Week 1 - Introduction, Open Science, and Power

Andrew Stewart

Andrew.Stewart@manchester.ac.uk

@ajstewart_lang

https://github.com/ajstewartlang

Week	Topic
1	Introduction, Open Science, and Power
2	Introduction to R
3	Data Wrangling and Visualisation
4	General Linear Model - Regression
5	General Linear Model - Regression
6	No Timetabled Lecture - Reading Week
7	Consolidation Lab
8	General Linear Model - ANOVA
9	General Linear Model - ANOVA
10	Tidy Thursday Data Wrangling & Visualisation Challenge
11	Reproducing your Computational Environment using Binder
12	Dynamic, Reproducible Presentations Using xaringan

Semester 1 Assignments

Data wrangling and visualisation – Due December 5th

ANOVA/ANCOVA – Due January 17th

This Unit

- Everything we cover in this Unit will be taught following the principles of Open Science.
- All of the statistical analyses you do will be conducted using the R open source data science language.
- We will cover core topics in Statistics for Psychology with an emphasis on reproducibility and transparency.
- You will learn how to produce reports in R Markdown - these reports contain your analysis code, your output and narrative describing what it all means.
- You will learn how to use GitHub and Binder for full computational reproducibility.
- You will lear how to use the `xaringan` package for reproducible presentations.
- This Unit provides a foundation for the Workshops Unit next Semester where we will look at more advanced statistical techniques incl. mixed models and Bayesian statistics.

- The sessions will be a blend of seminars and hands-on labs.
- If you have a laptop, I recommend you use that rather than the cluster PCs.
- Today we will cover the problems around low experimental power.
- Next week we'll look at the RStudio interface and you'll run through some R code in a script I've written.
- We'll then look at data wrangling and data visualisation (incl. animated graphs).
- We'll explore AN(C)OVA and regression in the context of the General Linear Model.
- We'll take part in a data wrangling challenge and you'll learn how to develop a fully reproducible workflow.

Assessment

- The assessment for this Unit is via two coursework assignments, each weighted 50%.
- The first is on data wrangling and visualisation, and the second on ANOVA.
- For both you'll need to write your assessment using R Markdown (which we'll cover in a week or two).

Replication and Reproducibility in Science

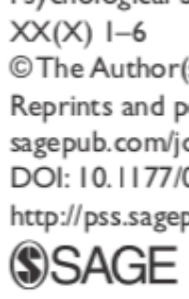
- Ionnidis (2005), *PLOS Medicine*, most published research findings are false.
- Prinz et al. (2011), *Nature Reviews Drug Discovery*, around 65% of cancer biology studies do not replicate.
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience.
- MacLeod et al. (2014), *Lancet*, 85% of biomedical research resources are wasted.
- Baker (2015), *Nature*, 90% of scientists recognise a ‘reproducibility crisis’.
- Nosek & Errington (2017), *eLife*, out of first 5 replication attempts of preclinical cancer biology work, only 2 have replicated.


SECURITY
STABILITY
OPPORTU

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

Psychological Science
XX(X) 1–6

DOI: 10.1177/0956797610383437
http://pss.sagepub.com



Dana R. Carney[1], Amy J.C. Cuddy[2], and Andy J. Yap[1]
[1]Columbia University and [2]Harvard University

Abstract

Humans and other animals express power through open, expansive postures, and they express powerlessness through closed, contractive postures. But can these postures actually cause power? The results of this study confirmed our prediction that posing in high-power nonverbal displays (as opposed to low-power nonverbal displays) would cause neuroendocrine and behavioral changes for both male and female participants: High-power posers experienced elevations in testosterone, decreases in cortisol, and increased feelings of power and tolerance for risk; low-power posers exhibited the opposite pattern. In short, posing in displays of power caused advantaged and adaptive psychological, physiological, and behavioral changes, and these findings suggest that embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices. That a person can, by assuming two simple 1-min poses, embody power and instantly become more powerful has real-world, actionable implications.

Power Posing - 2010 *vs.* 2016

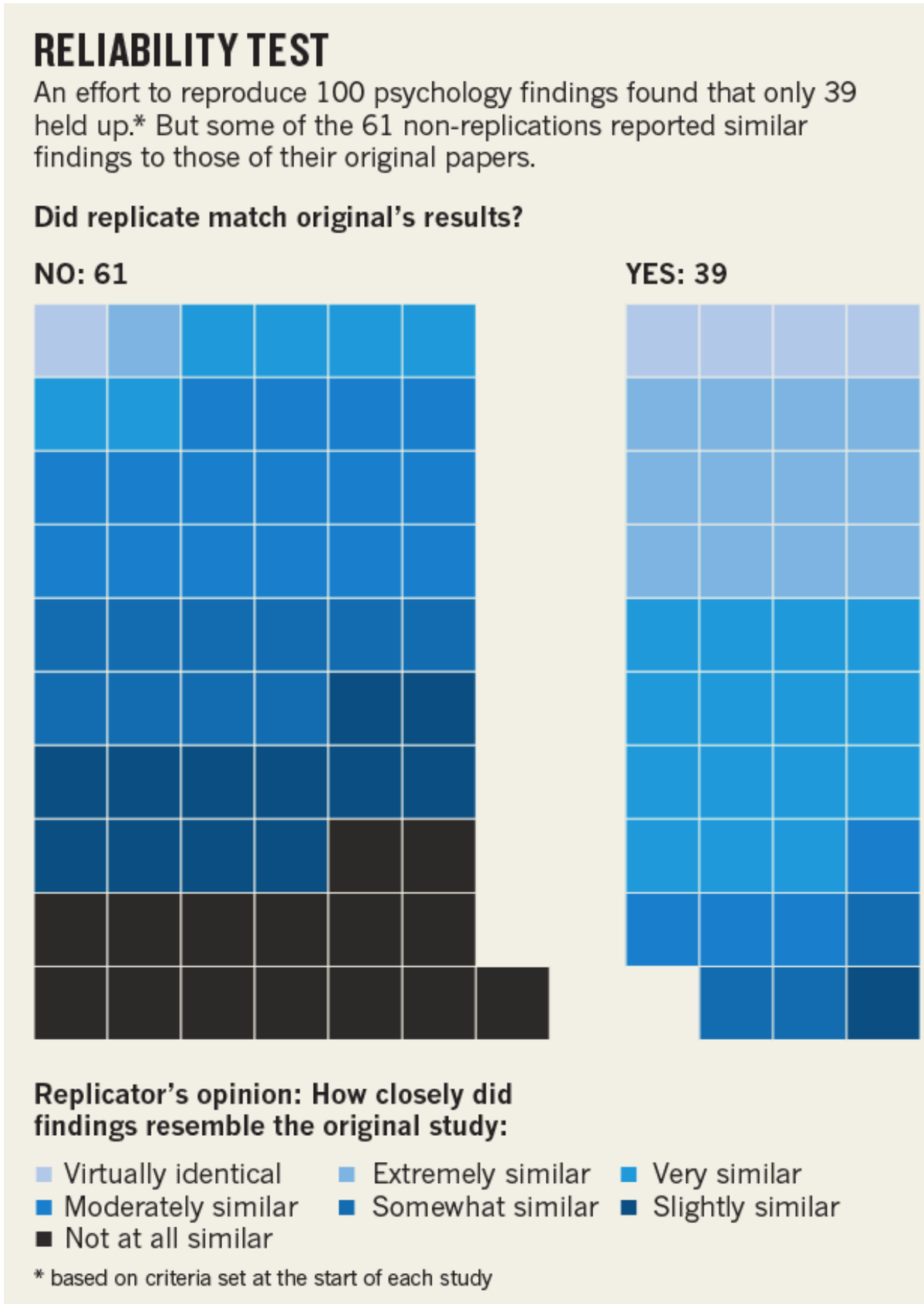
Appearance: Big … very big. Spread your hands and legs wide, argued the authors, and you will both exude power and – this was the new finding – feel great. Adopt a power pose and your testosterone rises and your stress levels fall. Or, as columnist David Brooks neatly put it: "If you act powerfully, you will begin to think powerfully."

And now? Well, that's the odd thing. One of the original report's three authors, Dana Carney, says it was all nonsense. "I do not believe that 'power pose' effects are real," she wrote in a blog that detailed the original research's methodological failings. Standing like John Wayne in a gunfight does not make you feel like a successful gunslinger. It just makes you look silly.

https://www.theguardian.com/politics/shortcuts/2016/sep/28/george-osbornes-power-pose-the-science-proves-feeble

(In)famous studies…

- Power posing
- Ego depletion
- Social priming
- Marshmallow test performance predicts future achievement
- Stanford prison experiment
- Growth mindset
- Learning styles
- Any others you know of?



270 authors tried to replicate 100 experiments drawn from high profile Psychology journals - *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience. Nord et al., (2017), *Journal of Neuroscience,* highlight wide heterogeneity in power in neuroscience studies.

Table 2. Median, maximum, and minimum power subdivided by study type

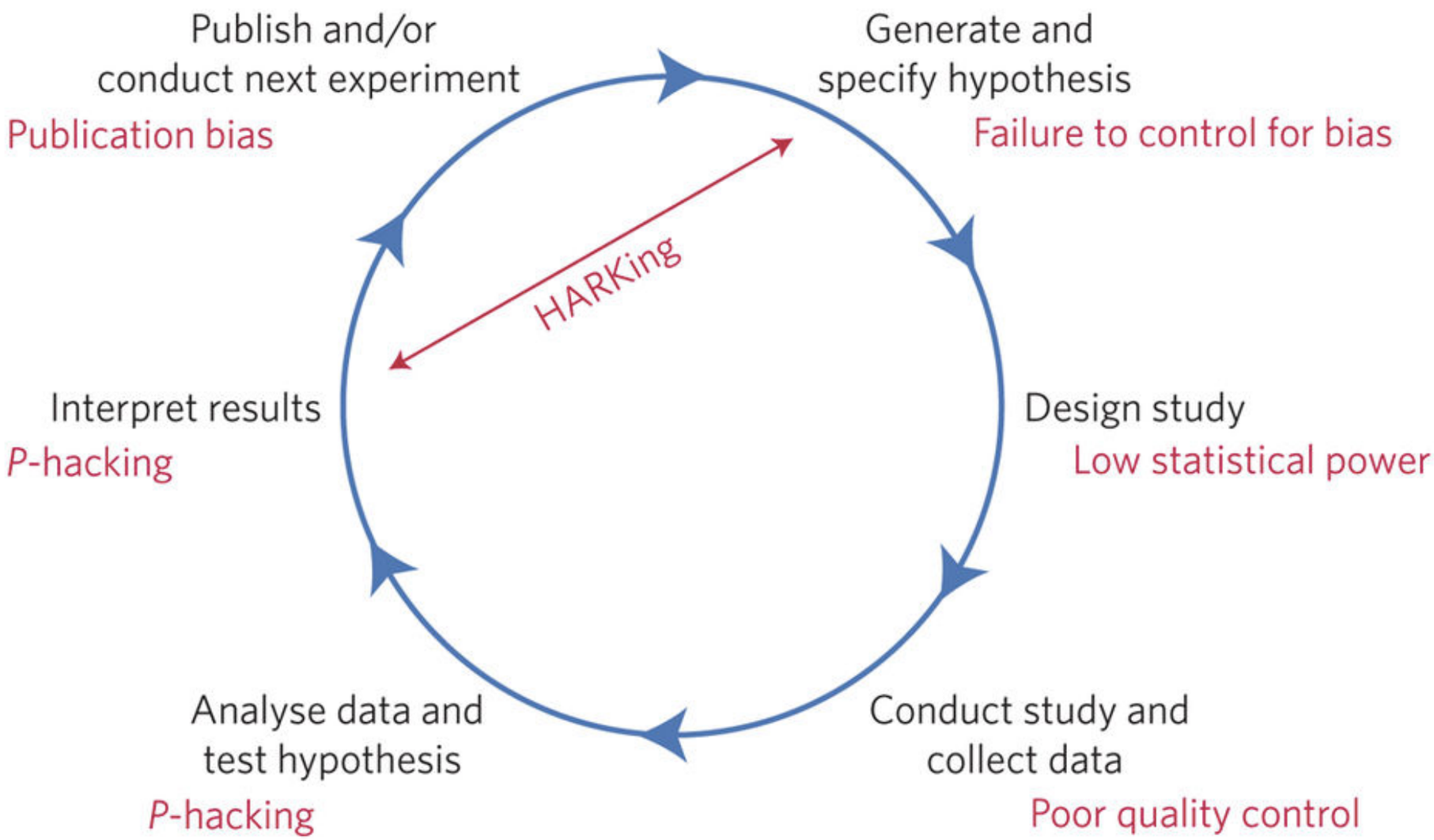
Group	Median power (%)	Minimum power (%)	Maximum power (%)	2.5th and 97.5th percentile (based on raw data)	95% HDI (based on GMMs)	Total N
All studies	23	0.05	1	0.05–1.00	0.00 – 0.72, 0.80 – 1.00	730
All studies excluding null	30	0.05	1	0.05–1.00	0.01 – 0.73, 0.79 – 1.00	638
Genetic	11	0.05	1	0.05 – 0.94	0.00 – 0.44, 0.63 – 0.93	234
Treatment	20	0.05	1	0.05–1.00	0.00 – 0.65, 0.91–1.00	145
Psychology	50	0.07	1	0.07–1.00	0.02 – 0.24, 0.28 – 1.00	198
Imaging	32	0.11	1	0.11–1.00	0.03 – 0.54, 0.71–1.00	65
Neurochemistry	47	0.07	1	0.07–1.00	0.02 – 0.79, 0.92–1.00	50
Miscellaneous	57	0.11	1	0.11–1.00	0.09 – 1.00	38

Is there not just “good science” and “bad science”?

