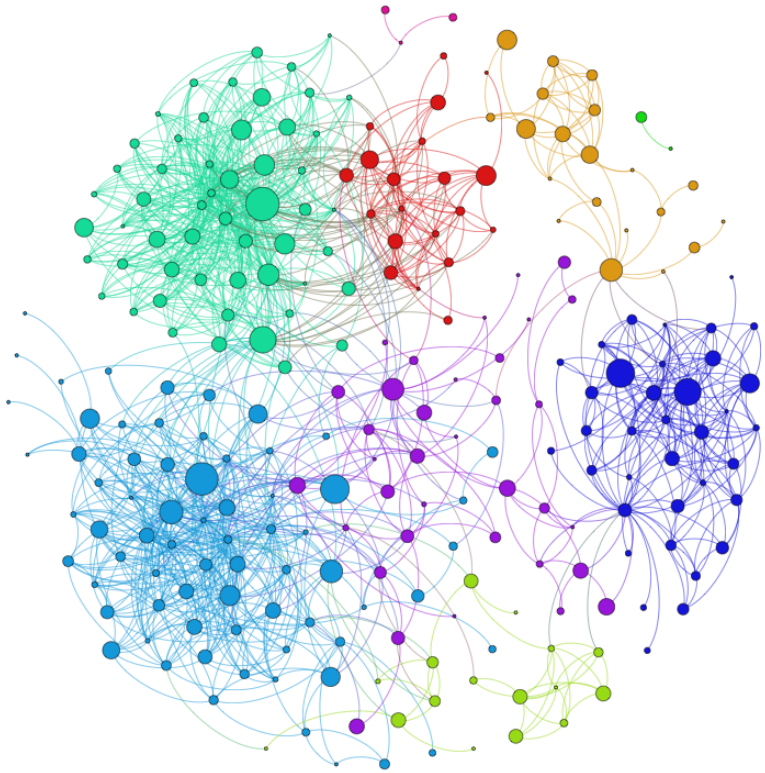
The Truth About P-Values

Jiaming Mao

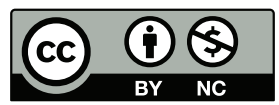
Xiamen University



This version: Spring 2019

Contact: jmao@xmu.edu.cn

Course homepage: jiamingmao.github.io/data-analysis



P-Value

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

– Wasserstein and Lazar (2016)

P-Value

- There has been a growing concern about issues of reproducibility and replicability of scientific conclusions.
- Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with the $p-$value.
- While the $p-$value can be a useful statistical measure, it is commonly misused and misinterpreted.

> *"Statistically speaking, science suffers from an excess of significance." – Hotz (2007)*

Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False," *PLoS Medicine*, 2(8).

Abstract

Go to:

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

The Statistical Crisis in Science

Andrew Gelman
Columbia University, New York, USA

Top journals in psychology routinely publish ridiculous, scientifically implausible claims, justified based on “p < 0.05.” And this in turn calls into question all sorts of more plausible, but not necessarily true, claims, that are supported by this same sort of evidence. To put it another way: we can all laugh at studies of ESP, or ovulation and voting, but what about MRI studies of political attitudes, or embodied cognition, or stereotype threat, or, for that matter, the latest potential cancer cure? If we can’t trust p-values, does experimental science involving human variation just have to start over? And what to we do in fields such as political science and economics, where preregistered replication can be difficult or impossible? Can Bayesian inference supply a solution? Maybe. These are not easy problems, but they’re important problems.

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

The Role of Statistical Inference

- Consider a hypothetical population described by random variables (x, y). Suppose we fit a simple linear model onto the population:

$$y = \beta x + e \tag{1}$$

  Fitting (1) $\Rightarrow \beta^*$.

- Now suppose the data we observe, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, is a random sample drawn from the population. Fitting (1) onto the observed data gives us $\widehat{\beta}$.

The Role of Statistical Inference

The goal of statistical inference is to form statements about β^* when we can only obtain $\widehat{\beta}$.

- In other words, statistical inference deals with the problem of uncertainty in our estimates due to sample variability.
- It does *not* deal with the problems of
  - whether (1) is a good model for predicting y based on x;
  - whether x has any causal effect on y and whether (1) is a good model for describing the causal effect (if exists).

Hypothesis Testing: What it is and What it is not

$\mathbb{H}_0 : \beta^* = 0$ vs. $\mathbb{H}_1 : \beta^* \neq 0$

Under H_0,

$$t\left(\widehat{\beta}\right) = \frac{\widehat{\beta}}{\widehat{se}\left(\widehat{\beta}\right)} \rightarrow^d \mathcal{N}(0,1) \tag{2}$$

$p-$value:

$$P\left(\widehat{\beta}\right) = \Pr\left(|t| \geq \left|t\left(\widehat{\beta}\right)\right| \middle| \mathbb{H}_0\right) \tag{3}$$

Based on (2), we can estimate $P\left(\widehat{\beta}\right)$ as

$$\widehat{P}\left(\widehat{\beta}\right) = 2\left(1 - \Phi\left(\left|t\left(\widehat{\beta}\right)\right|\right)\right) \tag{4}$$

, where Φ is the CDF of $\mathcal{N}(0,1)$.

Hypothesis Testing: What it is and What it is not

The rejection of $\mathbb{H}_0$ does *not* imply $|\beta^*|$ is significantly different from 0.

- Even if the rejection is correct, i.e. $\beta^* \neq 0$, it could be that $|\beta^*|$ is small and close to 0.
- To assess the magnitude of β^*, confidence intervals are more useful than $p-$values.
  - E.g., it is more informative to report a confidence interval of, say, $[.001, .009]$ than a $p-$value of .0124.

Hypothesis Testing: What it is and What it is not

The rejection of $\mathbb{H}_0$ does *not* mean x has a significant causal effect on y.

- It is never the case that statistical significance is the same as scientific, real-world significance. The most important variables are *not* those with the smallest p-values.

Hypothesis Testing: What it is and What it is not

In general, when there are p predictors, and we are testing the coefficient on x_j, the null hypothesis is not just "$\mathbb{H}_0 : \beta_j^* = 0$", but

$\mathbb{H}_0 : \beta_j^* = 0$ in a linear model that also includes predictors $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ and nothing else.

- The $t-$test can be thought of as checking whether adding x_j really improves predictions in a model that contains $\{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p\}$.
- Of course, adding more predictors never hurts the performance of a model on the training data. The $t-$test gauges whether the improvement in prediction is small enough to be due to random sampling.

Confidence Interval: What it is and What it is not

A 95% confidence interval $\text{CI} = \left[\widehat{\beta} - 1.96 \cdot \widehat{se}\left(\widehat{\beta}\right), \widehat{\beta} + 1.96 \cdot \widehat{se}\left(\widehat{\beta}\right)\right]$ does *not* mean that $\Pr\left(\beta^* \in \text{this particular CI}\right) = .95$.

- Correct interpretation: a 95% confidence interval for β^* means that if we estimate our model on many independent random samples drawn from the same population and construct $\text{CI}_m = \left[\widehat{\beta}_m - 1.96 \cdot \widehat{se}\left(\widehat{\beta}_m\right), \widehat{\beta}_m + 1.96 \cdot \widehat{se}\left(\widehat{\beta}_m\right)\right]$ on each sample, then 95% of $\{\text{CI}_m\}$ will contain β^*.

The Study and Target Populations

- Researchers often do not specify what the underlying population is that their observed sample is supposed to be drawn from – call it the **study population**, and the **target population** that they are trying to make inference on.
- Without a clear idea of what these populations are, hypothesis testing does *not* make sense.
- Questions we need to answer *before* conducting any analysis:
  1. What is our study population and what is our target population?
  2. How is the observed sample generated – is it, for example, a random draw from the study population?

The Study and Target Populations

"Psychology is the study of psychology students." – Anonymous

Vol 466|1 July 2010 nature

OPINION

Most people are not WEIRD

To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue **Joseph Henrich**, **Steven J. Heine** and **Ara Norenzayan**.

A 2008 survey of the top psychology journals found that 96% of subjects were from Western, educated, industrialized, rich and democratic (WEIRD) societies – particularly American undergraduates.

P-Value: What it is and What it is not

$p-$value is *not* the conditional probability of $\mathbb{H}_0$. When $p-$value $\leq \alpha$, it does *not* mean the probability of $\mathbb{H}_0$ being true, conditional on the observed data, is less than α.

$$\begin{aligned} \Pr\left(\mathbb{H}_0 | D\right) &= \frac{\Pr\left(D | \mathbb{H}_0\right) \Pr\left(\mathbb{H}_0\right)}{\Pr\left(D\right)} \\ &= \frac{\Pr\left(D | \mathbb{H}_0\right) \Pr\left(\mathbb{H}_0\right)}{\Pr\left(D | \mathbb{H}_0\right) \Pr\left(\mathbb{H}_0\right) + \Pr\left(D | \mathbb{H}_1\right) \Pr\left(\mathbb{H}_1\right)} \end{aligned} \tag{5}$$

, where $D \equiv \left(|t| \geq |\widehat{t}|\right)$, where $\widehat{t}$ is the t-statistic of the hypothesis test.

P-Value: What it is and What it is not

- (5) is the Bayes' rule.
  - In Bayesian terms, $\Pr(\mathbb{H}_0)$ is the **prior probability** of $\mathbb{H}_0$ being true, which represents our belief in $\mathbb{H}_0$ *before* we observe the evidence D.
  - $\Pr(\mathbb{H}_0|\, D)$ is the **posterior probability** of the $\mathbb{H}_0$ being true, which represents our belief in $\mathbb{H}_0$ after we observe D.
- We are interested in $\Pr(\mathbb{H}_0|\, D)$, but $p-$value gives us $\Pr(D|\, \mathbb{H}_0)$.
- In particular, if $\Pr(\mathbb{H}_0)$ is large, then even if $\Pr(D|\, \mathbb{H}_0)$ is small, $\Pr(\mathbb{H}_0|\, D)$ may not be small[1].

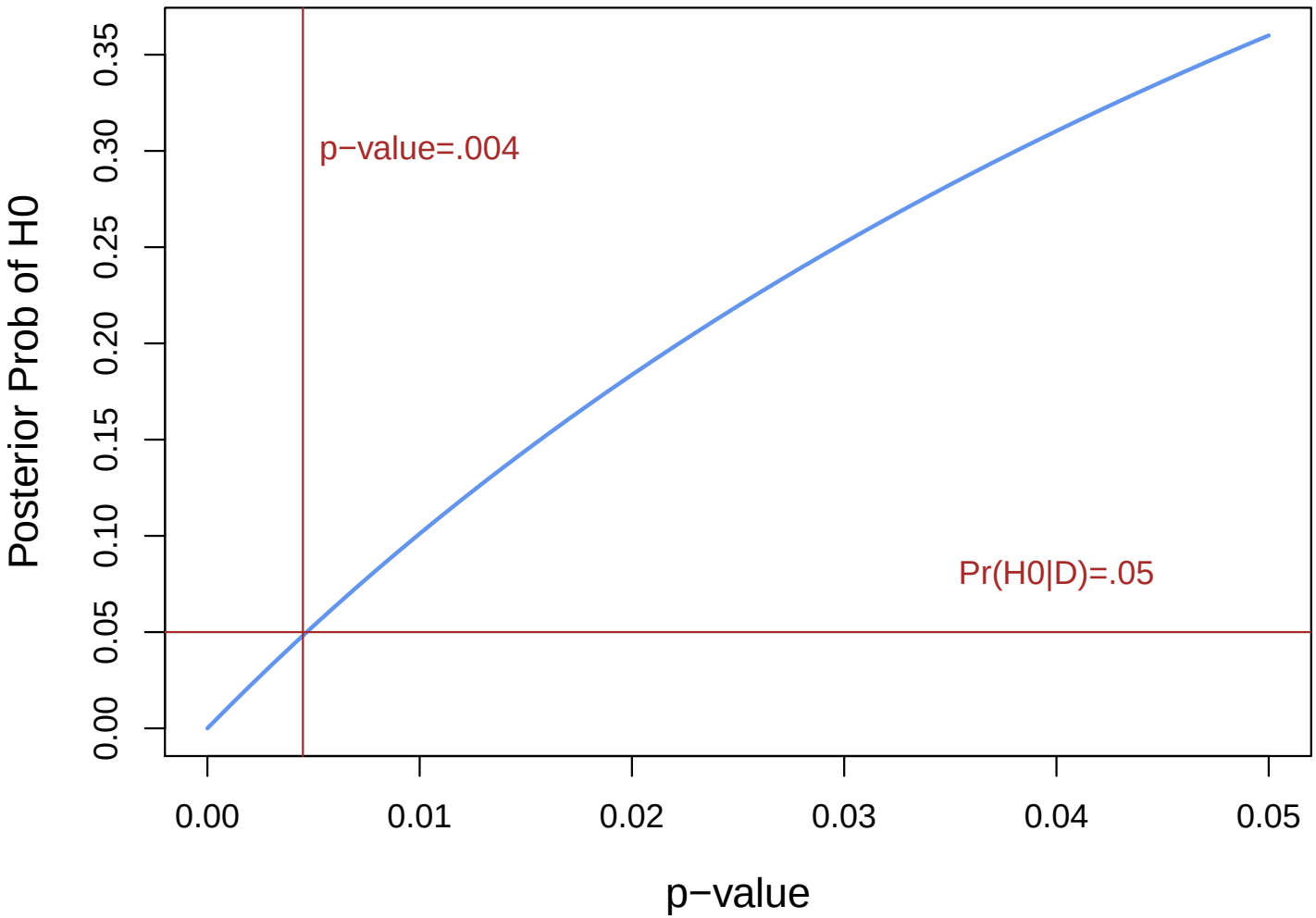
[1] This highlights the limitations of the frequentist approach – inability to incorporate prior knowledge.

