Hypothesis Testing: Do's and Don'ts

Vanessa LoBue

Jamil Bhanji

with a little help from Andy Field

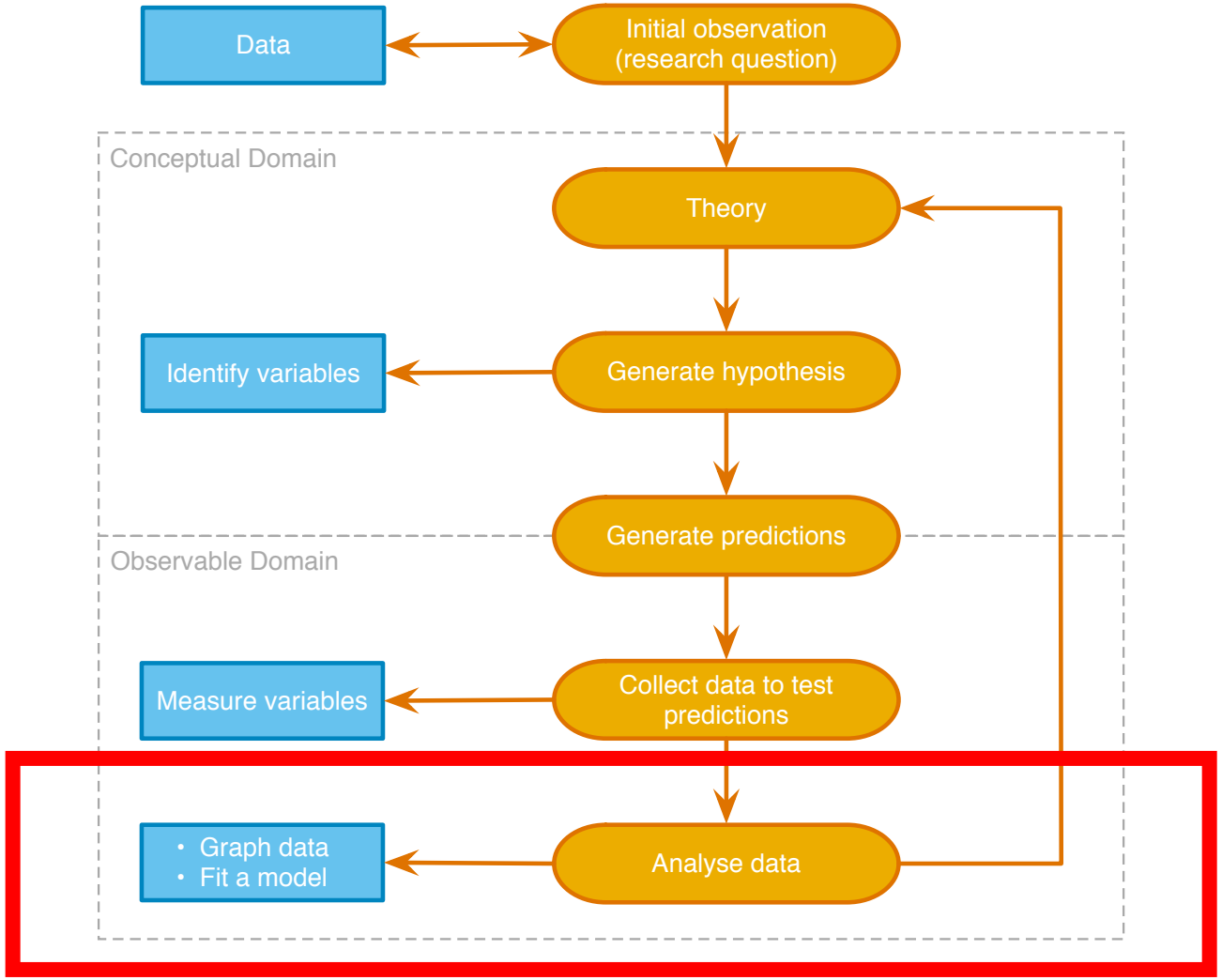
Reminders

- Submit your data set (Oct 13) – use the “Assignments” submission page
  - think about how you will organize the data dictionary

