PSY 503: Foundations of Statistics in Psychological Science

Effect Size and Power

Jason Geller, Ph.D. (he/him/his)

Princeton University

Last Updated: 2022-10-12

Knowledge Check

```
library(report)

df=tibble::tribble(
  ~Fake.Name, ~Own.Name,
         60L,       47L,
         78L,       63L,
         57L,       27L,
         64L,       42L,
         89L,       23L,
         79L,       75L,
         63L,       24L,
         81L,       44L,
         46L,       40L,
         50L,       44L
  )
```

Males completed a math test under their own name and a fake (movie stars)name

- What test should be run?
- What assumptions should be tested for this test?
- Run the test. What can we conclude?

03:00

Knowledge Check

```r
shapiro.test(df$Own.Name) # check norm
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Own.Name
## W = 0.91469, p-value = 0.3148
```

```r
shapiro.test(df$Fake.Name) # check norm
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Fake.Name
## W = 0.94695, p-value = 0.6326
```

```r
t_test=t.test(df$Fake.Name, df$Own.Name, paired = TRUE) # paired
```

Knowledge Check

```r
t_test=t.test(df$Fake.Name, df$Own.Name, paired = TRUE) # paired
```

- Effect sizes were labelled following Cohen's (1988) recommendations.

The Paired t-test testing the difference between df$Fake.Name and df$Own.Name (mean difference = 23.80) suggests that the effect is positive, statistically significant, and large (difference = 23.80, 95% CI [9.80, 37.80], t(9) = 3.85, p = 0.004; Cohen's d = 1.22, 95% CI [0.37, 2.03])

Last Class

- Two sample t-tests
    - Independent
    - Dependent (paired)
- Non-parametric tests
- Multiple comparisons

Today

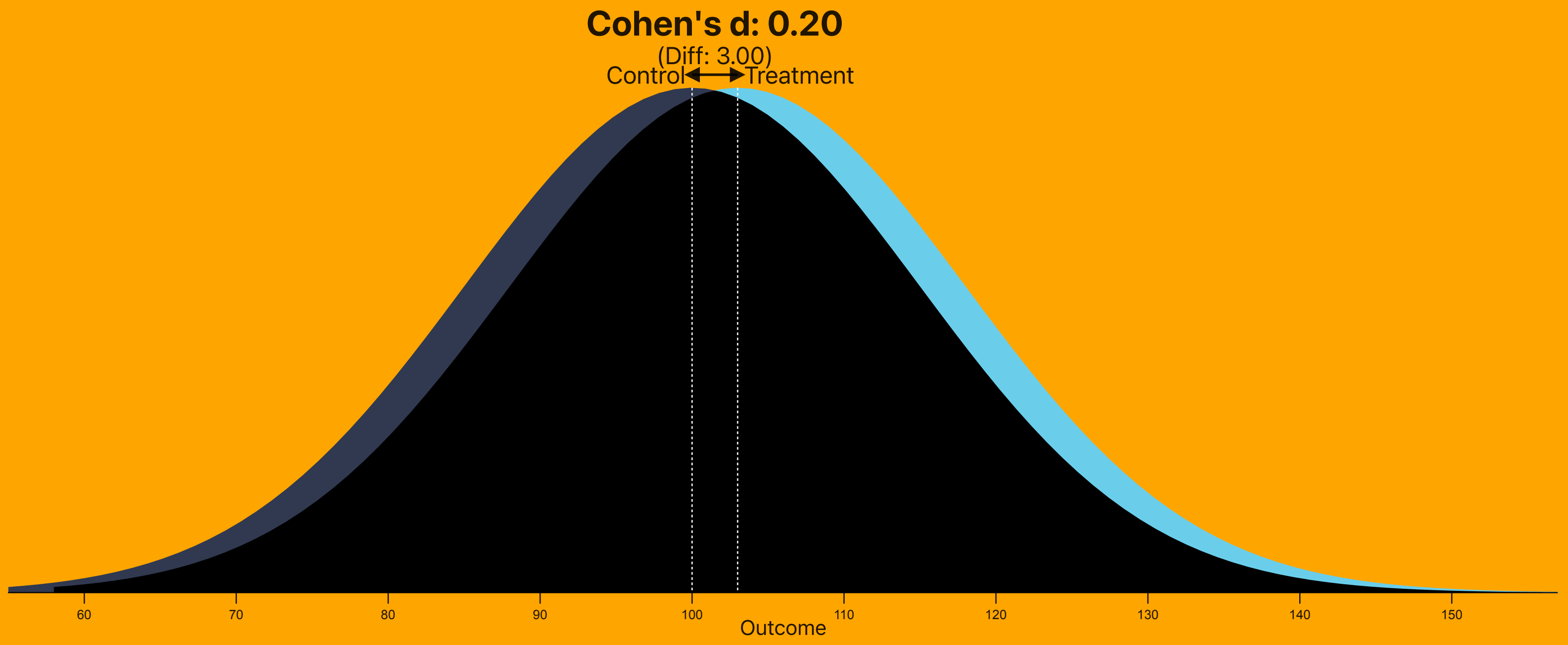
- Effect size
- Statistical Power
    - What is Power?
    - Why do we care about power?
    - Determining Power
        - R packages