P-Value: What it is and What it is not

```
p_H0 <- .9 # Pr(H0)
p_H1 <- 1 - p_H0 # Pr(H1)
p_D_H0 <- seq(.01, .05, by = .01) # Pr(D|H0), i.e. p-value
p_D_H1 <- .8 # Pr(D|H1)
p_H0_D <- (p_D_H0*p_H0)/(p_D_H0*p_H0 + p_D_H1*p_H1) # Pr(H0|D)
p_H0_D
```

```
## [1] 0.1011236 0.1836735 0.2523364 0.3103448 0.3600000
```

P-Value: What it is and What it is not



P-Value: What it is and What it is not

Moral

When the alternative hypothesis is highly unlikely (say, $\Pr(\mathbb{H}_1) < 0.1$), we need the $p-$value to be much smaller than the conventional threshold $\alpha = .05$ in order to "confidently" reject $\mathbb{H}_0$[a,b] (or we can just go Bayesian!)

---

[a] That is, we want to reject $\mathbb{H}_0$ when $\Pr(\mathbb{H}_0|D)$ is small.

[b] In other words, α should ideally be a function of $\Pr(\mathbb{H}_0)$ rather than a constant.

Journal's Paper on ESP Expected to Prompt Outrage

By BENEDICT CAREY
Published: January 5, 2011

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

Enlarge This Image

Heather Ainsworth for The New York Times

Work by Daryl J. Bem on extrasensory perception is scheduled to be published this year.

The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the paper, to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

The paper describes nine unusual lab experiments performed over the past decade by its author, Daryl J. Bem, an emeritus professor at Cornell, testing the ability of college students to accurately sense random events,

P-Values and Sample Size

- When N is small, (4) can be a poor estimator for $P\left(\widehat{\beta}\right)$, since it is based on the asymptotic distribution of $t\left(\widehat{\beta}\right)$[2,3].
- When N is very large ($N \to \infty$), $\widehat{\beta}$ converges to β^*: no need for hypothesis testing[4].

---

[2] The exact distribution of the $t-$statistic in finite samples can be derived under additional distributional assumptions. For example, if we assume a linear normal model

$$y = x'\beta + e, \quad e \sim \mathcal{N}\left(0, \sigma^2\right)$$

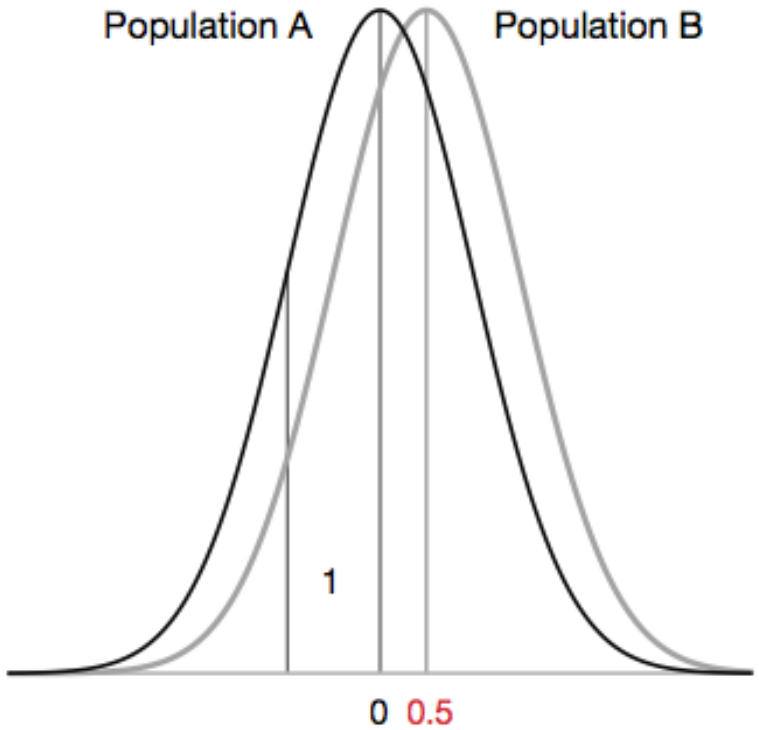
Then $t\left(\hat{\beta}_j\right) \sim t_{n-p-1}$, where x is $(p+1)$ dimensional. However, the assumptions of the linear normal model – gaussianity, homoskedasticity, and error independence – are seldomly satisfied.

[3] In this case one can sometimes use bootstrapping to obtain more accurate $p-$value estimates.

[4] In other words, hypothesis tests based on the asymptotic properties of test statistics are only valid for large samples, but only useful for samples that are not too large.

P-Values and Sample Size

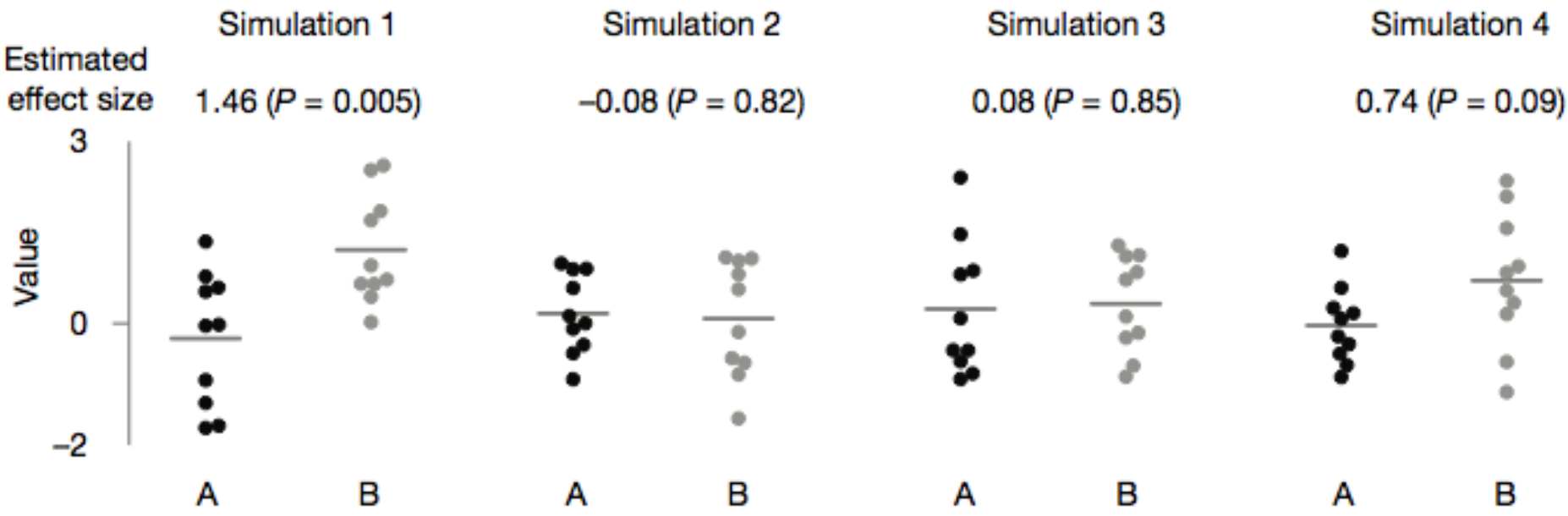
Simulation: draw data from the following populations:



