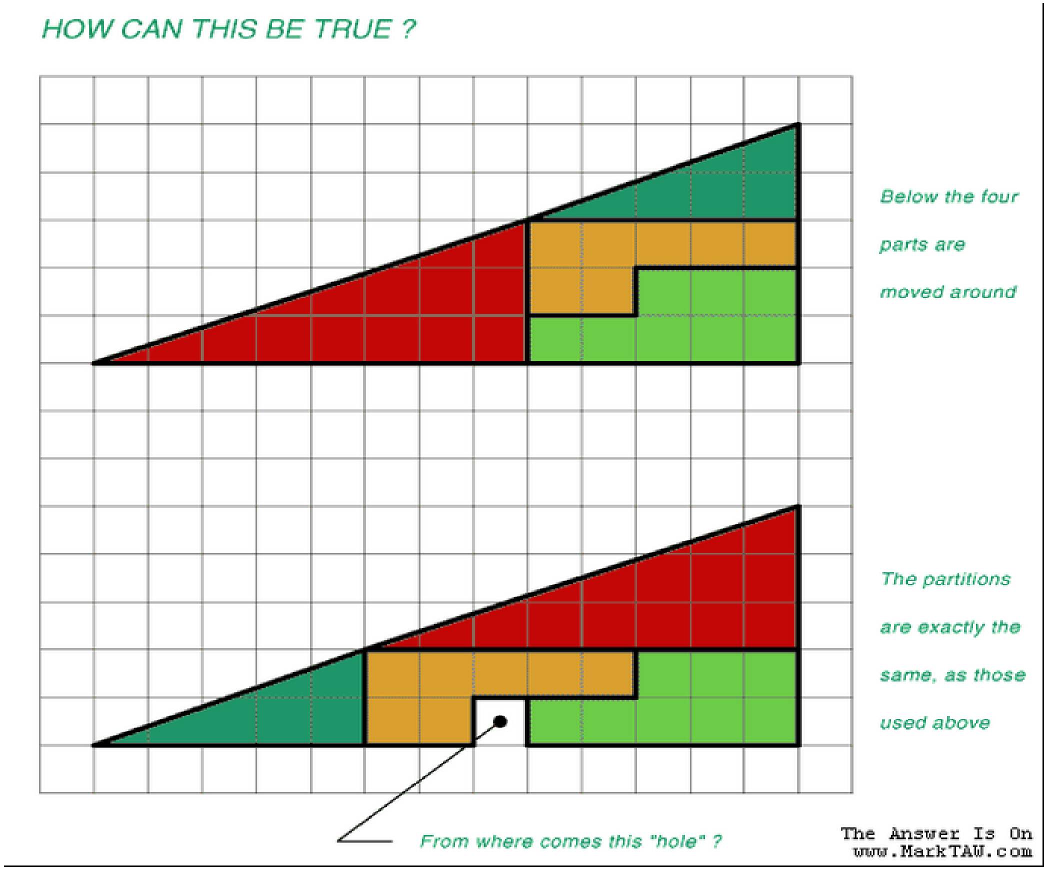
TIM KRASKA

LYING WITH STATISTICS AND VISUALIZATIONS



http://www.marktaw.com/blog/TheTriangleProblem.html

P-VALUE

P-VALUE HAS PROBLEMS!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

BAYESIAN TESTS

HYPOTHESIS TESTING

Suppose we have two models, H_0 and H_1.

Which model is better supported by the data?

The model that predicted the data best!

The ratio of predictive performance is known as the Bayes factor.

$$BF = \frac{p(data|H_0)}{p(data|H_1)}$$

Define hypothesis about (population) effect size $\boldsymbol{\delta}$

H_0: $\boldsymbol{\delta} = 0$ H_1: $\boldsymbol{\delta} > 0$

Collect data

Traditional

Compute "p-value":

p(data | H_0)

Interpretation:

If p is small (e.g., 0.05), data is rate under H_0, so we reject H_0 in favor of H_1.

Bayesian Statistics

Compute "Bayes Factor"

$$BF = \frac{p(data|H_0)}{p(data|H_1)}$$

Interpretation:

If $BF_{01} > 1$, data more likely under H_0

If $BF_{01} < 1$, data more likely under H_1

INTERPRETATION OF BAYES FACTOR

$$BF = \frac{p(data|H_0)}{p(data|H_1)}$$

Can directly index support for either H_0 or H_1

Interpretation.

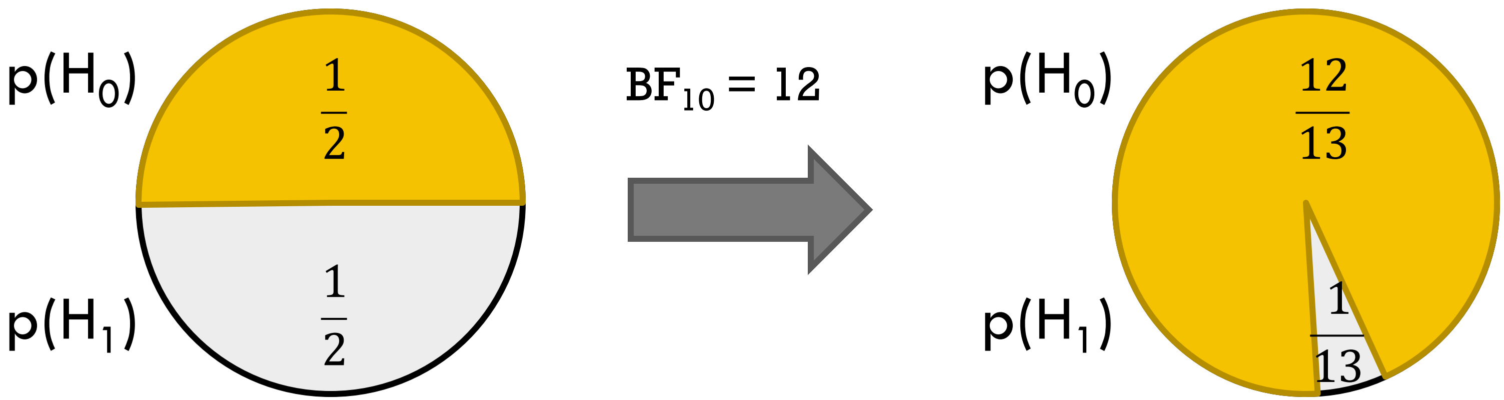
1) Relative predictive adequacy of models

