Statistical Significance, Practical Significance, and P-Hacking

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

Statistical Significance vs. Practical Significance

- Statistical significance does not let us determine: Is this difference or relationship meaningful? Is it a substantial, "large enough" difference?
- Very small differences can be statistically significant too
- Need to consider practical significance

Practical Significance

- Statistical significance tells you whether an effect exists (unlikely to be due to chance)
- Practical significance tells you whether the effect matters in the real world

Sample Size and Statistical Significance

- Sample size increases → SE decreases (SE=SD divided by the square root of N)
- That makes it much easier to find significant relationships (risk of Type II error decreases)
- With small samples, SE is very large → only very large effects will be statistically significant
- With large samples, SE is very small → even very small effects can be statistically significant

Effect Size

- Effect size: unlike statistical significance, does not depend on sample size (using SD not SE)
- When comparing 2 means:

$$ES = \frac{Mean_{t1} - Mean_{t2}}{\sqrt{(SD_{t1}^2 + SD_{t2}^2)/2}}$$

- This is the difference in means in SD units
- Example: ES=0.5 means difference between two groups is ½ of standard deviation

Interpretation Benchmarks

- This measure of effect size is called Cohen's D
- Interpretation:
  - < 0.1 = tiny
  - 0.1-0.2 = very small
  - 0.2-0.5 = small
  - 0.5-0.8 = medium
  - 0.8-1.2 = large
  - 1.2-2.0 = very large
  - > 2.0 = huge

Example

- Effectiveness of TV program for reading skills:
  - experimental group mean = 12.3 (SD = 2.1)
  - control group mean= 11.2 (SD = 1.8)
- T=2, p<.05 → statistically significant
- ES=(12.3-11.2)/sqrt((2.1^2 + 1.8^2)/2)=0.56
- 0.56 = medium effect size → practically significant

Coffee Example

- Does drinking coffee before an exam improve students' test scores?
- Randomly assigning 1000 college students:
  - **Group A (Coffee):** 1 cup of coffee 30 min before the test
  - **Group B (Decaf):** 1 cup of decaf coffee (placebo)
- Two independent sample means, t-test

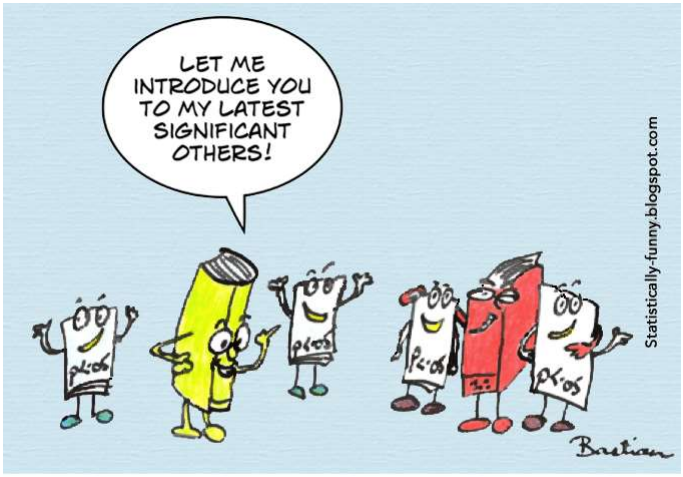
Group	n	Mean Score	Standard Deviation
A: Coffee	500	81.8	6.0
B: Decaf	500	81.0	6.2

Coffee Example Calculations

```
. ttesti 500 81.8 6.0 500 81 6.2

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       x |     500        81.8    .2683282           6    81.27281    82.32719
       y |     500          81    .2772724         6.2    80.45523    81.54477
---------+--------------------------------------------------------------------
Combined |   1,000        81.4    .1932431    6.110884    81.02079    81.77921
---------+--------------------------------------------------------------------
    diff |                .8    .3858497                .0428302     1.55717
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   2.0733
H0: diff = 0                                     Degrees of freedom =      998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9808         Pr(|T| > |t|) = 0.0384          Pr(T > t) = 0.0192

. *calculating effect size
. di 0.8/sqrt((6^2+6.2^2)/2)
.13112992
```