$$A \sim \mathcal{N}(0,1), B \sim \mathcal{N}(0.5,1)$$

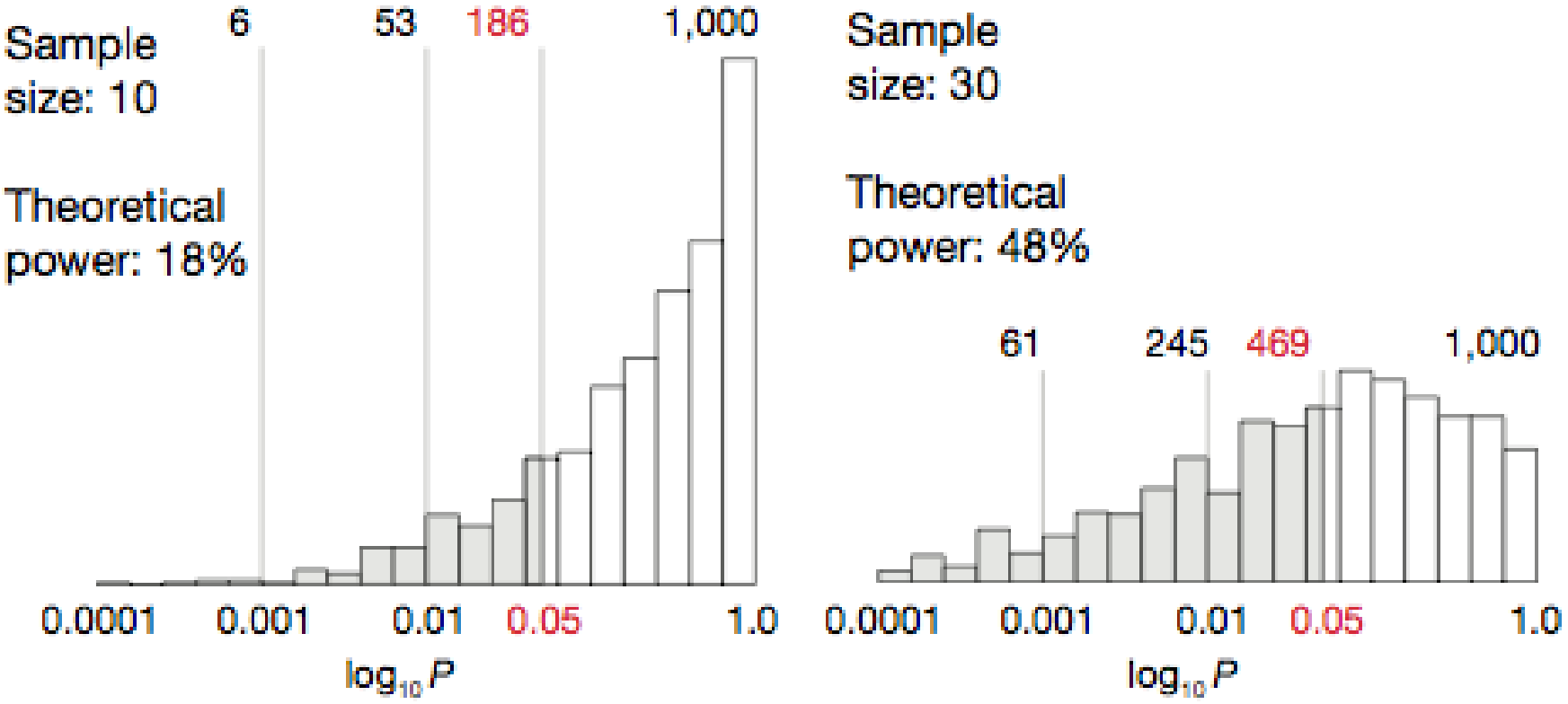
P-Values and Sample Size

$\mathbb{H}_0 : E(A) = E(B)$ vs. $\mathbb{H}_1 : E(A) \neq E(B)$



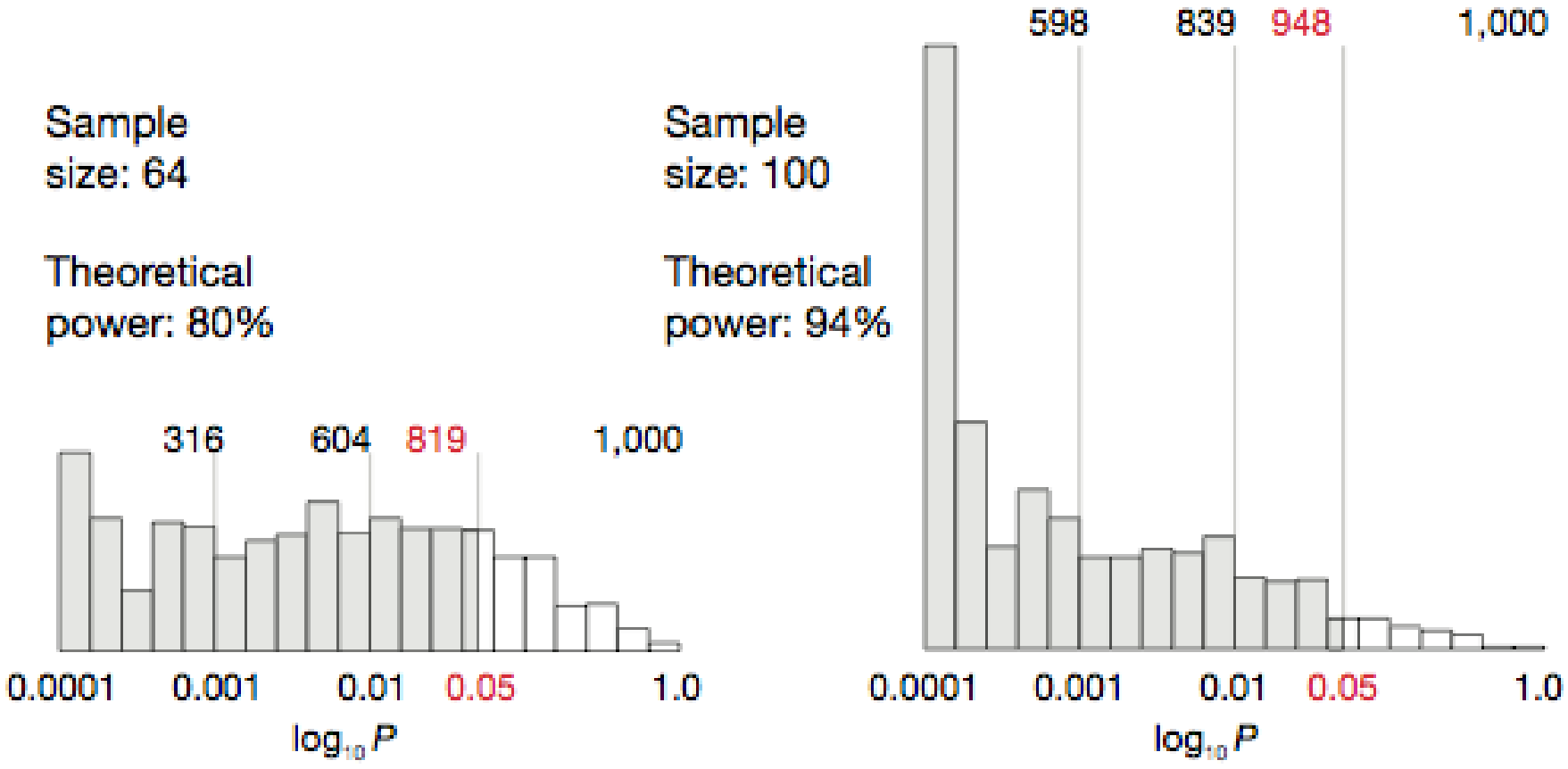
Each simulation draws 10 data points from populations A and B respectively and conducts a t-test of $\mathbb{H}_0 : E(A) = E(B)$ vs. $\mathbb{H}_1 : E(A) \neq E(B)$. Reported: effect size $(\widehat{E(B)} - \widehat{E(A)})$ and $p-$value. Source: Halsey et al. (2015)

P-Values and Sample Size



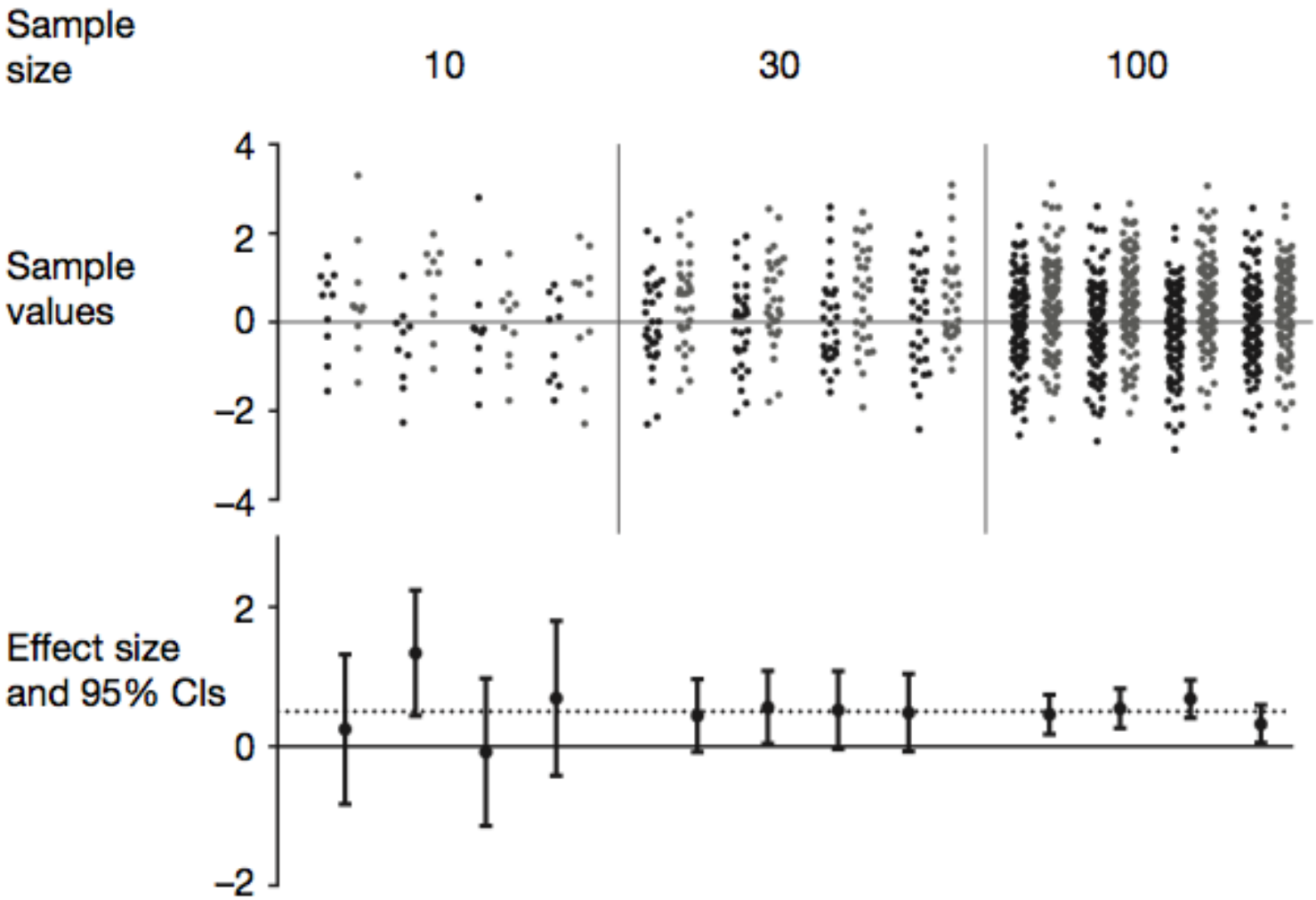
Distribution of $p-$values. Each plot is generated based on 1000 random samples. Source: Halsey et al. (2015)

P-Values and Sample Size



Distribution of $p-$values. Each plot is generated based on 1000 random samples. Source: Halsey et al. (2015)

P-Values and Sample Size



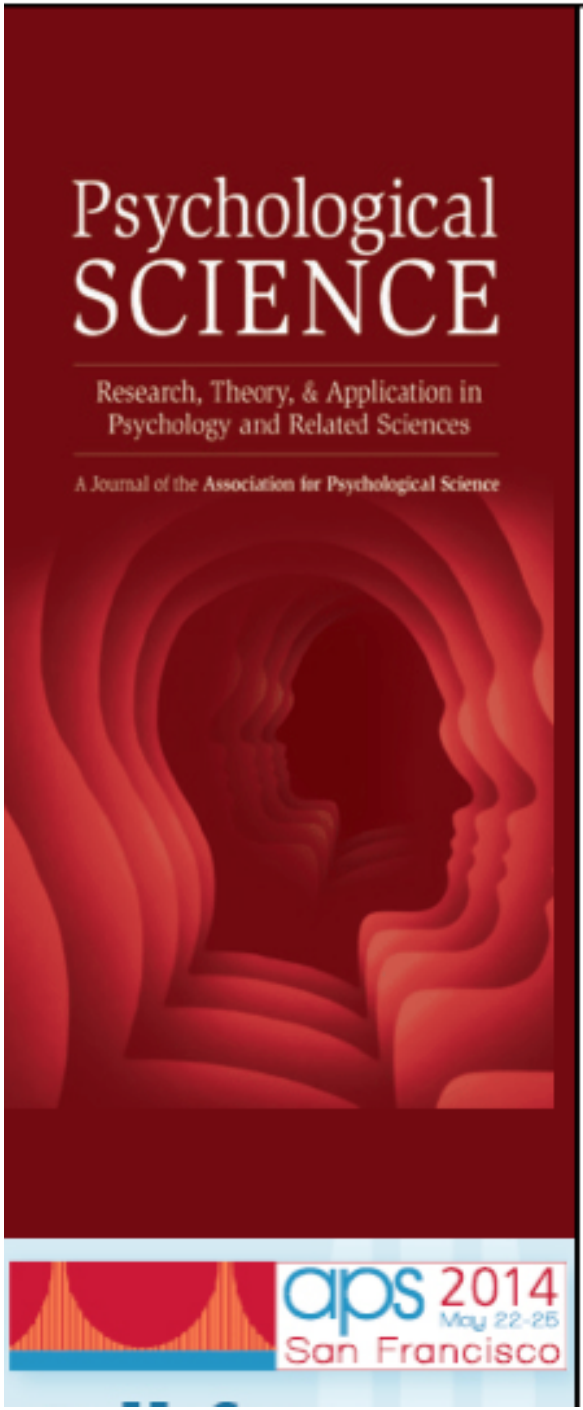
95% CI for $(\widehat{E(B)} - \widehat{E(A)})$. Source: Halsey et al. (2015)

P-Values and Sample Size

Moral

Do not trust the $p-$values calculated on small samples!

"The concept of $p-$values was originally developed by statistician Ronald Fisher in the 1920s in the context of his research on crop variance in Hertfordshire, England. Fisher offered the idea of $p-$values as a means of protecting researchers from declaring truth based on patterns in noise. In an ironic twist, $p-$values are now often used to lend credence to noisy claims based on small samples." – Gelman and Loken (2014)



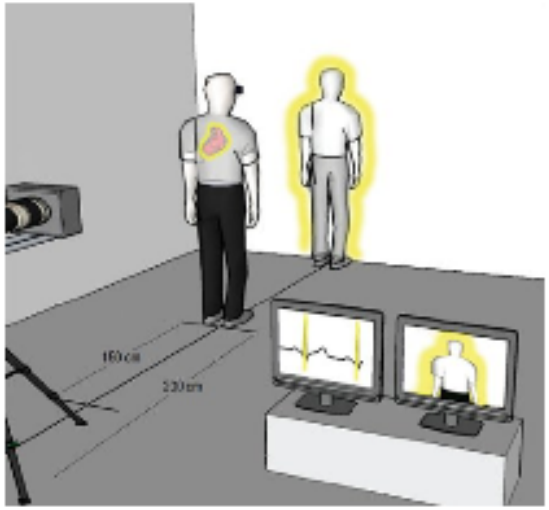
October 22, 2013

This Week in *Psychological Science* (TWiPS)

The links below take you to the journal via the APS website. If not already logged in, you will be redirected to log-in using your last name **(Gelman)** and Member ID **(8167)**.