Without realising it, good scientists have been engaging in questionable research practices (QRPs)…

Problems include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

From: A manifesto for reproducible science



Munafo et al. (2017), *Nature Human Behaviour*

Replicable Science ≠ Reproducible Science

Replicable Science is when someone else can run a study the same as or conceptually equivalent to your one, and find a similar pattern of effects.

Reproducible Science is when someone else can take your data and your analysis code, run it and then find the same effects that you have reported.

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKing is defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms of HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discuss several reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing's costs exceed its benefits is a complex one that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such as myself who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

ANNUAL REVIEWS

Annual Review of Psychology

Psychology's Renaissance

Leif D. Nelson,[1] Joseph Simmons,[2] and Uri Simonsohn[2]

[1]Haas School of Business, University of California, Berkeley, California 94720; email: Leif_Nelson@haas.berkeley.edu

[2]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: jsimmo@upenn.edu, urisohn@gmail.com

"the overwhelming majority of published findings are statistically significant (Fanelli 2012, Greenwald 1975, Sterling 1959). On the other hand, the overwhelming majority of published studies are underpowered and, thus, theoretically unlikely to obtain results that are statistically significant."

https://www.annualreviews.org/doi/full/10.1146/annurev-psych-122216-011836

ROBERT TAYLOR

Rein in the four horsemen of irreproducibility

Dorothy Bishop *describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.

In 1975, psychologist Anthony Greenwald noted that science is prejudiced against null hypotheses; we even refer to sound work supporting such conclusions as 'failed experiments'. This prejudice leads to publication bias: researchers are less likely to write up studies that show no effect, and journal editors are less likely to accept them. Consequently, no one can learn from them, and researchers waste time and resources be adequately powered. Other disciplines have yet to catch up.

I stumbled on the issue of *P*-hacking before the term existed. In the 1980s, I reviewed the literature on brain lateralization (how sides of the brain take on different functions) and developmental disorders, and I noticed that, although many studies described links between handedness and dyslexia, the definition of 'atypical handedness' changed from study to study — even within the same research group. I published a sarcastic note, including a simulation to show how easy it was to find an effect if you explored the data after collecting results (D. V. M. Bishop *J. Clin. Exp. Neuropsychol.* **12**, 812–816; 1990). I subsequently noticed similar phenomena in other fields: researchers try out many analyses but report only the ones that are 'statistically significant'.

This practice, now known as *P*-hacking, was once endemic to most branches of science that rely on *P* values to test significance of results, yet few people realized how seriously it could distort findings. That started to change in 2011, with an elegant, comic paper in which the authors crafted analyses to prove that listening to the Beatles could make undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). "Undisclosed flexibility," they wrote, "allows presenting anything as significant."

The term HARKing was coined in 1998 (N. L. Kerr *Pers. Soc. Psychol. Rev.* **2**, 196–217; 1998). Like *P*-hacking, it is so widespread that researchers assume it is good practice. They look at the data, pluck out a finding that looks exciting and write a paper to tell a story around this result. Of course, researchers should be free to explore their

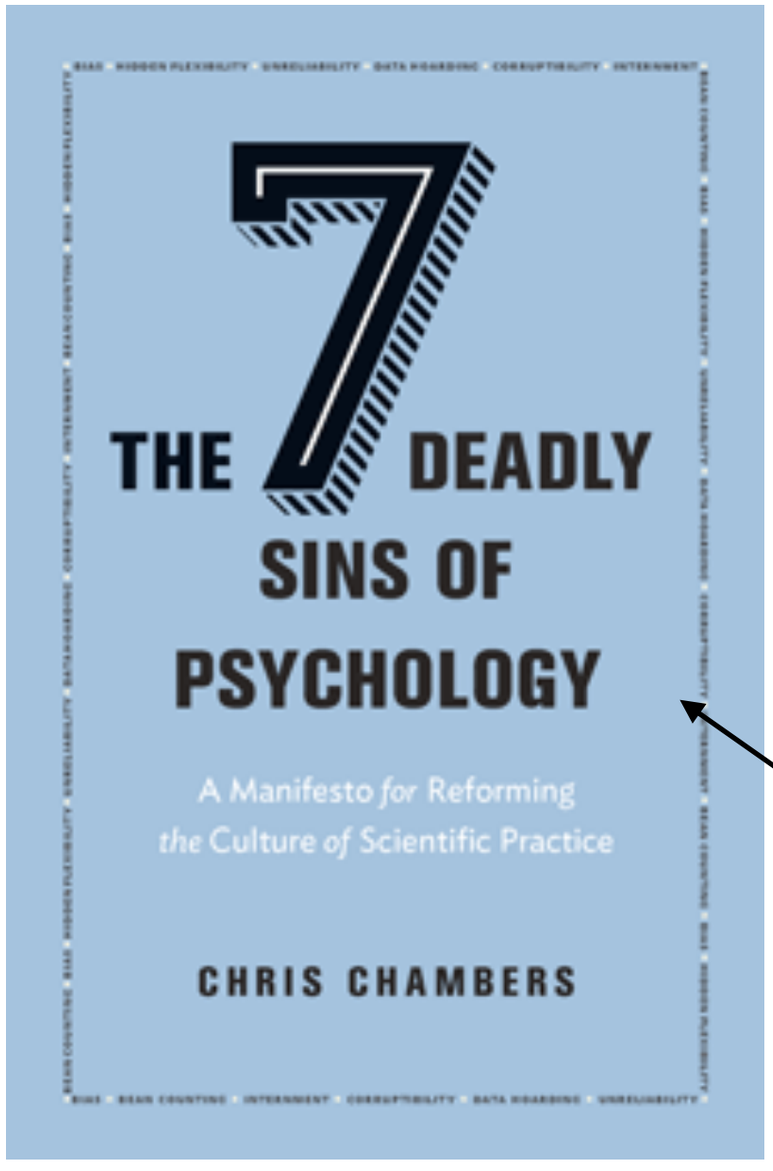
MANY RESEARCHERS PERSIST IN WORKING IN A WAY ALMOST GUARANTEED NOT TO DELIVER MEANINGFUL RESULTS.

https://www.nature.com/articles/d41586-019-01307-2

How do we make our science more replicable?

How do we make our science more reproducible?

A move towards open science…



Sins include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

You really should read this book!

http://www.stat.columbia.edu/~gelman/

Andrew Gelman gives the following recommendations to researchers:

- Analyze all your data.
- Present all your comparisons.
- Make your data public.
- Put in the effort to take accurate measurements (low bias, low variance, and a large enough sample size).
- Do repeated-measures comparisons where possible.

Open Science practices include…

- Pre-registering experiments.
- Registered reports.
- Using preprint servers (e.g., bioRxiv, PsyArXiv).
- Making data and analysis code freely available (e.g., via GitHub, OSF).
- Open access to journal articles.
- …and more.

Open Science recently recognised by G7 Science Ministers…

Focus: Incentives and the researcher ecosystem

Ambition: Foster a research environment in which career advancement takes into account Open Science activities, through incentives and rewards for researchers, and valuing the skills and capabilities in the Open Science workforce.

Recommendations:

At national levels: G7 nations should each engage with research stakeholders to identify and implement enhancements to research evaluation and reward systems that take into consideration the Open Science activities carried out by researchers and research institutions. Topics that could be discussed include:

- Recognizing Open Science practices during evaluation of research funding proposals, and research outcomes;
- Recognizing and rewarding research productivity and impact that reflect open science activities by researchers during career advancement reviews;
- Including credit for service activities such as reviewing, evaluating, and curation and management of research data; and,
- Developing metrics of Open Science practices.

REF 2019/02
January 2019

Panel criteria and working methods

200. The sub-panels welcome research practice that supports reproducible science and the application of best practice. Examples include registered reports, pre-registration, publication of data sets, experimental materials, analytic code, and use of reporting checklists for publication purposes and those relating to the use of animals in research. These contribute to the evaluation of rigour for submitted outputs. Replication studies may be submitted as outputs and will be evaluated on the extent to which they contribute significant new knowledge, improved methods, or advance theory or practice[1].

346...

Within the context of the institution's strategy, how the submitting unit is progressing towards an open research environment, including where this goes above and beyond the REF open access policy requirements, and wider activity to encourage the effective sharing and management of research data, as appropriate to the discipline. Consideration of reproducibility should also be included where relevant to the discipline.

is beginning to appear in tenure-track job adverts…

> Our Department embraces the values of open and reproducible science, and candidates are encouraged to address (in their statements and/or cover letter) how they have pursued and/or plan to pursue these goals in their work.

and is forming part of Universities' teaching manifestos.

Teaching with Open Science commitment:
To teach the practices and skills of open research and science in our undergraduate and postgraduate degree programmes

a. Promote open science in our teaching.
b. Design a Research Methods curriculum that teaches skills for open science and uses open science to enhance teaching (for example: teach R and use open data to practice analysis skills).
c. Learn about and adopt open educational practices in our teaching.
d. Produce and promote tools for helping student researchers adopt open practices, including training and guidance suitable to their level of study.
e. Author, share and use open educational resources to promote teaching with open science beyond our School and Institution.
f. Support our colleagues to learn the skills of teaching Open Science.

FAIR data

Findable

- The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier

- F2. Data are described with rich metadata (defined by R1 below)

- F3. Metadata clearly and explicitly include the identifier of the data they describe

- F4. (Meta)data are registered or indexed in a searchable resource

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Accessible

- Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Interoperable

- The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- I2. (Meta)data use vocabularies that follow FAIR principles

- I3. (Meta)data include qualified references to other (meta)data

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Reusable

- The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

How do you do Open and Reproducible Science?

Before Data Collection

- Specify your hypotheses and analysis plan.
- **Pre-register** your hypotheses and analysis plan at `osf.io`
- Consider data simulation so that you can write your analysis script before you have your real data.
- Consider submitting as a **registered report** - currently more than **200** journals now support this route. This involves acceptance in principle before you have even started collecting your data.

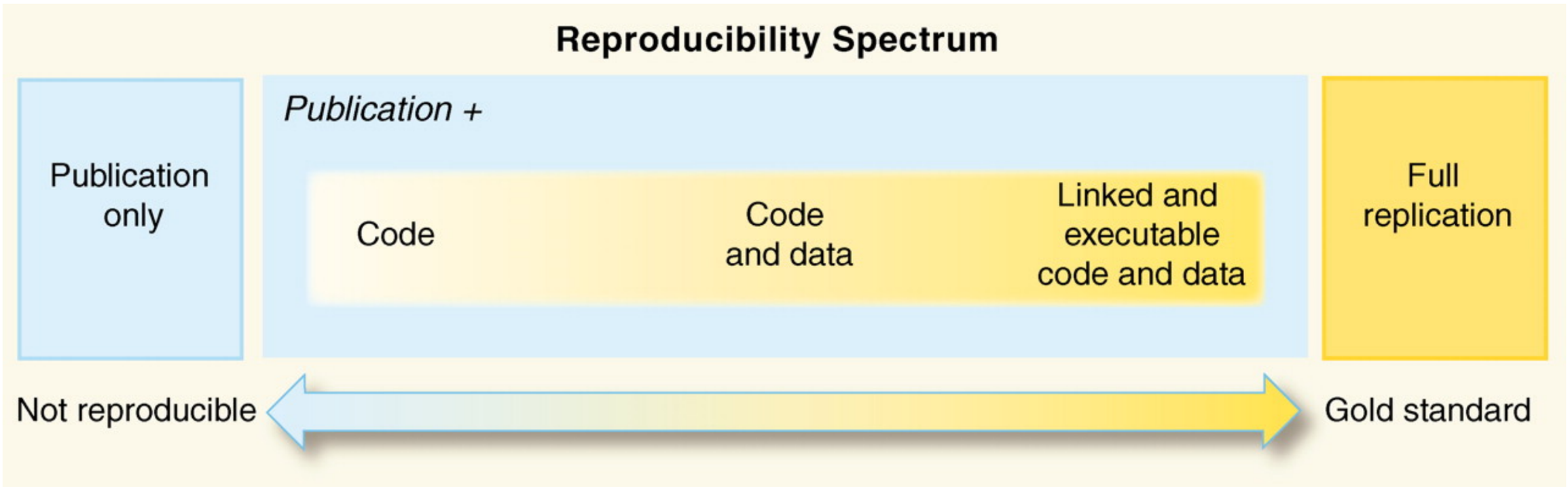
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