Example: $BF_{10} = 12$ → "The observed data are 12 times more likely under H_1 than H_0"

2) Updating factor

Example: $BF_{10} = 12$ → "After observing data, my prior odds for H_0 over H_1 have been increased by a factor of 12"

UPDATE FACTOR – EXAMPLE



Prior odds: 1:1
(without seeing the data)

Posterior odds: 12:1
(without seeing the data)

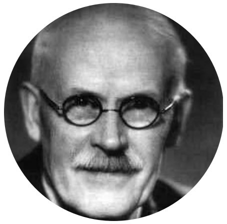
YOU CAN CONVERT T-STATISTICS

Let's assume the national average test score for a math-test is 50. After we tutored N= 65 students, we observed a mean test-score of 54.4 with SD=10. Does tutoring help?

Step 1: Convert our observed data to a test statistics

$$\boldsymbol{t = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{N}} = \frac{54.4-50}{10/\sqrt{65}} = 3.55}$$

Step 2: Convert t-score to Bayes factor



INTERPRETATION GUIDELINES

BF_{10}	Evidence	Direction
> 100	Extreme	In favor of H_1 over H_0
30 – 100	Very strong	In favor of H_1 over H_0
10 – 30	Strong	In favor of H_1 over H_0
3 – 10	moderate	In favor of H_1 over H_0
1 – 3	Anecdotal	In favor of H_1 over H_0
1	equal	Between H_1 and H_0
1 – 1/3	Anecdotal	In favor of H_0 over H_1
1/3 – 1/10	Moderate	In favor of H_0 over H_1
1/10 – 1/30	Strong	In favor of H_0 over H_1
1/30 – 1/100	Very strong	In favor of H_0 over H_1
< 100	Extreme	In favor of H_0 over H_1

ANOTHER EXAMPLE: AB TESTING

Sam wants to update his profile picture on his website to attract more junior students to enroll for 6.830 / 6.814. **He designs an a/b test to see if his new profile picture increases the enrollment.**

Current picture

New picture

More realistic example: ad-conversion rates based on title, image, etc.

PROBLEM WITH FREQUENTIST TESTING

After observing some data we find that the new model is **only slightly better (e.g., conversion rate of 10% vs 9.5% enrollment)** than the current model with a p-value of 0.11

→ proper procedure is to keep the current model.

PROBLEM WITH FREQUENTIST TESTING

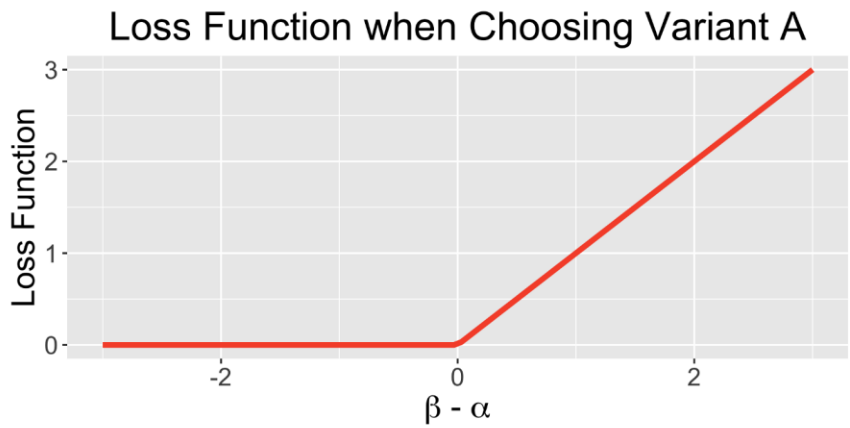
After observing some data we find that the new model is **only slightly better (e.g., conversion rate of 10% vs 9.5% enrollment)** than the current model with a p-value of 0.11

→ proper procedure is to keep the current model.

→ However, since the new model is making better predictions than the current model, this decision is very unsatisfying and potentially costly.

→ However, for this example even small improvements might matter. As we perform hundreds of experiments on the same handful of key business metrics, these marginal gains will accumulate on top of each other.

