Lecture 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 around the end of November

ANOVA – Due around mid-January

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.

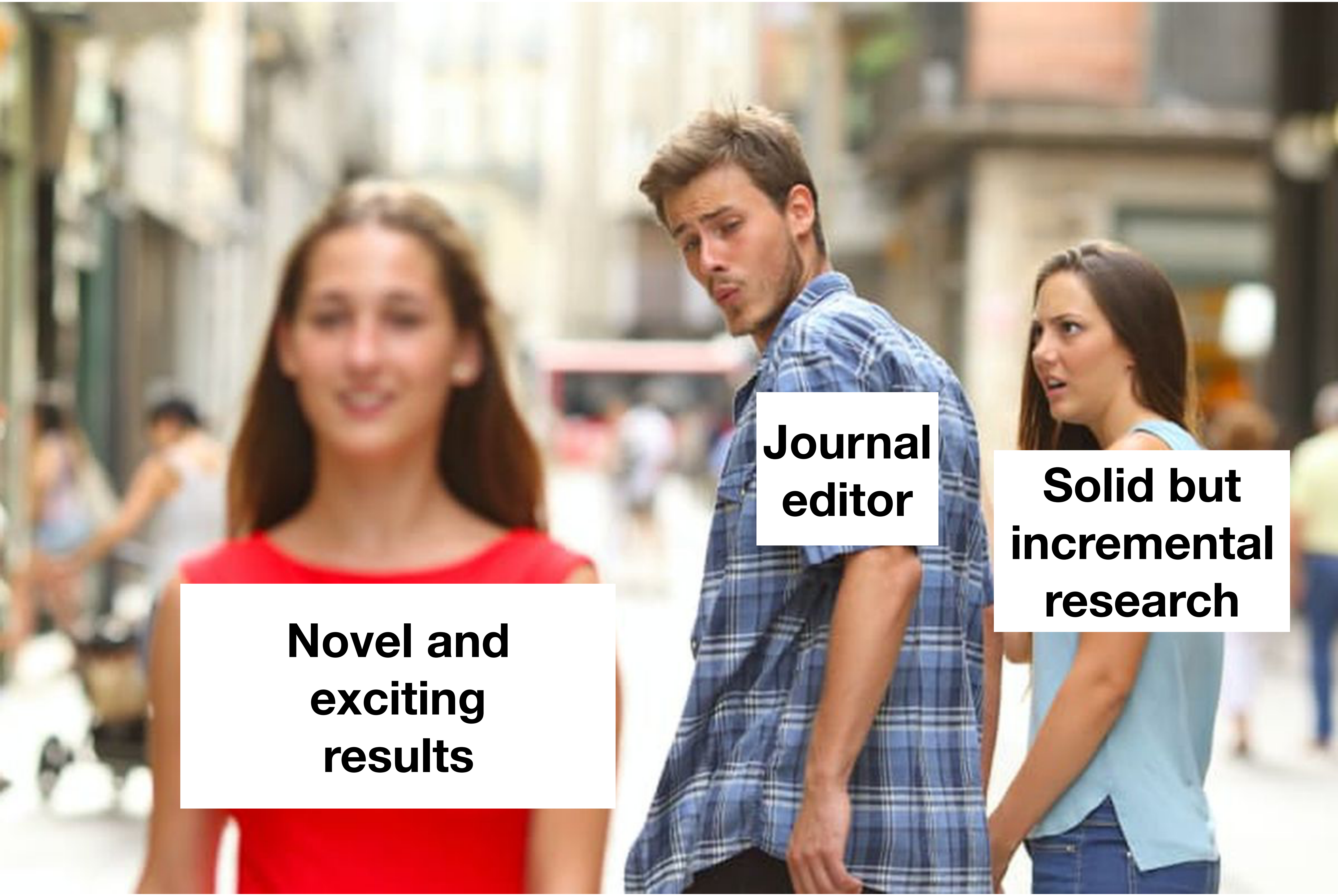
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience.

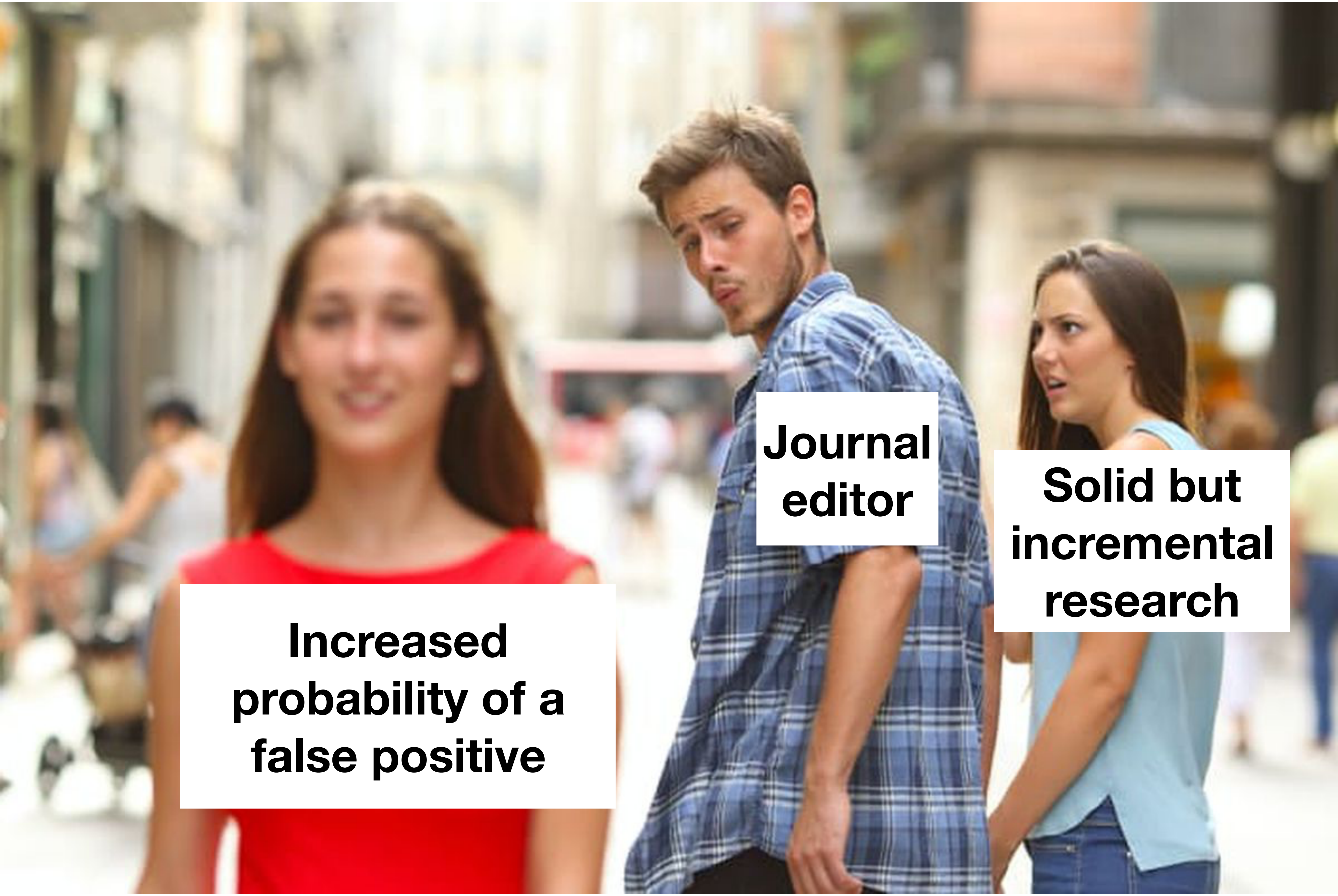
- MacLeod et al. (2014), *Lancet*, 85% of biomedical research resources are wasted.

- 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.