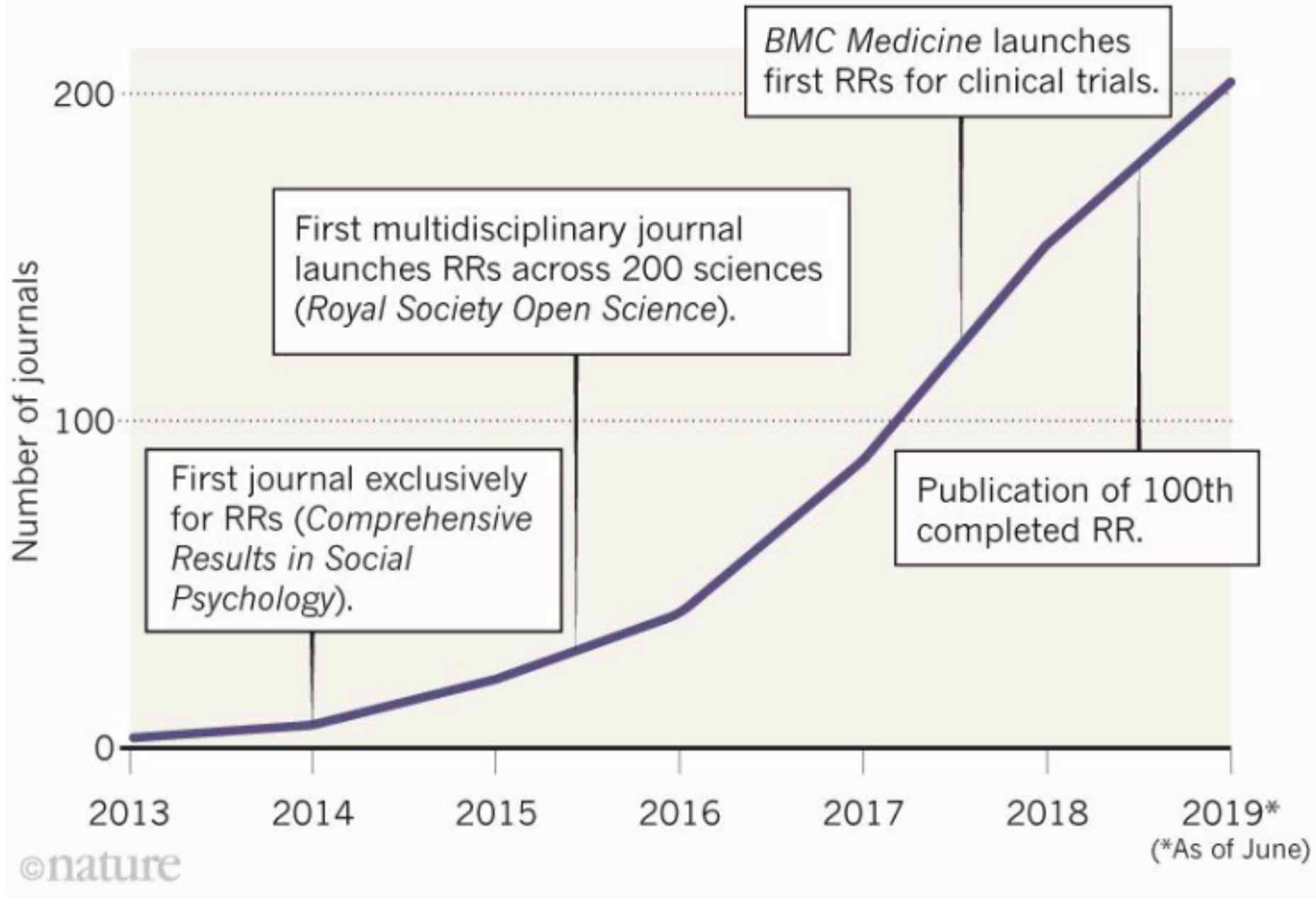
"*You can't do data science in a GUI*", Hadley Wickham

Registered Reports

https://cos.io/rr/?_ga=2.49773158.1336120275.1555407527-1361001319.1494339346

RAPID RISE

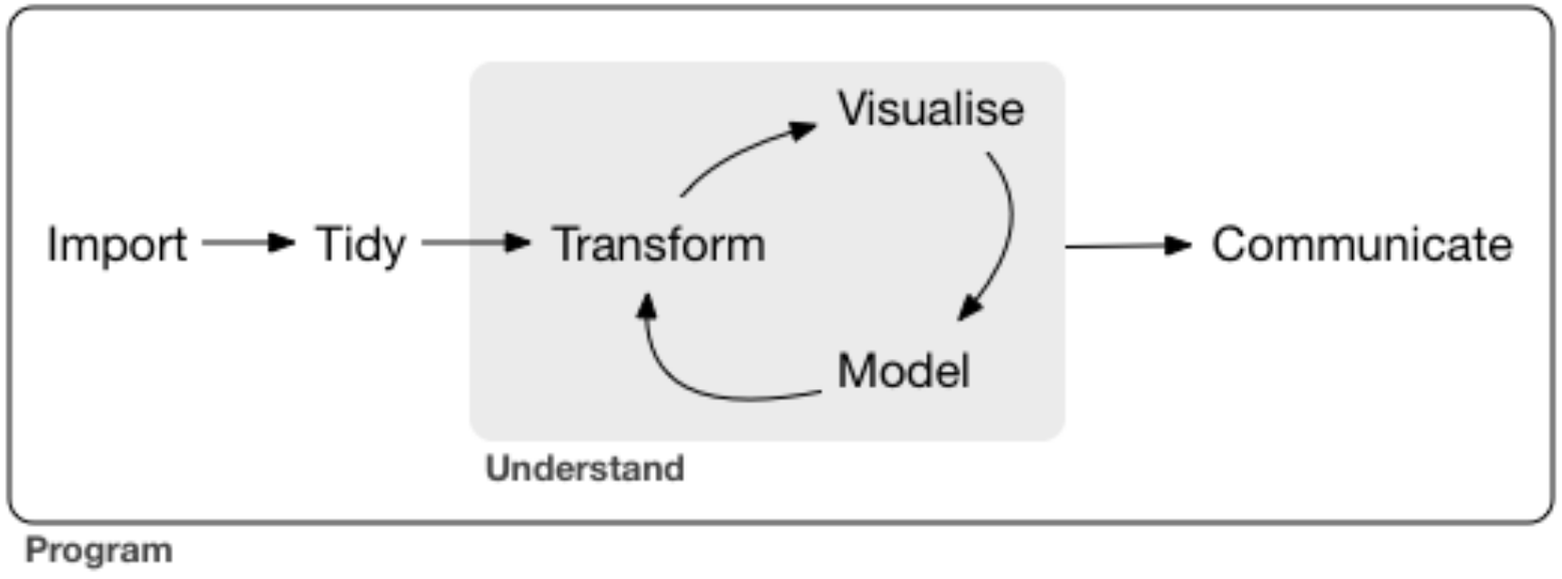
Since 2013, the number of journals offering Registered Reports (RRs) has risen to more than 200 titles.



https://www.nature.com/articles/d41586-019-02674-6

After Data Collection

- You need to use analysis software that allows for open sharing and reproducibility of the entire data wrangling/ analysis/write-up workflow.



Hadley Wickham and Garrett Grolemund

- You can share your data via osf.io, on GitHub etc.:

ajstewartlang / **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition**

Watch 0 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master ▾ **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition** / **RPs.csv**

Find file | Copy path

ajstewartlang Made consistent the labelling of factors in data files and in paper — 7b3b3b1 on 29 Mar 2017

0 contributors

1681 lines (1681 sloc) | 69.7 KB — Raw | Blame | History

Search this file...

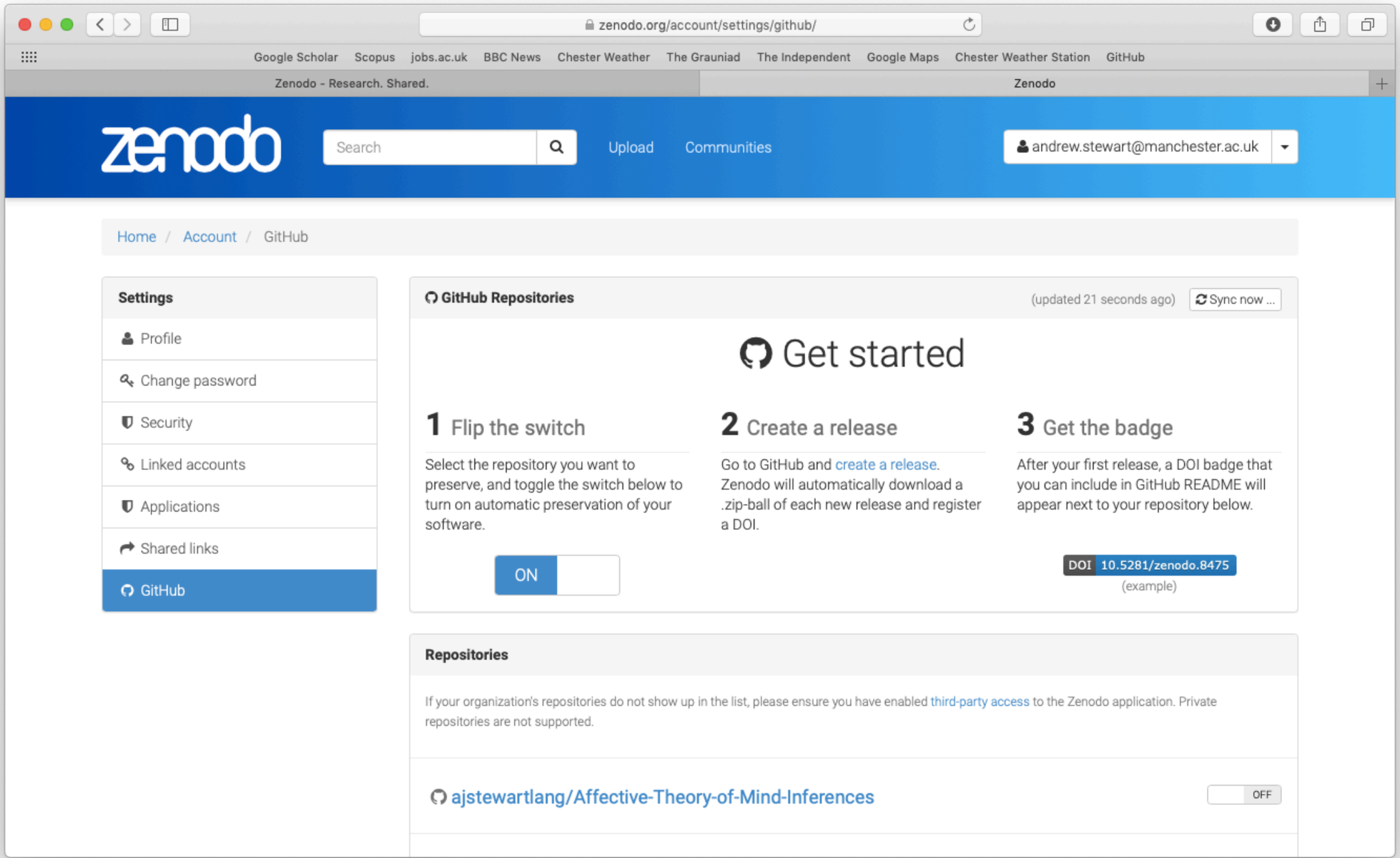
	P.s	Item	Condition	Probmanip	Speaker	statement	response	final	Meaning	Imposition
2	1	1	1	1708	302	1399	1867	1206	Indirect	High
3	1	2	2	1466	296	1377	1674	828	Indirect	Low
4	1	3	3	1393		1494	1950	1812	Direct	High
5	1	4	4	2463	530	1691	1866	965	Direct	Low
6	1	5	1	1552	267	1332	1477	1345	Indirect	High
7	1	6	2	1445	444	1004	1067	797	Indirect	Low
8	1	7	3	2159	501	739	1231	2240	Direct	High
9	1	8	4	1459		1086	946	978	Direct	Low
10	1	9	1	3302		1503	900	1736	Indirect	High

FaceTime

- alongside your analysis code

```
FPs$Meaning <- as.factor(FPs$Meaning)
FPs$Imposition <- as.factor(FPs$Imposition)

#this sets up the contrasts so that the intercept in the mixed LMM is the grand mean (i.e., the mean of all conditions)
my.coding <- matrix (c(.5, -.5))

contrasts (FPs$Meaning) <- my.coding
contrasts (FPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
fpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelprec)
lsmeans (fpmodelprec, pairwise~Meaning*Imposition, adjust="none")

fpmodelc <- lmer (statement ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=FPs, REML=T
summary (fpmodelc)
lsmeans (fpmodelc, pairwise~Meaning*Imposition, adjust="none")

fpmodelpostc <- lmer (response ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelpostc)
lsmeans (fpmodelpostc, pairwise~Meaning*Imposition, adjust="none")

#Regression Path Analysis
#Read in Regression Path data
RPs <- read.csv("~/RPs.csv")

RPs$Meaning <- as.factor(RPs$Meaning)
RPs$Imposition <- as.factor(RPs$Imposition)

contrasts (RPs$Meaning) <- my.coding
contrasts (RPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
rpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=RPs, REM
```

And preserve it with a DOI via Zenodo



Part of doing better science involves knowing how to build appropriate statistical models, and how to understand what those models are telling you (and what they are not…)

