Power Analysis

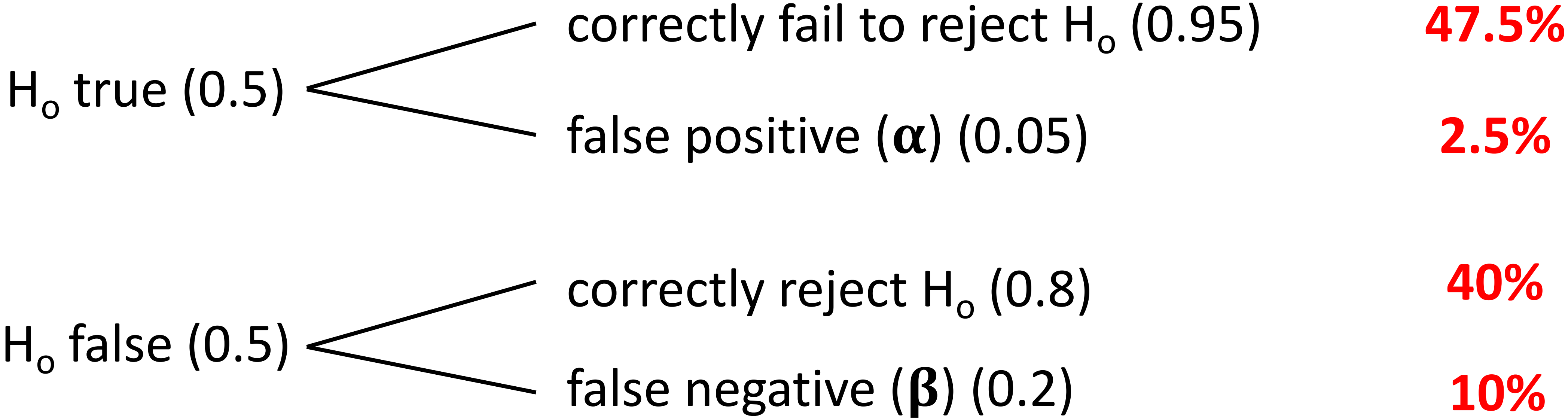
January 31, 2022

TRUTH

		No effect in population (H_o true)	True effect in population (H_o false)
Your statistical test	p > .05 Test not significant	😀	Type II error (**β**)
	p < .05 Test significant	Type I error (**α**)	😀 Power (1-**β**)

- In simulations we know whether H_o is true or not
- In reality we do not know and we are making an inference based on statistical test
- If you really don't know before you start your experiment, what's the most likely outcome? Say 50/50 chance H_o is false

5% significance level, 80% power



H_o true (0.5)

correctly fail to reject H_o (0.95) **47.5%**

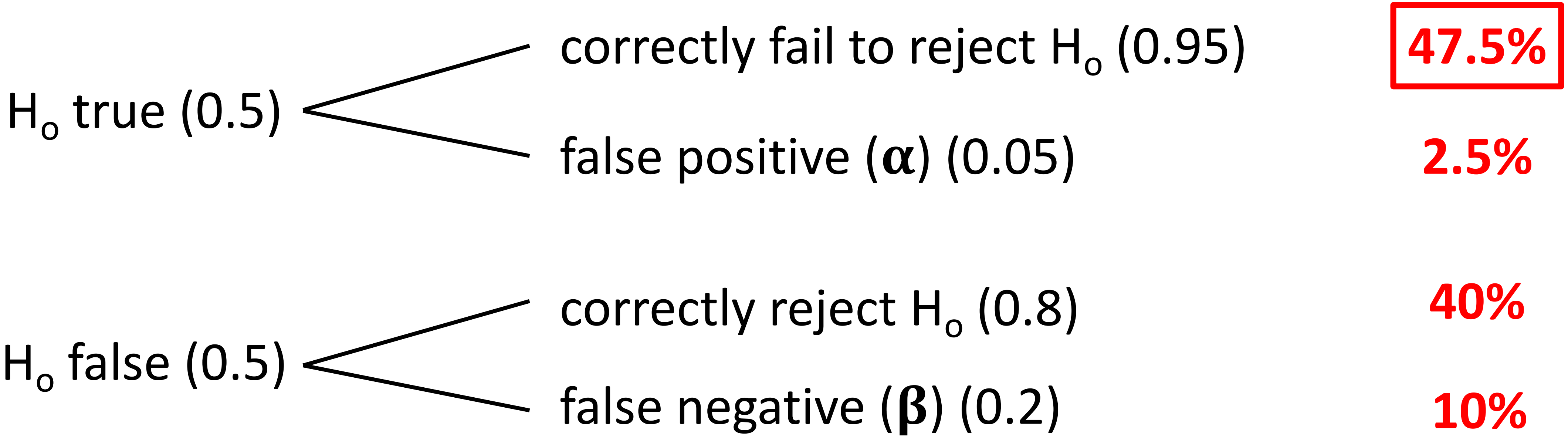
false positive (**α**) (0.05) **2.5%**

H_o false (0.5)

correctly reject H_o (0.8) **40%**

false negative (**β**) (0.2) **10%**

5% significance level, 80% power



H_o true (0.5)

correctly fail to reject H_o (0.95) **47.5%**

false positive (**α**) (0.05) **2.5%**

H_o false (0.5)

correctly reject H_o (0.8) **40%**

false negative (**β**) (0.2) **10%**

Most likely outcome is a null result

Power analysis in practice

- Typically we only do an experiment if we have a good reason to expect it will work
- Design experiment so that if there is an effect, our statistical test will return a "significant" p-value
- Requires experimenter to specify what effect they expect to see: **effect size**
- Key aspect of NIH "rigor and reproducibility": describe in grant application how you determined planned sample size and level of statistical power present in design