If we choose variant A when α is less than β, our loss is $\beta - \alpha$. If α is greater than β, we lose nothing. Our loss is the amount by which our metric decreases when we choose that variant



BAYESIAN A/B TESTING - PROCEDURE

Conversion=1 indicates a student entrolls and *conversion=0* indicates they did not. Binomial distribution

Conjugate prior distribution: beta distribution

```
Monte carlo simulation (using prior distribution):

S_control = sample_from_distr(control_dist, n=10000)

S_treatment = sample_from_distr(treatment_dist, n=10000)


/// Calculate proportion of treatment being better than control

probability_best = mean(int(samp_treatment > samp_control))


// Calculate expected loss- iterate over our samples and
calculate max(treat - control, 0)

loss = mean(argmax(s_treatment - s_control, 0))
```

P-HACKING (ALSO DATA DREDGING, DATA FISHING, DATA SNOOPING, DATA BUTCHERY)

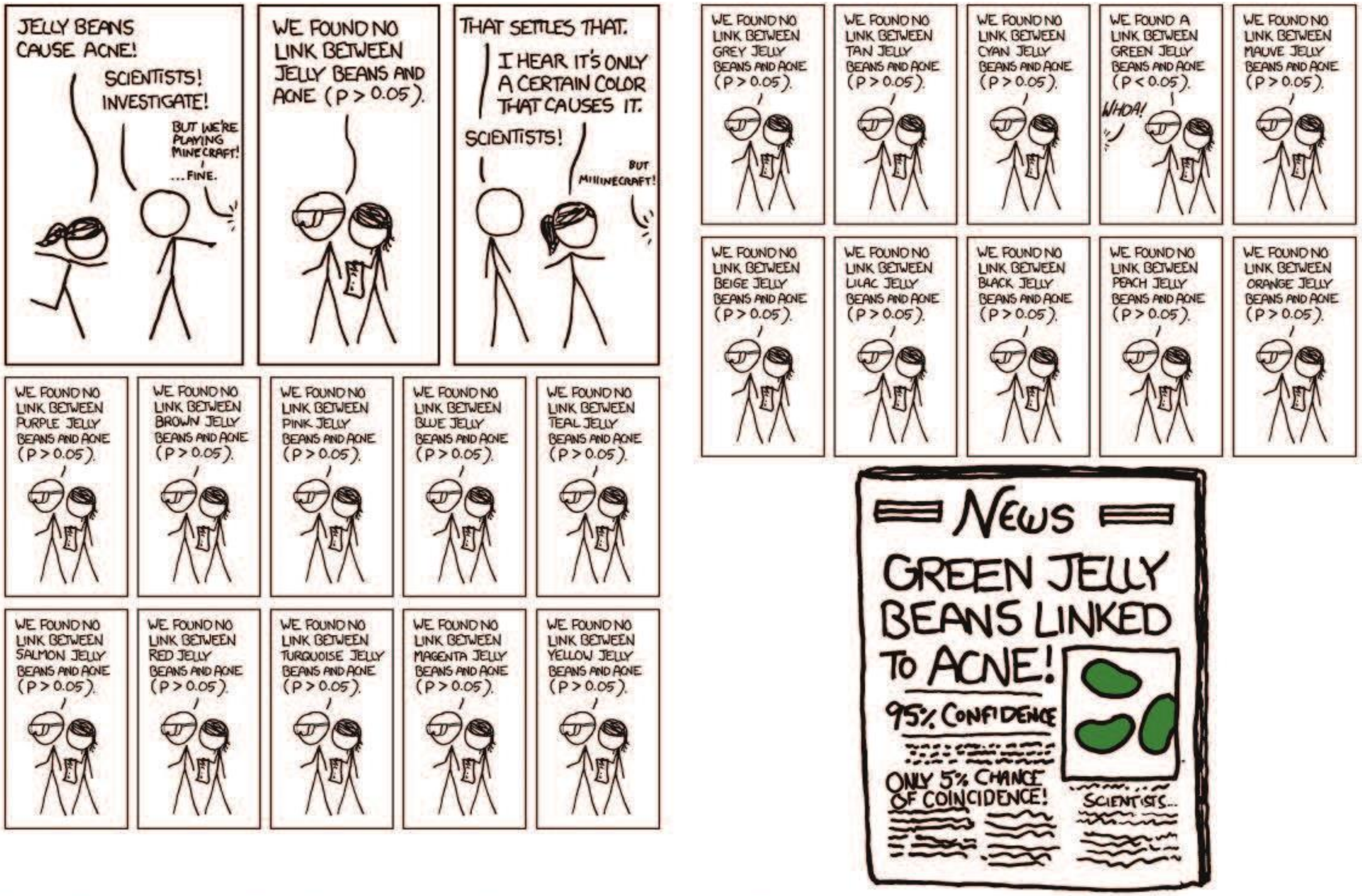


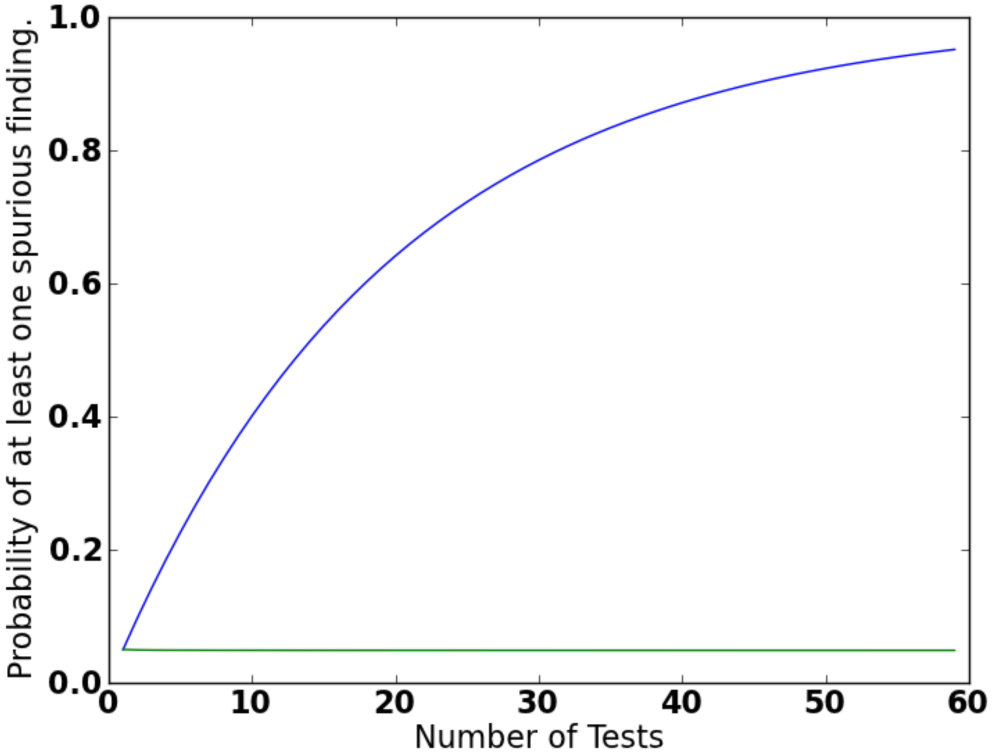
Figure 1. There is no overall effect of jelly beans on acne. Bummer. How about subgroups? Often subgroups are explored without alerting the reader to the number of questions at issue. Courtesy xkcd, `http://xkcd.com/882/`