Effect Size

- "The amount of anything that's of research interest" (Cumming & Calin-Jageman, 2017, p.111)

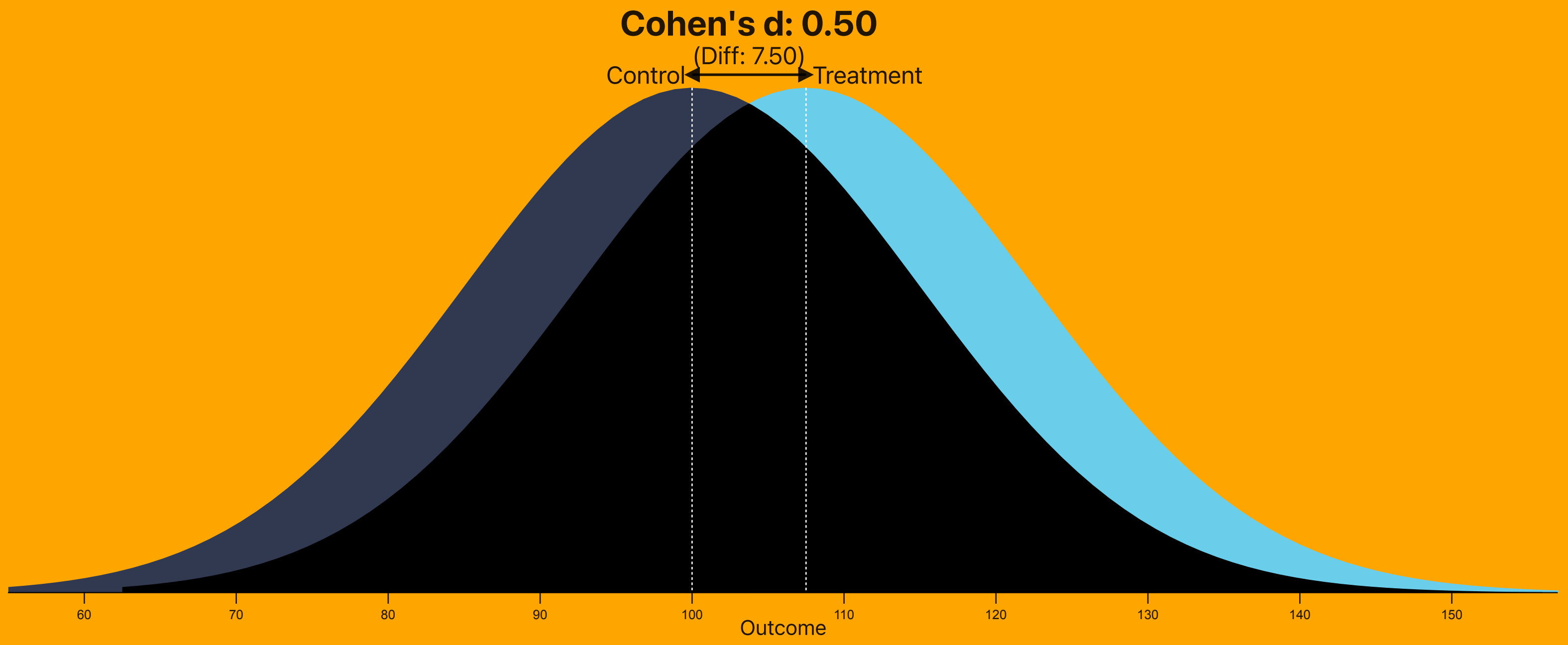
Effect Size

- Simple way to quantify the difference between two means / groups, by emphasizing the size of the difference rather than confounding with the sample size (like p-values)
- Small