Power analysis based on effect sizes

- Dimensionless values that are associated with particular kinds of comparisons
- Cohen's d

$$d = \frac{\bar{x}_1 - \bar{x}_2}{SD}$$

- eta-squared, Cohen's f

$$\eta^2 = \frac{Sum\ of\ squares_{effect}}{Sum\ of\ squares_{total}}$$

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

Power analysis based on effect sizes

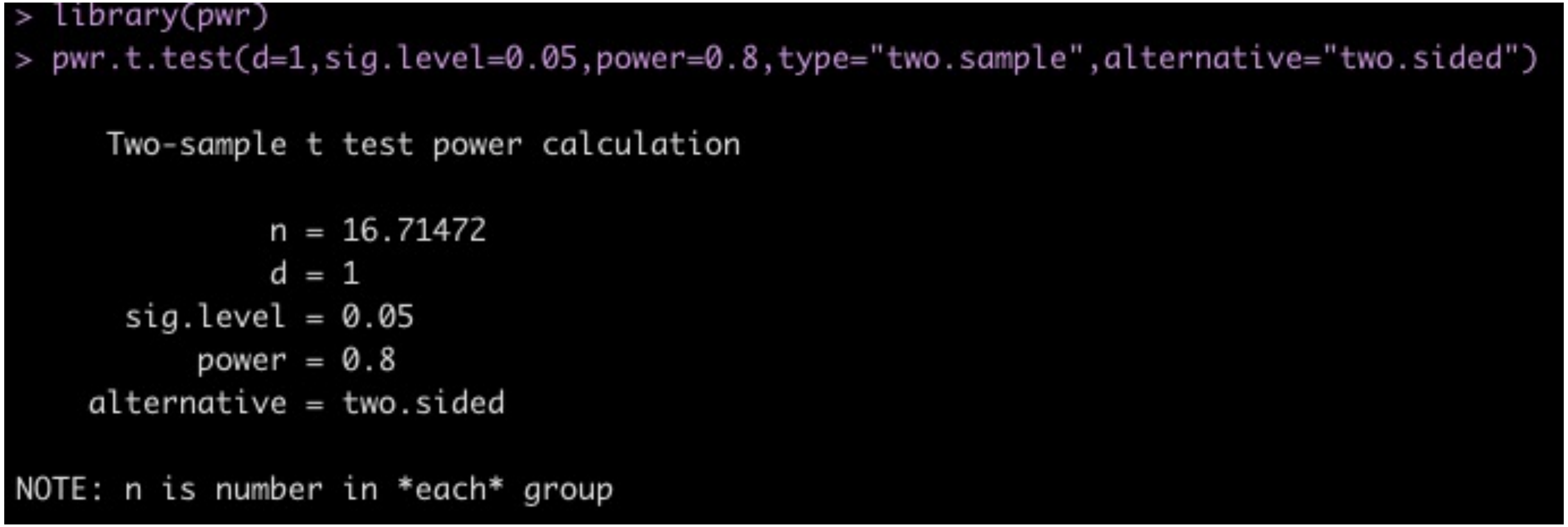
- You can look these up in tables (Cohen, 1988)
- Some general rules for these in psychology (d of 0.2 is "small" effect, 0.5 "moderate", 0.8 "large"; for f 0.1/0.25/0.4)
- Straightforward to calculate for some kinds of data and easy to determine power for standard statistical tests based on standardized effect size (G*Power, `pwr` library in R)

Power analysis based on effect sizes

- You collect pilot data on a novel object recognition memory task in mice and the mean is 65% correct and standard deviation is 5%.

- You have a drug that you think will enhance memory. You decide that a 5% difference would be meaningful (70% in the drug group, 65% in controls).