Coffee Example Conclusions

- t=2.07, p=0.038 → statistically significant at .05 level
- ES=0.13 → very small
- Not practically significant

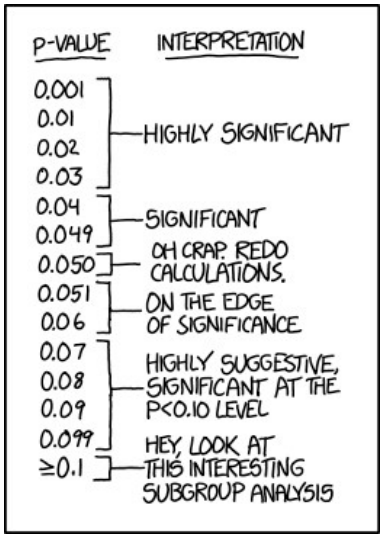


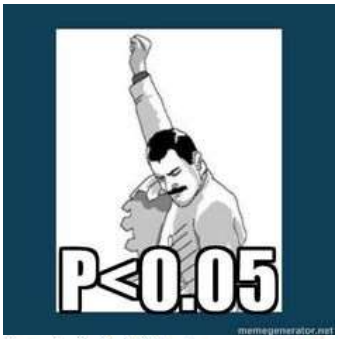
IT WAS GETTING HARDER AND HARDER TO FIND A TRULY MEANINGFUL RELATIONSHIP AT THE MEDICAL JOURNAL HAPPY HOUR.

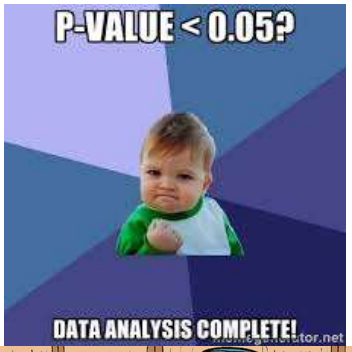
Debate on P-values and Significance

- https://www.nature.com/articles/s41562-017-0189-z
- "We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries."
- http://fivethirtyeight.com/features/statisticians-found-one-thing-they-can-agree-on-its-time-to-stop-misusing-p-values/
- "The p-value was never intended to be a substitute for scientific reasoning."

Excessive Focus on P-value Cutoffs







Reproducibility Crisis in Science

- https://www.wired.com/story/social-science-reproducibility/
- Reproducibility Project: https://en.wikipedia.org/wiki/Reproducibility_Project
- 100 psychology studies: only 36% successfully replicated
- 193 experiments from cancer studies: only 26% successfully replicated
- Similar issues found in economics, political science, sociology…

Key Reasons for Reproducibility Crisis

- Publication bias (journals prefer positive results) → "File drawer" problem
- Novelty is valued over replication
- Lack of transparency in methods and data
- Lack of study design pre-registration
- P-hacking

What is P-hacking?

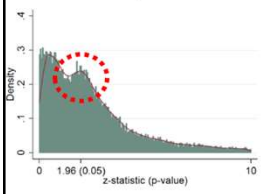
- Manipulating data analysis until you get p-value <.05
- Also known as "data dredging" or "fishing for significance"
- Driven by incentives to publish "positive" results
  - Dropping outliers that "ruin" significance (cognitive biases: "I am just cleaning the data")
  - Testing multiple variables or subsets of data, reporting only p<.05
  - Trying different models & analyses until p<.05 appears
  - Stopping data collection once p<.05 is achieved
- Result: The reported p-value ≠ true probability of the result under the null
- How Scientists Manipulate Research with P-values https://www.youtube.com/watch?v=kTMHruMz4Is

Some Evidence of P-hacking

Economics

Brodeur et al (AEJ:A, in press)

"Star Wars: The empirics strike back"