Turning Body and Self Inside Out: Visualized Heartbeats Alter Bodily Self-Consciousness and Tactile Perception

Jane Elizabeth Aspell, Lukas Heydrich, Guillaume Marillier, Tom Lavanchy, Bruno Herbelin, and Olaf Blanke



Studies of body perception have mostly focused on manipulations of exteroceptive cues (e.g., vision and touch); however, interoceptive cues (i.e., representations of internal bodily states) may be just as important for self-perception. Participants viewed a virtual body or a rectangle, each of which had a flashing outline that was synchronous or asynchronous with the participant's own heartbeat. Self-identification was stronger for people viewing the virtual body with the synchronous flashing outline than for those viewing the body with the asynchronous flashing outline or for those viewing the rectangles. This suggests that both interoceptive and exteroceptive cues play important roles in bodily self-perception.

Aging 5 Years in 5 Minutes: The Effect of Taking a Memory Test on Older Adults' Subjective Age

Matthew L. Hughes, Lisa Geraci, and Ross L. De Forrest

Subjective age -- how old people feel -- is related to psychological and physical well-being. In this study, the researchers examined whether common memory-testing procedures influence adults' subjective age. Older and younger adults rated their subjective age before and after taking a memory test. Older adults reported feeling older after taking the memory test, but younger adults did not. A follow-up study found that

This Week in Psychological Science

- "Turning Body and Self Inside Out: Visualized Heartbeats Alter Bodily Self-Consciousness and Tactile Perception" ($N = 17$)
- "Aging 5 Years in 5 Minutes: The Effect of Taking a Memory Test on Older Adults' Subjective Age" ($N = 57$)
- "The Double-Edged Sword of Grandiose Narcissism: Implications for Successful and Unsuccessful Leadership Among U.S. Presidents" ($N = 42$)
- "Beauty at the Ballot Box: Disease Threats Predict Preferences for Physically Attractive Leaders" ($N = 123, 156, 66$[5])

[5] 3 separate studies, 1 observational, 2 experimental.

More Problems with the Use of P-Values in Practice

- The Statistical Significance Filter (Publication Bias)
- P-Hacking (Data Snooping)

The Statistical Significance Filter

- Focusing on statistically significant results is an entrenched culture in scientific research and publishing: only significant results get published. Consequently, published empirical findings are not a representative sample of all empirical findings.

- This is called the **statistical significance filter** or **publication bias.**

The Statistical Significance Filter

Suppose β^* is unbiasedly estimated by $\widehat{\beta} \sim \mathcal{N}(\beta^*, 1)$. If we only consider statistically significant results (at the 5% level), then we will only consider cases in which $\left|\widehat{\beta}\right| > 2$.

- $E\left[\left|\widehat{\beta}\right| \middle| \left|\widehat{\beta}\right| > 2\right]$ is clearly an overestimate of $|\beta^*|$.
- In particular, if $|\beta^*| < 2$, then any statistically significant $\widehat{\beta}$ will *always* be too high in magnitude.

The Statistical Significance Filter

- Simulation: draw $M = 100$ random samples, each containing $N = 200$ data points, from the following population:

$$\begin{aligned} x &\sim \text{Bernoulli}\,(0.5) \\ y &\sim \mathcal{N}\,(1 + 0.1x, 1) \end{aligned} \tag{6}$$

- Given (6), we have:

$$y = \alpha^* + \beta^* x + e^* \tag{7}$$

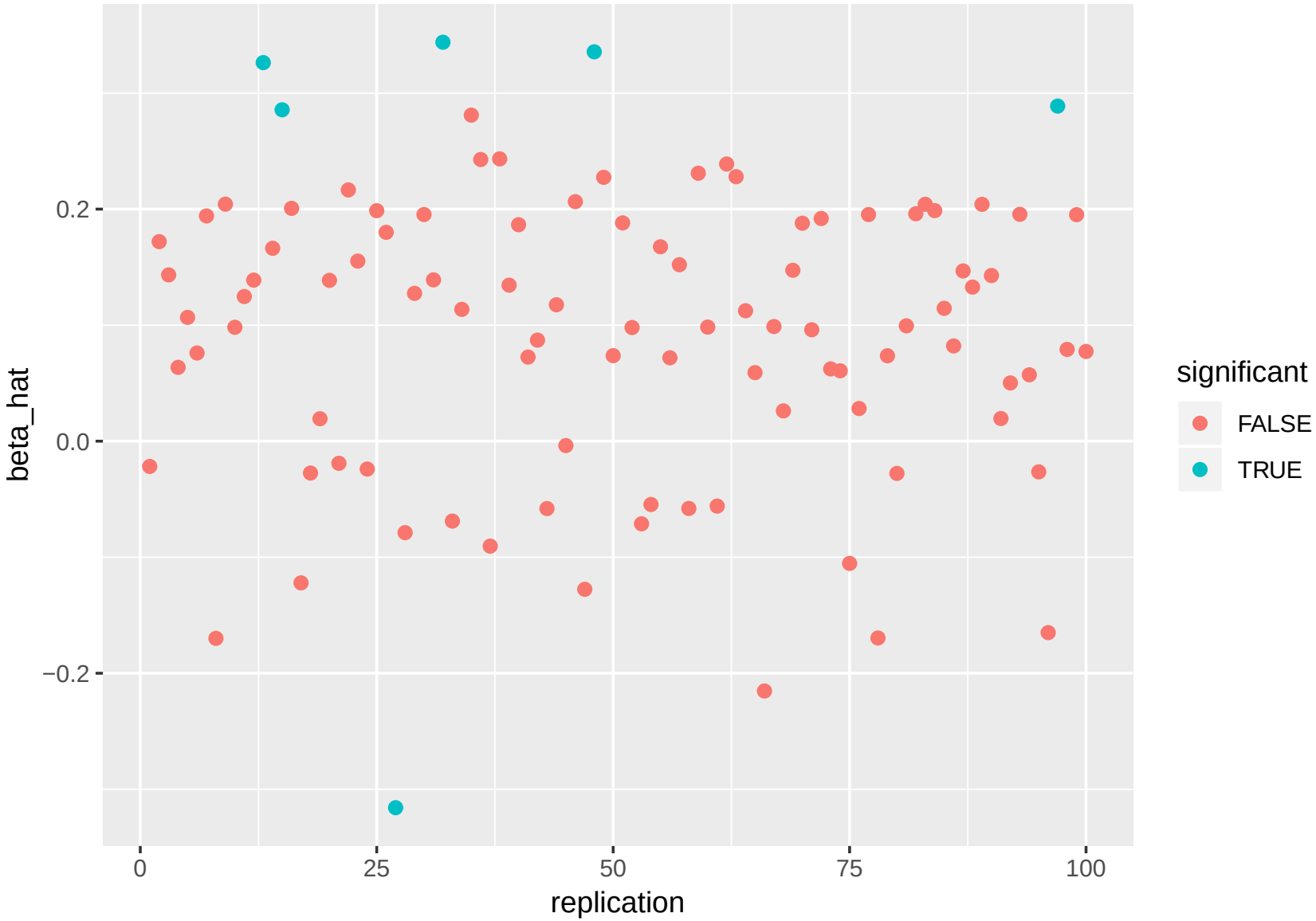
  where $\alpha^* = 1, \beta^* = 0.1$.

- Run regression on each data set m and obtain the OLS estimator $\widehat{\alpha}_m, \widehat{\beta}_m$.

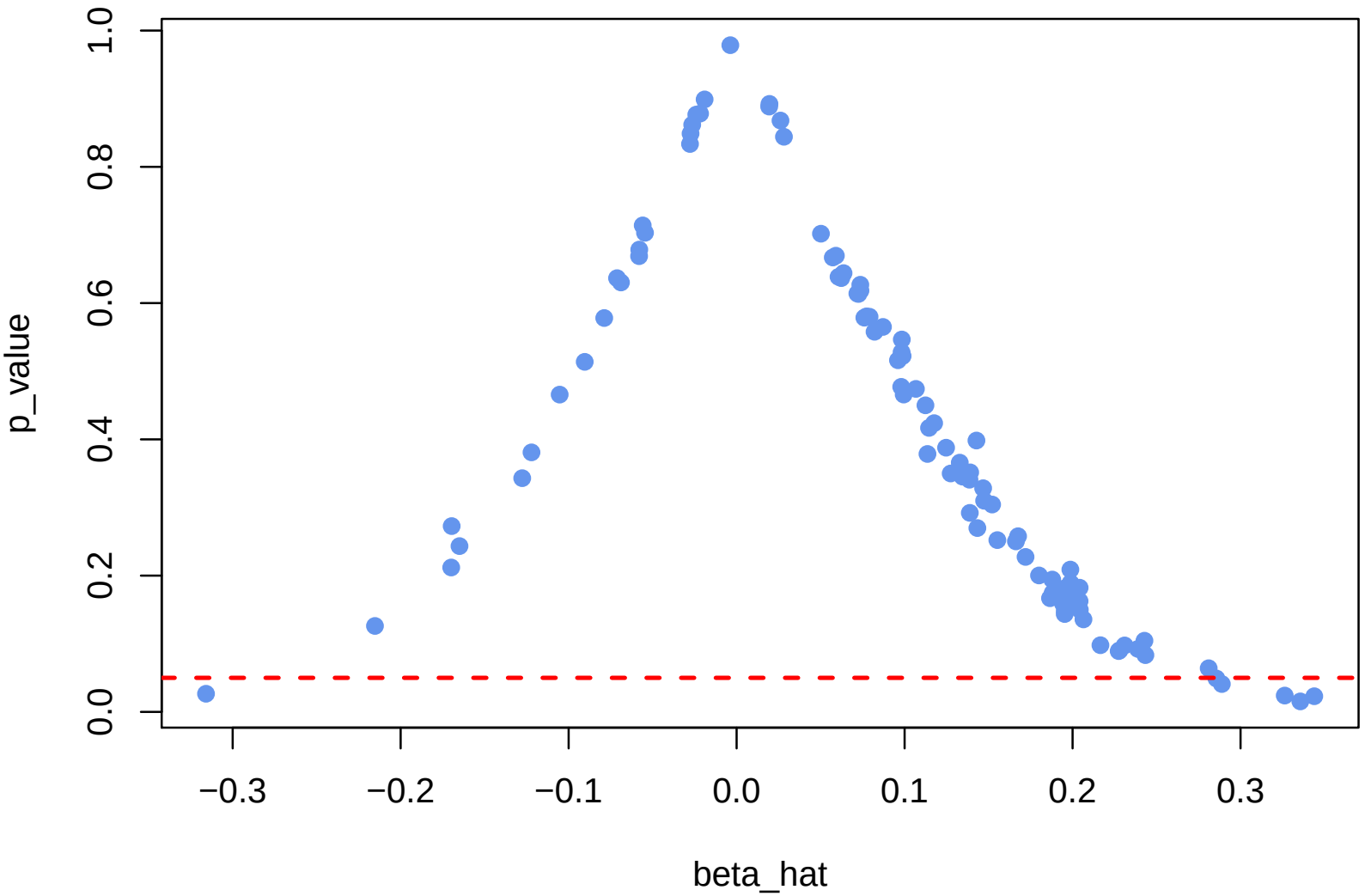
The Statistical Significance Filter

```r
# Simulation
require(dplyr)
require(broom)
M <- 100
N <- 200
beta_hat_dist <- replicate(M,{
  x <- rbinom(N,1,0.5)
  y <- 1 + 0.1 * x + rnorm(N)
  fit <- lm(y ~ x)
  coef <- tidy(fit) %>% filter(term == "x")
  c(beta_hat = coef$estimate, p_value = coef$p.value)
})

# E(beta_hat)
beta_hat_dist <- as.data.frame(t(beta_hat_dist))
beta_hat_dist %>% summarise (e_beta_hat = mean(beta_hat))

##   e_beta_hat
## 1 0.09559277
```

The Statistical Significance Filter



The Statistical Significance Filter



The Statistical Significance Filter

```
beta_hat_dist <- mutate(beta_hat_dist, significant = p_value <= .05)
beta_hat_dist %>% group_by(significant) %>%
  summarise(count = n(), e_beta_hat = mean(beta_hat))
```

```
## # A tibble: 2 x 3
##   significant count e_beta_hat
##   <lgl>       <int>      <dbl>
## 1 FALSE          94     0.0882
## 2 TRUE            6     0.211
```

- $E\left(\widehat{\beta}\middle|\text{significant}\right) \gg \beta^*$
- The power is low. The null hypothesis is false, but fails to be rejected about 90% of the time.