Effect Size

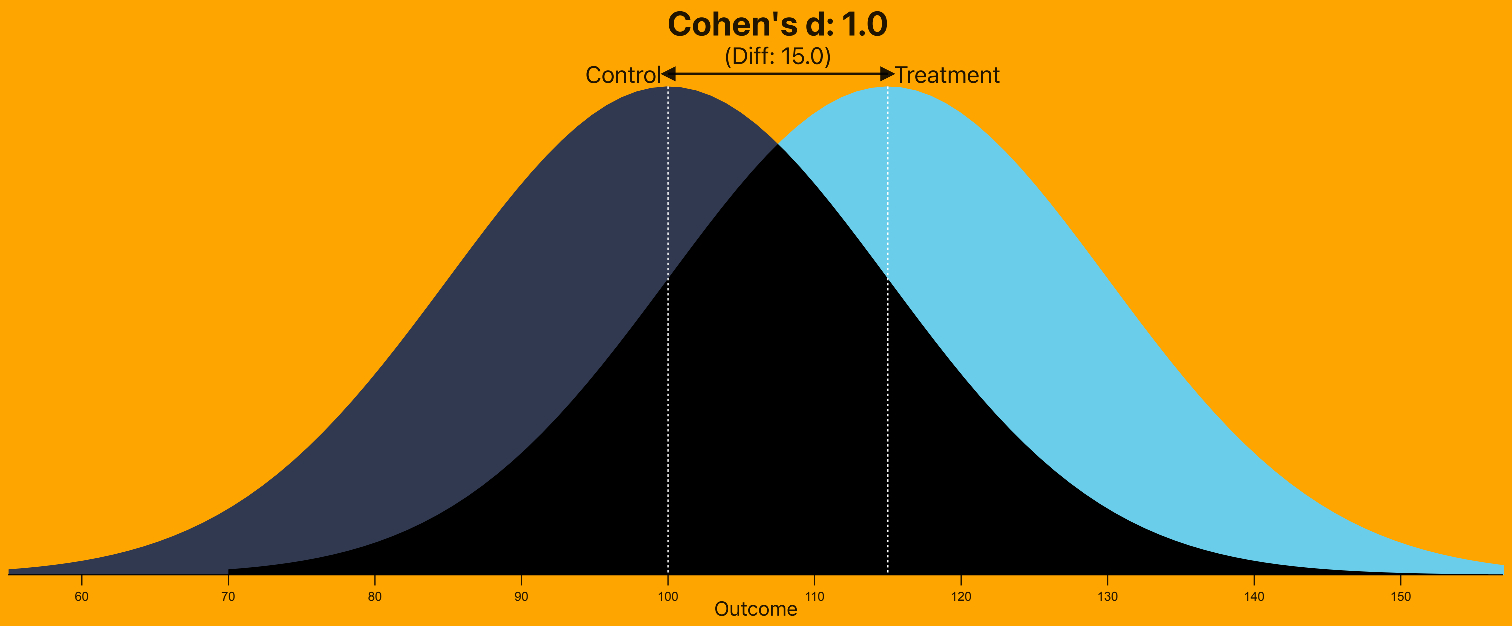
- Simple way to quantify the difference between two means / groups, by emphasizing the size of the difference rather than confounding with the sample size (like p-values)
- Medium



Effect Size

- Simple way to quantify the difference between two means / groups, by emphasizing the size of the difference rather than confounding with the sample size (like p-values)

- Large



Effect Size

- Effect size options:
  - Cohen's d
  - Hedges' g ($N < 20$)
  - Glass' Δ
  - r
- Always report an effect size with statistics!

Cohen's *d*: 1 Sample

$$d = \frac{\bar{X} - \mu}{s}$$

- d = effect size
- $\bar{X}$= sample mean
- μ_0= population mean
- s = sample standard deviation

Open R

- From lecture: A cookie company claims that there are 15 chocolate chips per cookie, but you aren’t convinced. You take 10 cookies and count the number of chocolate chips in each cookie. Here is the data: [13,14,15,17,18,19,21,20,19,20]. Are the number of chocolate chips significantly different from 15 (reminder: yes, p = 0.015)? Report the effect size.