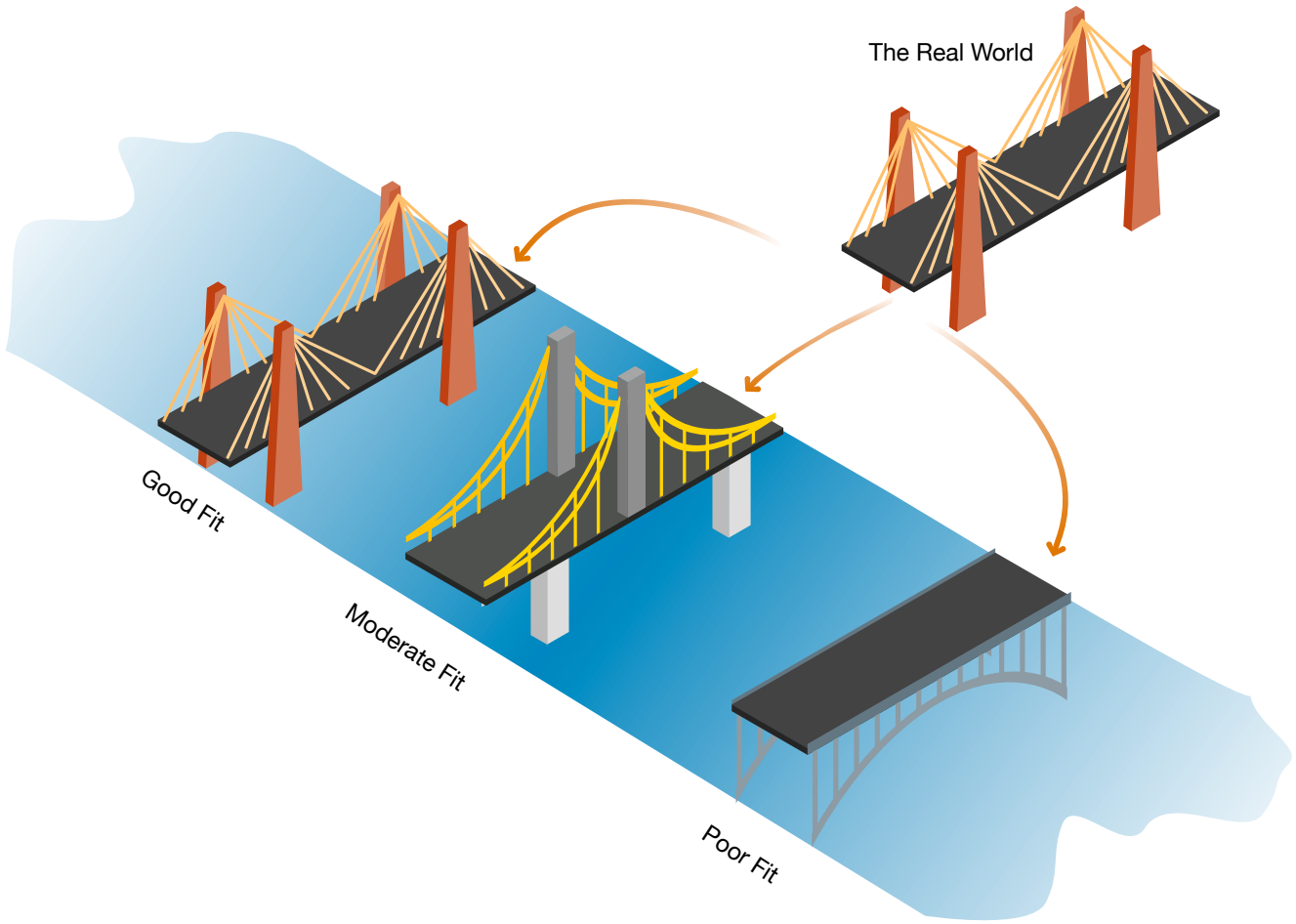
Aims

- Know what a statistical model is and why we use them
- Know what the ‘fit’ of a model is and why it is important.
- Distinguish models for samples and populations
- Problems with NHST and best practices

The Research Process



Building Statistical Models



Most Important Equation

$$\text{Outcome}_i = (\text{Model}) + \text{error}_i$$

A Simple Statistical Model

- In Statistics we fit models to our data (i.e. we use a statistical model to represent what is happening in the real world)
- The mean is a hypothetical value (i.e. it doesn't have to be a value that actually exists in the data set)
- As such, the mean is simple statistical model

Measuring the ‘fit’ of the model

- The mean is a *model* of what happens in the real world: the *typical* score
- It is not a perfect representation of the data
- How can we assess how well the mean represents reality?

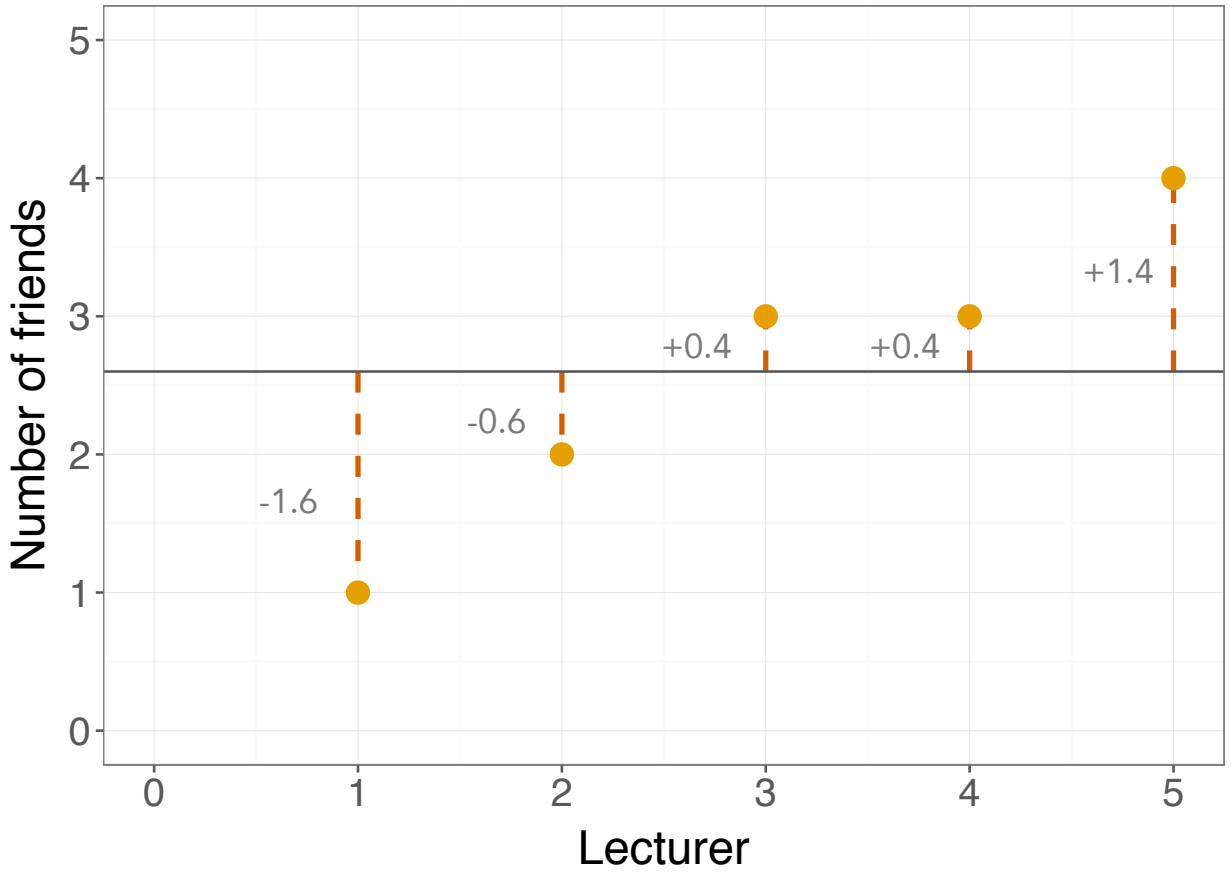
Calculating ‘Error’

- A deviation is the difference between the mean and an actual data point.