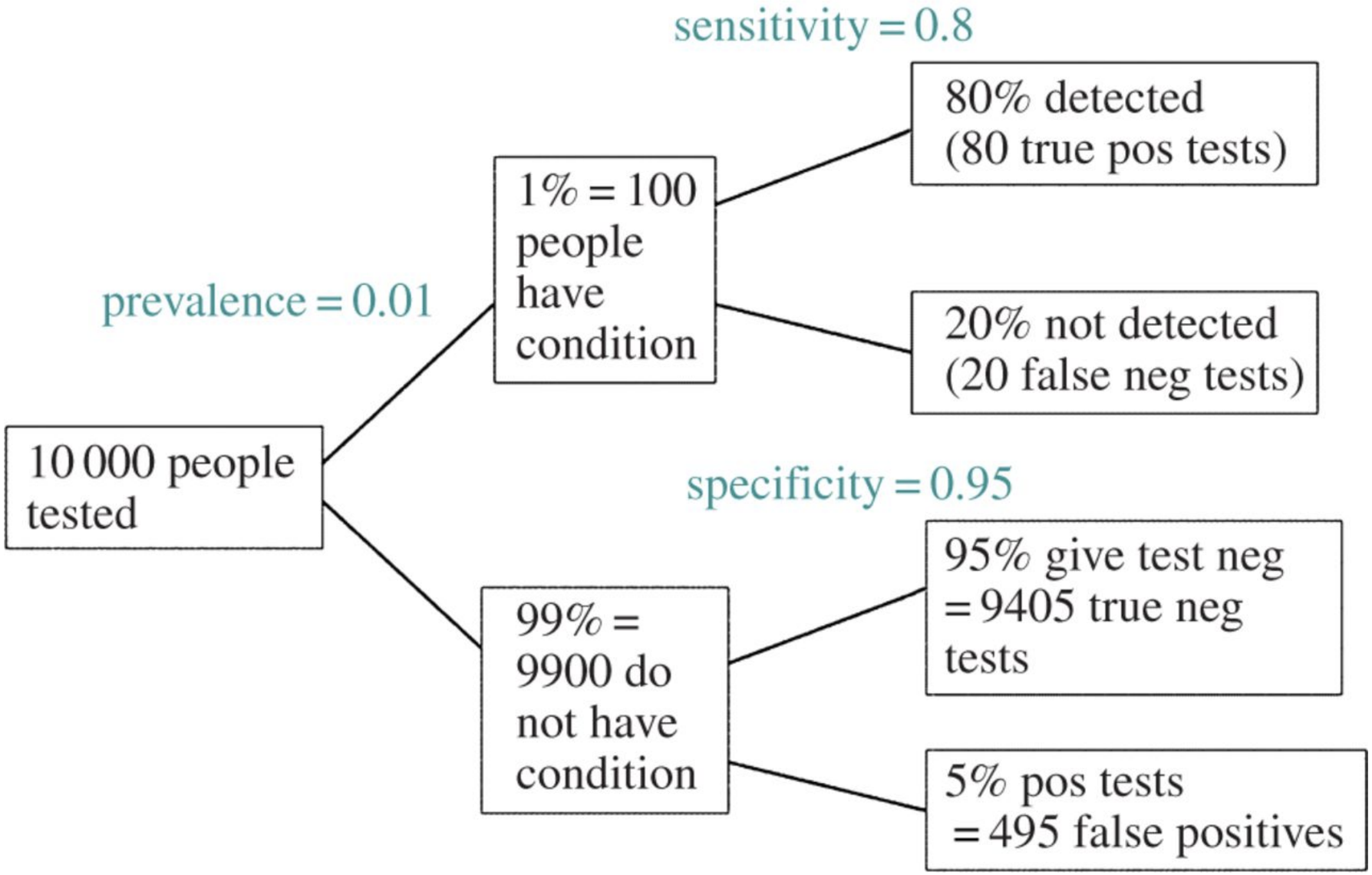
Why Understanding Statistics Matters…

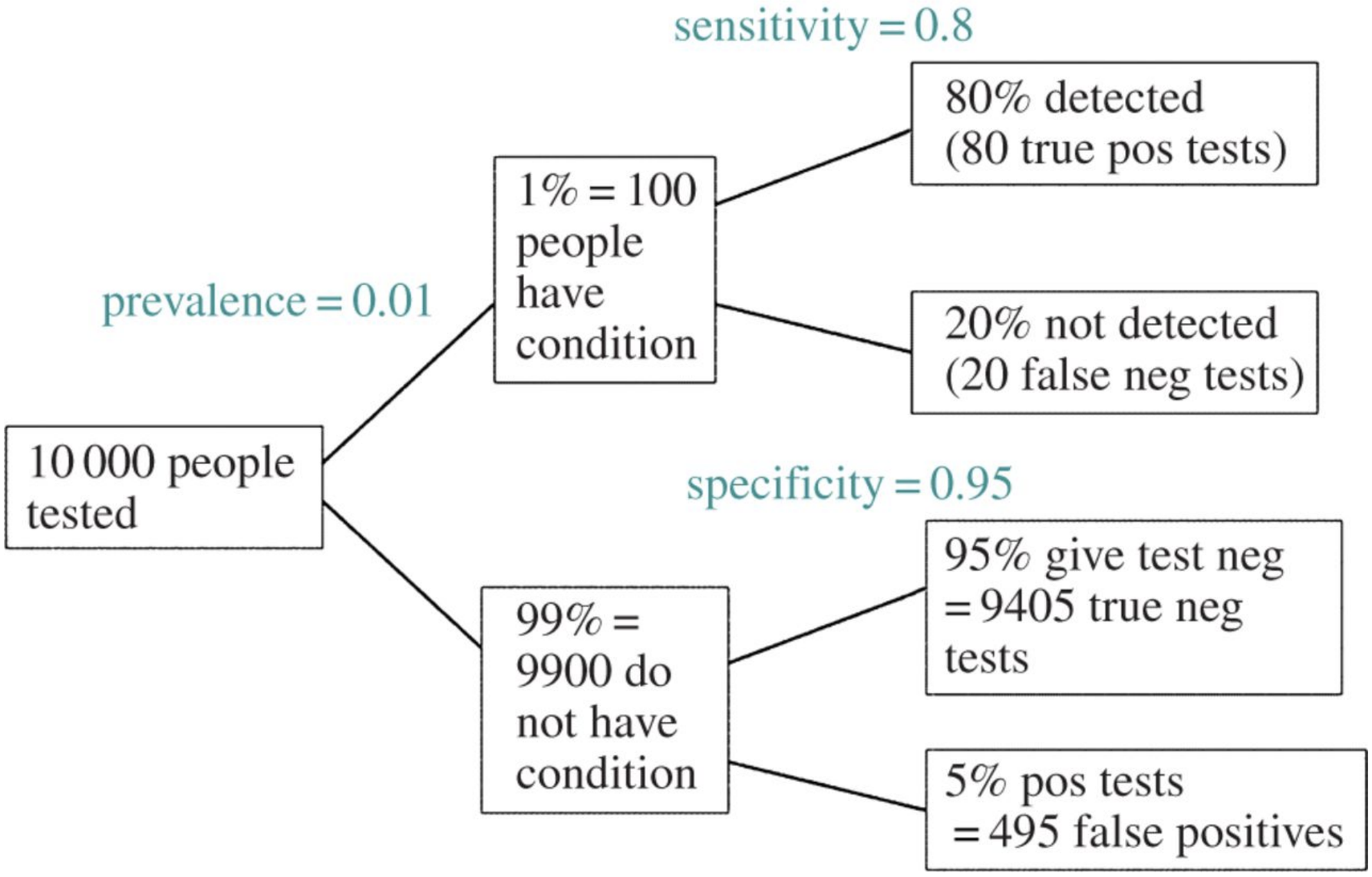
- Imagine a test in which 95% of people without a medical condition will be correctly diagnosed as not having it (specificity = 0.95).

- Imagine the test is able to correctly diagnose 4 out of the 5 people who **do** have the medical condition (sensitivity = 0.8).

- Imagine the prevalence of the medical condition in the population is 1%.

From Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. DOI: 10.1098/rsos.140216


sensitivity = 0.8
80% detected
(80 true pos tests)
1% = 100
people
have
condition
prevalence = 0.01
20% not detected
(20 false neg tests)
10 000 people
tested
specificity = 0.95
95% give test neg
= 9405 true neg
tests
99% =
9900 do
not have
condition
5% pos tests
= 495 false positives



- The results of the test suggest 575 people have the condition. But 495 of these are false positives! So 86% of the people who produced a positive result actually don't have the condition.

Understanding Statistics

- Appropriately powered studies, appropriately analysed (with corrections for multiple comparisons). Consider using Bayesian statistics where appropriate.
- Recognition that our research should focus on revealing *what effects are likely to be real*, rather than just statistical significance. We need to understand what significance is (and what it isn't).
- Registered reports allows pre-registration of planned experiments (hypotheses, N, analyses etc.):
  - https://osf.io/8mpji/wiki/home/

Some traditional basics…

- For a design with two experimental groups:
  - Null hypothesis (H0) - there is no statistically significant difference between those experimental groups.
  - Experimental hypothesis (H1) - there **is** a statistically significant difference between two experimental groups.

- We typically reject H0 that if we find that the result of a statistical test comparing the two experimental groups is $p < 0.05$ (also known as the alpha (α) level).

What is significance?

- Suppose that a treatment and a placebo are allocated at random to a group of people. We measure the mean response to each treatment, and wish to know whether or not the observed difference between the means is real (not zero), or whether it could plausibly have arisen by chance. If the result of a significance test is p=0.05, we can make the following statement:

If there were actually no effect (if the true difference between means were zero) then the probability of observing a value for the difference equal to, or greater than, that actually observed would be p=0.05. In other words there is a 5% chance of seeing a difference at least as big as we have done, by chance alone.

ASA Principles 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.

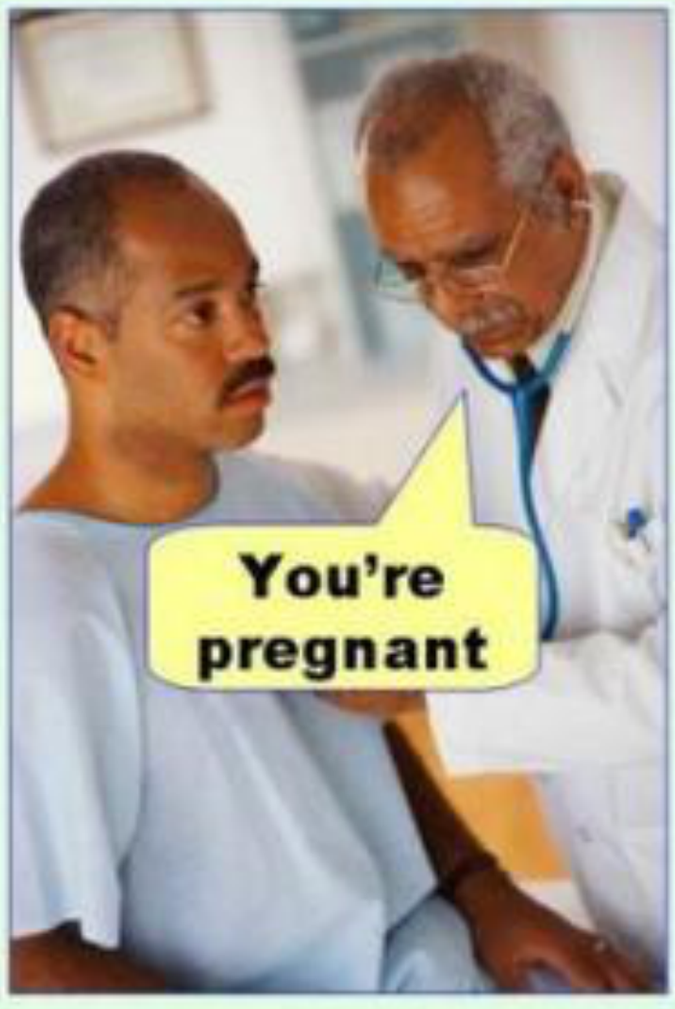
Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

Type I and Type II errors

- With an α level of 0.05, we have a 5% chance of falsely rejecting the null hypothesis (H0).
- Falsely rejecting H0 is known as a Type I error (i.e., thinking we have found a difference when these isn't one).
- There are also Type II errors which involve failing to find a difference when one is actually present.
- Most of what you have been taught at UG level will have involved trying to avoid Type I errors.

Type I and Type II errors

Type I error
(false positive)



Type II error
(false negative)

- Controlling for Type II errors is as important as controlling for Type I errors. The probability of a Type II error is known as Beta (β).

- The probability of arriving at a Type II error (not finding a difference where there is one) is related to the experimental power of your design.