- This equals a Cohen's d of 1. (70-65) / 5 = 1



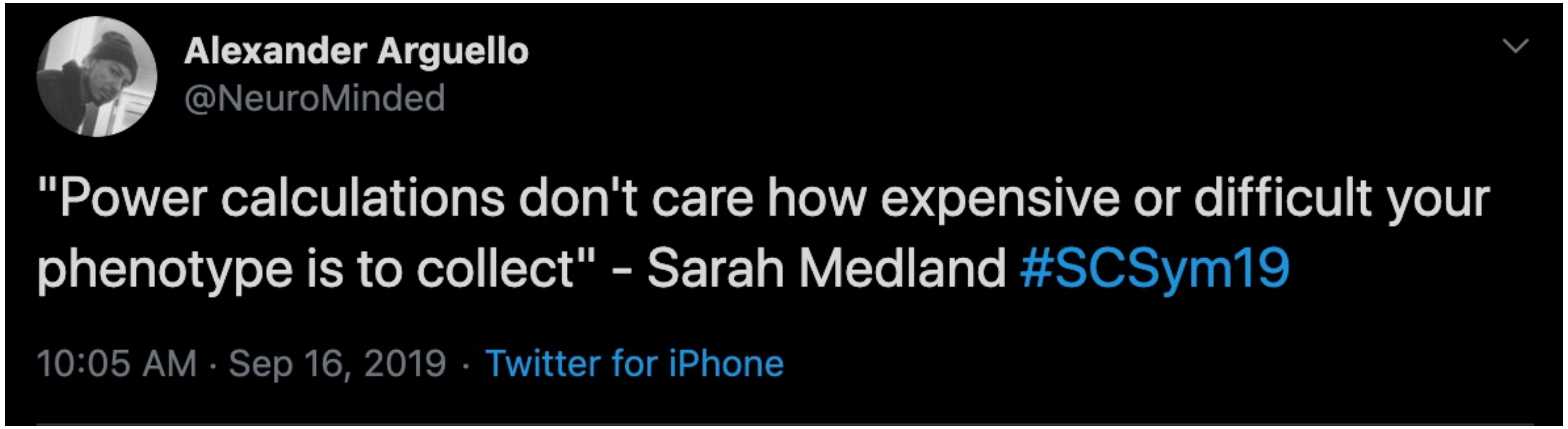
```
> library(pwr)
> pwr.t.test(d=1,sig.level=0.05,power=0.8,type="two.sample",alternative="two.sided")

     Two-sample t test power calculation

              n = 16.71472
              d = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power analysis based on effect sizes

- You collect pilot data on a novel object recognition memory task in mice and the mean is 65% correct and standard deviation is 5%.

- You have a drug that you think will enhance memory. You decide that a 5% difference would be meaningful (70% in the drug group, 65% in controls).

- "Based on a **power analysis**, 17 mice in each group provides greater than 80% power to detect an effect size of d=1.0 in a two-sample t-test at an alpha level of 0.05. This corresponds, based on pilot data, to a 5% difference in recognition between control and treatment groups (for example, 65% vs 70%).