- Deviations can be calculated by taking each score and subtracting the mean from it:

$$\text{deviance} = \text{outcome}_i - \text{model}_i$$

Calculating ‘Error’



Use the Total Error?

- We could just take the error between the mean and the data and add them.

Score	Mean	Deviation
1	2.6	-1.6
2	2.6	-0.6
3	2.6	0.4
3	2.6	0.4
4	2.6	1.4
	Total =	0

Sum of Squared Errors

- We could add the deviations to find out the total error.
- Deviations cancel out because some are positive and others negative.
- Therefore, we square each deviation.

- If we add these squared deviations we get the Sum of Squared Errors (*SS*).

Sum of Squared Errors

Score	Mean	Deviation	Squared Deviation
1	2.6	-1.6	2.56
2	2.6	-0.6	0.36
3	2.6	0.4	0.16
3	2.6	0.4	0.16
4	2.6	1.4	1.96
		Total	5.20

$$SS = \sum (X - \bar{X})^2 = 5.20$$

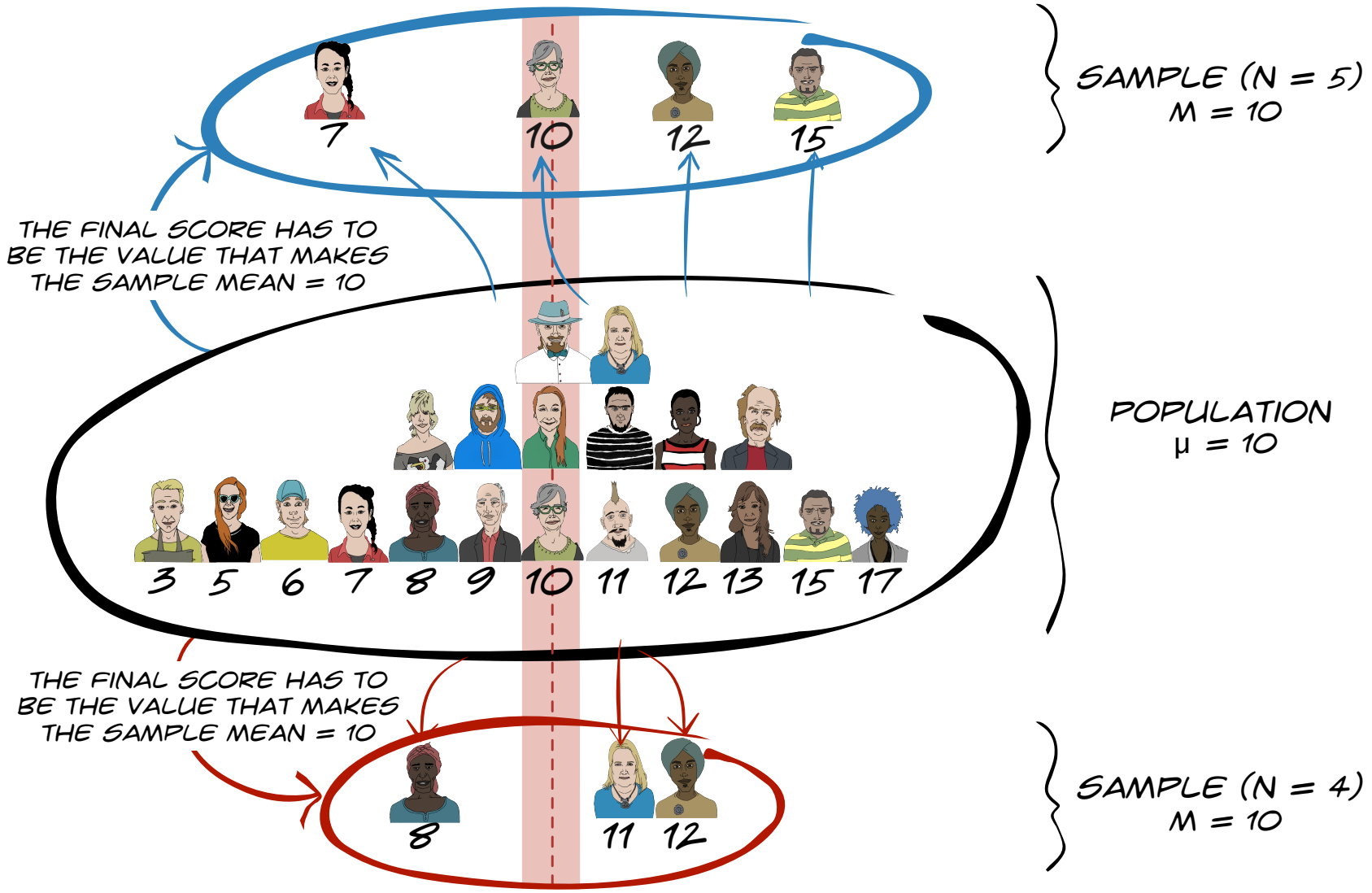
Mean Squared Error

- Although the SS is a good measure of the accuracy of our model, it depends on the amount of data collected. To overcome this problem, we use:

$$\text{mean squared error} = \frac{SS}{df} = \frac{\sum_{i=1}^{n} \left(\text{outcome}_i - \text{model}_i\right)^2}{N-1}$$

$$\text{mean squared error} = \frac{SS}{df} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.30$$

Degrees of Freedom



The Standard Error

- SD tells us how well the mean represents the sample data

- But, if we want to estimate this parameter in the population, then we need to know how well the mean represents the population