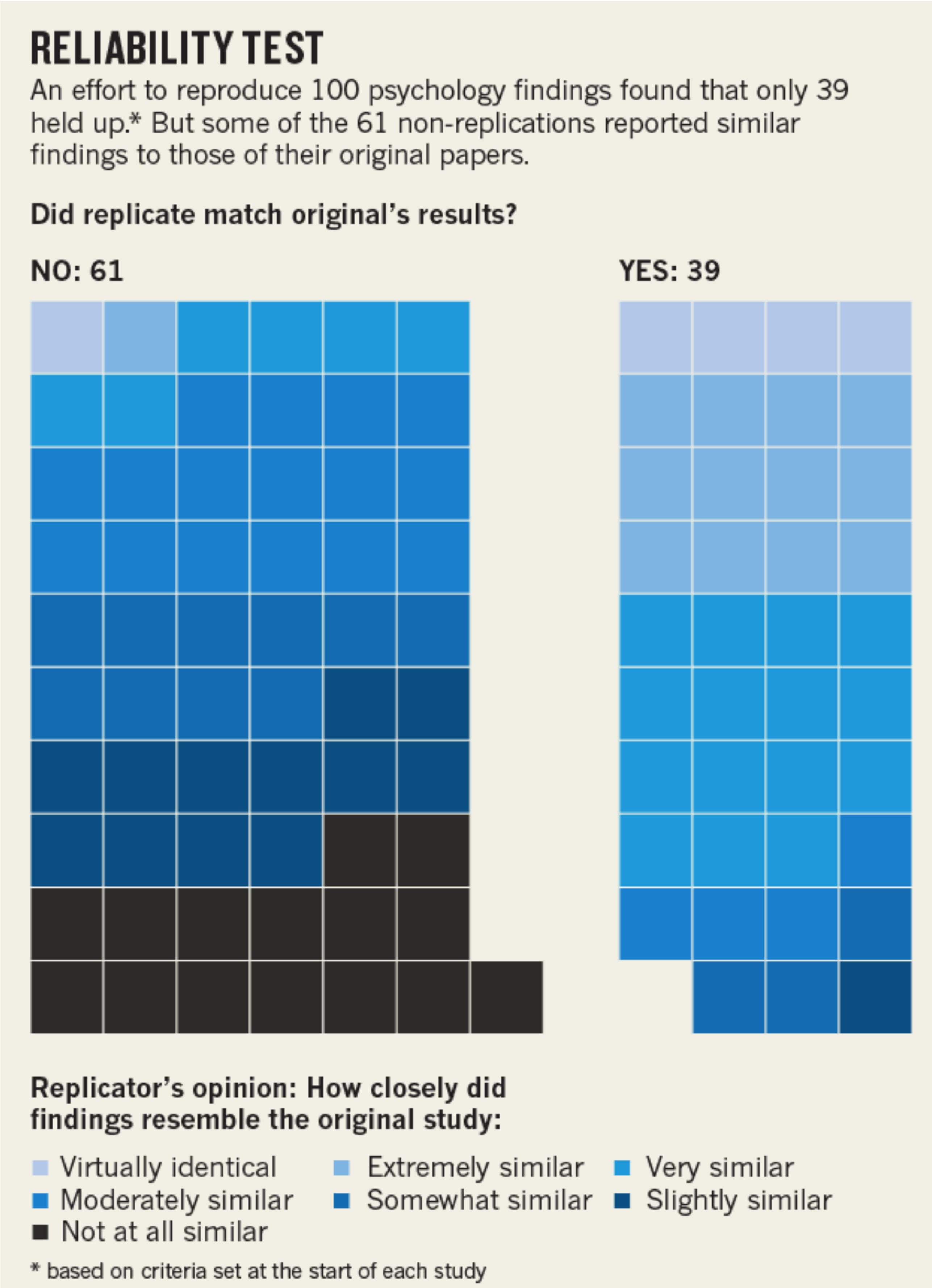
Note, I have uploaded a bunch of papers on the replication crisis to Blackboard.


Journal editor
Solid but incremental research
Novel and exciting results


Journal editor
Solid but incremental research
Increased probability of a false positive

(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

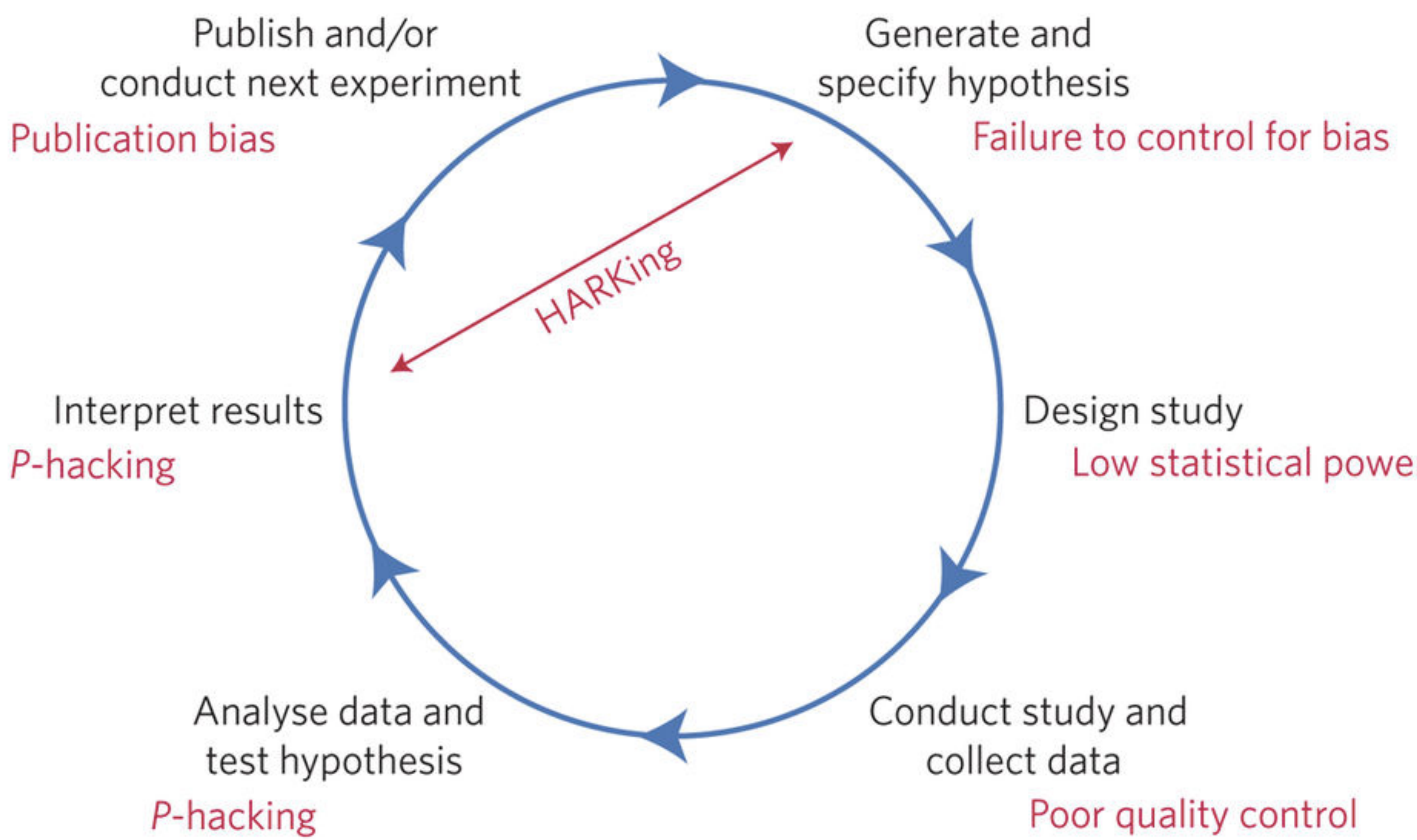
Group	Median power (%)	Minimum power (%)	Maximum power (%)	2.5^{th} and 97.5^{th} percentile (based on raw data)	95% HDI (based on GMMs)	Total N
All studies	23	0.05	1	0.05–1.00	0.00 – 0.72, 0.80 –1.00	730
All studies excluding null	30	0.05	1	0.05–1.00	0.01– 0.73, 0.79 –1.00	638
Genetic	11	0.05	1	0.05– 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

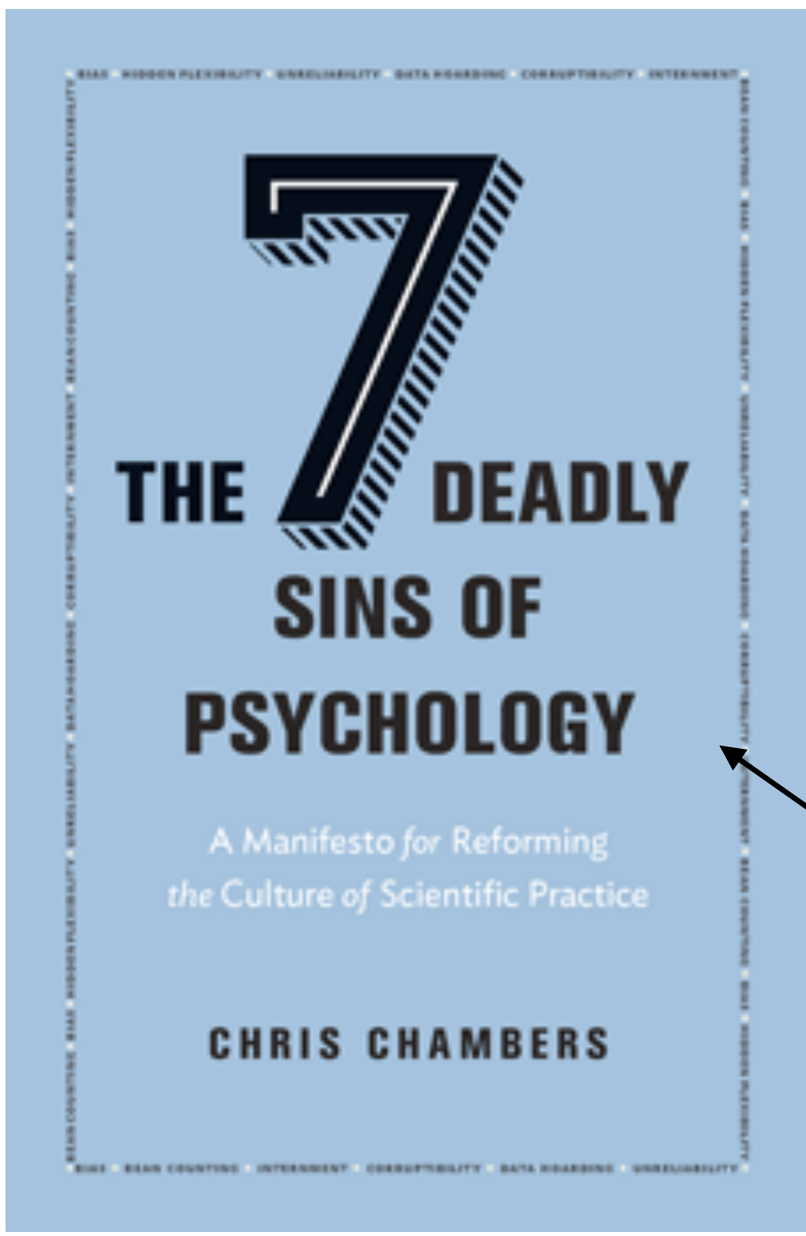
Replicable Science is when someone else can run a study the same as or conceptually equivalent to your one, and find a similar pattern of effects.