Alexander Arguello
@NeuroMinded

"Power calculations don't care how expensive or difficult your phenotype is to collect" - Sarah Medland #SCSym19

10:05 AM · Sep 16, 2019 · Twitter for iPhone

Power analysis based on effect sizes

- Can be difficult to interpret – what does a d of 1.0 mean in *your* experiment? What is an f of 0.35?

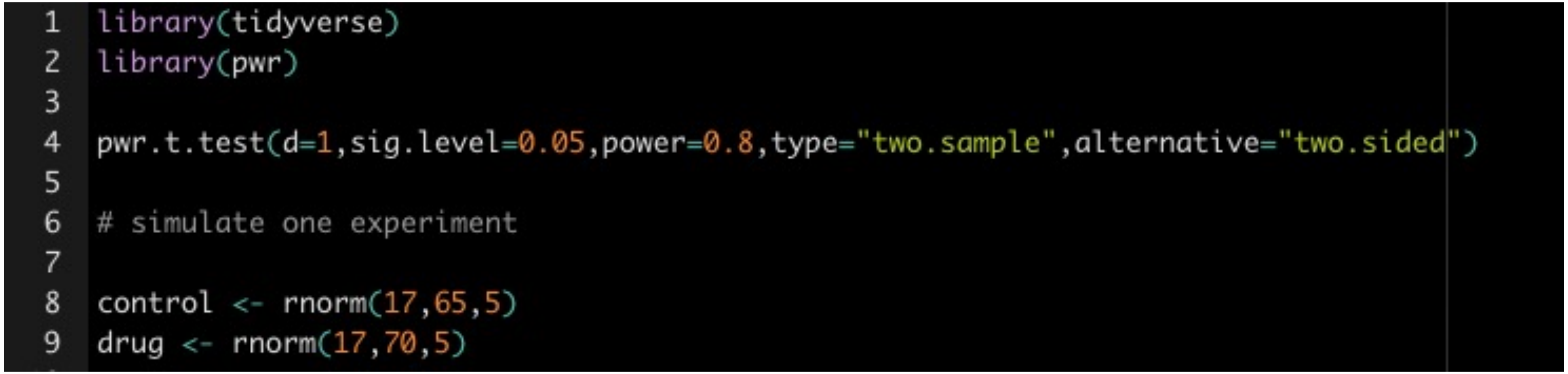
- Standarized effect sizes only work for relatively simple statistical tests and become complicated very quickly in more complex designs.
  - unequal variances
  - multi-factor regression/ANOVA

Power analysis based on simulations

- Simulate data from one experiment and run your statistical test on it
    - how are data distributed? what is relevant "effect size"?
    - what test are you going to use?

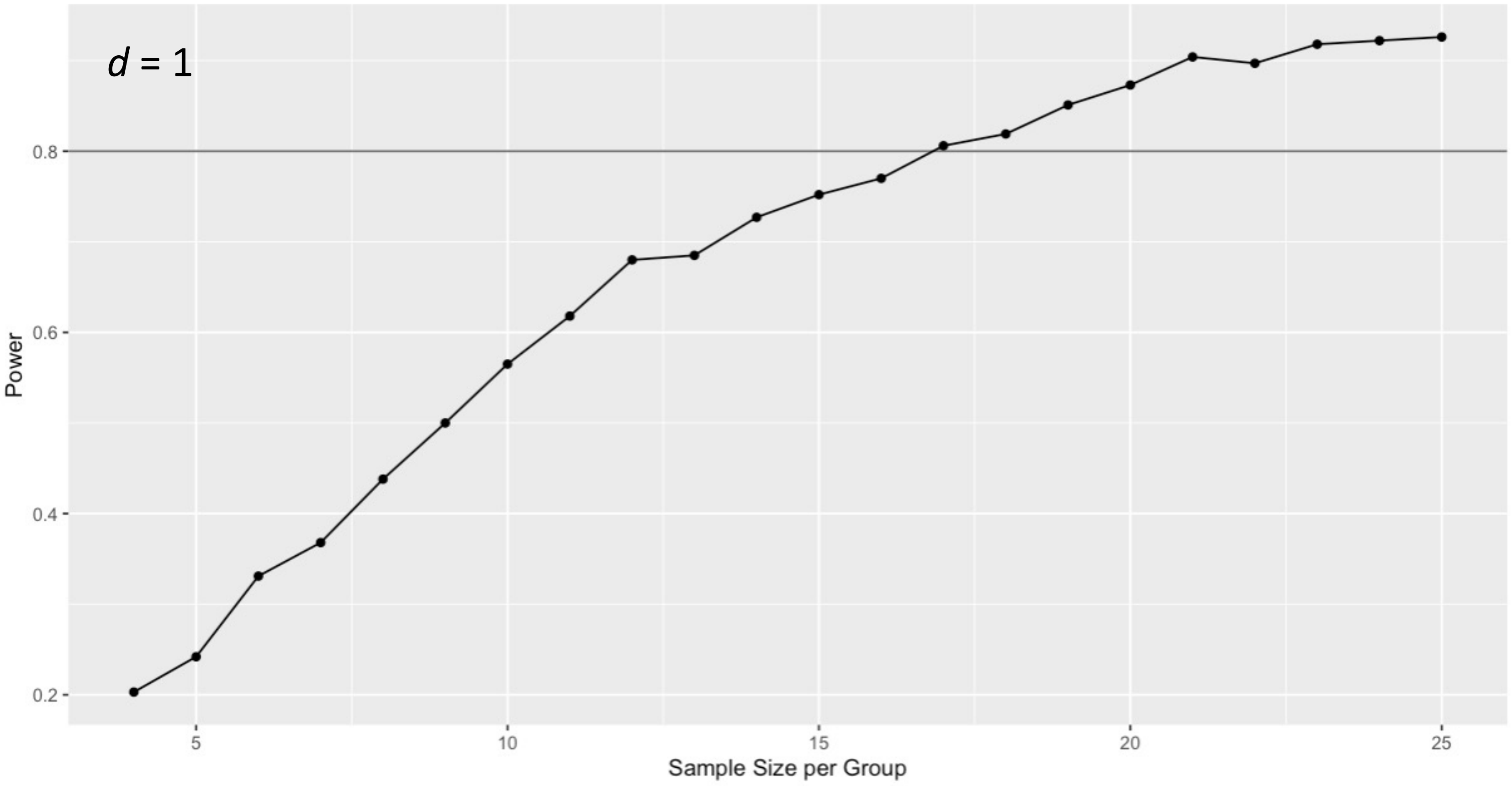
- Do this many times

- Power = number of "significant" tests over number of simulated experiments

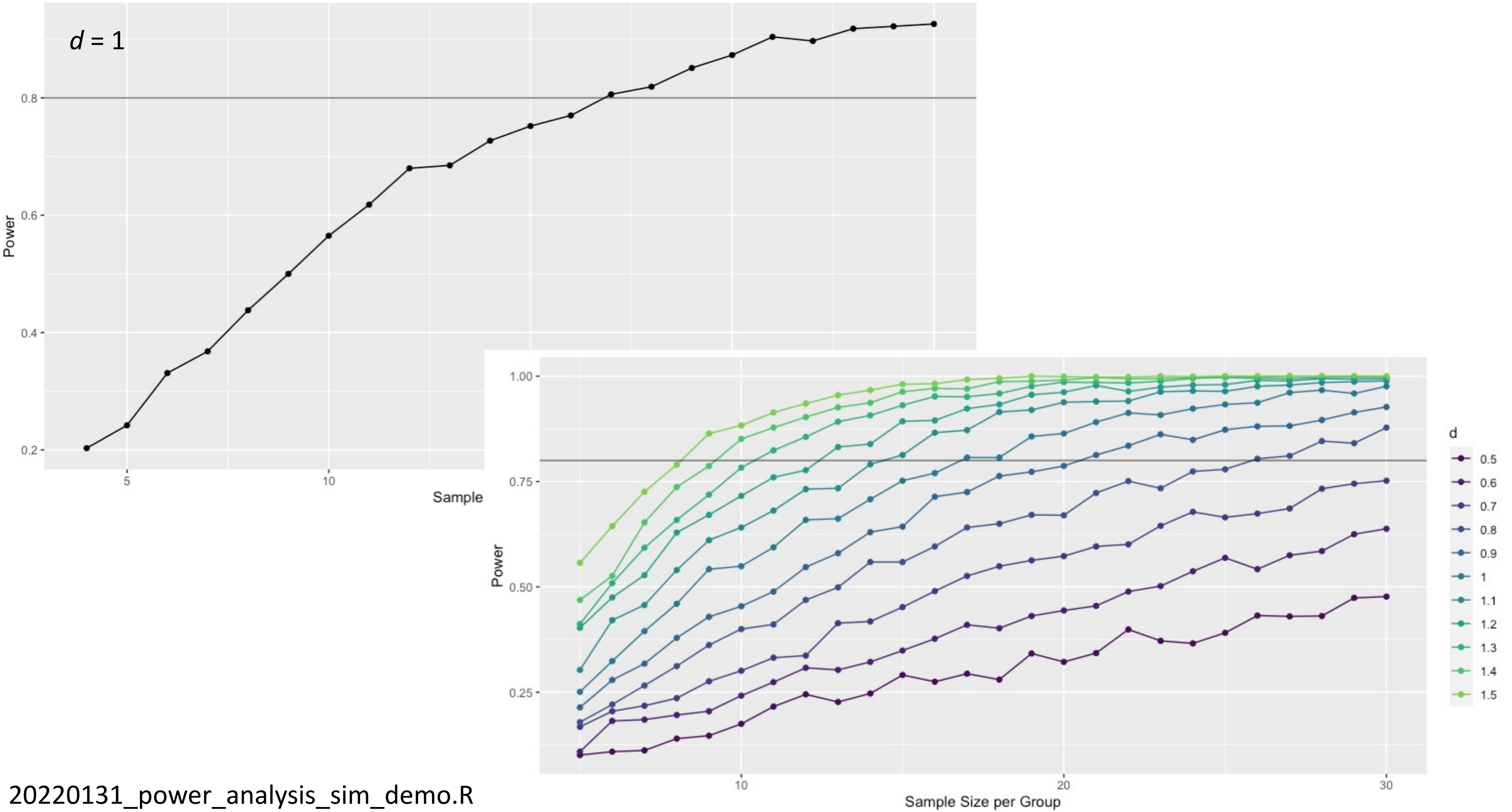