Populations and Samples

- Population
  - The collection of units (be they people, plankton, plants, cities, suicidal authors, etc.) to which we want to generalize a set of findings or a statistical model

- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population

Populations and Samples

- Sample
  - Mean and SD describe only the sample from which they were calculated
- Population
  - Mean and SD are intended to describe the entire population (very rare in Psychology)
- Sample to population:
  - Mean and SD are obtained from a sample, but are used to estimate the mean and SD of the population (very common in psychology)

The Standard Error

- SD tells us how well the mean represents the sample data

- SE tells us how far the sample mean is likely to be from the true population mean

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

Hypothesis Testing

- Null hypothesis, H_0
  - There is no effect.
  - E.g. Big Brother contestants and members of the public will not differ in their scores on personality disorder questionnaires

- The alternative hypothesis, H_1
  - AKA the experimental hypothesis
  - E.g. Big Brother contestants will score higher on personality disorder questionnaires than members of the public

Test Statistics

- A Statistic for which the frequency of particular values is known.

- Observed values can be used to test hypotheses.

$$\text{Test statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence rep-

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716

NHST and wider problems in science

- Incentive structures and publication bias
- Researcher degrees of freedom
- *p*-hacking and HARKing

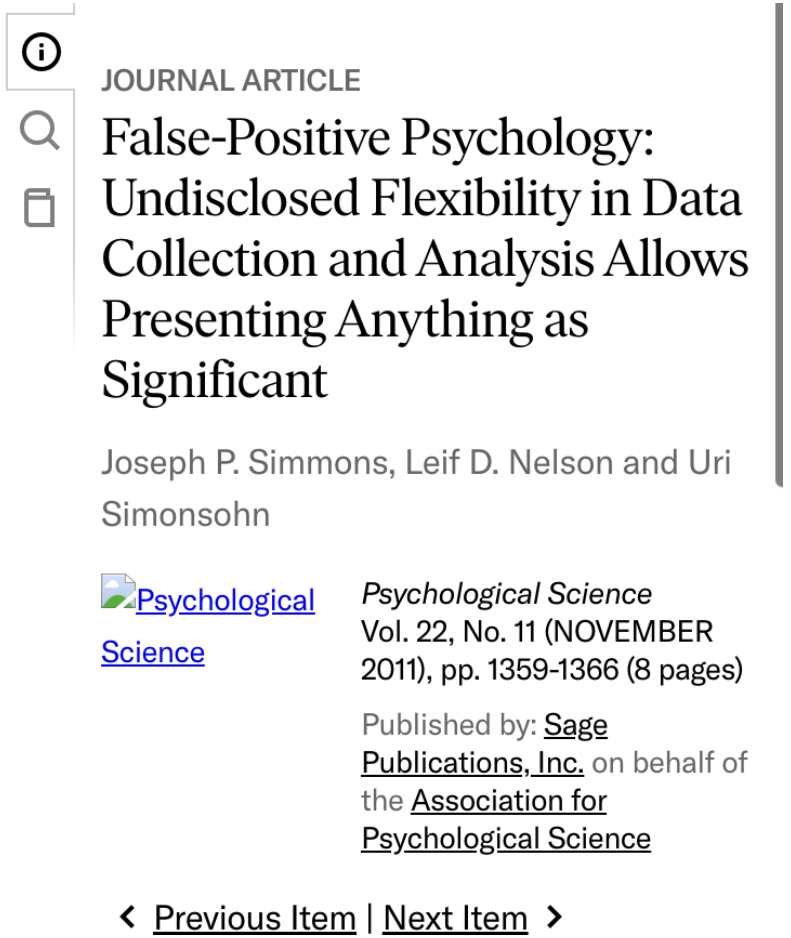
Incentive Structures

- File drawer effect—null results don’t get published
- Shrinking job market
- High impact journals favor “flashy” findings

Researcher degrees of freedom

- A scientist has many decisions to make when designing and analysing a study.
  - The alpha level, the level of power, how many participants should be collected, which statistical model to fit, how to deal with extreme scores, which control variables to consider, which measures to use, and so on

- Researchers might use these researcher degrees of freedom to shed their results in the most favourable light



JOURNAL ARTICLE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons, Leif D. Nelson and Uri Simonsohn

Psychological Science

Psychological Science
Vol. 22, No. 11 (NOVEMBER 2011), pp. 1359-1366 (8 pages)

Published by: Sage Publications, Inc. on behalf of the Association for Psychological Science

‹ Previous Item | Next Item ›

p-hacking

- *p*-hacking—researcher degrees of freedoms that lead to the selective reporting of significant *p*-values

    - Trying multiple analyses/measuring multiple *outcomes* but reporting only the significant results
    - Stopping data collection at a point other than when the pre-determined sample size is reached
    - Including (or not) data based on the effect they have on the *p*-value

p-hacking

- *p*-hacking—researcher degrees of freedoms that lead to the selective reporting of significant *p*-values
    - Including (or excluding) variables in an analysis based on how those variables affect the *p*-value
    - Merging groups of variables or scores to yield significant results
    - Transforming, or otherwise manipulating scores to yield significant *p*-values

HARKing

- HARKing
  - The practice in research articles of presenting a hypothesis that was made *after* data collection as though it were made *before* data collection

Throw Away NHST?

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–10

DOI: 10.1177/1745691620958012
www.psychologicalscience.org/PPS
SAGE

The Practical Alternative to the *p* Value Is the Correctly Used *p* Value

Daniël Lakens
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

Abstract

Because of the strong overreliance on *p* values in the scientific literature, some researchers have argued that we need to move beyond *p* values and embrace practical alternatives. When proposing alternatives to *p* values statisticians often commit the "statistician's fallacy," whereby they declare which statistic researchers really "want to know." Instead of telling researchers what they want to know, statisticians should teach researchers which questions they can ask. In some situations, the answer to the question they are most interested in will be the *p* value. As long as null-hypothesis tests have been criticized, researchers have suggested including minimum-effect tests and equivalence tests in our statistical toolbox, and these tests have the potential to greatly improve the questions researchers ask. If anyone believes *p* values affect the quality of scientific research, preventing the misinterpretation of *p* values by developing better evidence-based education and user-centered statistical software should be a top priority. Polarized discussions about which statistic scientists should use has distracted us from examining more important questions, such as asking researchers what they want to know when they conduct scientific research. Before we can improve our statistical inferences, we need to improve our statistical questions.