- For any experiment, Power = 1 - β

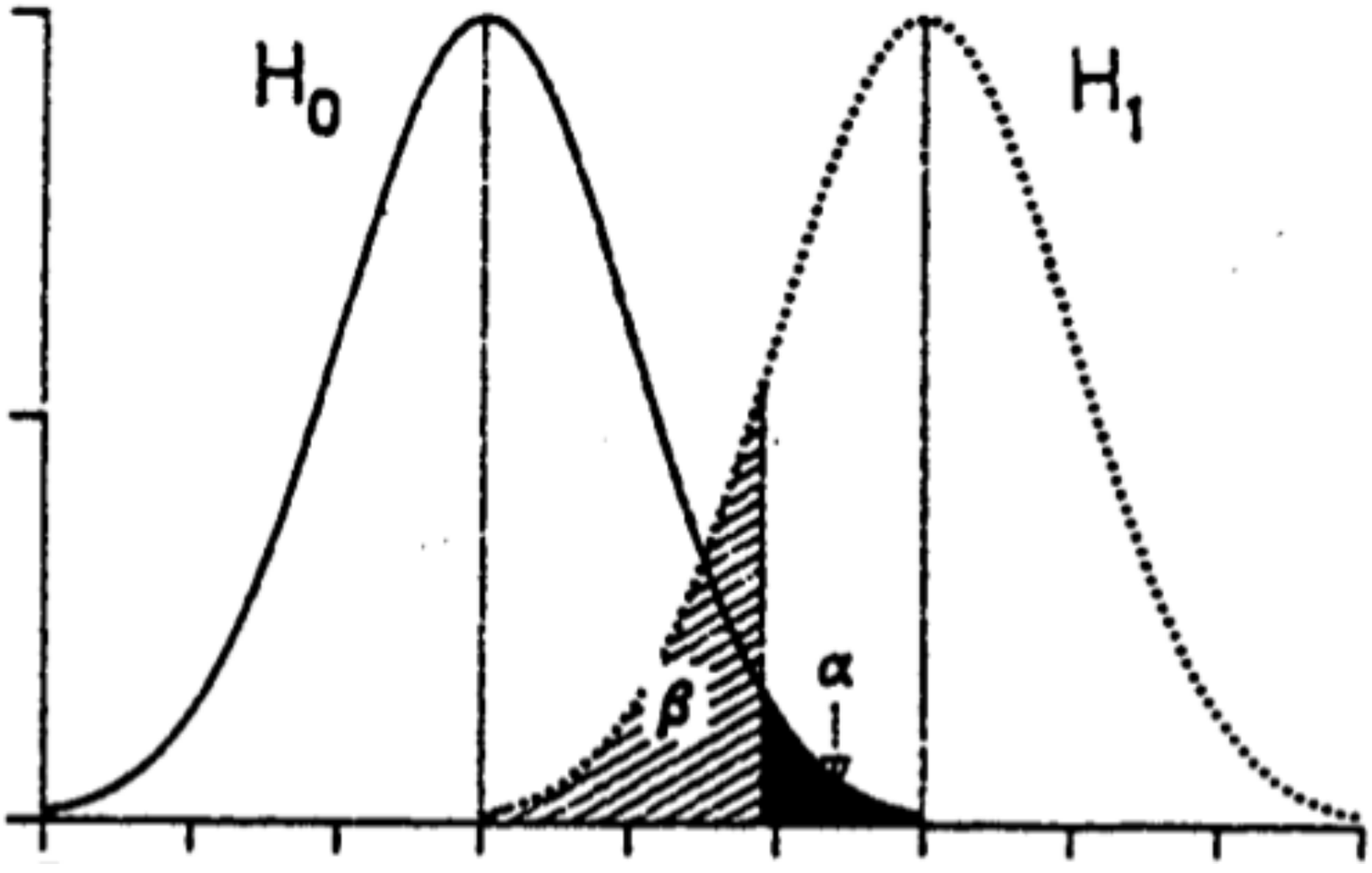
Is Power That Big a Deal?

- Cohen (1992) describes why power *is* such a big deal (and what can happen if experiments do not have sufficient power). Low powered studies have a lowered chance of finding a real effect, and also a higher chance of suggesting an effect is present when it is not.

- Reports the results of a review of 1960 volume of Journal of Abnormal and Social Psychology that he conducted at the time and the results of a Sedlmeier and Gigerenzer (1989) review of a 1984 volume of of the same journal.

- In 1960, the average power of the experiments reported in JASP to detect medium effect sizes was 0.48. In 1984, it was 0.25 (in other words only a 25% chance of finding an effect even if it was there!)

Power as a function of α



- If we were to increase α, we would increase power (by reducing β) but would risk a rise in the probability of a Type I error.

Calculating Power

- Power (1-β) is related to:
    - sample size (i.e., N)
    - effect size
        - α
- Cohen (1992) proposes that a reasonable level of Power to aim for should be around 0.8
- Power of 0.8 (with a β of 0.20), alongside an α of 0.05 results in a β:α ratio of 4:1 in terms of the risk associated with respective errors

Effect Sizes

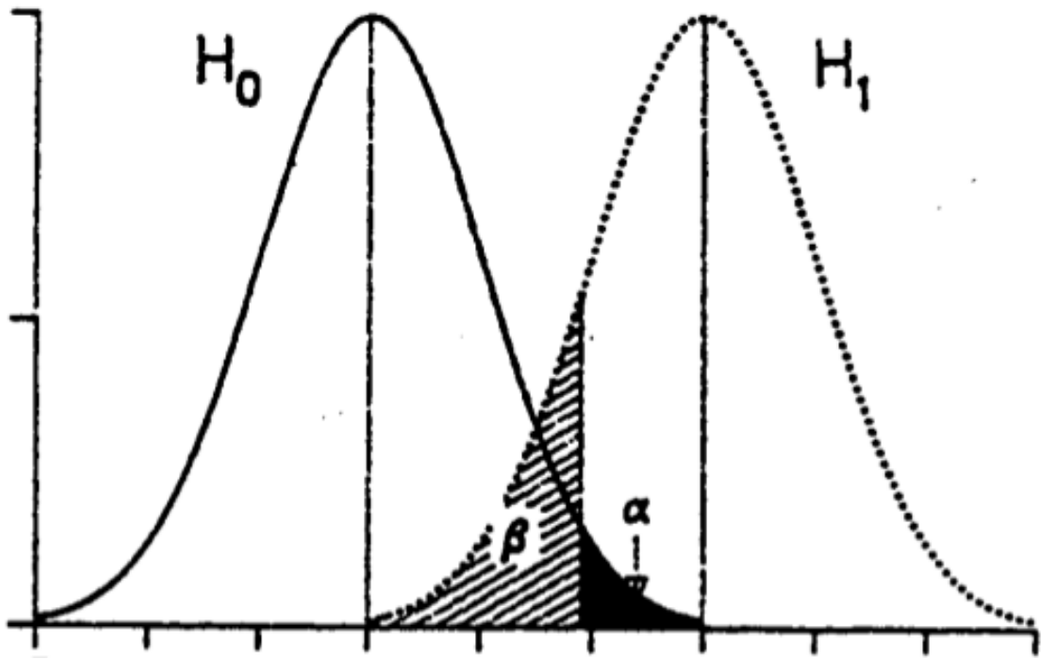
- We can measure the size of an experimental effect in an objective, standardised manner.
- The two most common measures of effect size are Cohen's d and Pearson's r.

	Small	Medium	Large
d	0.2	0.5	0.8
r	0.1	0.3	0.5

Calculating Cohen's d

$$d = \frac{\mu 1 - \mu 0}{\sigma}$$

μ0= mean where H0 is true
μ1= mean where H1 is true
σ= standard deviation



Where does sample size fit with all this?

$$\delta = d\sqrt{n}$$

By using this standardized equation we can combine effect size and sample size and use standardized Power table to calculate Power by hand.

- A clinical psychologist wants to test hypothesis (H1) that people who seek treatment have higher IQs than general population. She wants to use IQs of 25 randomly sample patients and is interested in a difference of 5 points between the mean of the general population and the mean of her client population.

So, $\mu_0 = 100$, $\mu_1 = 105$, $\sigma = 15$

$$d = \frac{105\text{-}100}{15} = 0.33$$

$$\delta = 0.33 \sqrt{25} = 0.33\ (5) = 1.65$$

Power as a function of δ for α=0.05

δ	α = 0.05
1.4	0.29
1.5	0.32
1.6	0.36
1.7	0.4
1.8	0.44
1.9	0.48
2	0.52
2.1	0.56

So, for δ=1.65, power is about 0.38

- So, with power = 0.38, if H0 is false and μ1 = 105, only about 38% of the time will the clinician find a statistically significant difference between her sample mean and the mean specified by H0. In other words, 62% of the time the clinician will be making a Type II error (i.e., failing to find a difference when there is one present).

- So, how would you increase the power of this experiment?

- Remember, Power (1-β) is related to:
    - sample size (i.e., N)
    - effect size
    - α

- Can’t do anything about the effect size.
- If you change the α level, you do increase the power but also the probability of a Type I error.

- You can increase the sample size.....

For $\alpha = 0.05$, at power = 0.8, $\delta = 2.8$

δ	$\alpha = 0.05$
2.6	0.74
2.7	0.77
2.8	0.8
2.9	0.83
3	0.85
3.1	0.87

So, if we know that $\delta = 2.80$, and we know that $d = 0.33$

$$\delta = d\sqrt{n}$$

$$n = \left(\frac{\delta}{d}\right)^2 = \left(\frac{2.80}{0.33}\right)^2 = 8.48^2$$

$$= 71.91$$

Rounding up, that gives us 72 participants.

- In the previous example, we wanted to calculate the power of a study looking at whether the mean of a particular sample (i.e., people who seek clinical help) differed from the mean of the population. This is also known as the one-sample t-test.

- How about testing to see whether two independent sample means differ from each other (e.g., independent samples t-test)?

Power calculations for differences between two independent means

To calculate Cohen's d, we want the difference between two mean (μ1-μ2) under H1 minus the difference (μ1-μ2) under H0, divided by σ. Under H0 though, (μ1-μ2) is zero (because there is no difference between the means under the null hypothesis) so,

$$d = \frac{(\mu 1 - \mu 2) - 0}{\sigma} = \frac{(\mu 1 - \mu 2)}{\sigma}$$

An example

- Imagine the case where we want to test the difference between two group means. Imagine also that we expect the difference to be about 5 points. From past research, we know that the standard deviation (σ) is about 10.

$$d = \frac{(\mu 1 - \mu 2)}{\sigma} = \frac{5}{10} = 0.5$$

d = 0.5 is a moderate effect size.

What is the power of our experiment with 25 people in each of our two groups?

For two-sample designs, we define δ as:

$$\delta = d\sqrt{\frac{n}{2}}$$

Where n is the number of cases in any one sample. We are assuming equal sample sizes here btw.

$$\delta = (0.50)\sqrt{\frac{25}{2}} = 0.50\ \sqrt{12.5} = 0.50\ (3.54)$$

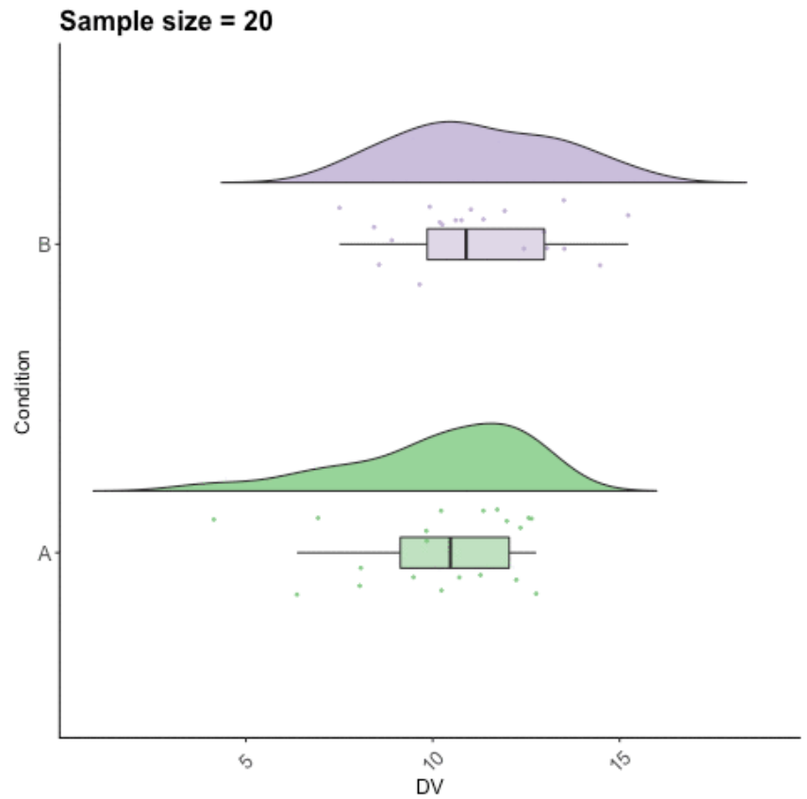
$$= 1.77$$

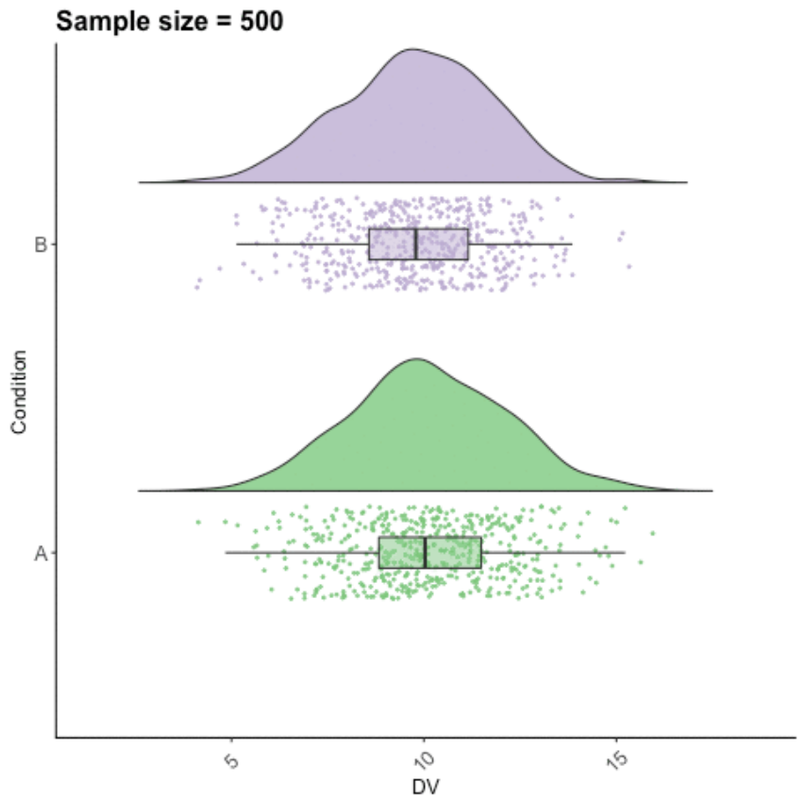
With $\alpha = 0.05$ and $\delta = 1.77$, Power is about 0.43

This means there is a 57% chance of failing to find a difference (even though one might be present).