```r
library(tidyverse)
library(pwr)

pwr.t.test(d=1,sig.level=0.05,power=0.8,type="two.sample",alternative="two.sided")

# simulate one experiment

control <- rnorm(17,65,5)
drug <- rnorm(17,70,5)
```

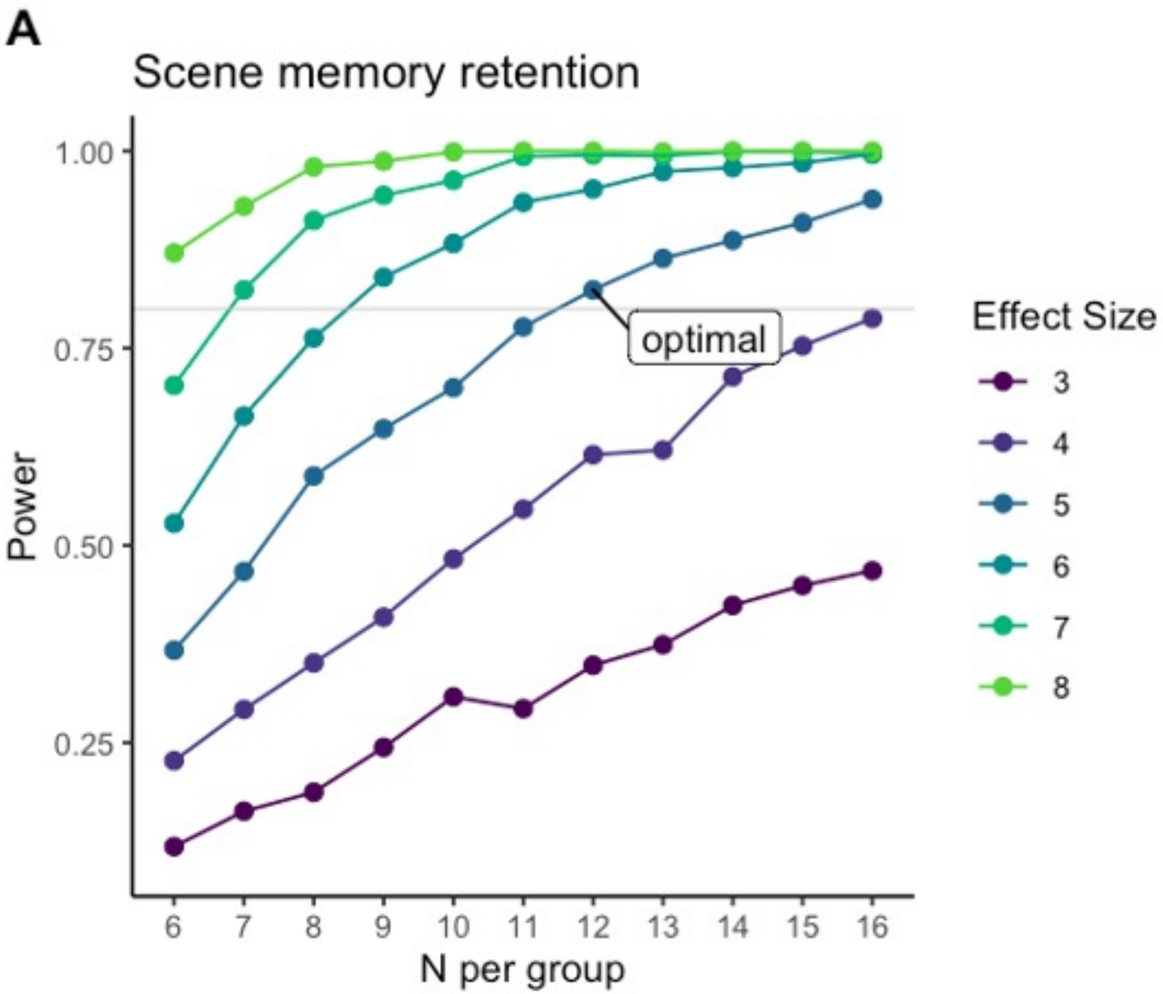
20220131_power_analysis_sim_demo.R

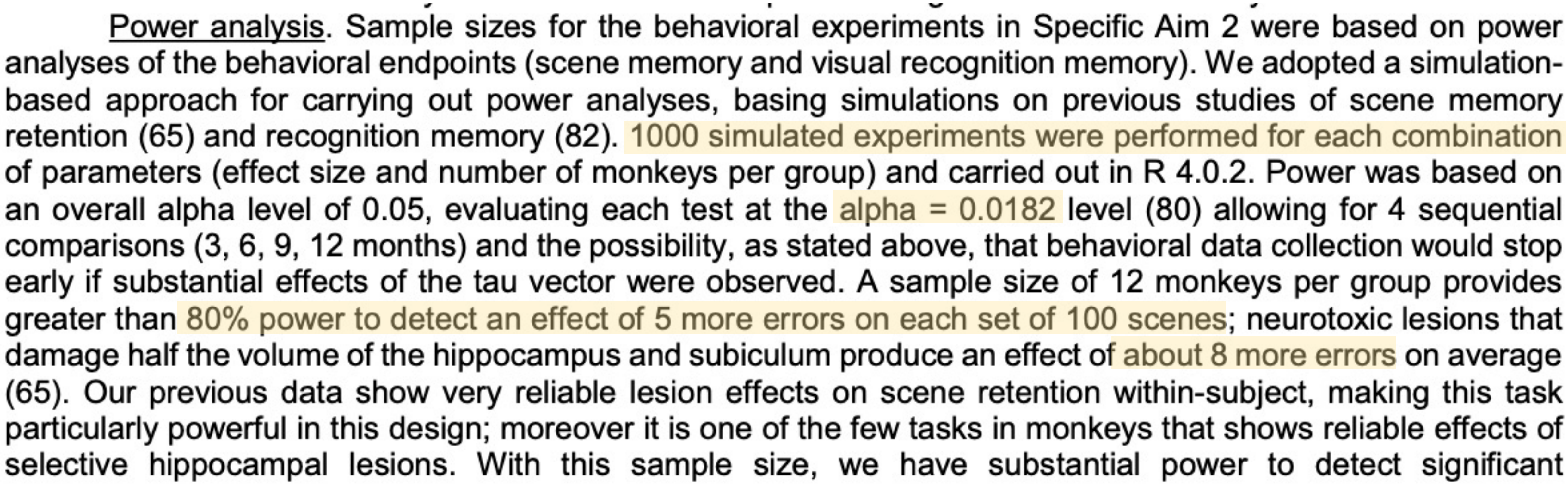


20220131_power_analysis_sim_demo.R

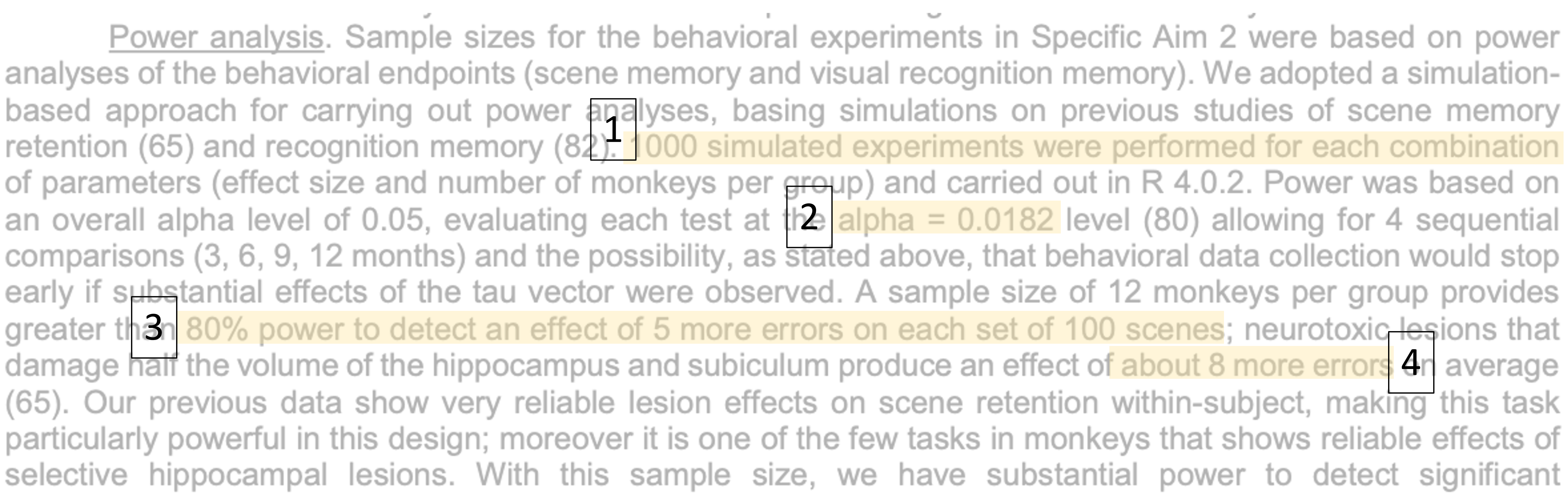


20220131_power_analysis_sim_demo.R


A
Scene memory retention
Power
1.00
0.75
0.50
0.25
N per group
6
7
8
9
10
11
12
13
14
15
16
optimal
Effect Size
3
4
5
6
7
8



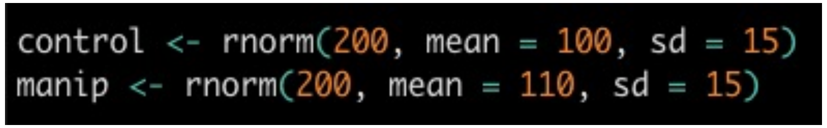
Power analysis. Sample sizes for the behavioral experiments in Specific Aim 2 were based on power analyses of the behavioral endpoints (scene memory and visual recognition memory). We adopted a simulation-based approach for carrying out power analyses, basing simulations on previous studies of scene memory retention (65) and recognition memory (82). 1000 simulated experiments were performed for each combination of parameters (effect size and number of monkeys per group) and carried out in R 4.0.2. Power was based on an overall alpha level of 0.05, evaluating each test at the alpha = 0.0182 level (80) allowing for 4 sequential comparisons (3, 6, 9, 12 months) and the