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.

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.

In REF2021 UoA Environment…

29. The revised template will also include a **section on 'open research'**, detailing the submitting unit's open access strategy, 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. The panels will set out further guidance on this in the panel criteria.

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

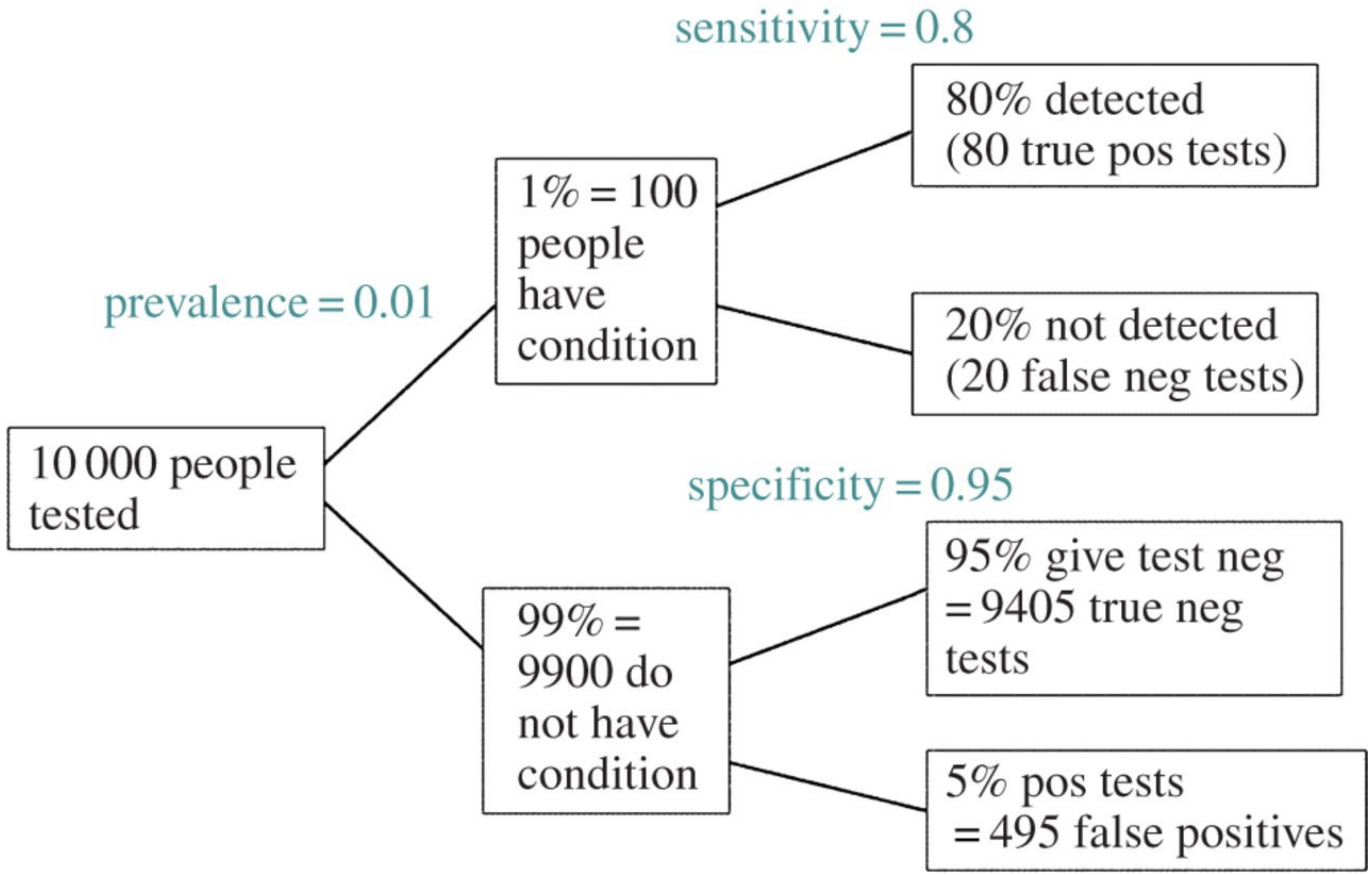
a. Promote open science in our teaching.
b. Design a Research Methods curriculum that teaches skills for open science and uses open science to enhance teaching (for example: teach R and use open data to practice analysis skills).
c. Learn about and adopt open educational practices in our teaching.
d. 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.

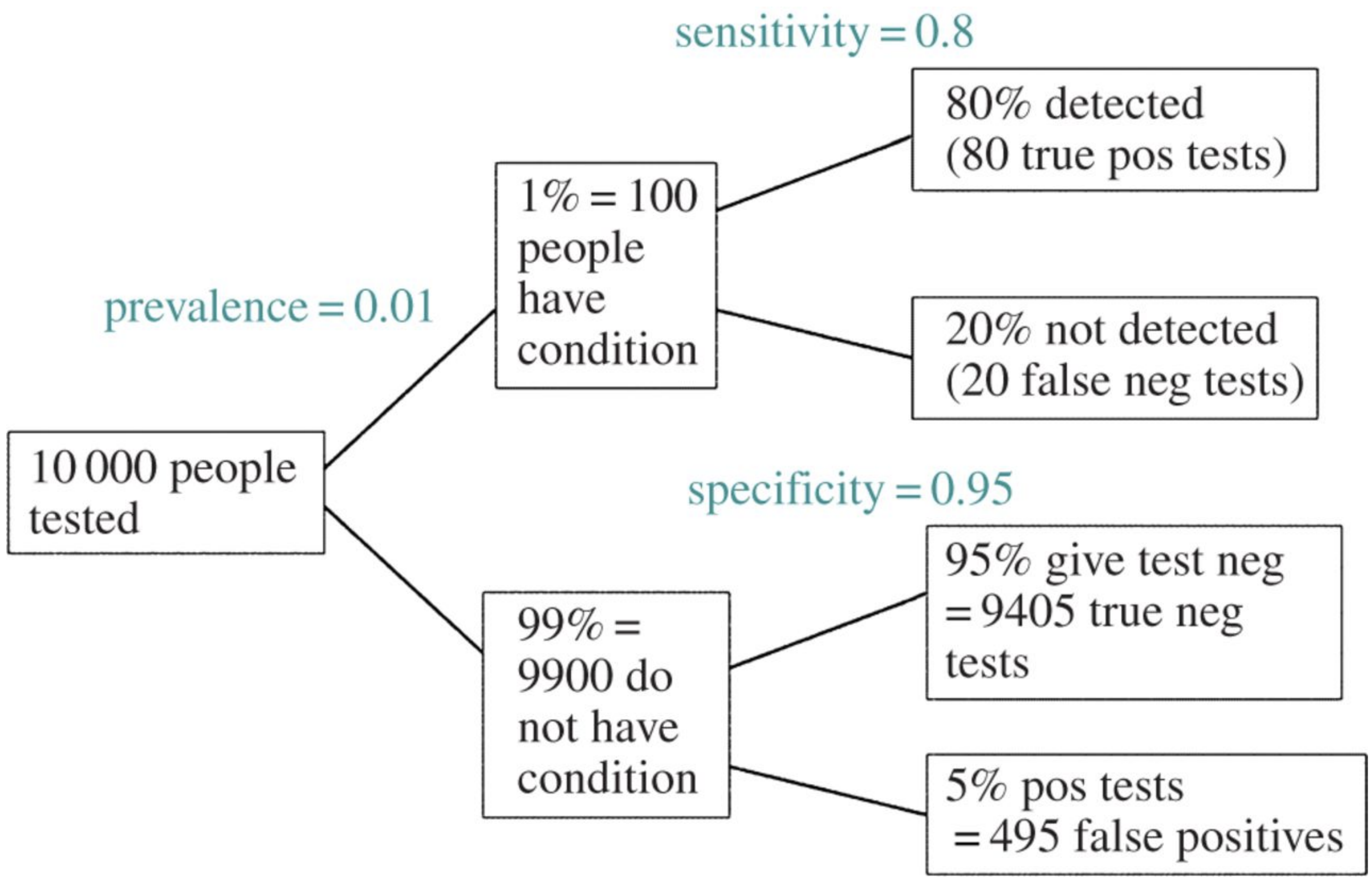
Part of doing better science involves knowing how to build appropriate statistical models, and how to understand what those models are telling you (and what they are not…)