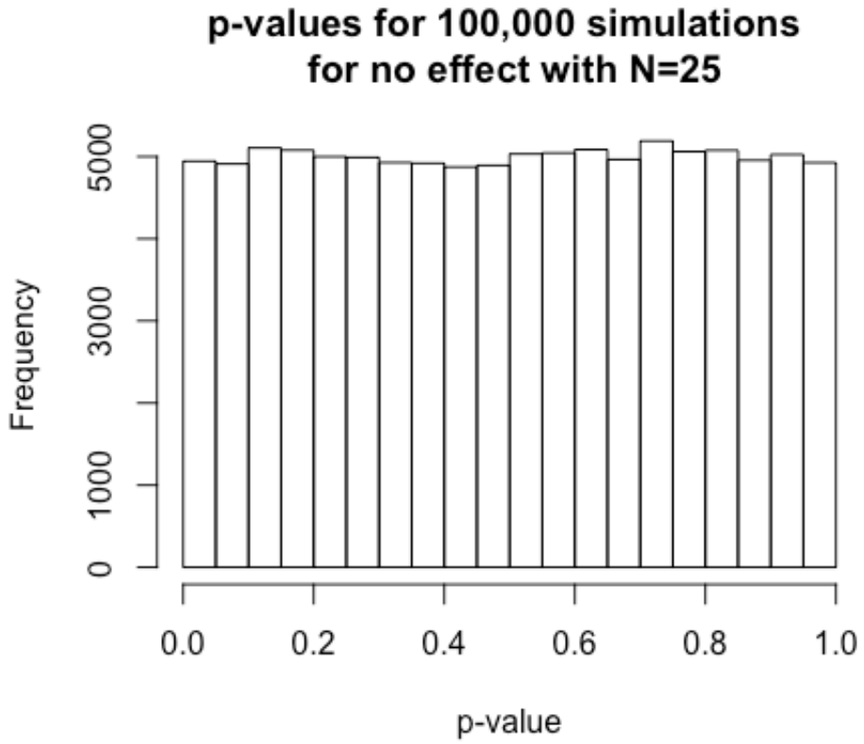
δ	$\alpha = 0.05$
1.4	0.29
1.5	0.32
1.6	0.36
1.7	0.4
1.8	0.44
1.9	0.48

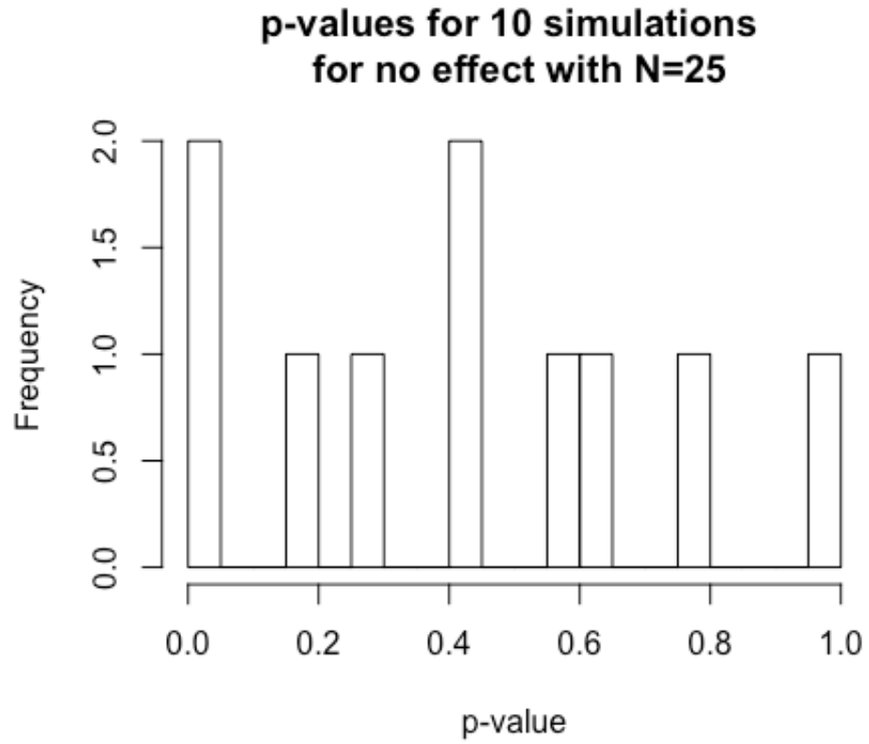
The Problem of Sampling Bias

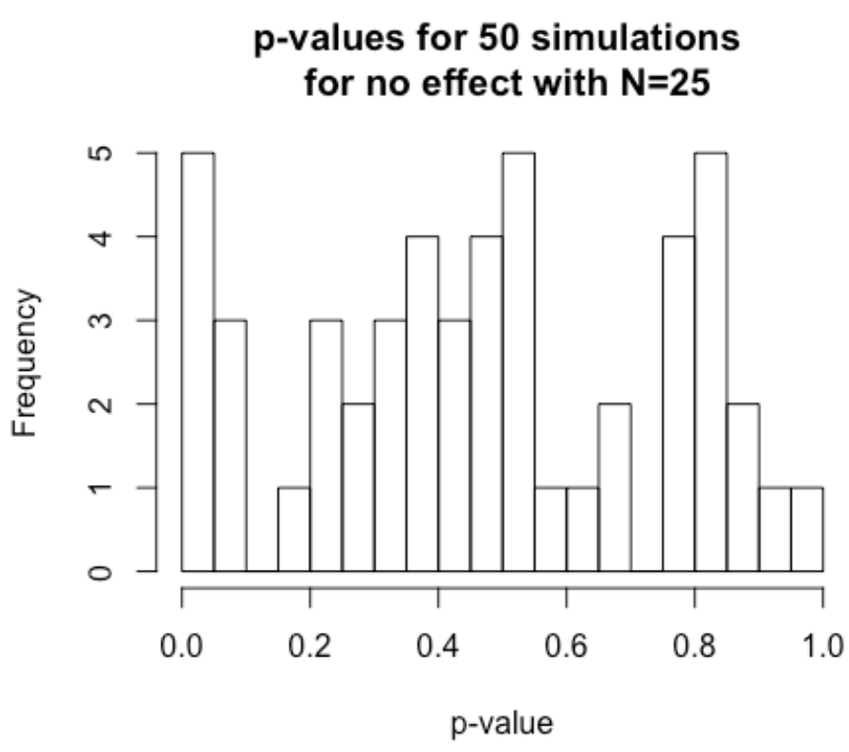




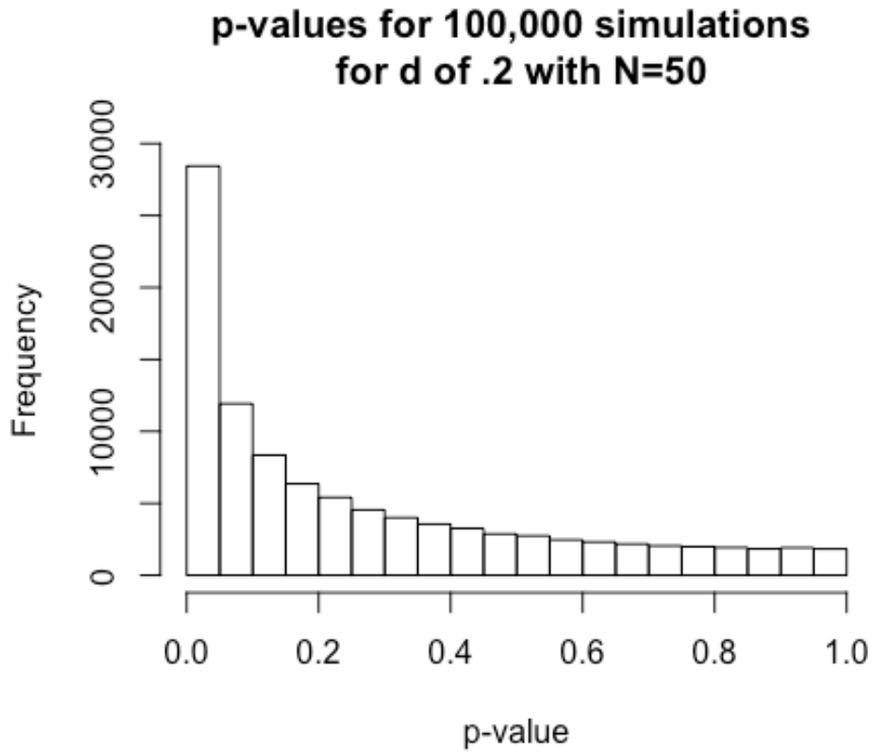
Samples for conditions A and B are drawn from the same population. Due to sampling error, with small samples (e.g., N=20) we might sometimes conclude there is a difference between A and B where there isn't one (as you can see with the N=500 samples).


p-values for 100,000 simulations
for no effect with N=25
Frequency
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
p-value

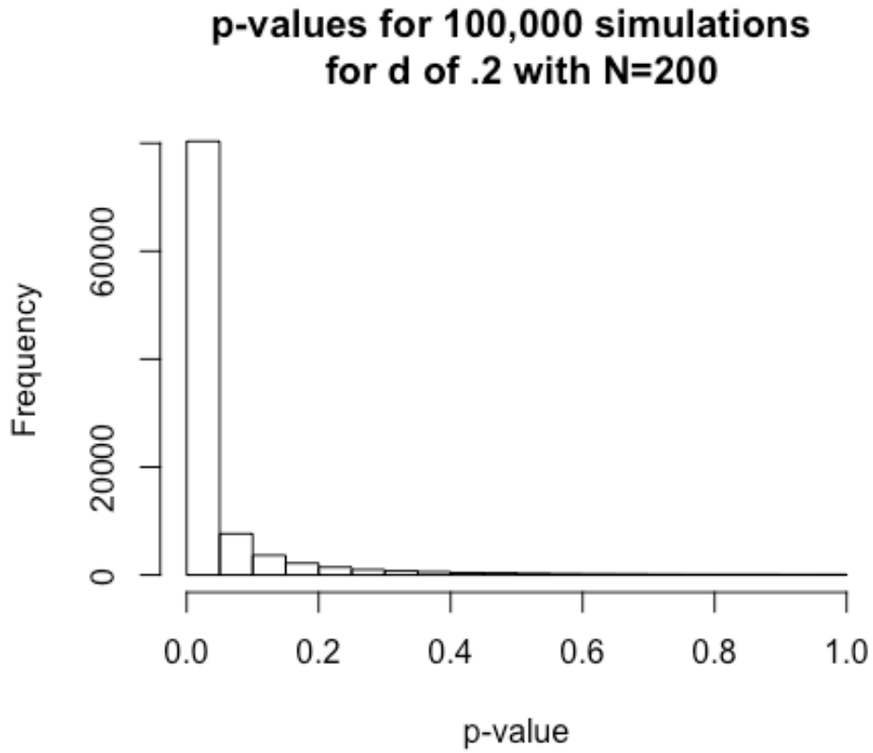

p-values for 10 simulations
for no effect with N=25
Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
p-value


p-values for 50 simulations
for no effect with N=25
Frequency
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
p-value

Real effects will <u>not</u> always replicate.



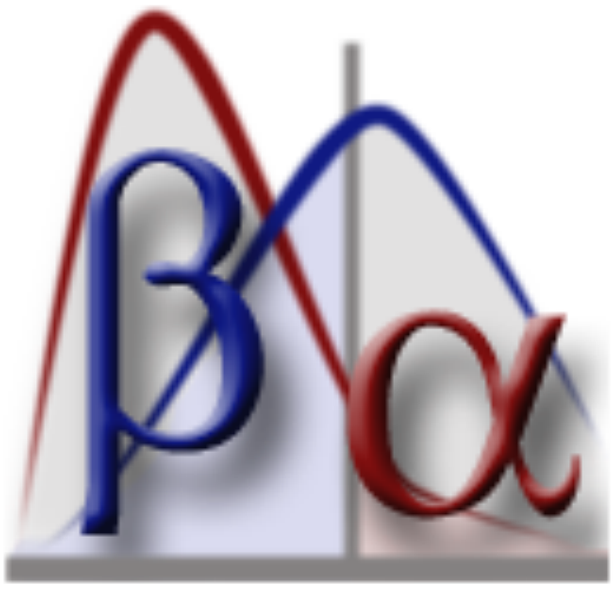
Assuming $p < .05$ alpha, N=50 gives us around 30% power, which means that 70% of the time we'll miss the effect (even though it is present).



N=200 gives us around 80% power, which means that 20% of the time we'll miss finding the effect (even though it is present).