possibility, as stated above, that behavioral data collection would stop early if substantial effects of the tau vector were observed. A sample size of 12 monkeys per group provides greater than 80% power to detect an effect of 5 more errors on each set of 100 scenes; neurotoxic lesions that damage half the volume of the hippocampus and subiculum produce an effect of about 8 more errors on average (65). Our previous data show very reliable lesion effects on scene retention within-subject, making this task particularly powerful in this design; moreover it is one of the few tasks in monkeys that shows reliable effects of selective hippocampal lesions. With this sample size, we have substantial power to detect significant



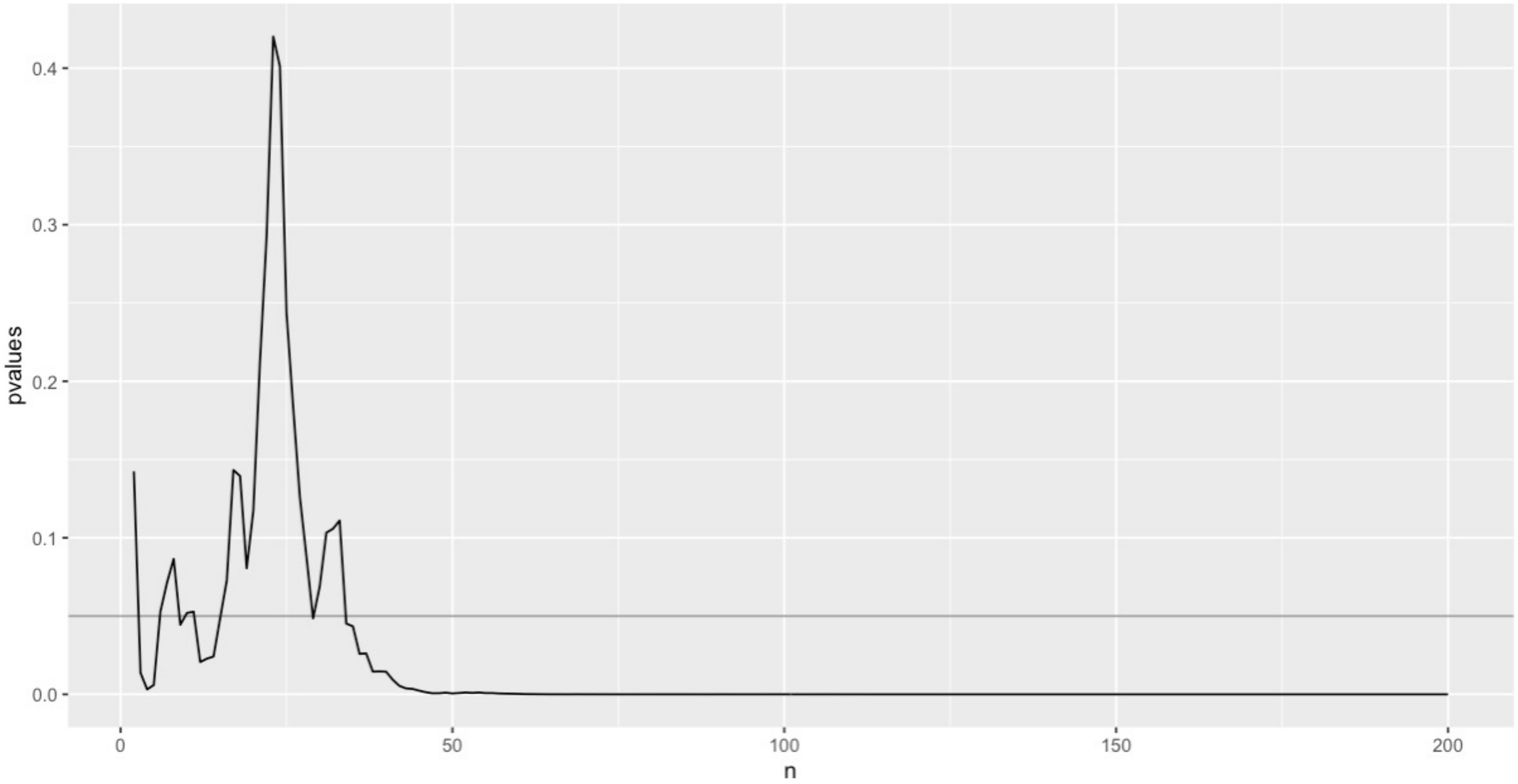
Power analysis. Sample sizes for the behavioral experiments in Specific Aim 2 were based on power analyses of the behavioral endpoints (scene memory and visual recognition memory). We adopted a simulation-based approach for carrying out power analyses, basing simulations on previous studies of scene memory retention (65) and recognition memory (82). 