P(detecting an effect when there is none) = α = 0.05

P(not detecting an effect when there is none) = $1 - \alpha$

P(not detecting an effect when there is none, on every experiment) = $(1 - \alpha)^k$

P(detecting an effect when there is none on at least one experiment) = $1 - (1 - \alpha)^k$

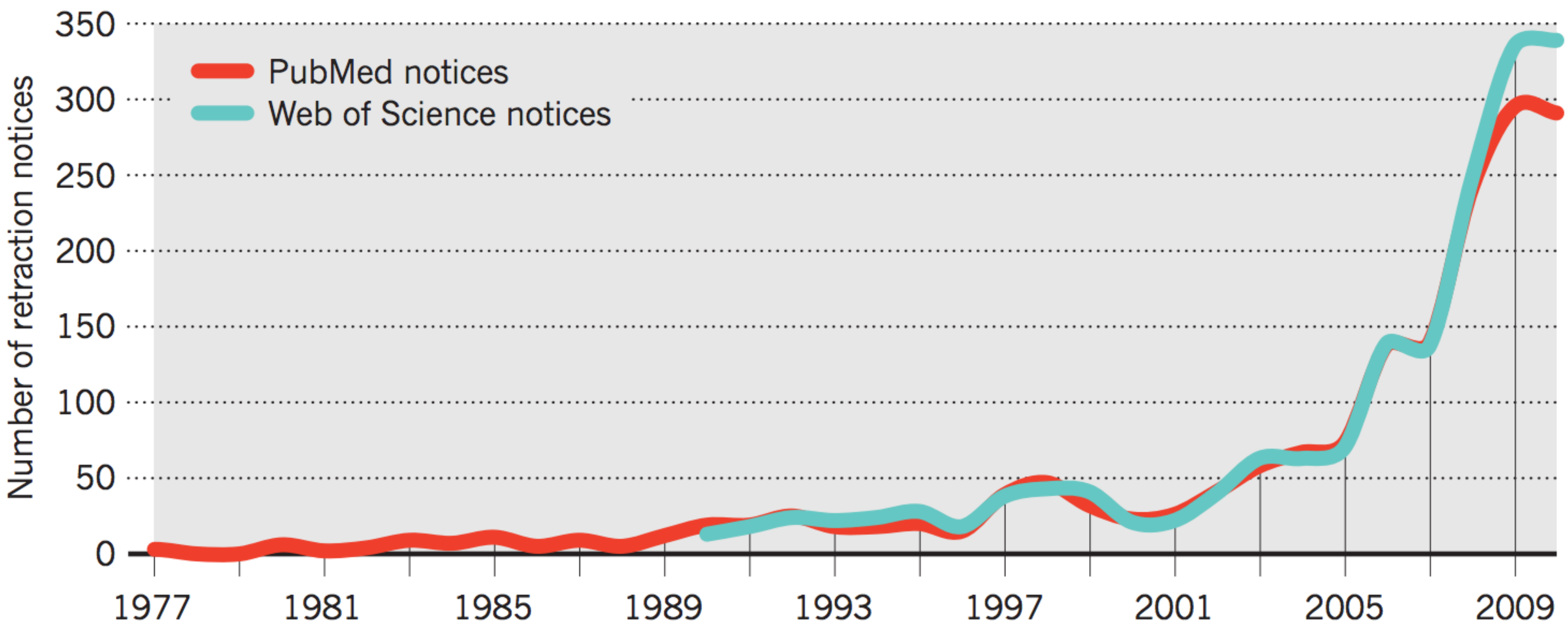


$\alpha = 0.05$

"Familywise Error Rate"