For other power analyses, more involved equations are required. Luckily we don't have to be proficient in using them in order to do power analyses. A powerful (and free) program is available.

G*Power
www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/



- G*Power covers statistical power analyses for many different statistical tests such as: t test, F test, χ2-test. z test and some exact tests.

- G*Power offers five different types of statistical power analysis:

    - A priori (sample size N is computed as a function of power level 1-β, significance level α, and the to-be-detected population effect size)

    - Compromise (both α and 1-β are computed as functions of effect size, N, and an error probability ratio q = β/α)

    - Criterion (α and the associated decision criterion are computed as a function of 1-β, the effect size, and N)

    - Post-hoc (1-β is computed as a function of α, the population effect size, and N)

    - Sensitivity (population effect size is computed as a function of α, 1-β, and N)

- G*Power is available for Mac OS X and Windows. **G*Power is free.**

- G*Power covers statistical power analyses for many different statistical tests such as: t test, F test, χ2-test. z test and some exact tests.

- G*Power offers five different types of statistical power analysis:

  - A priori (sample size N is computed as a function of power level 1-β, significance level α, and the to-be-detected population effect size)

  - Compromise (both α and 1-β are computed as functions of effect size, N, and an error probability ratio q = β/α)

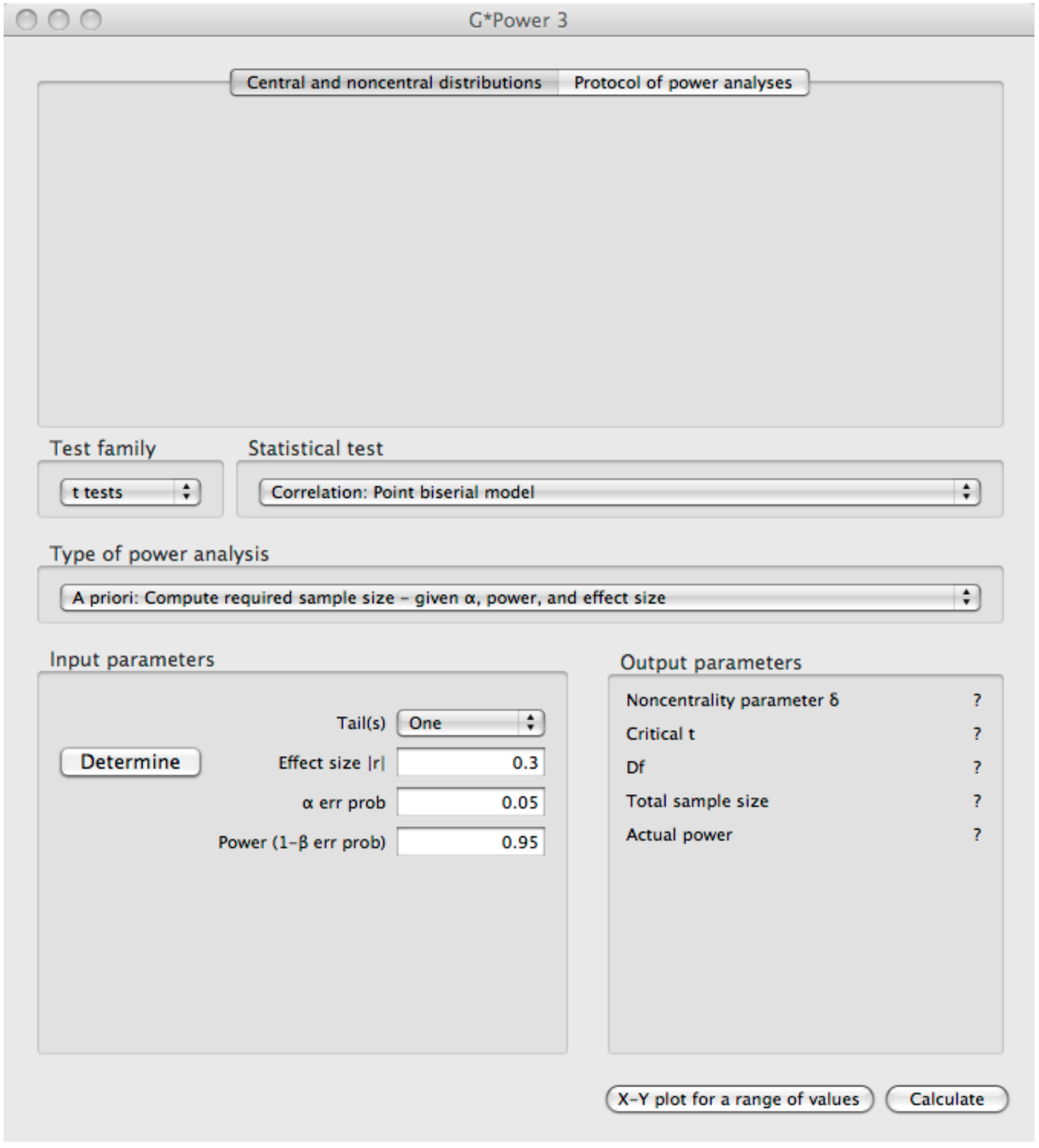
  - Criterion (α and the associated decision criterion are computed as a function of 1-β, the effect size, and N)

  - Post-hoc (1-β is computed as a function of α, the population effect size, and N)

  - Sensitivity (population effect size is computed as a function of α, 1-β, and N)

- G*Power is available for Mac OS X and Windows. **G*Power is free.**

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

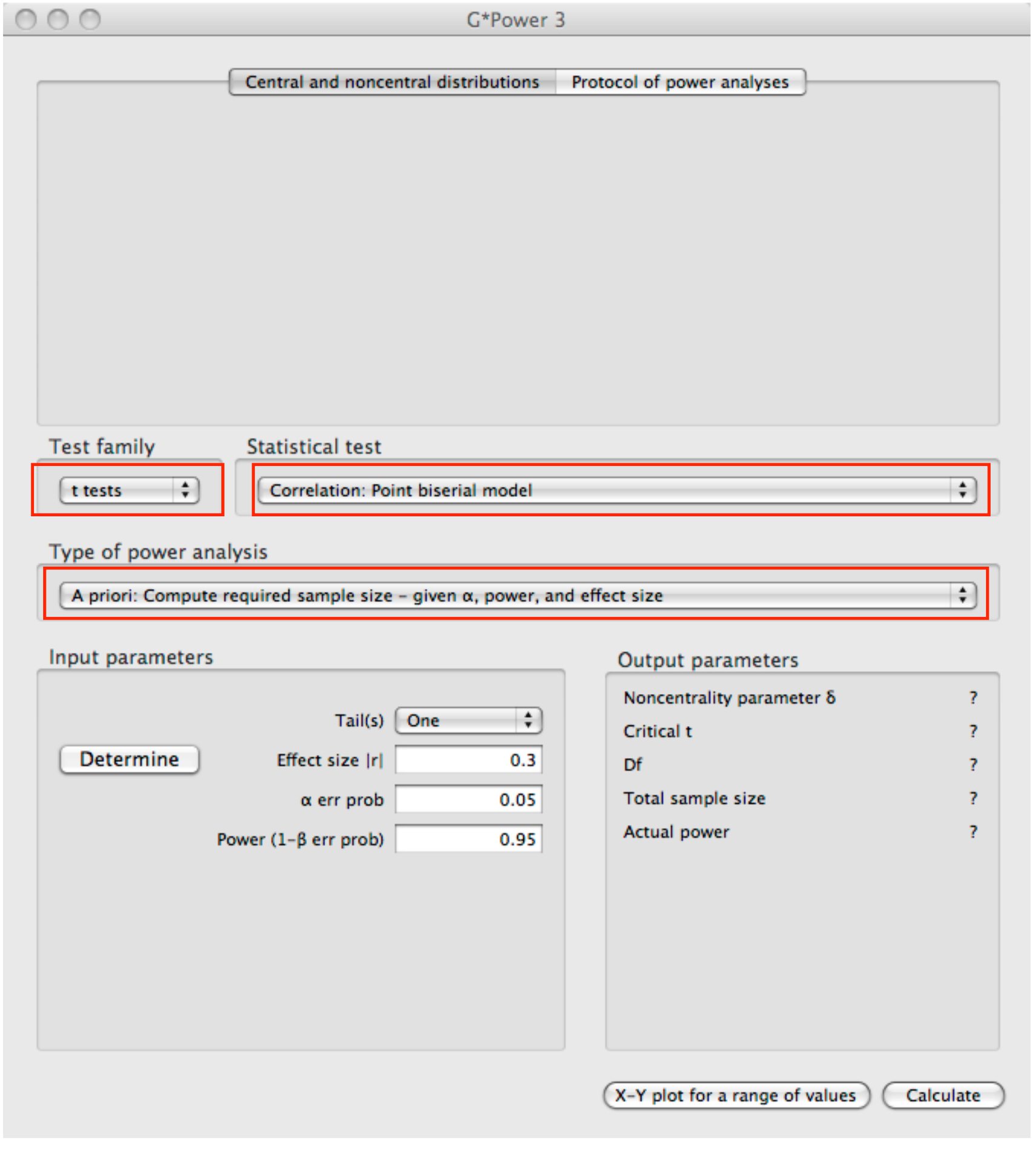


Let’s return to our earlier example.

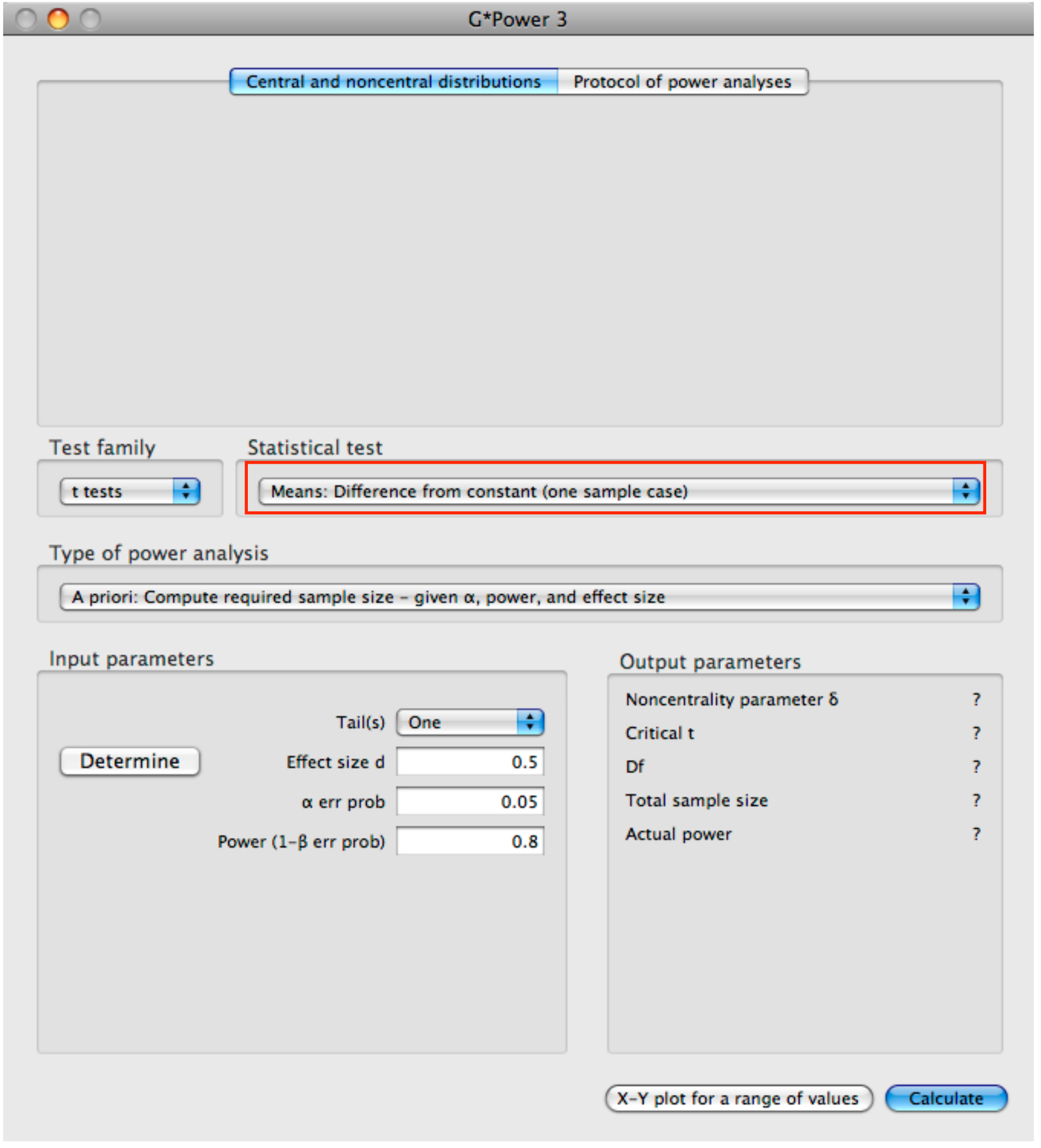
A clinical psychologist wants to test hypothesis (H1) that people who seek treatment have higher IQs than general population. She wants to use IQs of 25 randomly sampled patients and is interested in a difference of 5 points between the mean of the general population and the mean of her client population.

So, $\mu_0 = 100$, $\mu_1 = 105$, $\sigma = 15$

First we need to select the Test Family,Type of Test and Type of Power Analysis.