1000 simulated experiments were performed for each combination of parameters (effect size and number of monkeys per group) and carried out in R 4.0.2. Power was based on an overall alpha level of 0.05, evaluating each test at the alpha = 0.0182 level (80) allowing for 4 sequential comparisons (3, 6, 9, 12 months) and the possibility, as stated above, that behavioral data collection would stop early if substantial effects of the tau vector were observed. A sample size of 12 monkeys per group provides greater than 80% power to detect an effect of 5 more errors on each set of 100 scenes; neurotoxic lesions that damage half the volume of the hippocampus and subiculum produce an effect of about 8 more errors on average (65). Our previous data show very reliable lesion effects on scene retention within-subject, making this task particularly powerful in this design; moreover it is one of the few tasks in monkeys that shows reliable effects of selective hippocampal lesions. With this sample size, we have substantial power to detect significant

1) Describe how you carried out your power analysis (simulations or based on a standard effect size calculation)
2) Specify your alpha level (do not assume reviewers will assume you are using .05)
3) Specify how much power you have and what effect size you're powered to detect
4) Place your effect size in context of real data if you can

"March of the p-values"

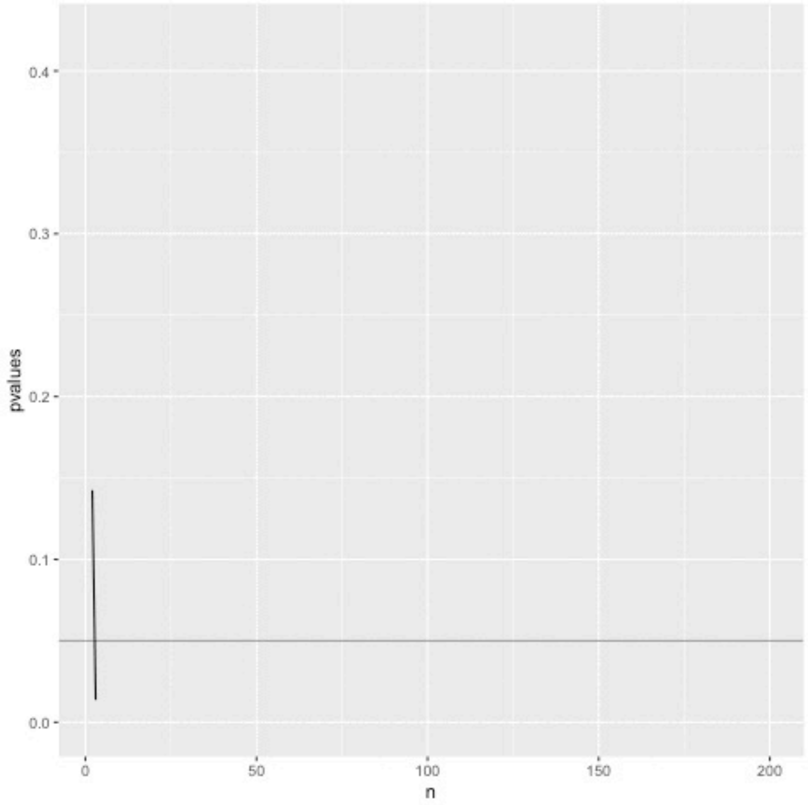