MISTAKES AND FRAUD

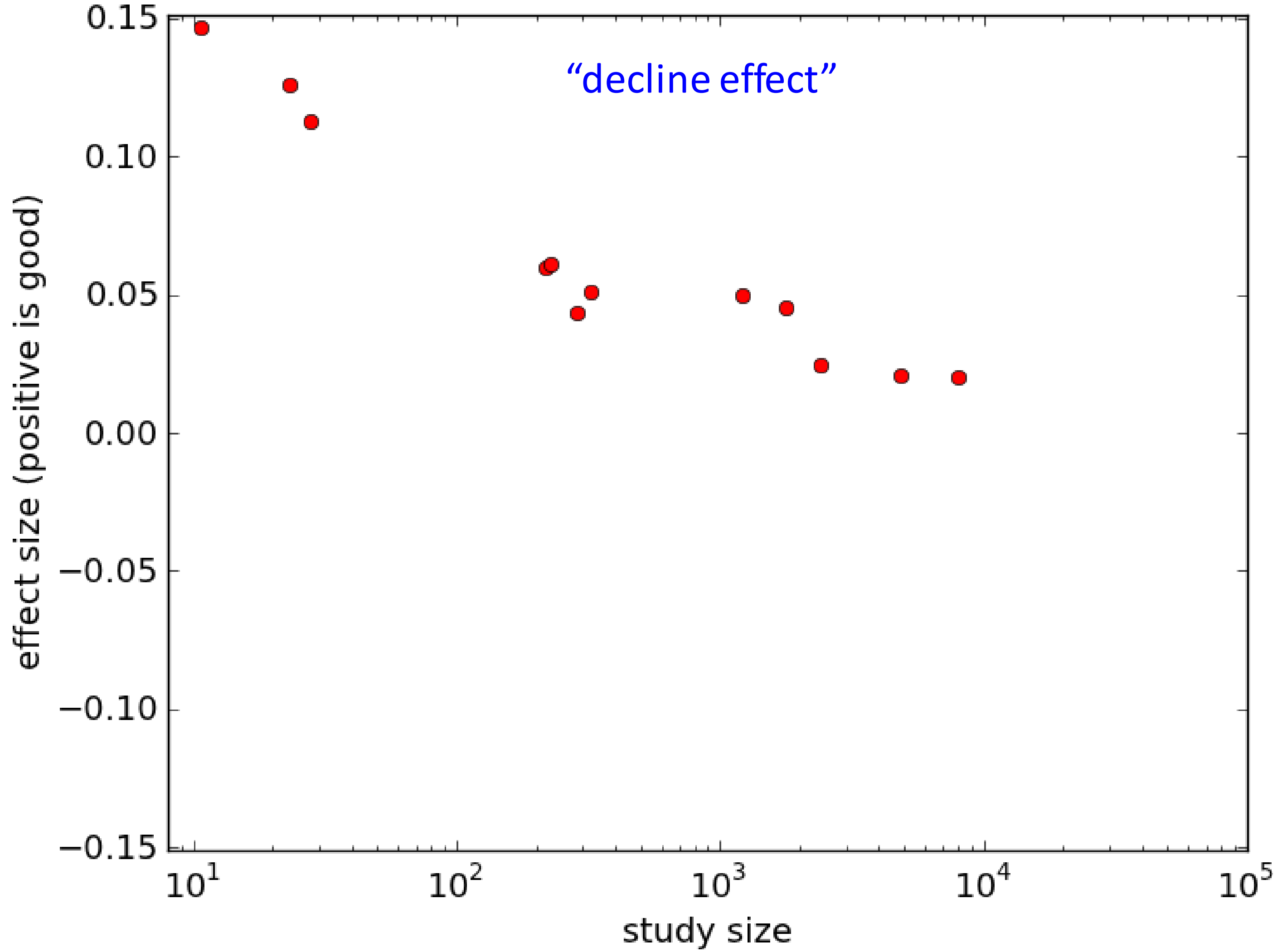
2001 – 2011:
- 10X increase in retractions
- only 1.44X increase in papers