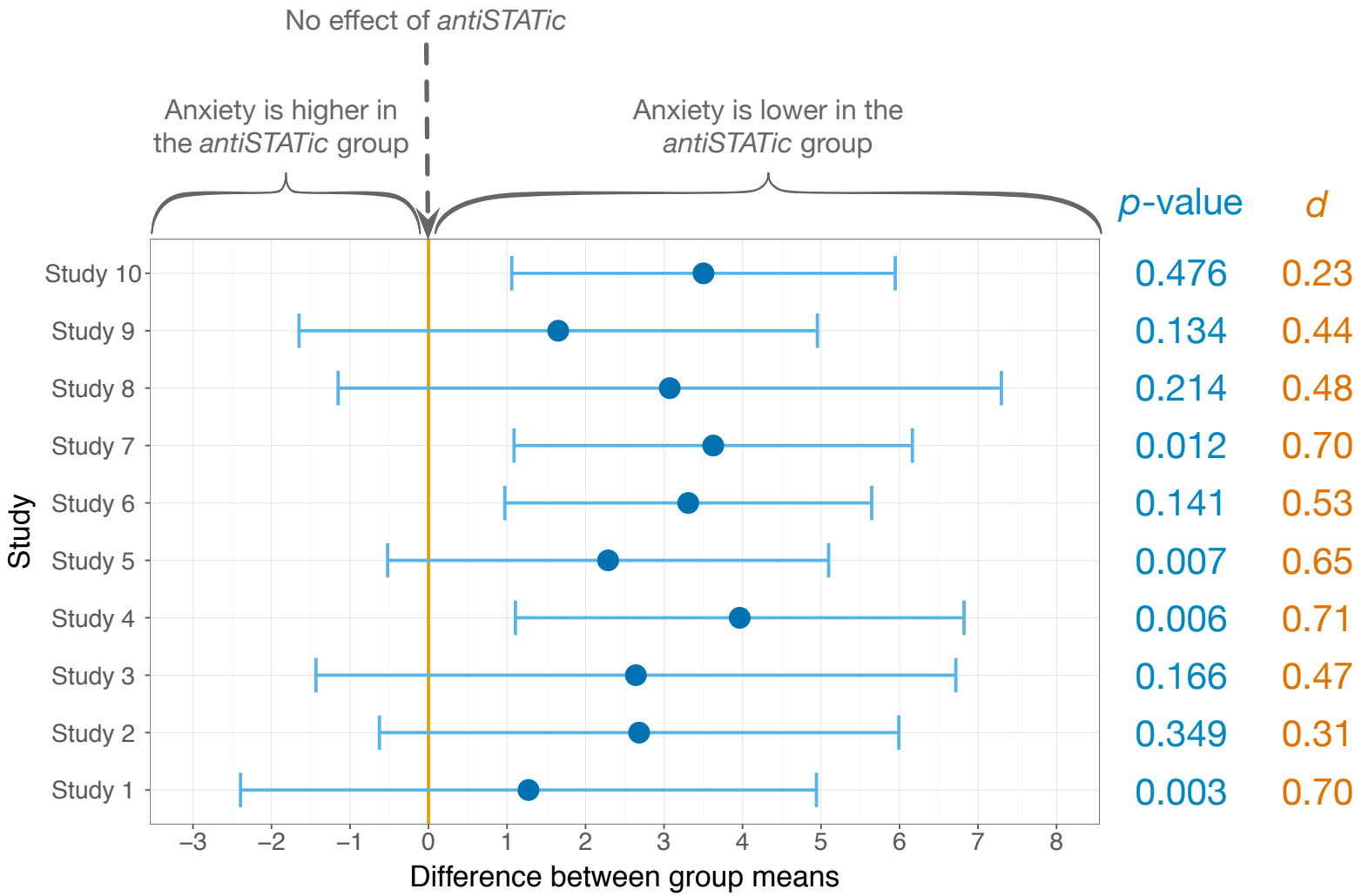
Throw Away NHST?

“I do not think that it is useful to tell researchers what they want to know. Instead, we should teach them the possible questions they can ask…Unless we examine which questions researchers ask, depending on the goals they have when they perform a study, the phase of the research line, the knowledge that already exists on the topic, and the philosophy of science that researchers subscribe to, it is impossible to draw conclusions about the statistical approach that gives the most useful answer.” Lakens (2021)