```
control <- rnorm(200, mean = 100, sd = 15)
manip <- rnorm(200, mean = 110, sd = 15)
```

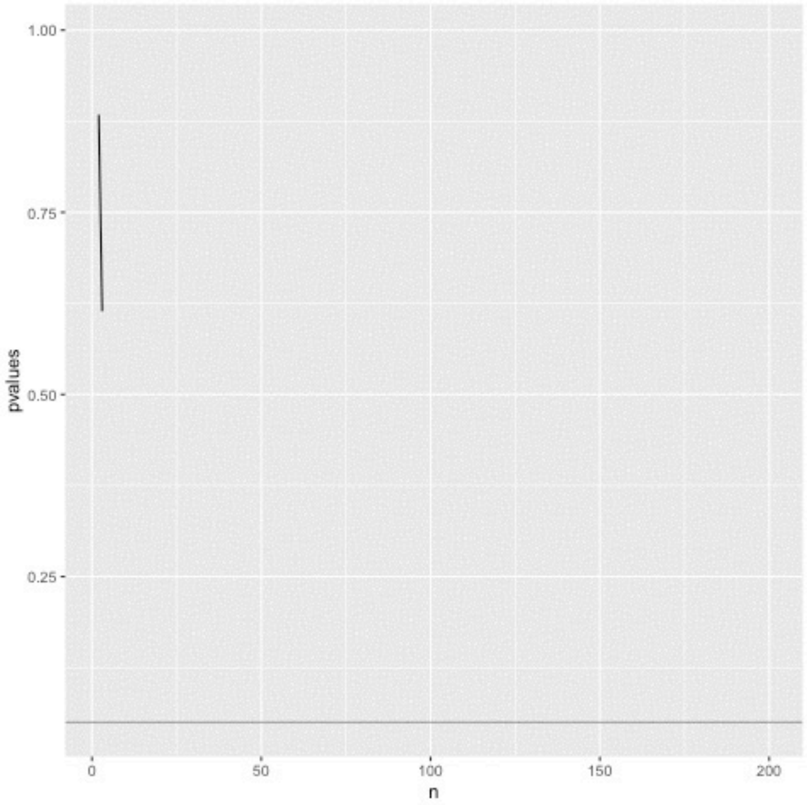


20220131_march_of_the_p_values.R

"March of the p-values"



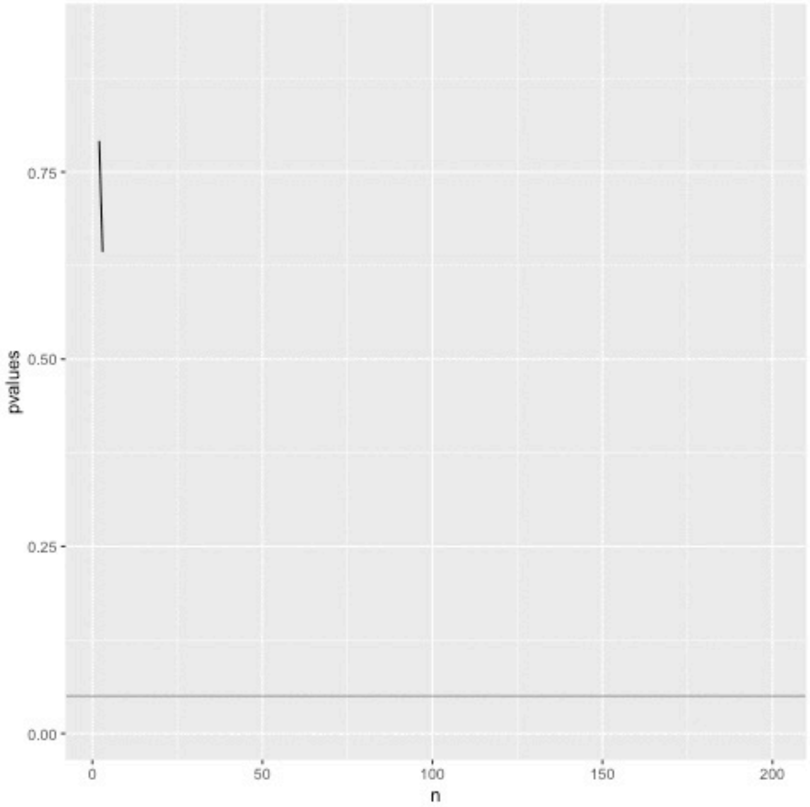
mean diff = 10 in population



mean diff = 0 in population

20220131_march_of_the_p_values.R

"March of the p-values"



a type 1 error in the wild

mean diff = 0 in population

20220131_march_of_the_p_values.R

"Post hoc" power analysis

- Also "retrospective" power analysis, "observed power"
- This is not a thing

- Lenth: "You’ve got the data, did the analysis, and did not achieve 'significance.' So you compute power retrospectively to see if the test was powerful enough or not. This is an empty question. Of course it wasn’t powerful enough – that’s why the result isn’t significant. Power calculations are useful for design, not analysis."

https://homepage.divms.uiowa.edu/~rlenth/Power/index.html