03:15

Open R

```r
data1 = c(13,14,15,17,18,19,21,20,19,20)
X = mean(data1)
s = sd(data1)
mu = 15
d = (X - mu) / s
abs(d) # report positive value
```

```
## [1] 0.9431191
```

Effect Size Packages

```
library(effectsize) # easystats package
library(report) # report results
library(MOTE) # erin package effect size

cookie=t.test(data1, mu=15) # perform ttest

effectsize::cohens_d(cookie, data = data1, mu=15) # get cohens_d
```

```
## Cohen's d |       95% CI
## -------------------------
## 0.94      | [0.17, 1.68]
##
## - Deviation from a difference of 15.
```

```
effectsize::hedges_g(cookie, data = data1, mu=15)
```

```
## Hedges' g |       95% CI
## -------------------------
## 0.86      | [0.16, 1.54]
##
## - Deviation from a difference of 15.
```

```
# use when N < 20 bias corrected
```

- Effect sizes were labelled following Cohen's (1988) recommendations.

The One Sample t-test testing the difference between data1 (mean = 17.60) and mu = 15 suggests that the effect is positive, statistically significant, and large (difference = 2.60, 95% CI [15.63, 19.57], t(9) = 2.98, p = 0.015; Cohen's d = 0.94, 95% CI [0.17, 1.68])

Cohen's *d*: 2 Sample

$$d = \frac{\bar{X}_1 - \bar{X}_2}{SD_{pooled}}$$

- d = effect size
- $\bar{X}$= sample mean
- μ_0= population mean
- s = sample standard deviation

Cohen’s *d*: 2 sample

From last lecture: A math test was given to 300 17 year old students in 1978 and again to another 350 17 year old students in 1992

Group 1: X1 = 300.4, s1 = 34.9, n = 300 Group 2: X2 = 306.7, s2 = 30.1, n = 350

- What is the effect size?

00:15

Open R

```r
group1 <- rnorm(300, mean = 300.4, sd=34.9)

group2 <- rnorm(350, mean = 306.7, sd=30.1)

r=t.test(group1, group2)

effectsize::cohens_d(r, data = data1, pooled_sd = TRUE)#nonwelch
```

```
## Cohen's d |         95% CI
## --------------------------
## -0.21     | [-0.37, -0.06]
##
## - Estimated using un-pooled SD.
```

```r
effectsize::cohens_d(r, data = data1, pooled_sd = FALSE)#welch
```

```
## Cohen's d |         95% CI
## --------------------------
## -0.21     | [-0.37, -0.06]
##
## - Estimated using un-pooled SD.
```

Reporting

- Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference between group1 and group2 (mean of x = 300.06, mean of y = 306.75) suggests that the effect is negative, statistically significant, and small (difference = -6.69, 95% CI [-11.62, -1.77], t(592.03) = -2.67, p = 0.008; Cohen's d = -0.21, 95% CI [-0.37, -0.06])

Cloak vs. No Cloak Data: Effect Size

- What is the effect size?

```
longdata <- read_csv("https://raw.githubusercontent.com/doomlab/statsofdoom-files/master/graduate/R%20F
```

Cloak vs. No Cloak Data: Effect Size

```r
library(MOTE)# erin effect size package
library(rio)

# extract means and sd from cloak df
 M <- longdata  %>%
   group_by(Cloak) %>%
   dplyr::summarize(mean=mean(Mischief), sd=sd(Mischief), N =12)

# run indep t test using MOTE
effect <- d.ind.t(m1 = M$mean[1], m2 = M$mean[2],
                  sd1 = M$sd[1], sd2 = M$sd[2],
                  n1 = 12, n2 = 12, a = .05)
effect$d # get effect size
```

```
## [1] 0.6995169
```

```r
# using effect size package
library(effectsize) # effect size package
cohens_d(Mischief~Cloak, data = longdata)
```

```
## Cohen's d |        95% CI
## -------------------------
## 0.70      | [-0.13, 1.52]
##
## - Estimated using pooled SD.
```

Cloak vs. No Cloak: Effect Size

- While our statistical test indicated no differences, the effect size indicates a medium difference between means
- This difference in interpretation is likely due to low power with a small sample size

Cohen's *d*: Dependent

- Lots of different ones
    - State which one you are using!
- D_{avg} - looks at both SDs without controlling for r

$$d_{avg} = \frac{\overline{M_1} - \overline{M_2}}{\sqrt{\frac{(\sigma_1^2 + \sigma_2^2)}{2}}}$$