The Statistical Significance Filter

Run the same simulation with $\beta^* = .2, .3, .4,$ and $.5$ respectively.

Results for $\beta^* = .2, .3$:

```
## [[1]]
## # A tibble: 2 x 3
##   significant count e_beta_hat
##   <lgl>       <int>      <dbl>
## 1 FALSE          70      0.122
## 2 TRUE           30      0.352
##
## [[2]]
## # A tibble: 2 x 3
##   significant count e_beta_hat
##   <lgl>       <int>      <dbl>
## 1 FALSE          49      0.181
## 2 TRUE           51      0.398
```

The Statistical Significance Filter

Run the same simulation with $\beta^* = .2, .3, .4,$ and $.5$ respectively.

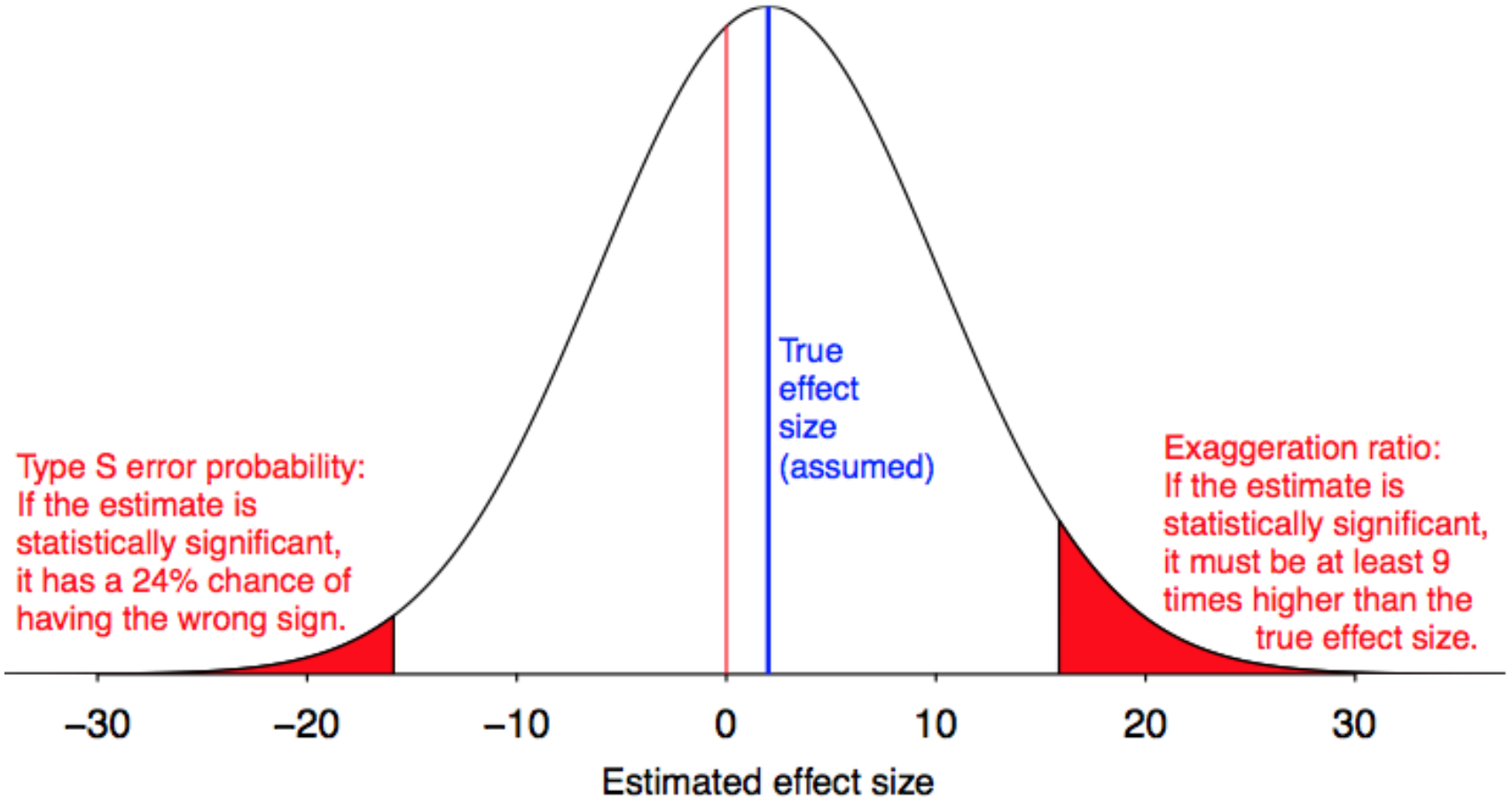
Results for $\beta^* = .4, .5$:

```
## [[1]]
## # A tibble: 2 x 3
##   significant count e_beta_hat
##   <lgl>       <int>      <dbl>
## 1 FALSE          13      0.210
## 2 TRUE           87      0.461
##
## [[2]]
## # A tibble: 2 x 3
##   significant count e_beta_hat
##   <lgl>       <int>      <dbl>
## 1 FALSE           5      0.274
## 2 TRUE           95      0.521
```

The Statistical Significance Filter

- The bigger $|\beta^* - \beta_{\mathbb{H}_0}|$ is, where $\beta_{\mathbb{H}_0}$ denotes the hypothesized value under $\mathbb{H}_0$, the greater the power of the test.
  - Power increases with N and $|\beta^* - \beta_{\mathbb{H}_0}|$.
- Power $\uparrow \Rightarrow \left| E\left(\widehat{\beta} \middle| \text{significant}\right) - \beta^* \right| \downarrow$

The Statistical Significance Filter



Lower power leads to high exaggeration ratios.

The Statistical Significance Filter

Moral

Kenkel (2016) offers the following suggestions:

- Assume the magnitudes of published results are exaggerated and adjust our own beliefs accordingly.
- Collect new data to replicate published findings and adjust our beliefs in the direction of the replication results.
- When writing our own papers, don't throw results away just because they're "insignificant."
- When reviewing others' papers, don't judge on the basis of significance. Assess whether the research design is well suited to address the question at hand, not whether it turned up results that are desired, or interesting, or surprising, etc.

In Search of Statistical Significance

Many specifications in search of statistical significance. Source.

In Search of Statistical Significance

One specification (here: Sir Perceval) achieves statistical significance. Source.

In Search of Statistical Significance

- The best practice for data analysis is to fix the model before seeing the training data and keep a separate test data set for assessing the performance of the estimated model.
- In practice, however, many researchers will try many specifications after seeing the data, until they get their desired, i.e. statistically significant, results.
- Some will also manipulate the data collection and processing stage by, for example, estimating a model on different subsets of the data in search of statistical significance, or stop data collection as soon as $p < 0.05$.

In Search of Statistical Significance

- Indeed, a dataset can be analyzed in many different ways, with the choices being not just what models to use, but also decisions on what measures to study, what data to include or exclude, etc.[6], that it can be easy to find a statistically significant result even if nothing is going on, as long as you look hard enough.

  "If you torture the data long enough, it will confess." – Ronald Coase

- Such practices of *data-dependent analyses* are called $p-$**hacking** or **data snooping**[7].

[6] This has been called the "researcher degrees of freedom."

[7] Also called specification search, data dredging, fishing . . . – you get the idea.

In Search of Statistical Significance

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Multiple Testing

- The problems of $p-$hacking are essentially the problems of **multiple testing** (also called **multiple comparisons**).
- If you perform multiple hypothesis tests, the probability of at least one producing a statistically significant result at the significance level α *purely due to chance*, is necessarily greater than α.
- Assuming each test is independent, under the $\mathbb{H}_0$ of all tests,

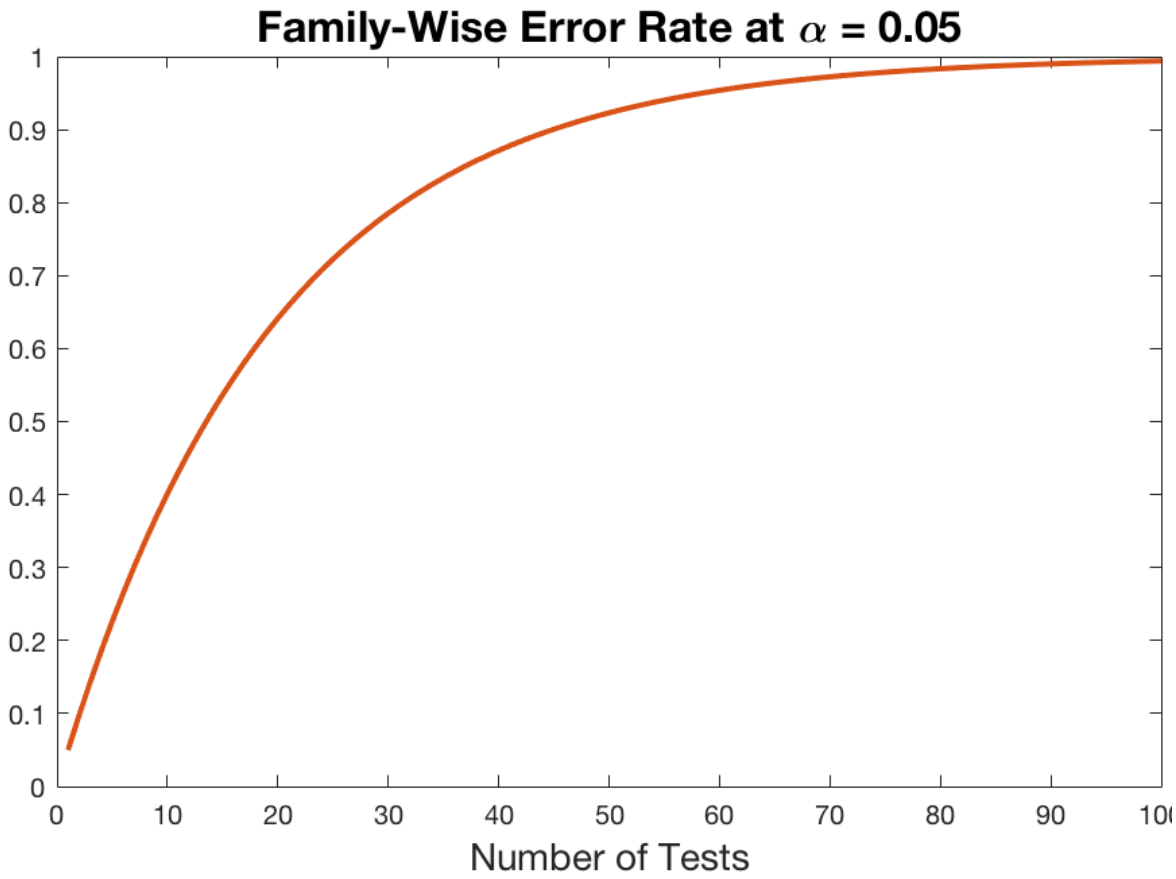
$$\Pr(\text{at least one is (falsely) positive}) = 1 - (1 - \alpha)^n$$

, where n is the number of tests conducted[8].

[8] Try Jerry Dallal's demo: 100 Independent 0.05 Level Tests For An Effect Where None Is Present.