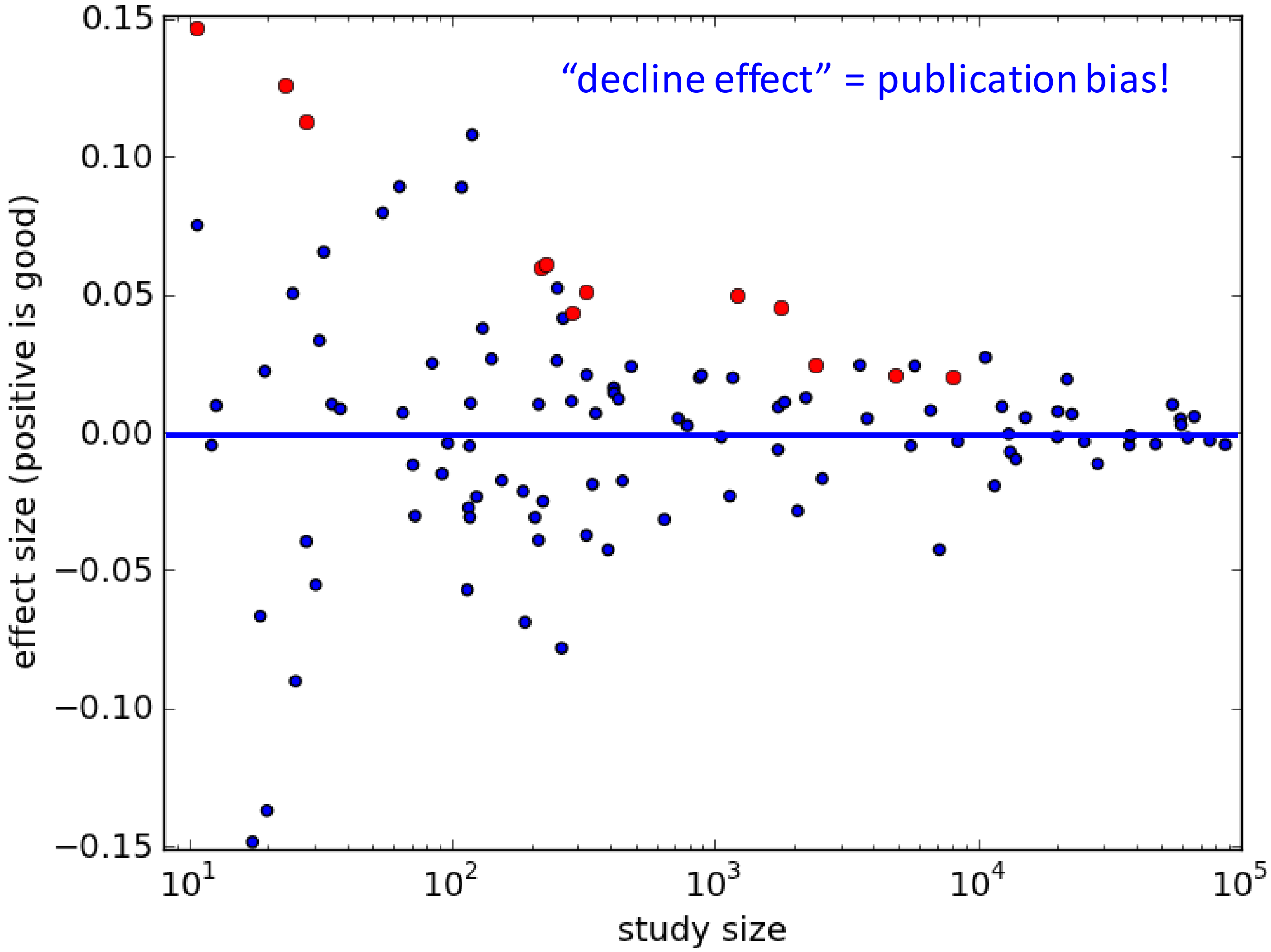


Richard Van Noorden, 2011, Nature 478
The Rise of the Retractions

http://www.nature.com/news/2011/111005/pdf/478026a.pdf

PUBLICATION BIAS





“decline effect” = publication bias!
effect size (positive is good)
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
10^1
10^2
10^3
10^4
10^5
study size

MANY ANALYSTS, ONE DATA SET

MANY ANALYSTS, ONE DATA SET

Variations in Analytic Choices Affect Results

Abstract:

"**Twenty-nine teams involving 61 analysts used the same data** set to address the same research question: **whether soccer referees are more likely to give red cards to dark-skin-toned players** than to light- skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units.

MANY ANALYSTS, ONE DATA SET

Variations in Analytic Choices Affect Results

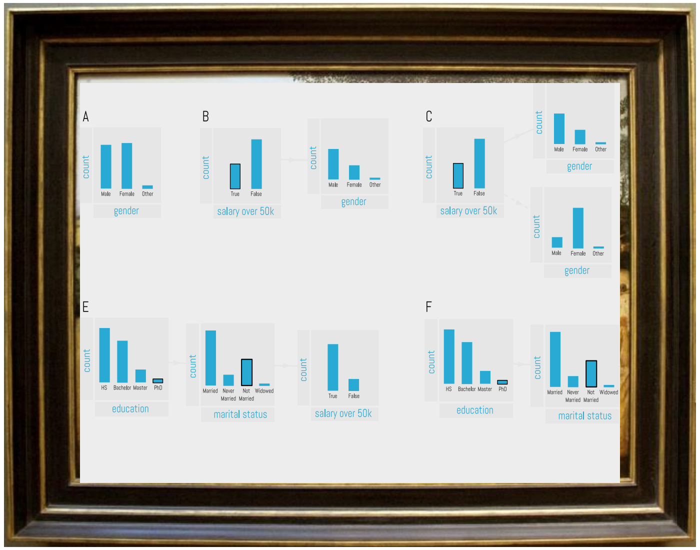
Abstract:

"**Twenty-nine teams involving 61 analysts used the same data** set to address the same research question: **whether soccer referees are more likely to give red cards to dark-skin-toned players** than to light- skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. **Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship**. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. …. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results."

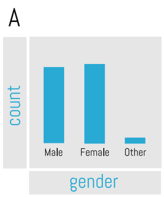
WHY VISUALIZATIONS CONTRIBUTE TO THE PROBLEM

If a visualization provides any insight, it is an hypothesis test (just one where you not necessarily know if it is statistical significant)

Otherwise, visualizations have just to be taken as pretty pictures about (potentially) random facts



IF VISUALIZATIONS ARE USED TO FIND SOMETHING INTERESTING, THE USER IS DOING MULTIPLE HYPOTHESIS TESTING



RUNNING EXAMPLE: SURVEY ON AMAZON MECHANICAL TURK

Project Name: randdb-survey This name is not displayed to Workers.

Survey about demographics, habits and opinions

Requester: Zheguang Samuel Zhao **Reward:** $2.00 per HIT **HITs available:** 0 **Duration:** 2 Days

Qualifications Required: Masters has been granted

HIT Preview

49. Your first guess of "Stonebraker" is?

- A Simpsons character
- A type of stone
- An antient Egyptian profession
- A Turing-award winner