- d_z - we would overestimate the effect size

$$d_z = \frac{t}{\sqrt{n}}$$

Dependent: Effect Size (Lakens, 2013)

- d_{rm} - looks at both SDs and controls for r

$$d_{rm} = \frac{M_1 - M_2}{\sqrt{(SD_1^2 + SD_2^2) - (2 \times r \times SD_1 \times SD_2)}} \times \sqrt{2 \times (1 - r)}$$

- You do not normally have to calculate all of these, just showing how these are different

R

Using the cloak data, calculate *d* for dependent samples

Use:

- `MOTE` d.dep.t.avg(m1, m2, sd1, sd2, n, a = 0.05) (d_{avg})
- `MOTE` d.dep.t.rm(m1, m2, sd1, sd2, r, n, a = 0.05) (d_{rm})
- What Cohen's *d* measure is used in the `effectsize` package?

05:15

d_{avg}

```r
effect2 <- d.dep.t.avg(m1 = M$mean[1], m2 = M$mean[2],
                  sd1 = M$sd[1], sd2 = M$sd[2],
                  n=12, a = .05)
effect2$d
```

```
## [1] 0.7013959
```

d_{rm}

```
#MOTE
effect2 <- d.dep.t.rm(m1 = M$mean[1], m2 = M$mean[2],
                      sd1 = M$sd[1], sd2 = M$sd[2],
                      n=12, a = .05,  r= .7)

effect2$d
```

```
## [1] 0.6909434
```

d_z

```r
#MOTE

diff <- longdata$Mischief[longdata$Cloak == "Cloak"] - longdata$Mischief[longdata$Cloak == "No Cloak"]

effect2.1 = d.dep.t.diff(mdiff = mean(diff, na.rm = T),
                         sddiff = sd(diff, na.rm = T),
                         n = length(diff), a = .05)
effect2.1$d
```

```
## [1] 1.098244
```

Effect Size Interpretation

- Cohen (1988)
    - d < 0.2 - Very small
    - 0.2 <= d < 0.5 - Small
    - 0.5 <= d < 0.8 - Medium
    - d >= 0.8 - Large