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


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

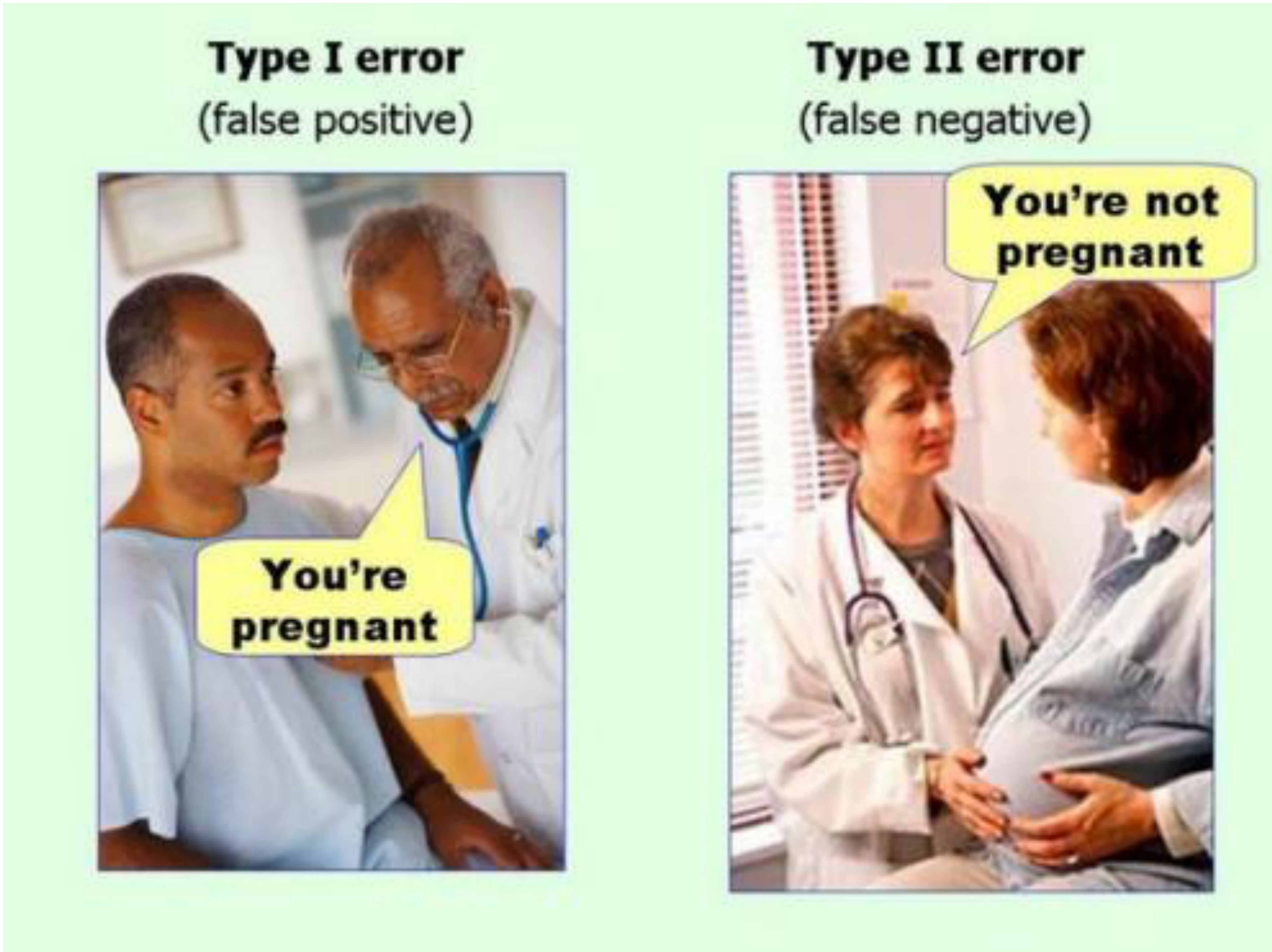
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

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



- 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

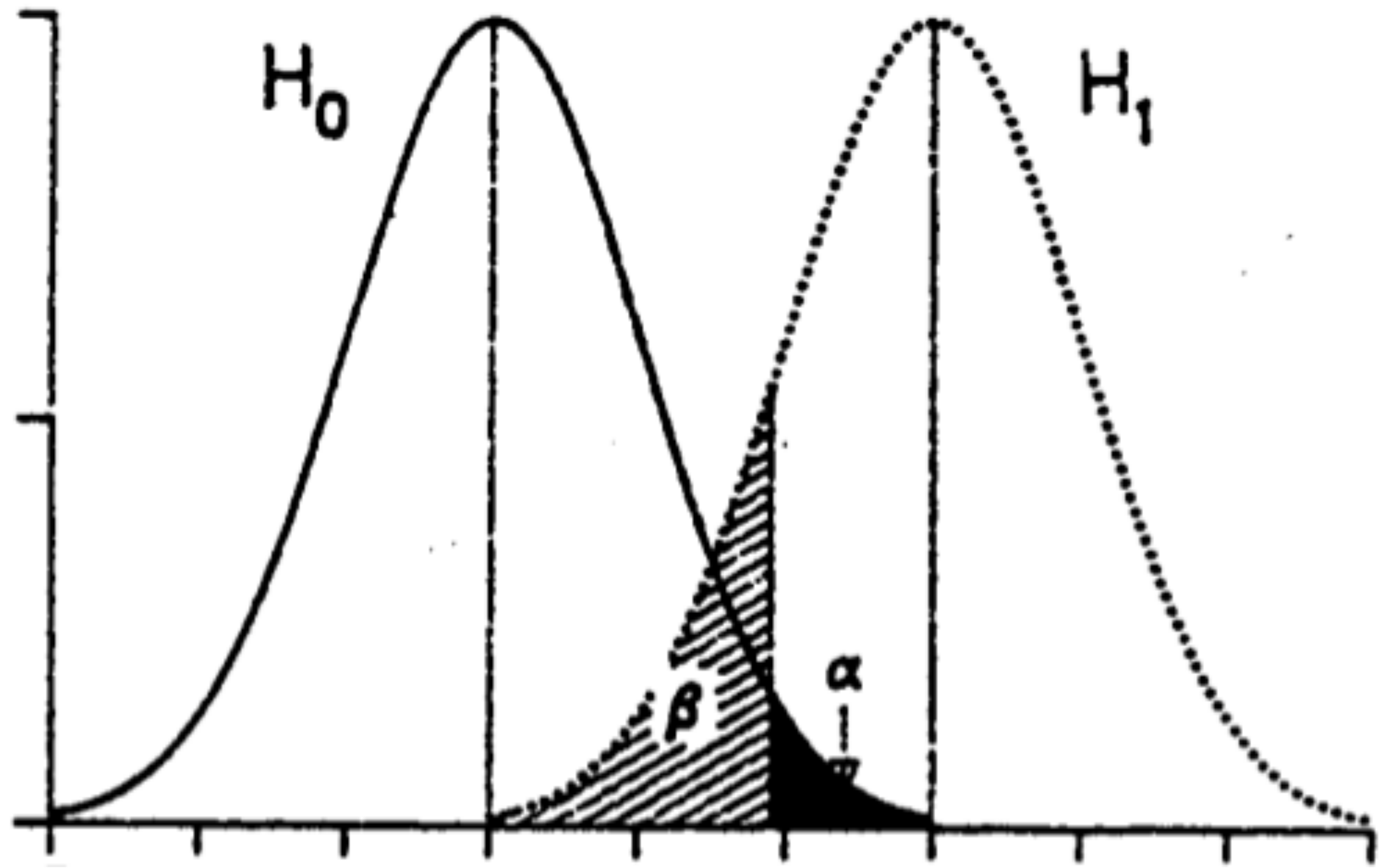
Is Power That Big a Deal?

- 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

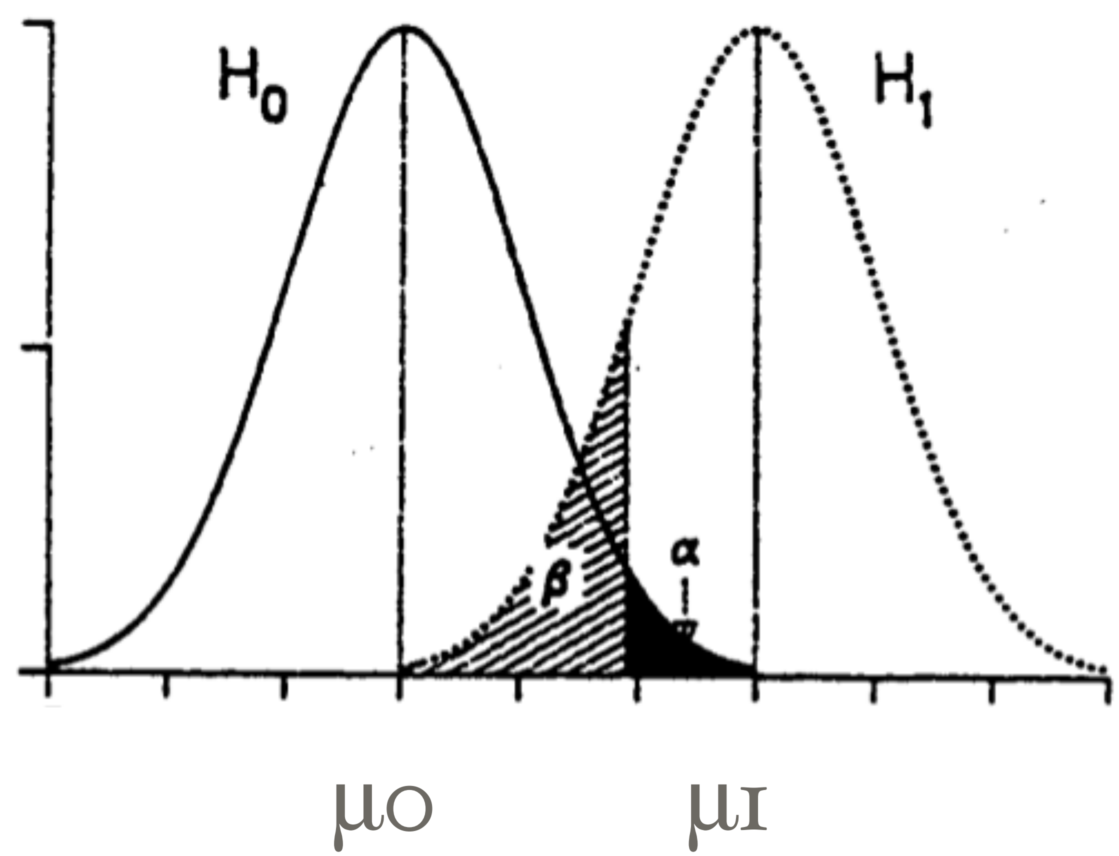
- We can measure the size of an experimental effect in an objective, standardised manner.
- The two most common measures of effect size are Cohen's d and Pearson's r.