50. Can you jump on one foot for 5 minutes non-stop?

- Yes
- No

51. Which smartphone operating system do you prefer?

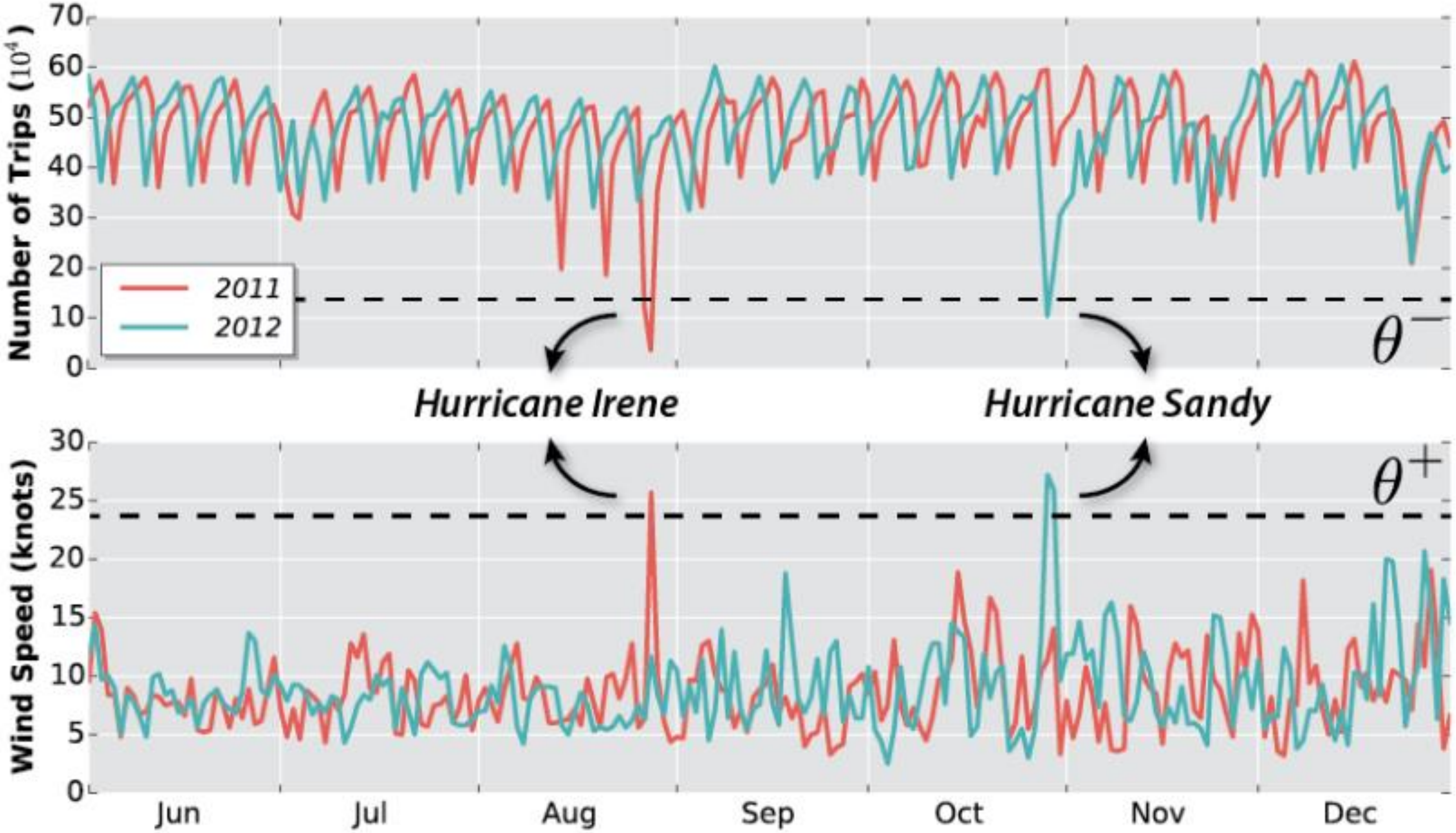
- Apple iOS
- Android

OUR GOAL: TO FIND GOOD INDICATORS (CORRELATIONS) THAT SOMEBODY KNOWS WHO MIKE STONEBRAKER IS.

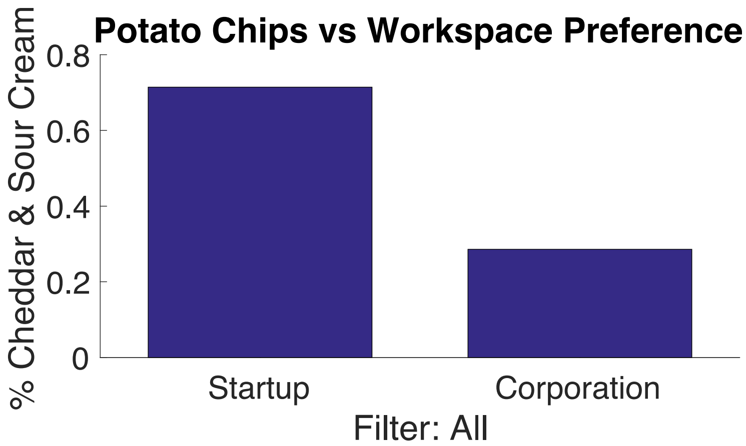
AND AFTER SEARCHING FOR A BIT, ONE OF MY FAVORITES

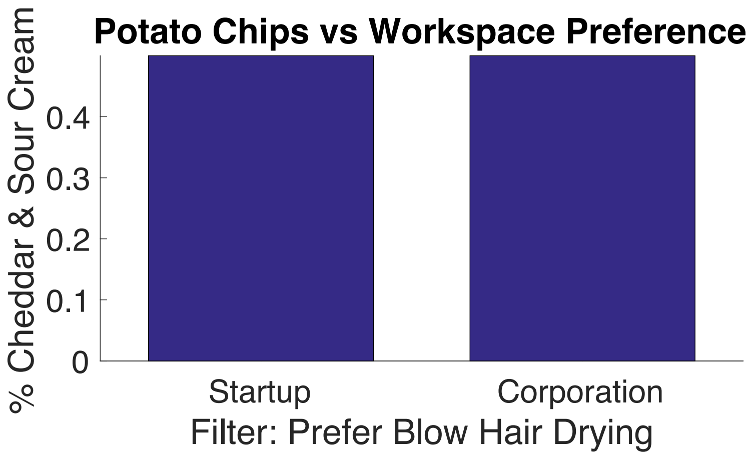
Pearson correlation significance-level $p < 0.05$