Multiple Testing

N	Pr (false positive)
1	0.05
2	0.0975
5	0.2262
10	0.4013
50	0.9231
100	0.9941



Multiple Testing

"Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading." – Professionalism Guideline 8, Ethical Guidelines for Statistical Practice, American Statistical Association, 1997

Multiple Testing

Simulation: draw $N = 100$ data points from the following population:

$$\begin{aligned} \text{treatment} &\sim \text{Bernoulli}\,(0.5) \\ \text{male} &\sim \text{Bernoulli}\,(0.5) \\ y &\sim U\,(0,1) \end{aligned} \tag{8}$$

```
N <- 100
treatment <- rbinom(N,1,0.5)
male <- rbinom(N,1,0.5)
y <- runif(N)
```

Multiple Testing

```
# Regression on the entire sample
require(AER)
fit_all <- lm(y ~ treatment)
coeftest(fit_all)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.576674   0.045399 12.7024   <2e-16 ***
## treatment   -0.067942   0.059611 -1.1397   0.2572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Testing

```
# Regression on the male subsample
fit_male <- update(fit_all, subset = male == 1)
coeftest(fit_male)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.610265   0.064848  9.4107 1.78e-12 ***
## treatment   -0.175482   0.083719 -2.0961  0.04137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Testing

```
# Regression on the female subsample
fit_female <- update(fit_all, subset = male == 0)
coeftest(fit_female)

##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.546138   0.062031  8.8043 1.379e-11 ***
## treatment   0.041826   0.082892  0.5046    0.6162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Testing

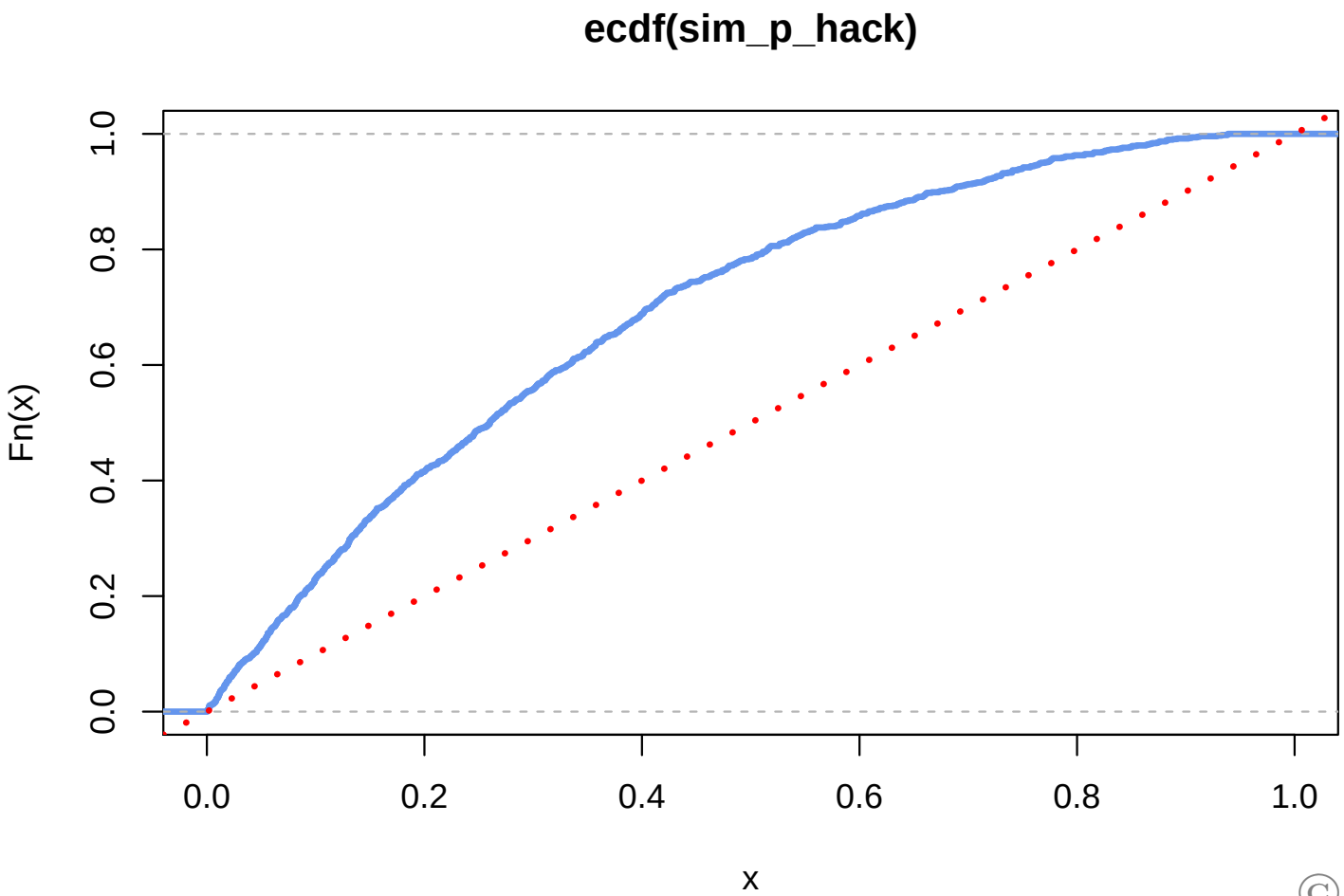
Simulation: draw $M = 1000$ random samples, each containing $N = 100$ data points, from the population specified in (8).

```
require(dplyr)
require(broom)
M <- 1000
extract_p <- function(fitted_model) {
     tidy(fitted_model) %>% filter(term == "treatment") %>%
     select(p.value) %>% as.numeric()}
sim_p_hack <- replicate(M, {
        treatment <- rbinom(N,1,0.5)
        male <- rbinom(N,1,0.5)
        y <- runif(N)
        fit_all <- lm(y ~ treatment)
        fit_male <- update(fit_all, subset = male == 1)
        fit_female <- update(fit_all, subset = male == 0)
        p_all <- extract_p(fit_all)
        p_male <- extract_p(fit_male)
        p_female <- extract_p(fit_female)
        min(p_all, p_male, p_female)
})
```

Multiple Testing