	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, $\mu 0 = 100$, $\mu 1 = 105$, $\sigma = 15$

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

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

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 $\delta=1.65$, 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 ($1-\beta$) 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 ($\mu1-\mu2$) under H1 minus the difference ($\mu1-\mu2$) under H0, divided by σ. Under H0 though, ($\mu1-\mu2$) is zero (because there is no difference between the means under the null hypothesis) so,

$$d = \frac{(\mu1-\mu2) - 0}{\sigma} = \frac{(\mu1-\mu2)}{\sigma}$$

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.

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

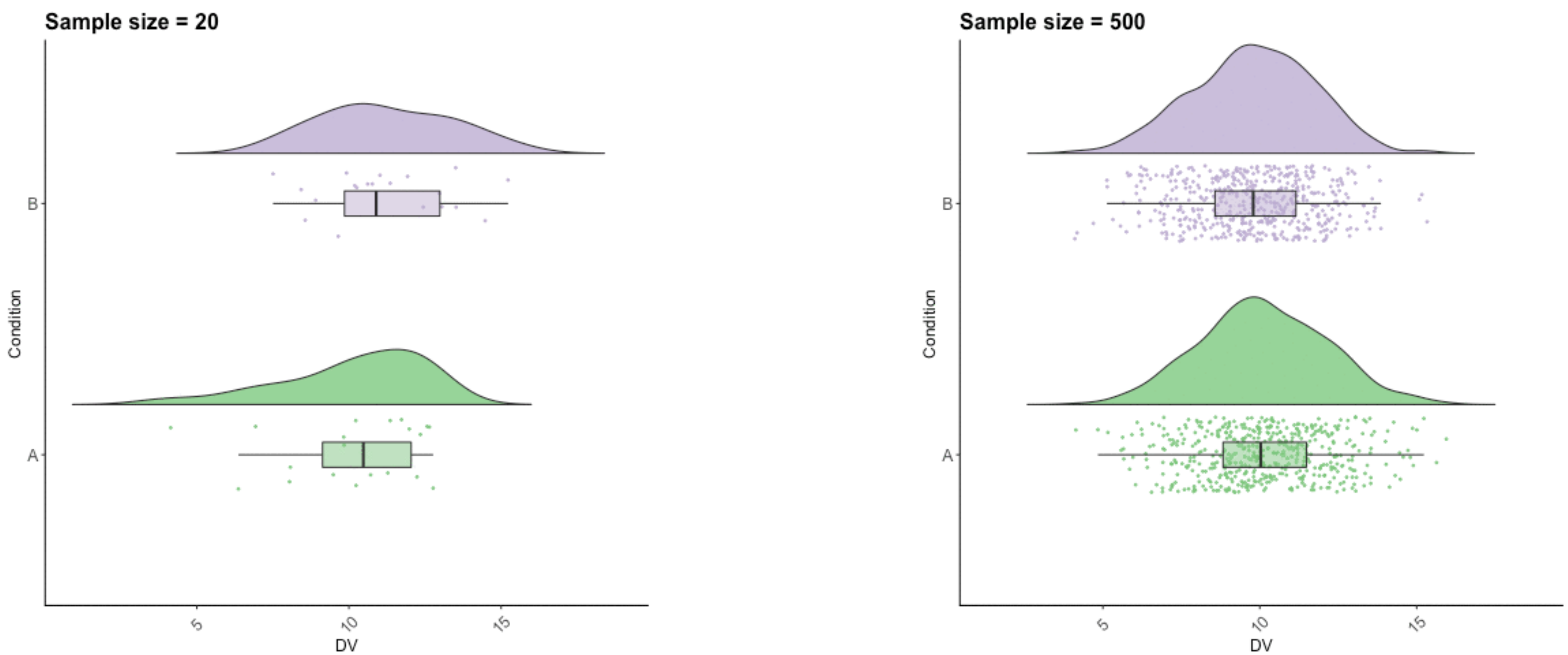
$$= 1.77$$

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

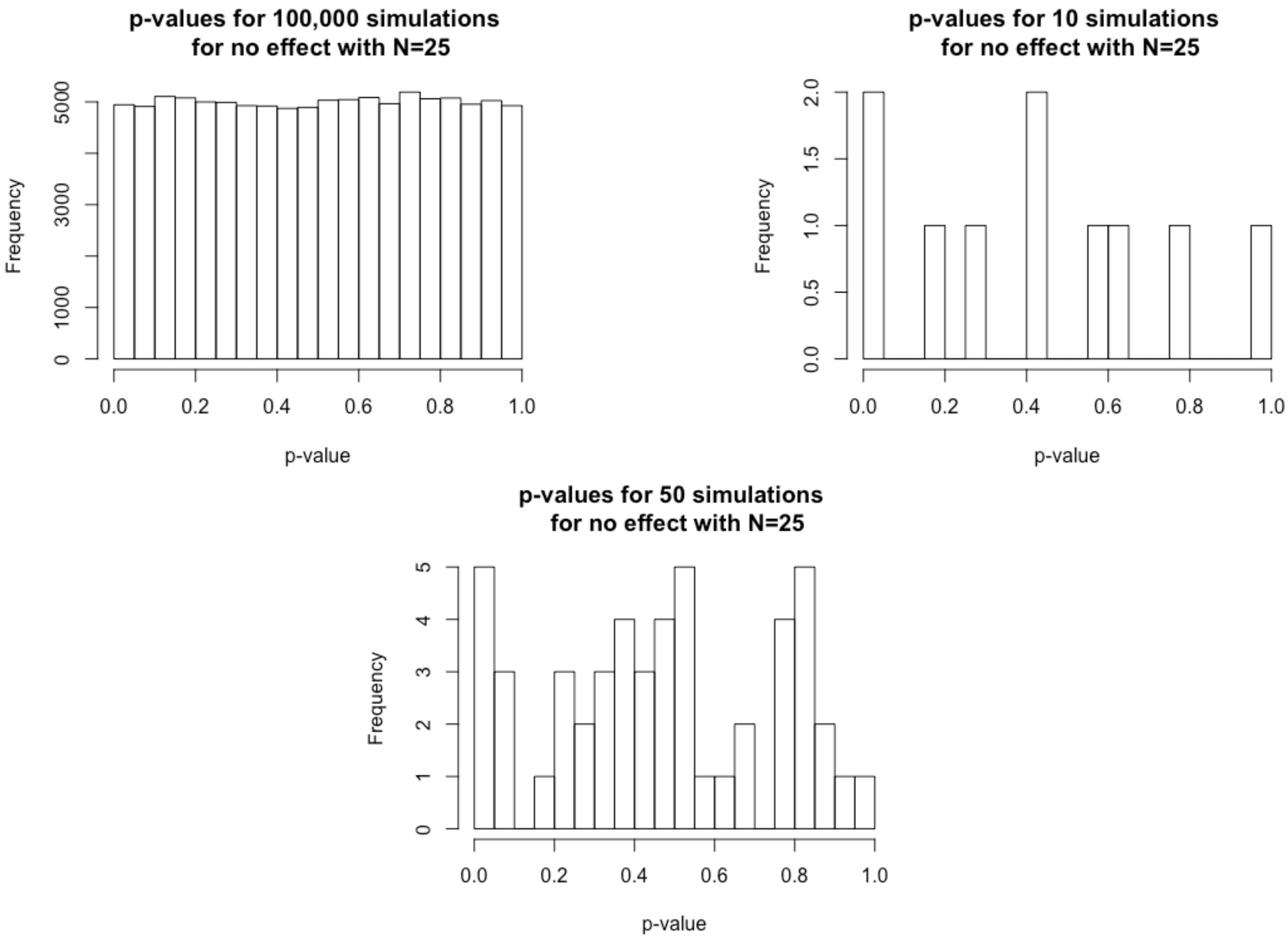
With $\alpha = 0.05$ and $\delta = 1.77$, Power is about 0.43