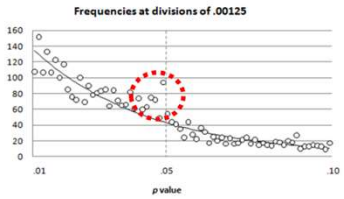


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

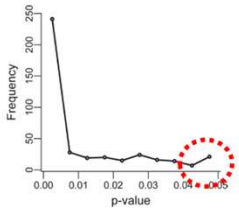
"A peculiar prevalence of p values just below .05"

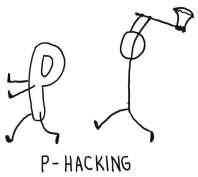


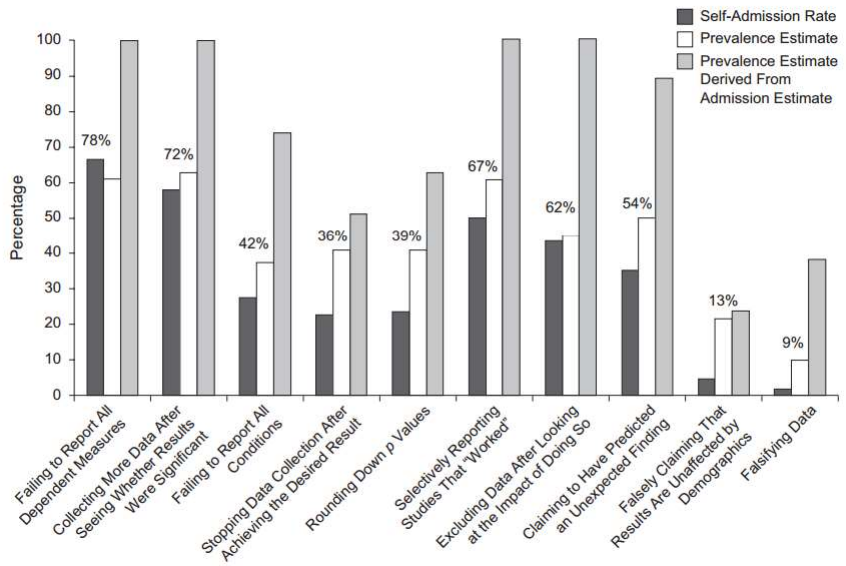
Biology

Head et al (PLOS Biology 2015)

"Extent and Consequences of P-Hacking in Science"







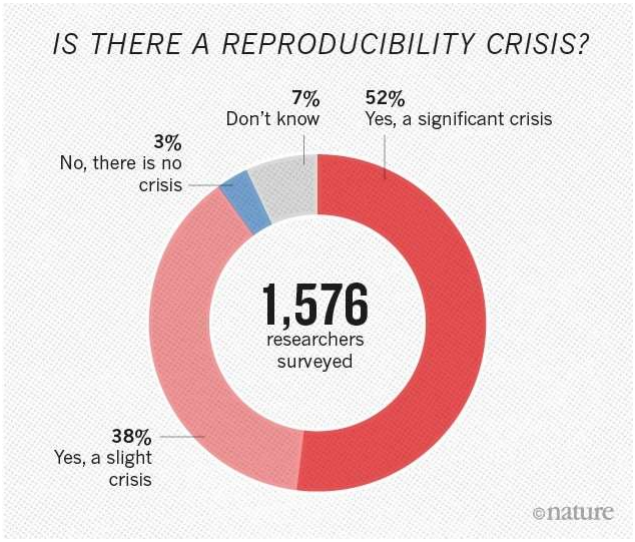
- John, L.K. et al. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." (anonymous survey of 2000 psychologists). https://journals.sagepub.com/doi/pdf/10.1177/0956797611430953

Possible Solutions

- Pre-registration of hypotheses, methods, analysis plans
- Open data and open code for transparency
- "Registered reports" publication model
- Replication incentives (funding, publication)
- Better statistical education
- If using multiple comparisons/variables, apply statistical corrections
- Just report p-values and effect sizes – no p-value cutoffs, no "statistical significance"

Moving to a World Beyond "p<.05"

- https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913
- "A label of statistical significance adds nothing to what is already conveyed by the value of *p*; in fact, this dichotomization of *p*-values makes matters worse."



- From: https://www.nature.com/articles/533452a