Select one sample t-test



Click on ‘Determine’ to get the additional parameters window

Power Plot

Mean H0 100

Mean H1 105

SD σ 15

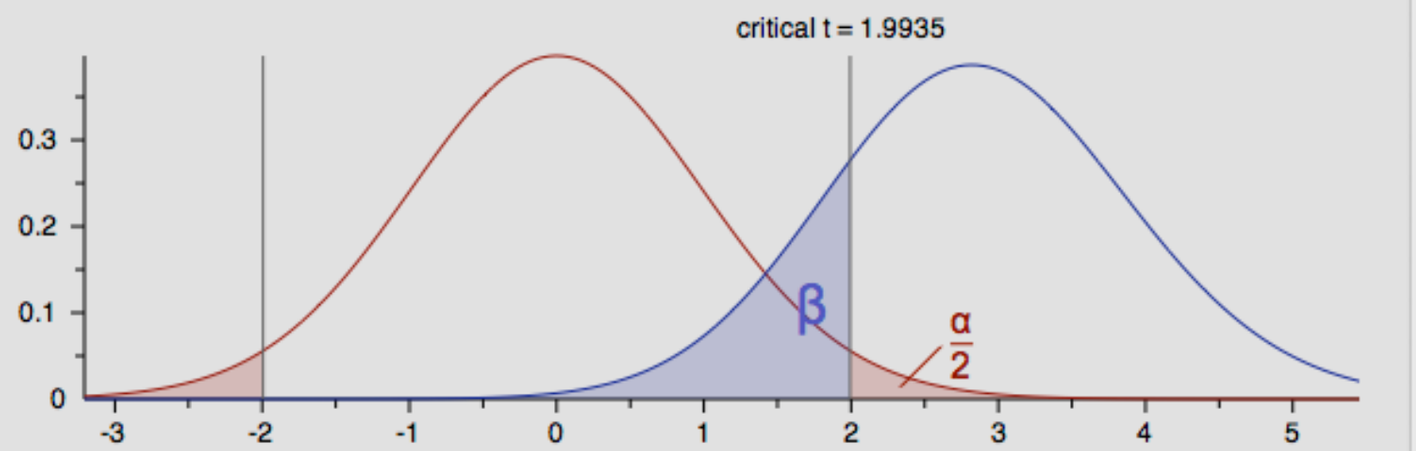
Calculate Effect size d 0.3333333

Calculate and transfer to main window

Close effect size drawer

G*Power 3

Central and noncentral distributions | Protocol of power analyses



Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input parameters

Determine

Tail(s): Two

Effect size d: 0.3333333

α err prob: 0.05

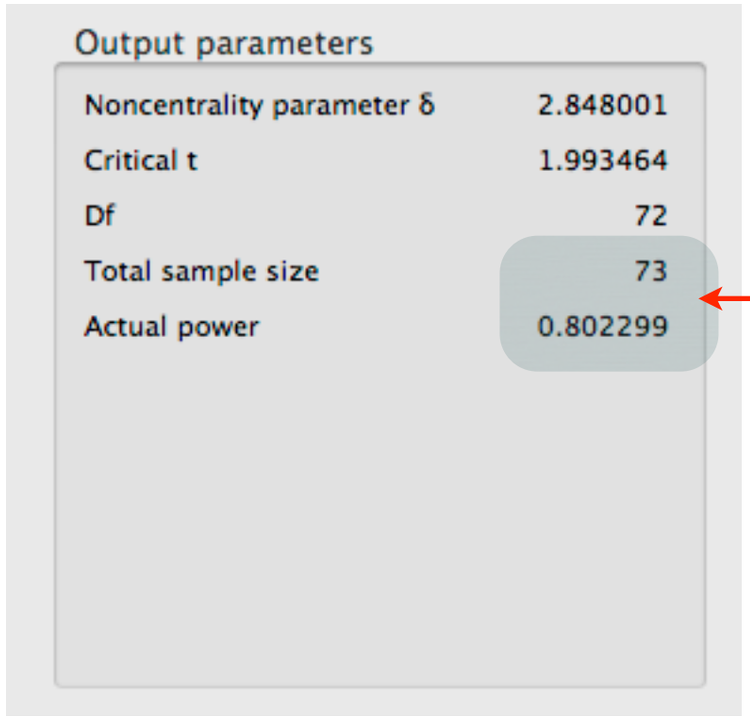
Power (1–β err prob): .8

Output parameters

Noncentrality parameter δ	2.848001
Critical t	1.993464
Df	72
Total sample size	73
Actual power	0.802299

X-Y plot for a range of values

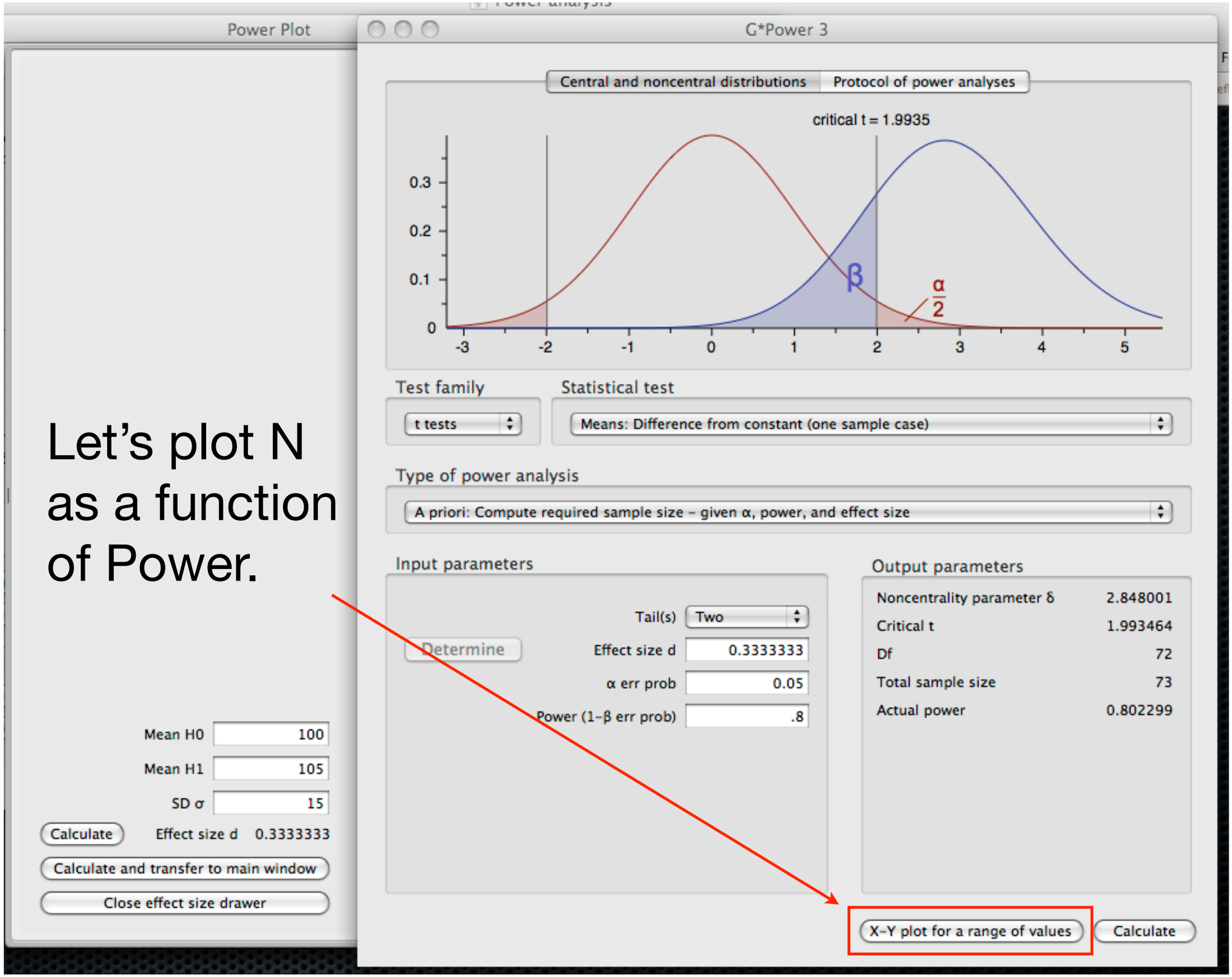
Calculate



So according to G*Power, we need a sample size of 73 for a Power of just greater than 0.8

When we calculated sample size manually, we worked it out to be 72. Why the difference? This is because of rounding error. In G*Power, if we select the Power level to be 0.795 (which we would round up to 0.8), it calculates our sample size as 72.

Let's plot N as a function of Power.

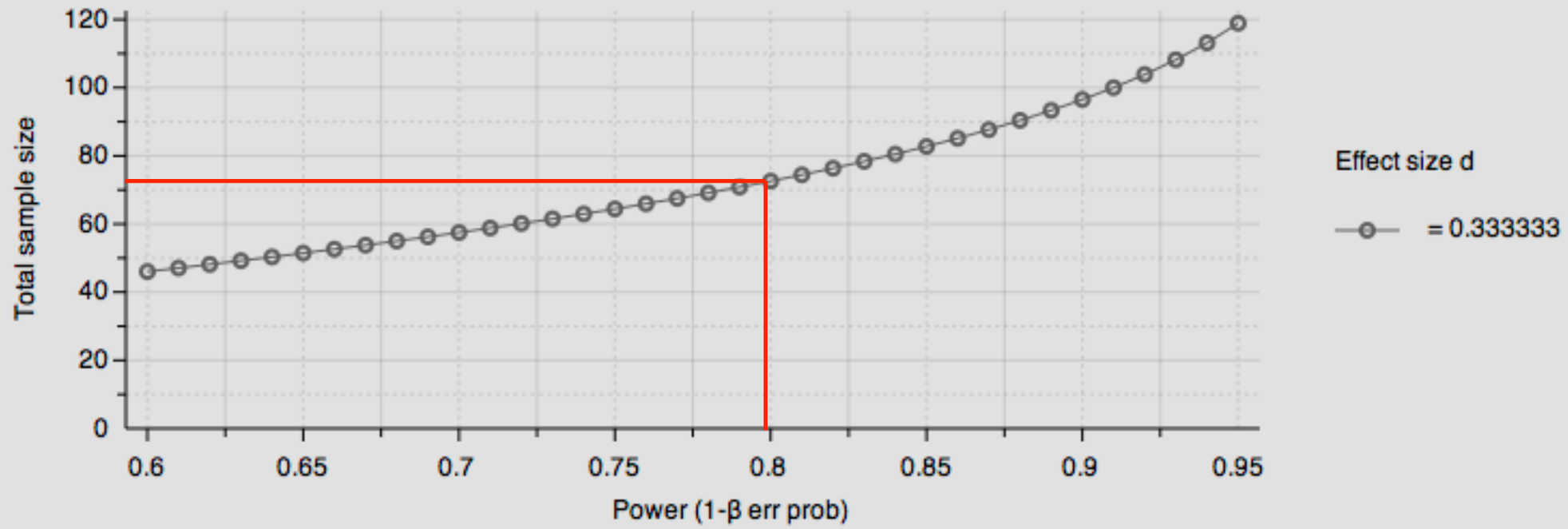


Power Plot

Graph | Table

t tests - Means: Difference from constant (one sample case)

Tail(s) = Two. α err prob = 0.05. Effect size d = 0.333333



Parameters

Plot (on y axis) Total sample size ☑ with markers ☐ displaying the values in the plot

as a function of Power (1-β err prob) from 0.6 in steps of 0.01 through to 0.95

Plot 1 graph(s) interpolating points

with Effect size d at 0.3333333

and α err prob at 0.05

Draw plot

Calculate by hand

Imagine we have two independent groups (20 Ss in each group). We want to know what the power is of a particular two-tailed experiment (with an α level of 0.05) where the difference between the group means is 2, and σ is 5.

δ	α = 0.05
1	0.17
1.1	0.2
1.2	0.22
1.3	0.26
1.4	0.29
1.5	0.32

For the same experiment, how many total Ss would we need to have a design with a Power of 0.8?

Please work this out long hand first of all. Then use G*Power to check your answer.

δ	α = 0.05
2.6	0.74
2.7	0.77
2.8	0.8
2.9	0.83
3	0.85
3.1	0.87

For more complex designs, or with more complex analyses you may be better off simulating your data - for example, simulating 10,000 "experiments" to determine whether the statistical model you want to build is able to detect the effect size that you are looking for.

You need to have an idea as to what kind of effect size will be theoretically important (which is not always easy to determine in the case of some theories).

References on Experimental Power

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112,* 155-159.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Howell, D.C. (2007). Statistical Methods for Psychology, Sixth Edition. Thompson.

Before next week - Online introductory guide to R, RStudio, and R Markdown

This is a very clear and focused introduction to R, RStudio, and R Markdown. Please read the first four chapters before next week…

http://rbasics.netlify.com

There is also a folder on Blackboard called “R cheatsheets” which contains lots of useful resources related to R.

Before next week - Installing R and RStudio

I highly recommend you bring a laptop (if you have one) to future MRes classes. Although R and RStudio are both installed on the PC cluster machines, it will be a lot easier for you to complete the R work on your own laptops.

You can install R from here:

https://www.stats.bris.ac.uk/R/

And RStudio from here:

https://www.rstudio.com/products/rstudio/download/#download