```
mean(sim_p_hack <= .05)
```

```
## [1] 0.117
```



Ovulation and Voting

Psychol Sci. 2013 Jun;24(6):1007-16. doi: 10.1177/0956797612466416. Epub 2013 Apr 23.

The fluctuating female vote: politics, religion, and the ovulatory cycle.

Durante KM[1], Rae A, Griskevicius V.

Author information

Abstract

Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulation-induced changes in political orientation mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships.

Ovulation and Voting

Sample: participants were 275 women with a mean age of 27.95 years (SD = 6.05, range = 18–44 years) who had regular monthly menstrual cycles (25–35 days) and were not using hormonal contraception.

Fertility: we created a high-fertility group (cycle days 7–14, n = 78) and a low- fertility group (cycle days 17–25, n = 85). For our main analyses, we did not include women on cycle days 15 and 16 ... We also did not include women at the beginning of the ovulatory cycle (cycle days 1–6) or at the end of the ovulatory cycle (cycle days 26–28).

Relationship status: participants who indicated that they were engaged, living with a partner, or married were classified as being in a committed relationship (n = 82); all others (e.g., not dating or dating) were classified as single (n = 81).

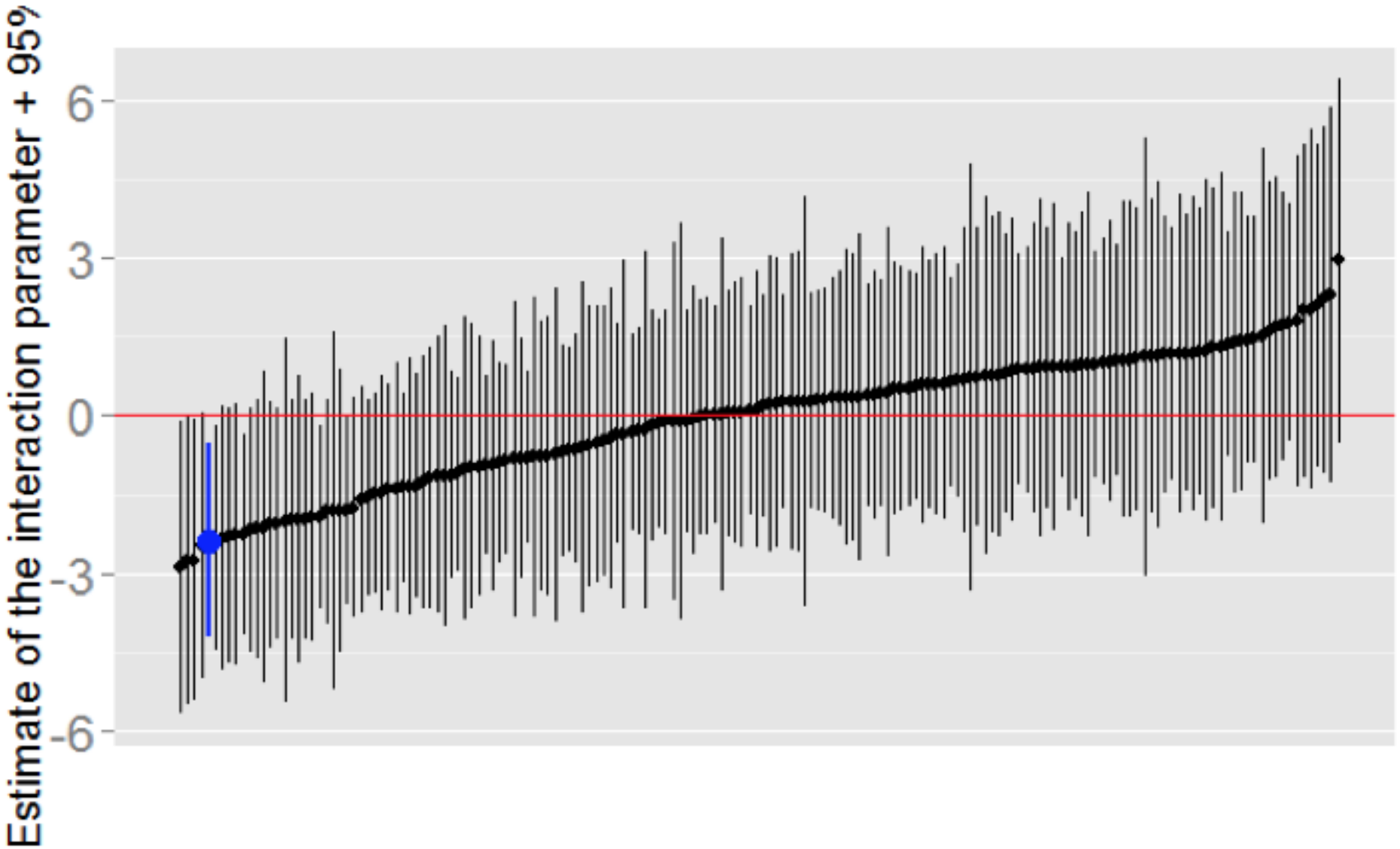
Ovulation and Voting

Researcher degrees of freedom:

- Exclusion criteria based on cycle length (3 options)
- Exclusion criteria based on "How sure are you?" response (2)
- Cycle day assessment (3)
- Fertility assessment (4)
- Relationship status assessment (3)

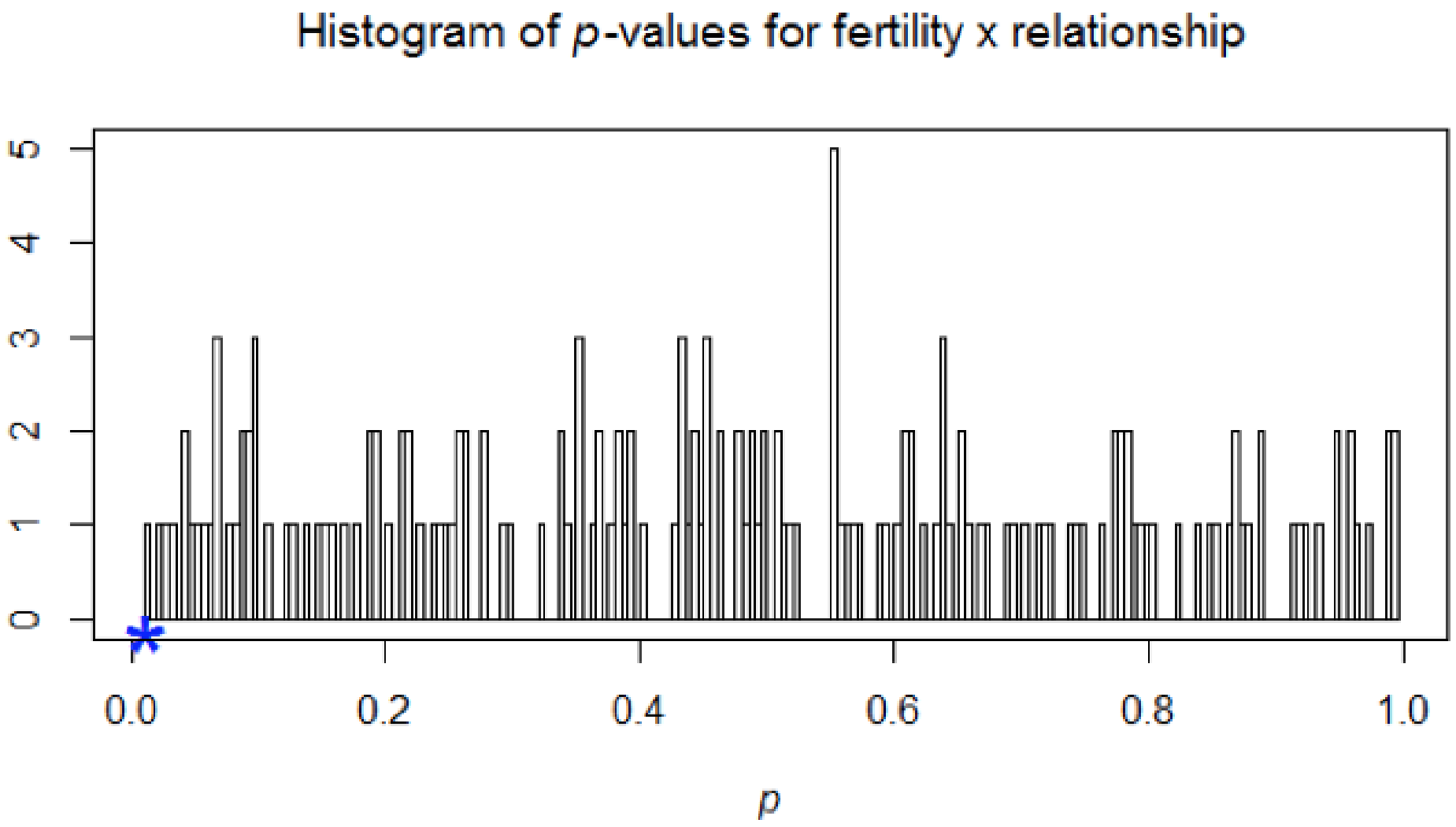
Altogether: 168 possibilities

Ovulation and Voting



Gelman (2016)

Ovulation and Voting



Gelman (2016)

Ovulation and Voting

Psychological Science
1–3

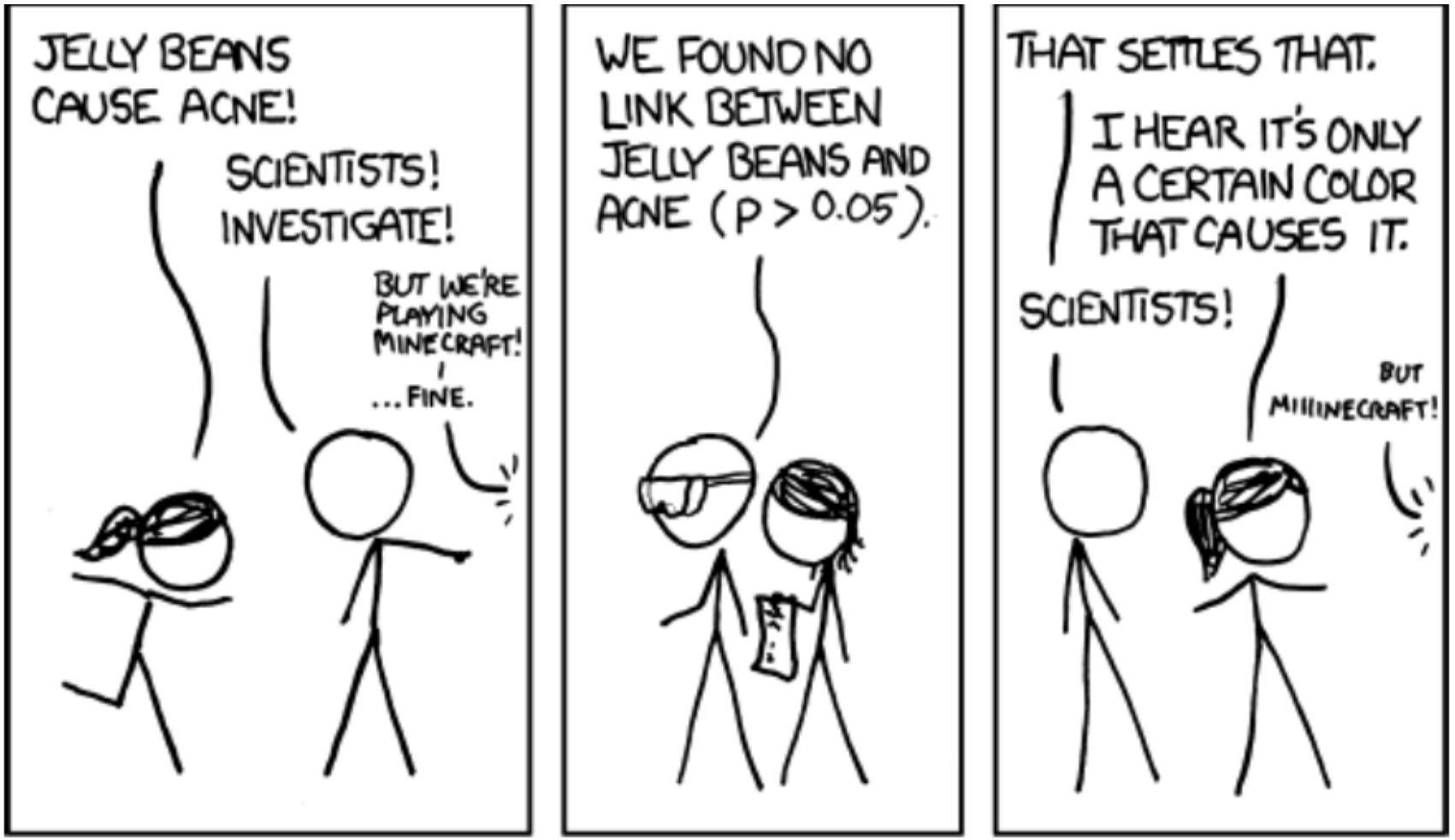
DOI: 10.1177/0956797613520236
pss.sagepub.com
SAGE

Women Can Keep the Vote: No Evidence That Hormonal Changes During the Menstrual Cycle Impact Political and Religious Beliefs

Christine R. Harris[1] and Laura Mickes[2]
[1]University of California, San Diego, and [2]Royal Holloway, University of London

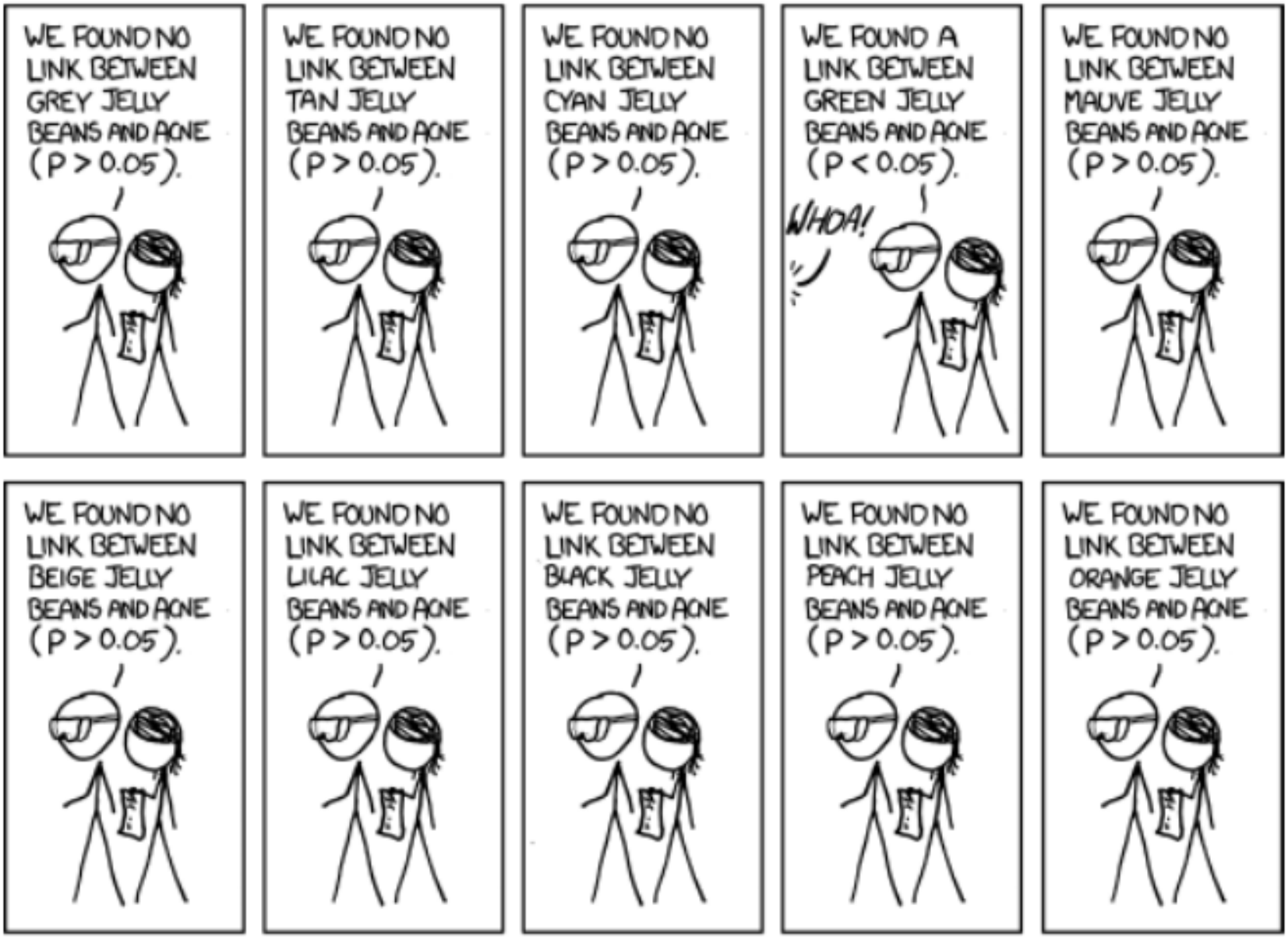
Harris and Mickes (2014): "We attempted to directly replicate the findings of Durante et al. ... We unequivocally failed to confirm two of the three key findings from the research reported by Durante et al. ... This study adds to a growing number of failures to replicate several menstrual cycle effects on preferences and attraction ... which invites concerns that this literature as a whole may have a false-positive rate well above the widely presumed 5%."

Fishing for Significance

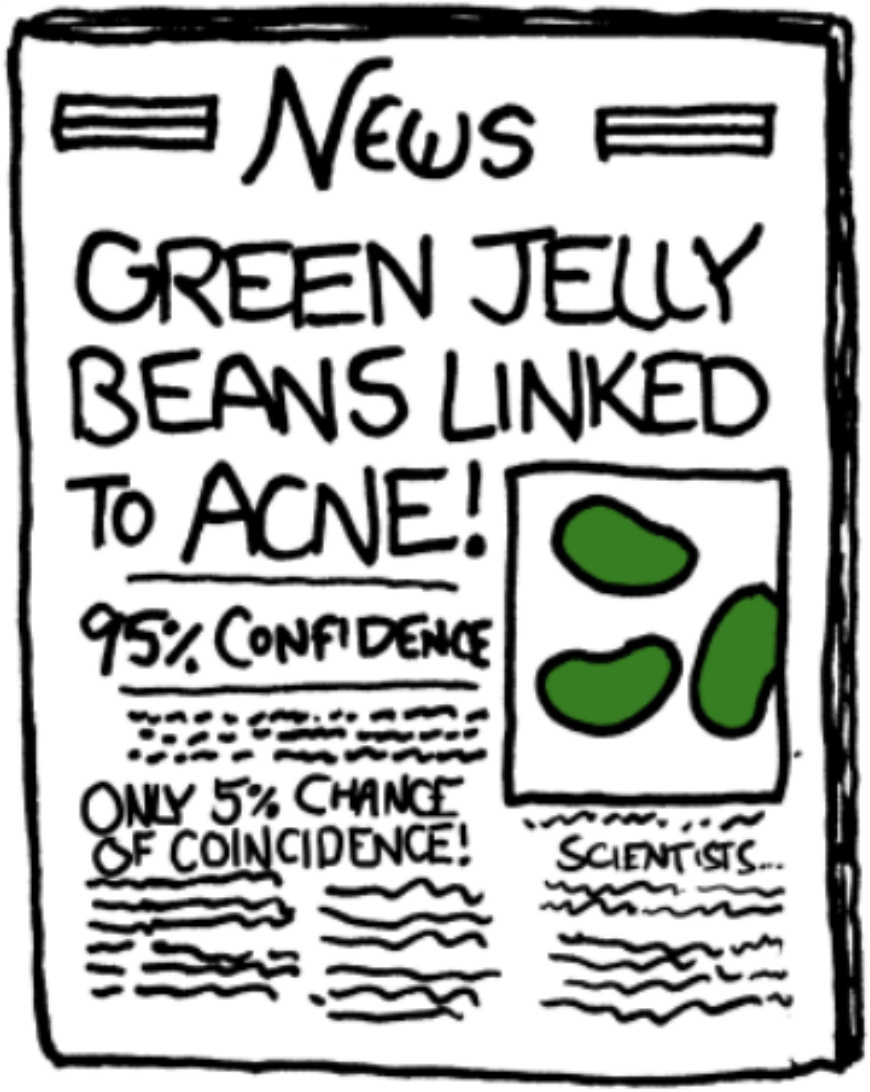


xkcd

Fishing for Significance



Fishing for Significance



The Mind-Reading Post-Mortem Salmon

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

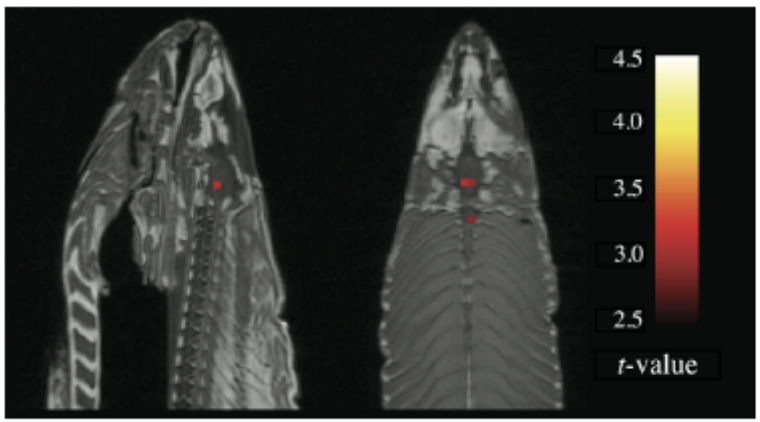
INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

GLM RESULTS



A *t*-contrast was used to test for regions with significant BOLD signal change

Bennett, et al. (2009)

The Bonferroni Correction

- The probability of making at least one type I error when simultaneously performing n hypothesis tests is called the **joint type I error rate**, also called the **family-wise error rate** (**FWER**).
- The Bonferroni correction bounds the FWER at below α by setting the significance threshold for each individual test at α/n[9]:

$$1 - \left(1 - \frac{\alpha}{n}\right)^n \leq \alpha$$

- The Bonferroni correction is *conservative*: it is derived under the assumption of independent tests. Using the Bonferroni correction when the number of tests is large leads to a significant loss of power[10].

[9] For example, if 10 hypothesis tests are performed, then the Bonferroni corrected significance level is 0.005 for each individual test in order for the FWER at below 0.05.

[10] Recall that power $\downarrow$ as $\alpha \downarrow$. Thus while the Bonferroni correction reduces the number of false positive findings, it does so at the expense of our ability to reject the null when it should have been.

Data Snooping

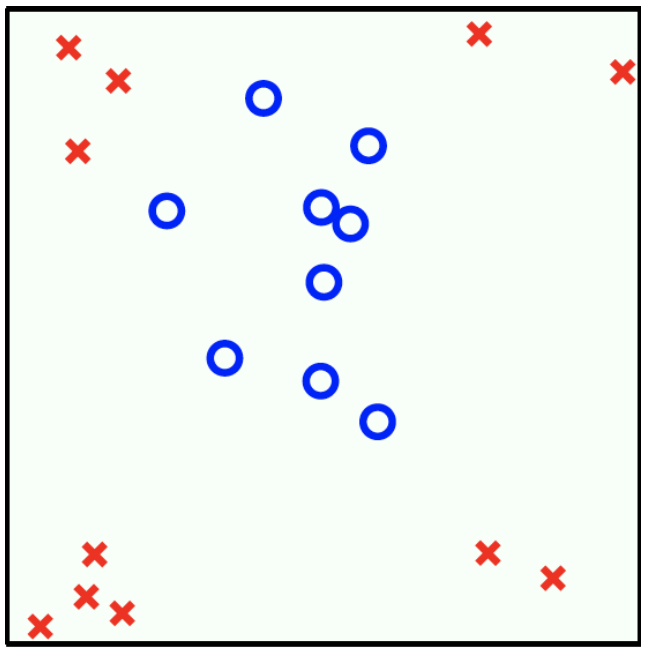
More generally, we have the following principle:

> If a data set has affected any step in the learning process, it cannot be fully trusted in assessing the outcome.

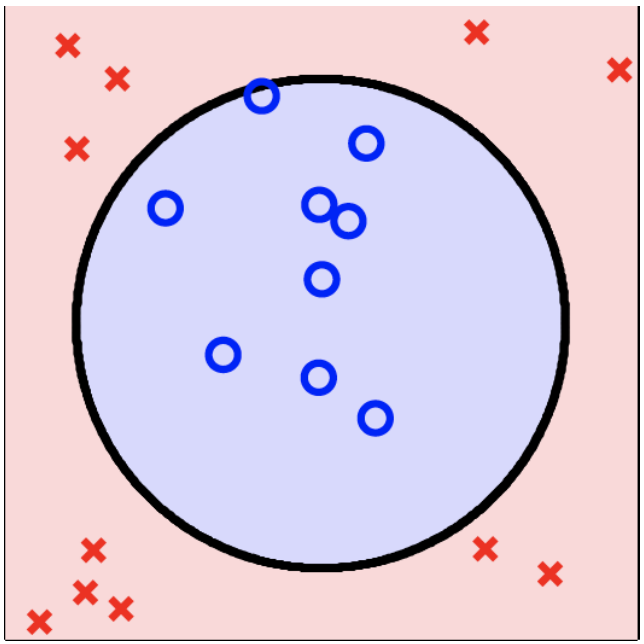
- Choosing a model based on the particular features of a data set invalidates the VC generalization bounds calculated based on the VC dimension of the final estimated model.
- If your approach is try a series of models on your data set before choosing one, then the *effective* VC dimension should be the VC dimension of the entire union of models that you would consider in this specification search process.

Data Snooping

The specification search process doesn't have to be explicit. Sometimes we do this in our head by *looking* at the data before choosing a model. Doing so is also an act of data snooping.



eye test



Data Snooping

- In general, the more we are willing to let a particular data set dictate our model choice, the poorer our result will generalize out of sample.

- The smaller the sample size, the more data snooping is a problem. When the sample size is large, multiple testing and data snooping are less of a concern, as $N \uparrow \Rightarrow$ generalization bound for $|E_{out} - E_{in}| \downarrow$.