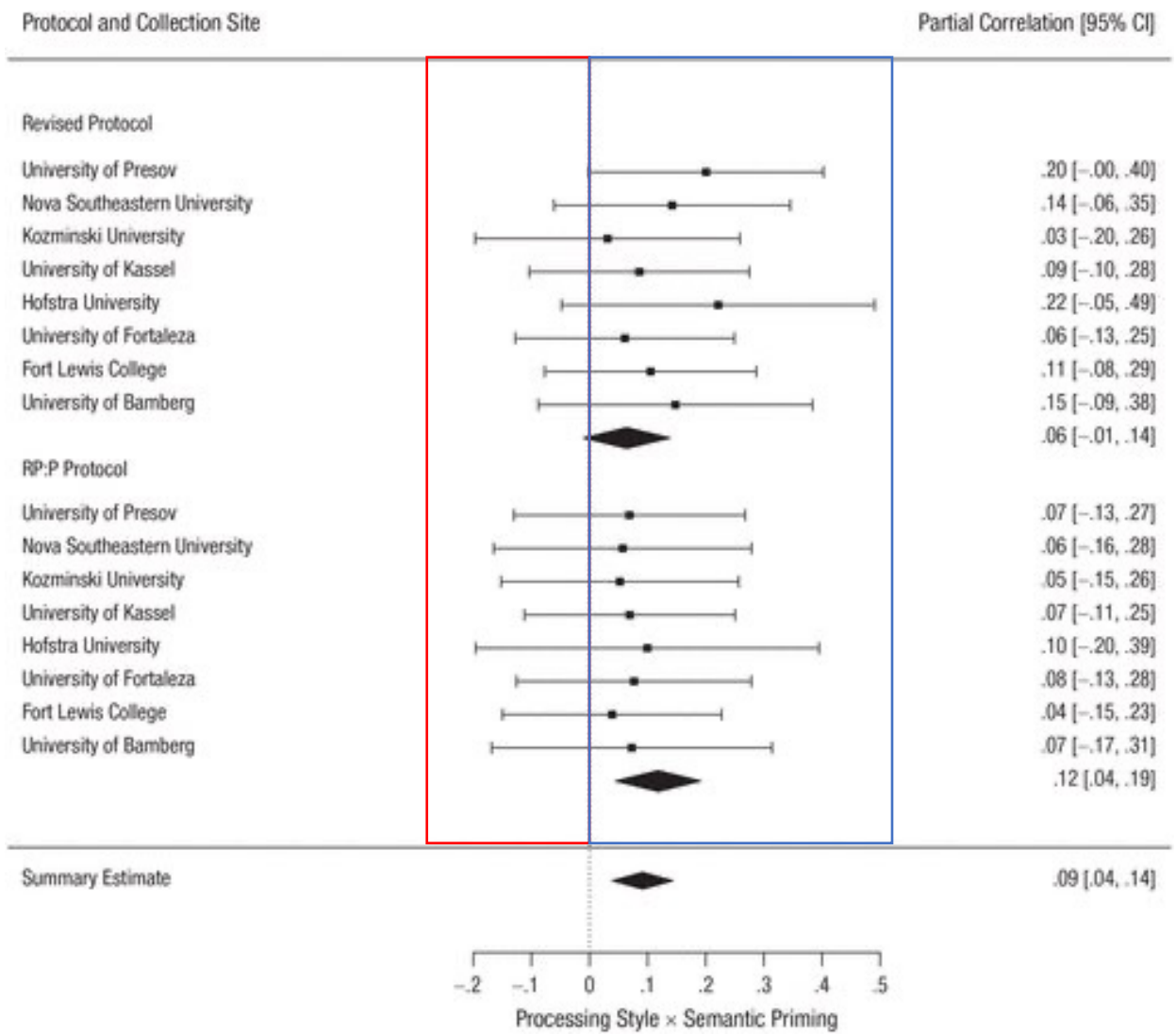
Misconceptions around *p*-values

- **Misconception 1: A significant result means that the effect is important**
  - No, because significance depends on sample size.
- **Misconception 2: A non-significant result means that the null hypothesis is true**
  - No, a non-significant result tells us only that the effect is not big enough to be found (given our sample size), it doesn't tell us that the effect size is zero.
- **Misconception 3: A significant result means that the null hypothesis is false?**
  - No, it is logically not possible to conclude this.

Avoid All-or-Nothing Thinking



ManyLabs 5


Protocol and Collection Site
Partial Correlation [95% CI]
Revised Protocol
University of Presov
.20 [−.00, .40]
Nova Southeastern University
.14 [−.06, .35]
Kozminski University
.03 [−.20, .26]
University of Kassel
.09 [−.10, .28]
Hofstra University
.22 [−.05, .49]
University of Fortaleza
.06 [−.13, .25]
Fort Lewis College
.11 [−.08, .29]
University of Bamberg
.15 [−.09, .38]
.06 [−.01, .14]
RP:P Protocol
University of Presov
.07 [−.13, .27]
Nova Southeastern University
.06 [−.16, .28]
Kozminski University
.05 [−.15, .26]
University of Kassel
.07 [−.11, .25]
Hofstra University
.10 [−.20, .39]
University of Fortaleza
.08 [−.13, .28]
Fort Lewis College
.04 [−.15, .23]
University of Bamberg
.07 [−.17, .31]
.12 [.04, .19]
Summary Estimate
.09 [.04, .14]
−.2
−.1
0
.1
.2
.3
.4
.5
Processing Style × Semantic Priming

Power

- Type I error
  - Occurs when we believe that there is a genuine effect in our population, when in fact there isn't.
  - The probability is the α-level (usually 0.05)

- Type II error
  - Occurs when we believe that there is no effect in the population when, in reality, there is.
  - The probability is the β-level (often 0.2)

Power

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error (α) Common values: 0.05, 0.01	Power (1-β) Common values: 0.8, 0.9
Fail to reject null hypothesis	Confidence Interval (1-α) Common values: 0.95, 0.99	Type II error (β)