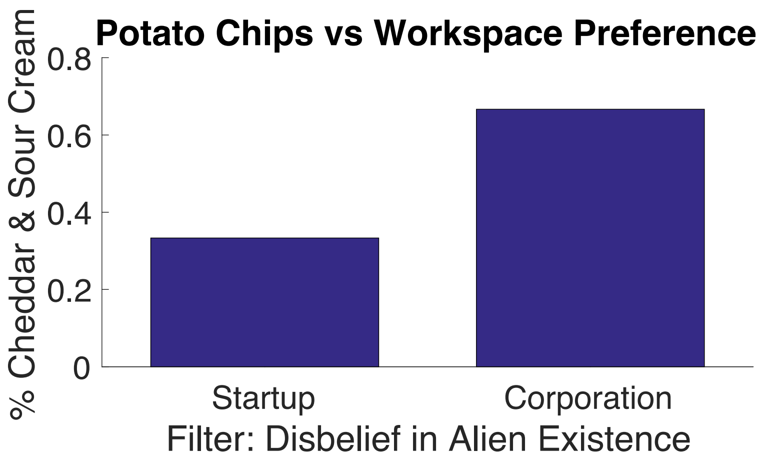
REAL HYPOTHESIS GENERATORS (DATA POLYGAMY AS AN EXAMPLE)

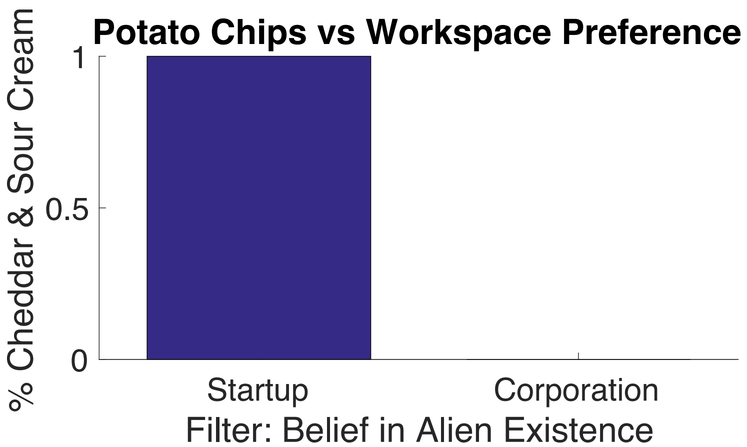


SEEDB ON OUR SURVEY DATA









My suggestions, papers should include in the future a a warning like

WARNING

After using the tool, **throw away the data**. It is not safe![1]

[1]To be more precise: you do not have to throw it all away, but you can not use the same data anymore for significance testing

What is needed is a multi-hypothesis control techniques

- Hold-out data set / Additional Tests
- Family-wise error (e.g., Bonferroni correction)
- False Discovery Rate (e.g., alpha-investing
- Permutation-based techniques
- Bayesian techniques (e.g., Bayesian FDR)
- Uniform Convergence and (Structural) Risk Minimization (more on that later)

FAMILY-WISE ERROR RATE CORRECTIONS

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

- **Either requires to know the number of tests k upfront (Bonferroni) or acceptance threshold decreases exponentially**
- **Significantly decreases the power of the test**

HOLD-OUT DATASET

- Hypothesis is tested on both D_1 (exploration dataset) and D_2 (hold-out dataset)
- Type 1 error is reduced to α^2 (as tested on both D_1 and D_2). E.g., 0.05 becomes 0.025 (assuming a single test)
- Requires multi-hypothesis control on the hold-out (for multiple tests)
- Reduces significantly the power of the test (Power of large numbers)

FALSE DISCOVERY RATE

$$\mathrm{FDR} = \mathrm{E}\left[\frac{V}{R}^{*}\right]$$

False discoveries

All discoveries

FDR-controlling procedures are designed to control the expected ratio of false discoveries among all discoveries returned by a procedure.

- Under complete null hypothesis, controlling FDR at level α guarantees also "weak control" over FWER.

$$FWER = P(V \geq 1) = E\left(\frac{V}{R}\right) = \mathrm{FDR} \leq \alpha.$$

- Not true if true discoveries exists (strong control)
- Increased power

** We define FDR to be zero when R = 0*

FALSE DISCOVERY RATE

$$\mathrm{FDR} = \mathrm{E}\left[\frac{V^*}{R}\right]$$

False discoveries

All discoveries

Benjamini-Hochberg procedure(BH)

1. Sort all p-values such that $p_1 < p_2 < \ldots. < p_m$
2. Determine the maximum k, such that $p_k < \frac{k}{m} \cdot \alpha$
3. Reject the null hypotheses corresponding to the p-values $p_1, p_2, \ldots, p_k$

** We define FDR to be zero when R = 0*

CLOSING THOUGHTS

"It is easy to lie with statistics,
but it is easier to lie without them."

attributed to Frederick Mosteller (1916-2006)

REFERENCES

- **How to lie with Statistics - Darrell Huff**
- **How to lie with Maps - Mark Monmonier**
- **http://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/**
- **Nuzzo R: Scientific method: statistical errors. Nature. 2014 Feb 13;506(7487)**
- **Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.**
- **Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13.**