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

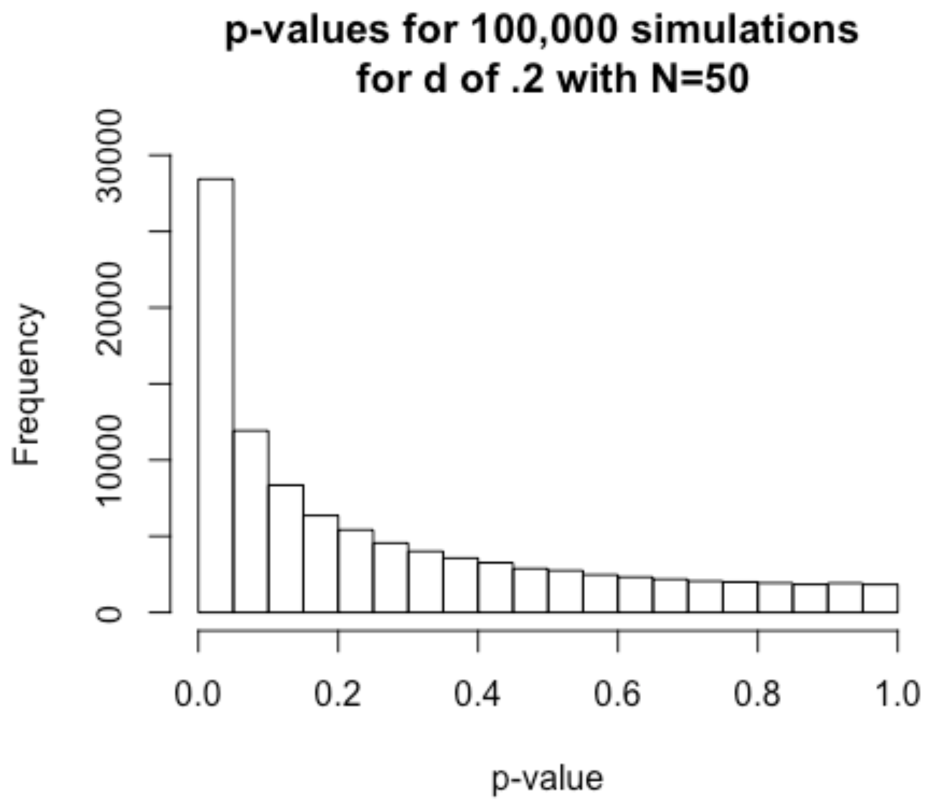
The Problem of Sampling Bias



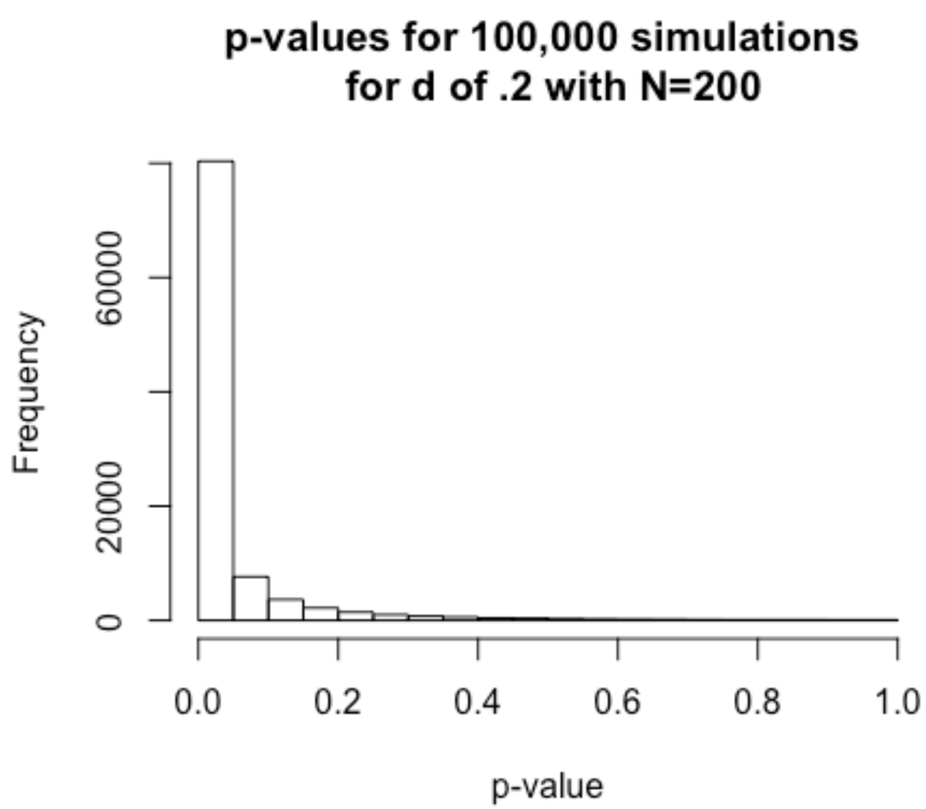
Samples for conditions A and B are drawn from the same population. Due to sampling error, with small samples (e.g., N=20) we might sometimes conclude there is a difference between A and B where there isn't one (as you can see with the N=500 samples).


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)

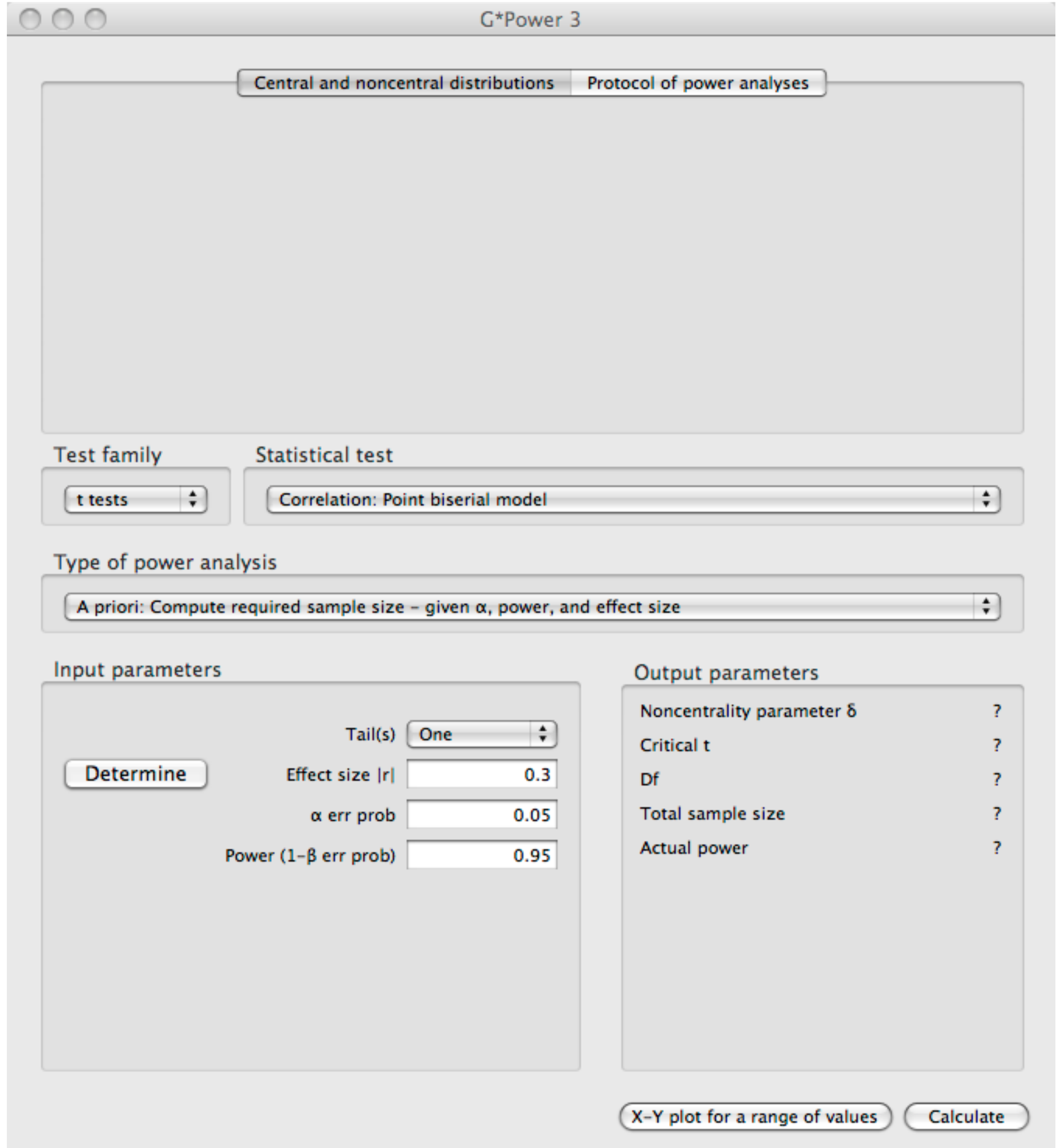
  - Criterion (α and the associated decision criterion are computed as a function of 1-β, the effect size, and N)

  - 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.