Data Snooping

- A type of data snooping that is more difficult to avoid involves the reuse of the same data set by different people. This occurs, for example, when researchers work on the same public data set.

- When working with a public data set, it is common for a researcher to read about what others have done using the same data before formulating her model. In doing so, her model choice is already affected by the data set, since it is based what others have shown to work well or not well on *that particular* data set.

Data Snooping

- When researchers work on the same data set, while each formulating new hypotheses based on the work of others, the *effective* VC dimension corresponds to a much larger model space than the model chosen by any individual researcher – the model space contains all hypotheses that have been considered (and mostly rejected) by every researcher in this adaptive analysis process.

- This is a particular problem for social scientific research, where the ability to generate new data is limited, and many, mutually dependent, studies are based on the same datasets[11].

[11] This partly explains why social science models tend to have poorer generalization ability (predictive power) than natural science models.

Moral

When reading others' research...

- Be aware of the multiple testing and data snooping that might have been going on behind published results, and adjust our beliefs about their generalization performance accordingly.
- Be aware of the inherent data snooping problem in adaptive analysis when reading published results based on data sets that have been used by many others, and adjust our beliefs about their generalization performance accordingly.

Moral

When conducting your own research...

- Formulate the research question and decide on what model to use *before* seeing the data.
  - If possible, generate new data based on your research design.
- If you *intend to* engage in data snooping and choose a model based on the data, then you should decide on the *set* of models you are going to choose from *before* seeing the data, and account for the data snooping in your analysis by
  - adjusting the significance level of your hypothesis tests by, for example, using the Bonferroni correction;
  - using a test data set to evaluate the performance of your final estimated model. The test set should be allocated at the beginning and only used at the end. Once a data set has been used, it should be treated as *contaminated* as far as testing the performance is concerned.

Moral

When reporting your research...

- Aim for honesty and transparency.
- Clearly state your research question, the research design, and the reasoning behind your model choice.
- Clearly state if your analysis involves data snooping and how you have accounted for it.
- Report every hypothesis test you have performed relevant to the research question and highlight results that are robust across tests.
- Include a limitations section and point out any limitations and uncertainties in the analysis.

Appendix: The ASA Statement on P-Values

The ASA Statement on P-Values

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.

Acknowledgement

Part of this lecture is adapted from the following sources:

- Abu-Mostafa, Y. S., M. Magdon-Ismail, and H. Lin. 2012. *Learning from Data*. AMLBook.
- Gelman, A. *The Statistical Crisis in Science*. Talk at the Chief Economists' workshop, Bank of England, London, May 20, 2016. [link]
- Kenkel, B. *The Statistical Crisis in Science, or: How I Learned to Stop Worrying and Love Insignificant Results*. Lecture at Vanderbilt University, January 28, 2016. [link]
- Shalizi, C. R. 2019. *Advanced Data Analysis from an Elementary Point of View*. Manuscript.
- Smith, M. K. *Common ~~Misteaks~~ Mistakes in Using Statistics: Spotting and Avoiding Them.* Online writing, retrieved on 2018.01.01. [link]

Reference I

- Bennett, C. M., M. B. Miller, and G. L. Wolford. 2009. "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction," *Neuroimage*, 47.
- Carey B. "Journal's Paper on ESP Expected to Prompt Outrage," *The New York Times*, Jan. 5, 2011.
- Durante K. M., A. Rae, and V. Griskevicius. 2013. "The fluctuating female vote: politics, religion, and the ovulatory cycle," *Psychological Science*, 24(6).
- Gelman, A. and E. Loken. 2014. "The Statistical Crisis in Science," *American Scientist*, 102(6).
- Halsey, L. G., D. Curran-Everett, S. L. Vowler, and G. B. Drummond. 2015. "The fickle P value generates irreproducible results," *Nature Methods*, 12(3).
- Harris, C. R. and L. Mickes. 2014. "Women Can Keep the Vote: No Evidence That Hormonal Changes During the Menstrual Cycle Impact Political and Religious Beliefs," *Psychological Science*, 25(5).

Reference II

- Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "Most people are not WEIRD," *Nature*, 466(29).
- Hotz, R. L. "Most Science Studies Appear to Be Tainted By Sloppy Analysis," *The Wall Street Journal*, Sep. 14, 2007.
- Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False," *PLoS Medicine*, 2(8).
- Ioannidis, J. P. A. 2008. "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 19 (5).
- Wasserstein, R. L. and N. A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70(2).
- Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. 2019. "Moving to a World Beyond 'p < 0.05'," *The American Statistician*, 73(1).