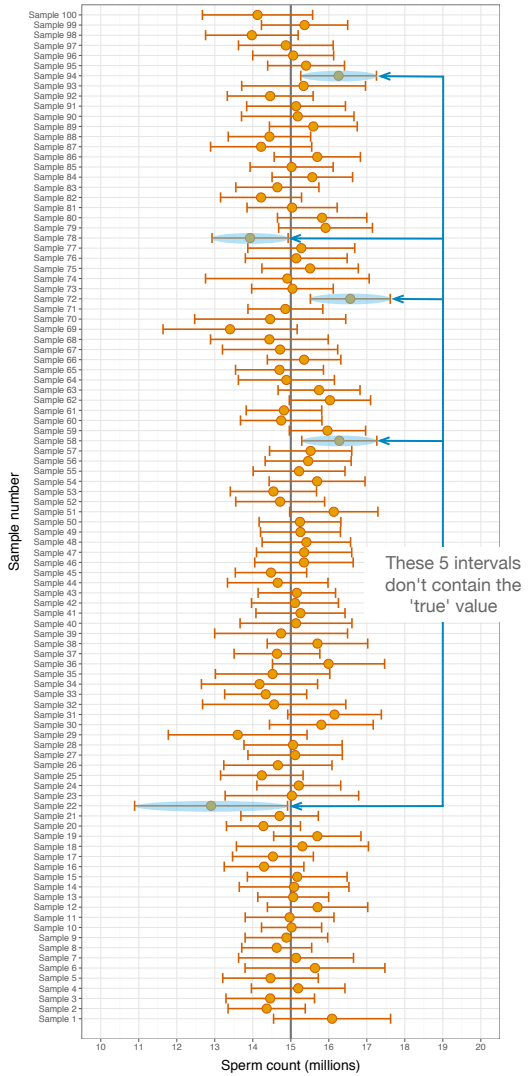
Type I error (α) - probability of finding an effect that is not there

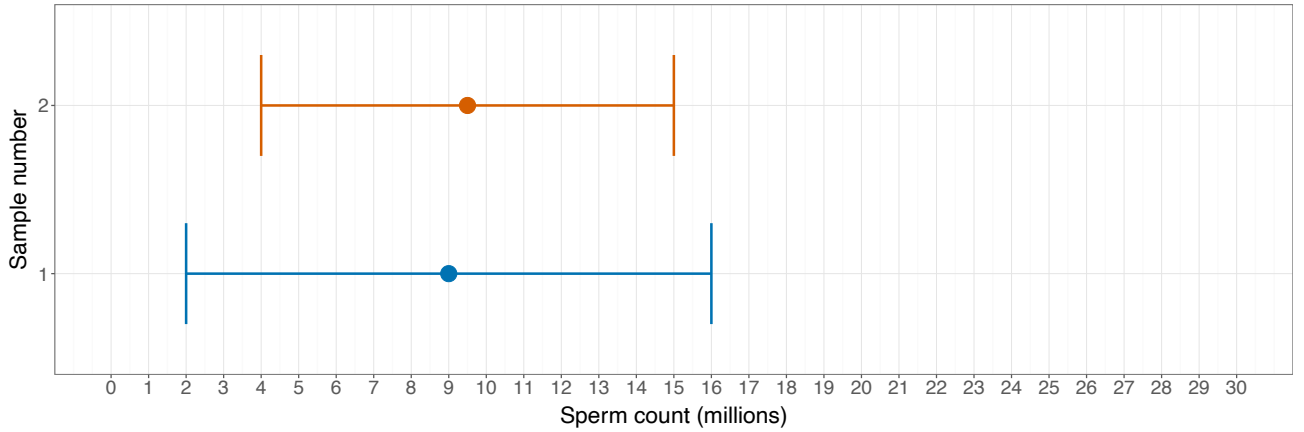
Type II error (β) - probability of not finding an effect that is there