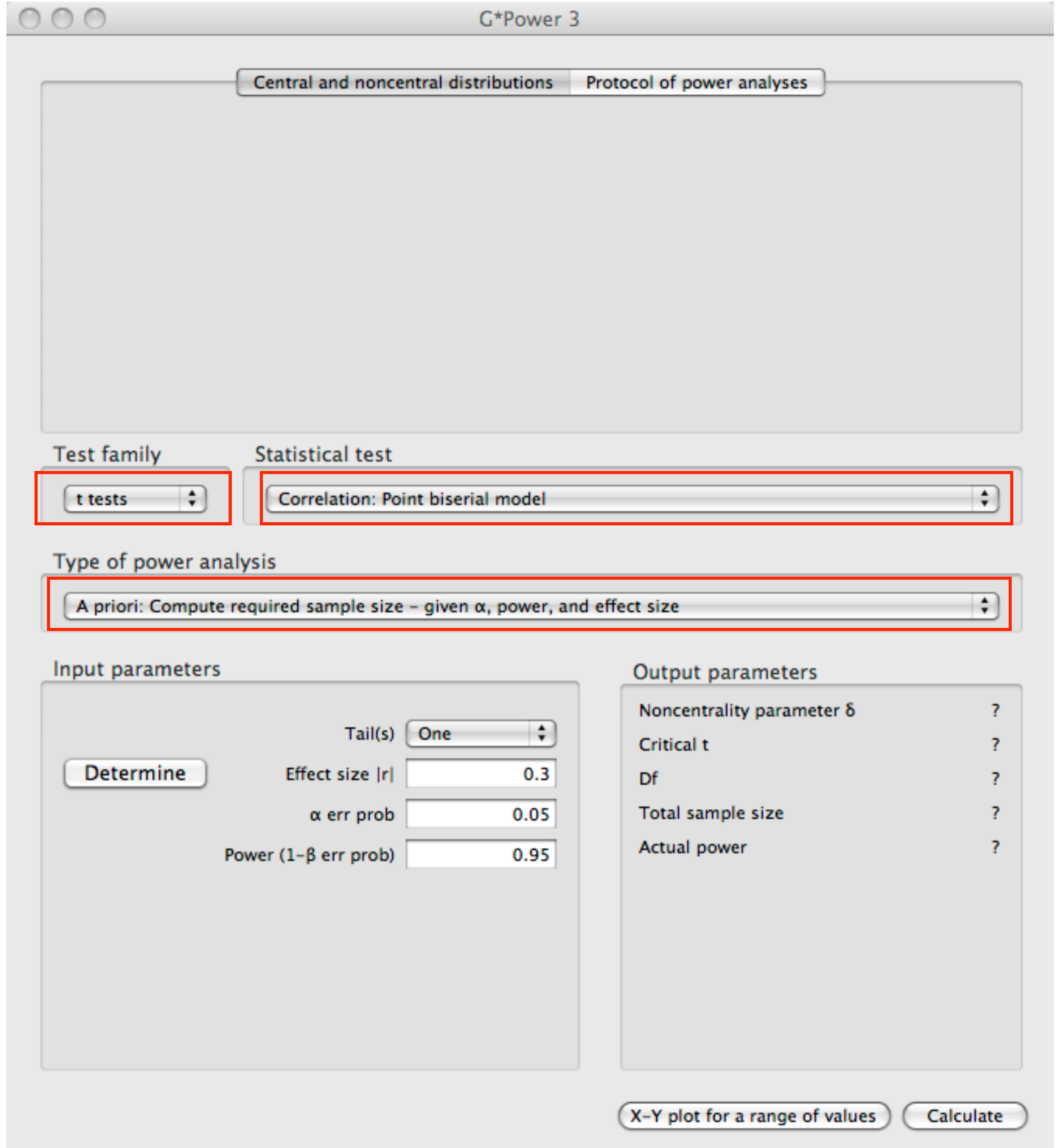


Let’s return to our earlier example.

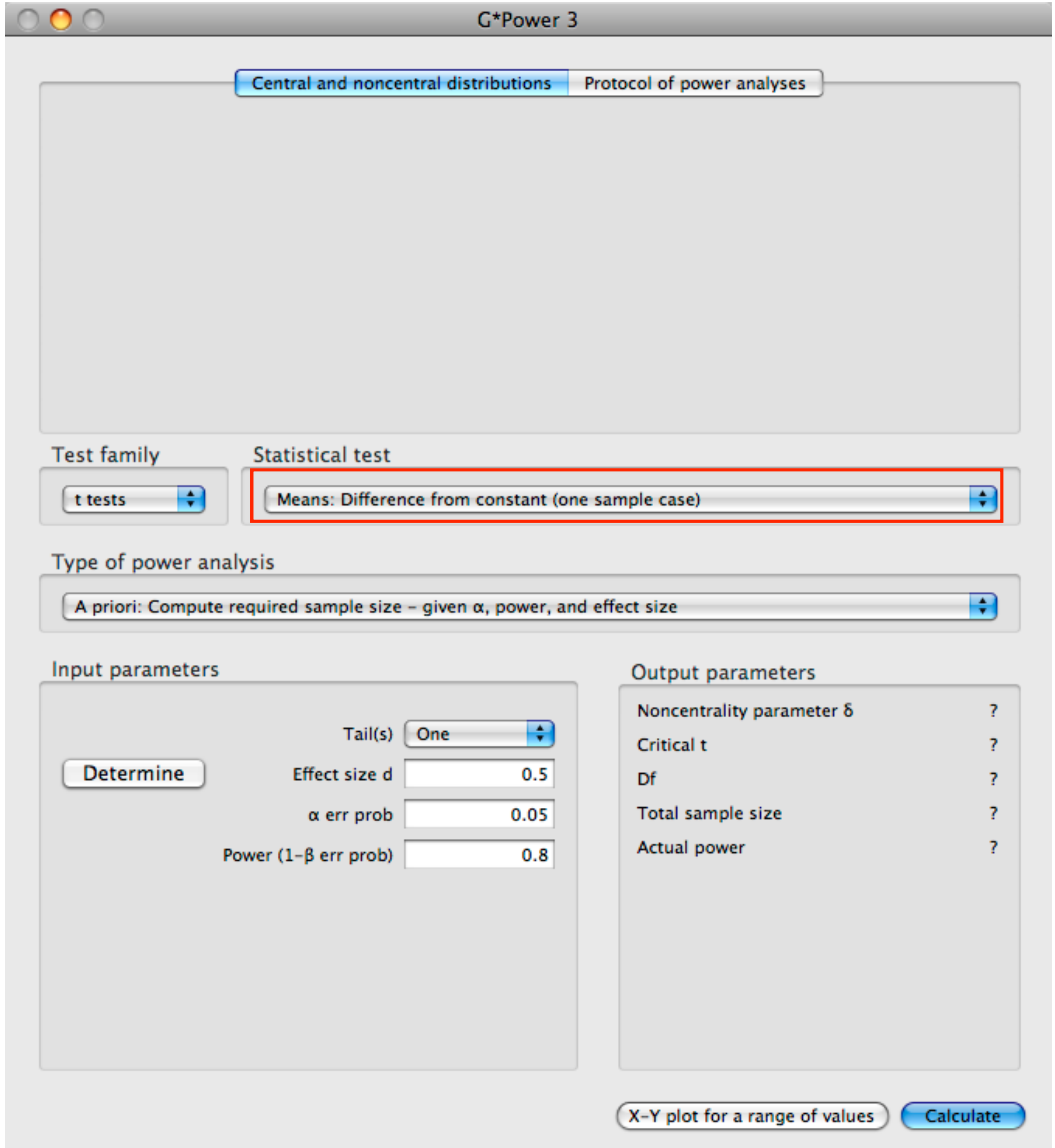
A clinical psychologist wants to test hypothesis (H1) that people who seek treatment have higher IQs than general population. She wants to use IQs of 25 randomly sampled patients and is interested in a difference of 5 points between the mean of the general population and the mean of her client population.

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

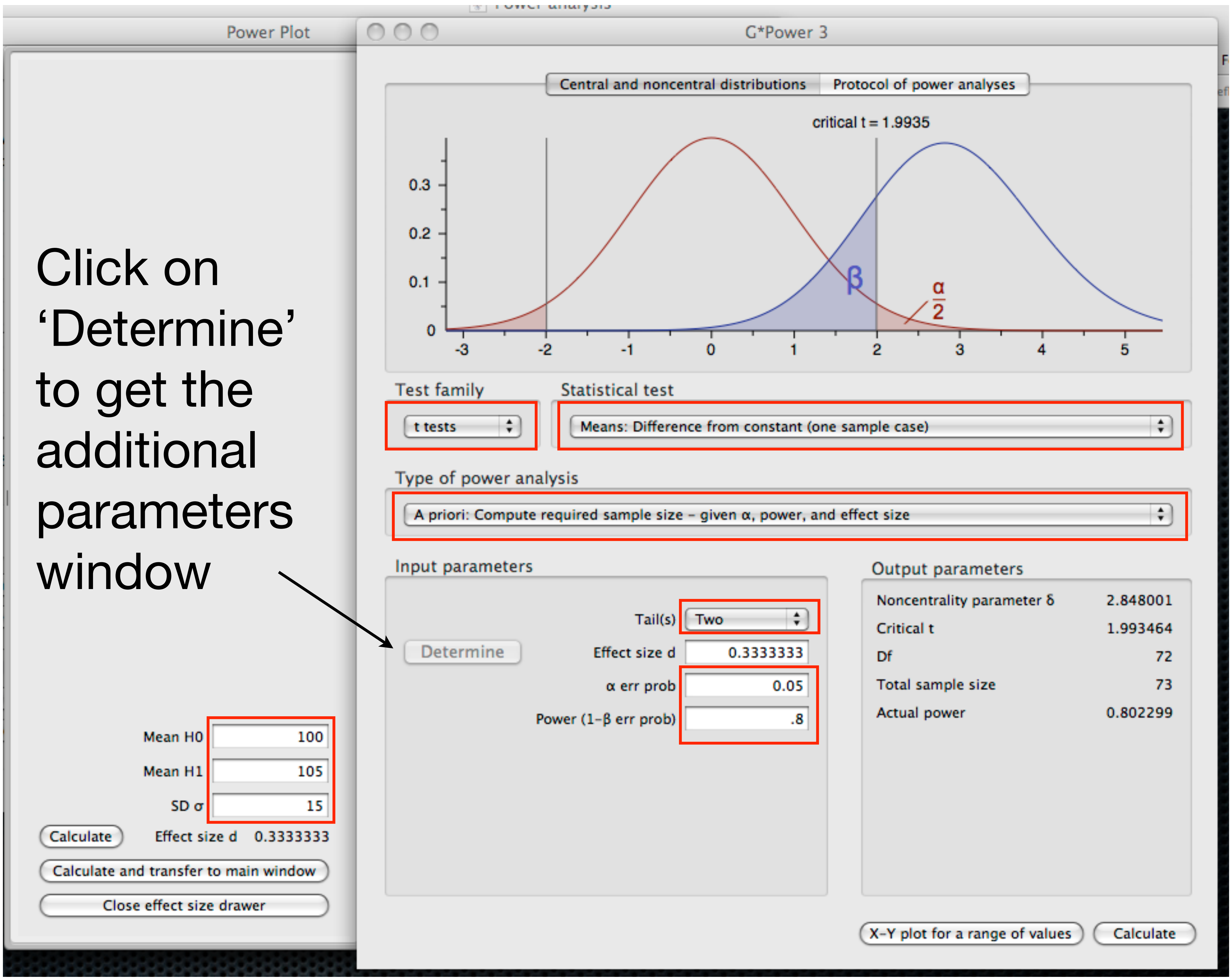
First we need to select the Test Family,Type of Test and Type of Power Analysis.