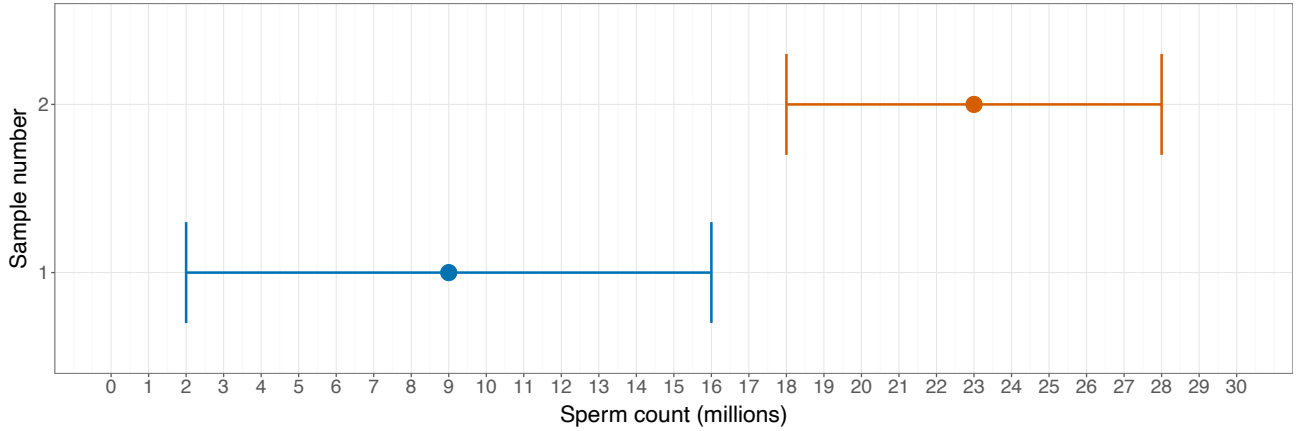
Conducting a Power Analysis

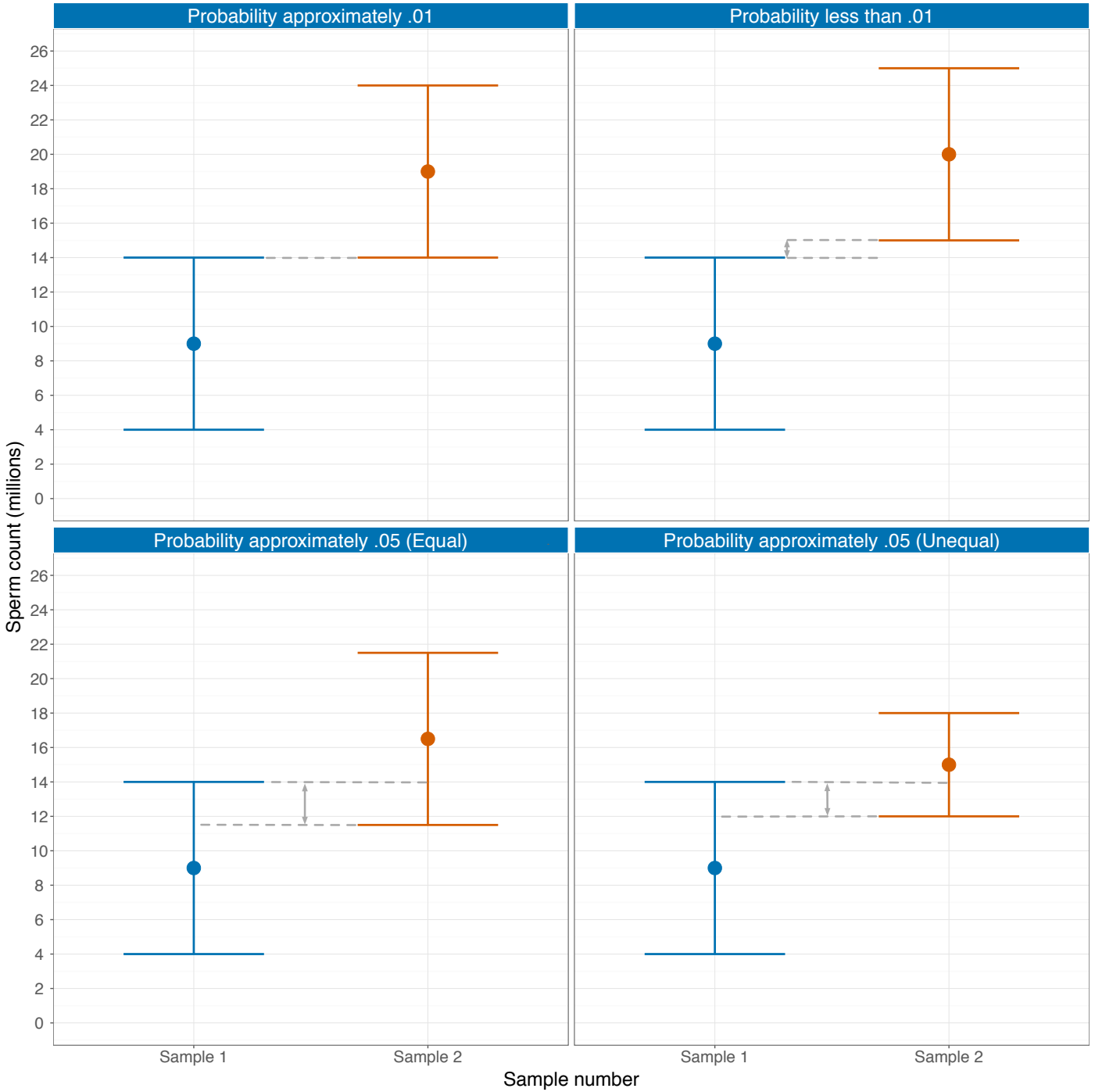
- Literature Review
  - Analyze information from papers with similar methods to get estimate of effect size
- Pilot Study
  - Allow you to get rough estimate of effect size
- Cohen's Recommendation
    - Not super reliable but common
    - Cohen d of .2 is small, .5 is medium, and .8 is large

Showing Confidence Intervals Visually









Confidence intervals and statistical significance

Report Effect sizes

- An effect size is a standardized measure of the size of an effect:
  - Standardized = comparable across studies
  - Not (as) reliant on the sample size
  - Allows people to objectively evaluate the size of observed effect.
  - They encourage interpreting effects on a continuum and not applying a categorical decision rule such as ‘significant’ or ‘not significant’

Effect Size Measures

- There are several effect size measures that can be used:
  - Cohen's *d*
  - Pearson's *r*
  - Odds Ratio/Risk rates

$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

Effect Size Measures

- *r* = .1, *d* = .2 (small effect):
  - the effect explains 1% of the total variance.
- *r* = .3, *d* = .5 (medium effect):
  - the effect accounts for 9% of the total variance.
- *r* = .5, *d* = .8 (large effect):
  - the effect accounts for 25% of the variance.

- Beware of these ‘canned’ effect sizes though:
  - The size of effect should be placed within the research context.

Preregistration

- Pre-registration of research
  - The practice of making all aspects of your research process (rationale, hypotheses, design, data processing strategy, data analysis strategy) publicly available before data collection begins
  - Registered reports in an academic journal
    - If the protocol is deemed to be rigorous enough and the research question novel enough, the protocol is accepted by the journal typically with a guarantee to publish the findings no matter what they are
  - Public websites (e.g., the Open Science Framework, AsPredicted)

Best Practices

- The ASA statement on *p*-values (Wasserstein & American Statistical Association, 2016).
  - The ASA points out that *p*-values *can* indicate how incompatible the data are with a specified statistical model (e.g., how incompatible the data are with the null hypothesis. You are at liberty to use the degree of incompatibility to inform your own beliefs about the relative plausibility of the null and alternative hypotheses, as long as you don't interpret *p*-values as a measure of the probability that the hypothesis in question is true. They are also not the probability that the data were produced by random chance alone.
  - Scientific conclusions and policy decisions *should not* be based only on whether a *p*-value passes a specific threshold.

Best Practices

- The ASA statement on *p*-values (Wasserstein & American Statistical Association, 2016).
  - Don’t *p*-hack. Be fully transparent about the number of hypotheses explored during the study, and all data collection decisions and statistical analyses.
  - Don’t confuse statistical significance with practical importance. A *p*-value does not measure the size of an effect and is influenced by the sample size, so you should never interpret a *p*-value in any way that implies that it quantifies the size or importance of an effect.
  - ‘By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.’

Why make your work reproducible?

- Avoid making mistakes!
- Easier to write up results
- Helps reviewers see it your way
- Easier to remember what you did (continuity of your work)
- Helps build your reputation

Aims

- Know what a statistical model is and why we use them
- Know what the ‘fit’ of a model is and why it is important.
- Distinguish models for samples and populations
- Problems with NHST and best practices