Select one sample t-test



Click on 'Determine' to get the additional parameters window



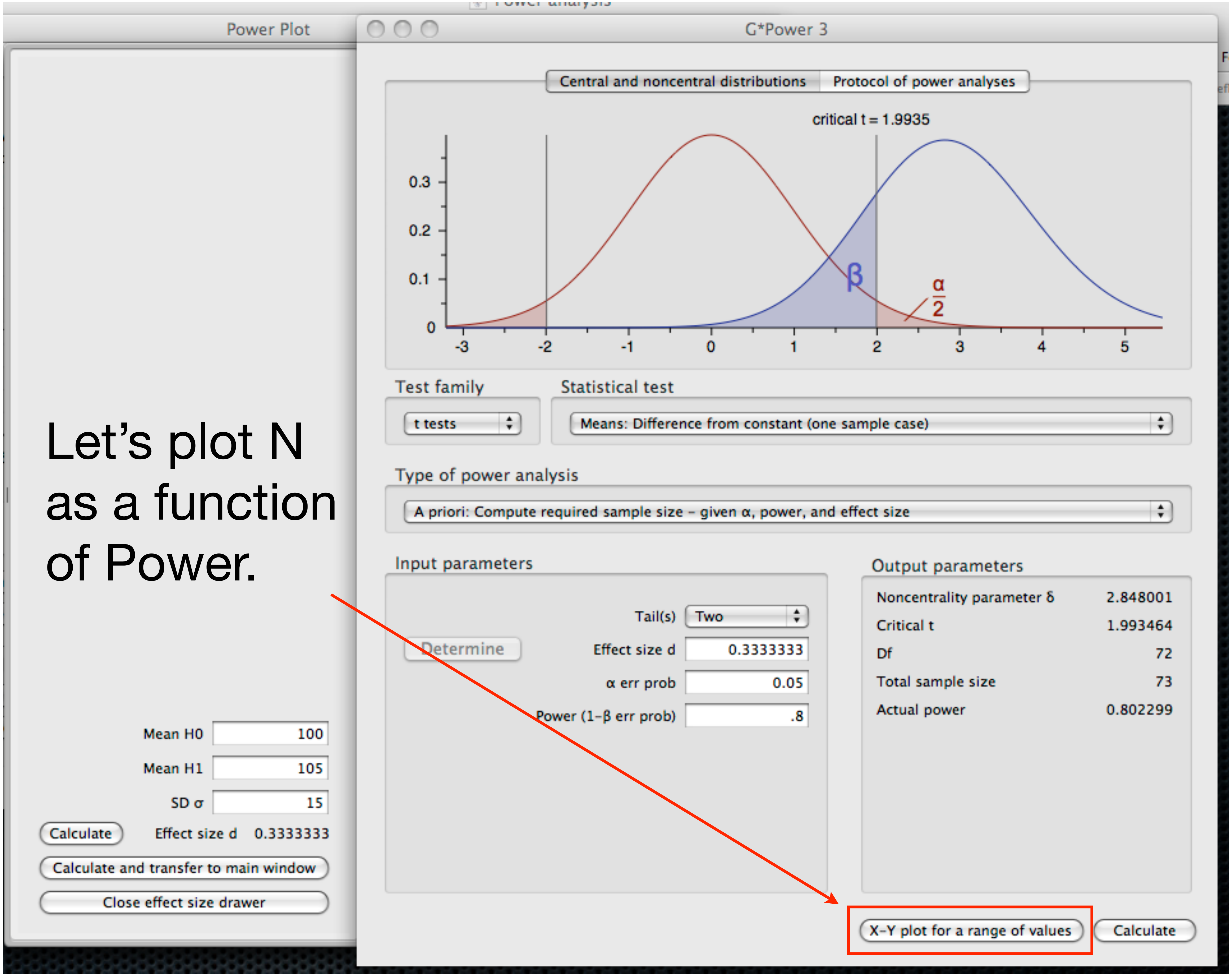
Output parameters

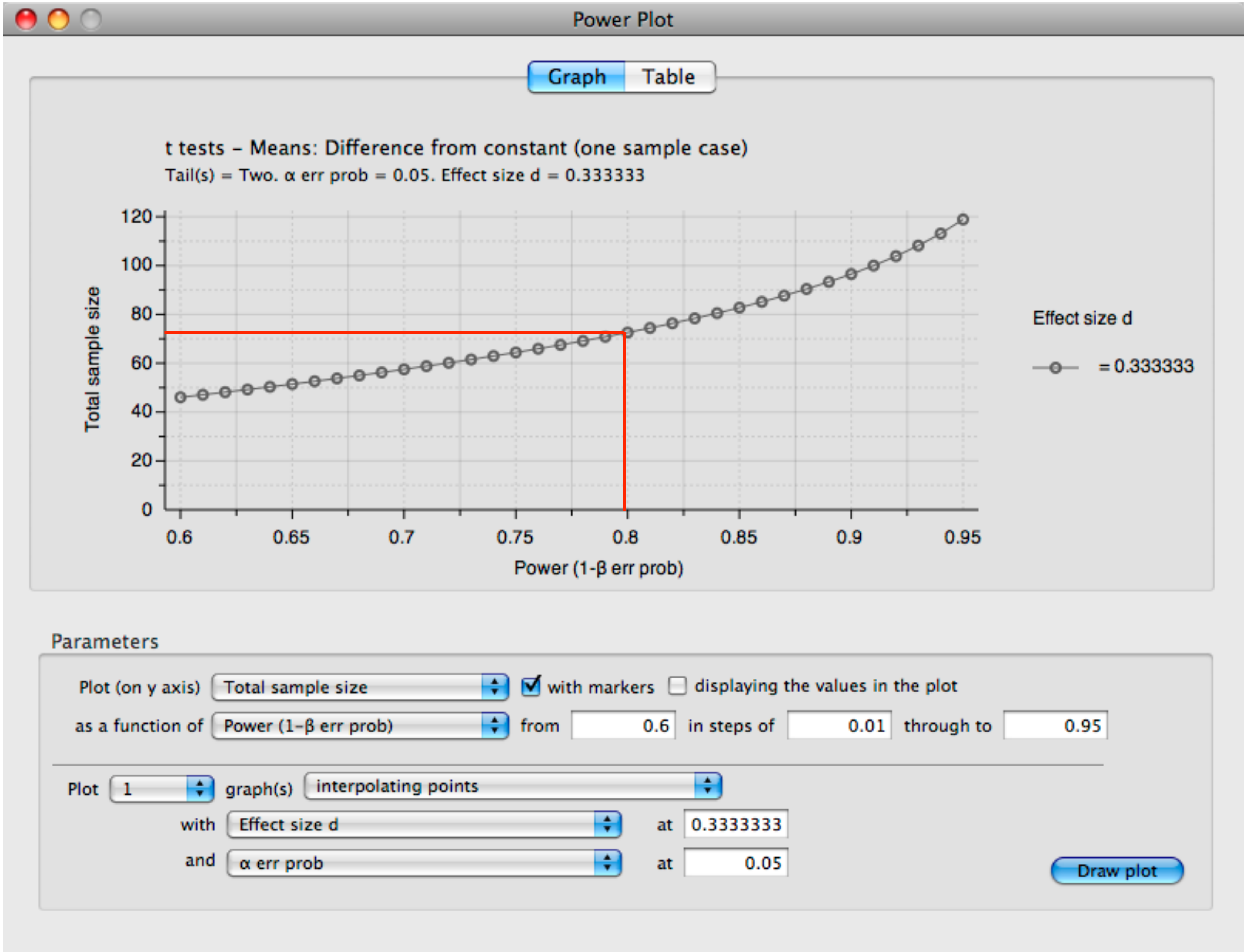
Noncentrality parameter δ	2.848001
Critical t	1.993464
Df	72
Total sample size	73
Actual power	0.802299

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
Total sample size
120
100
80
60
40
20
0
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Power (1-β err prob)
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	$\alpha = 0.05$
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