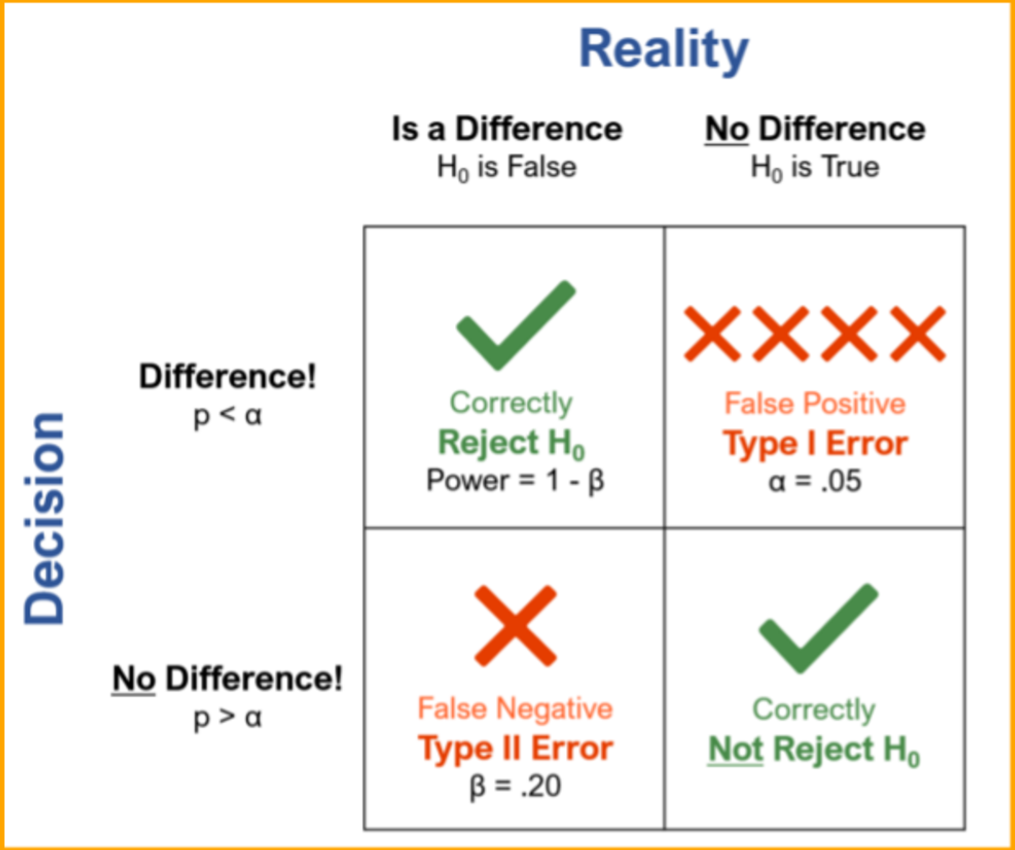
Power

Recap of NHST

- Do invisibility cloaks increase mischeivous behavior?
  - H_0:There is no effect of cloaks on behavior
  - H_1: There is an effect of cloaks on behavior

Recap of NHST

- A world in which H_1 exists
  - Two types of errors:



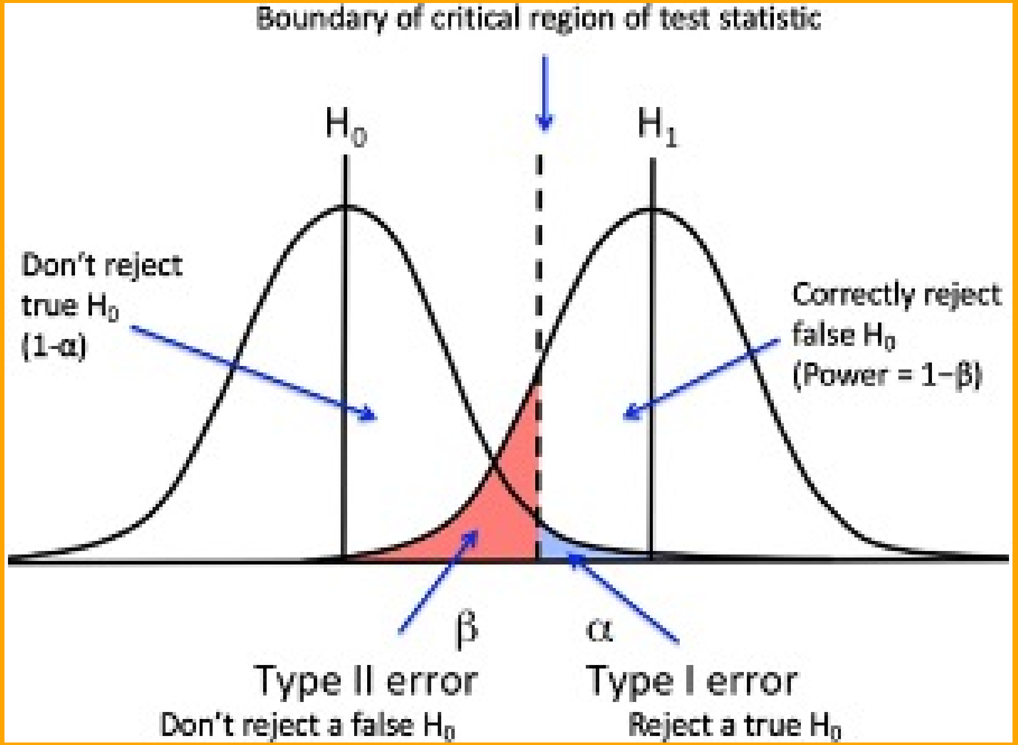
Reality

Decision

	Is a Difference H_0 is False	**No Difference** H_0 is True
Difference! $p < \alpha$	✓ Correctly **Reject H_0** Power = 1 - β	✗✗✗✗ False Positive **Type I Error** α = .05
No Difference! $p > \alpha$	✗ False Negative **Type II Error** β = .20	✓ Correctly **Not Reject H_0**

Power

- Power $1 - \beta$: Probability rejecting null when it is false
- Detecting the effect when it really exists


Boundary of critical region of test statistic
H_0
H_1
Don't reject true H_0 (1-α)
Correctly reject false H_0 (Power = 1−β)
β
α
Type II error
Don't reject a false H_0
Type I error
Reject a true H_0

What is common?

In psychology:

- $\beta = .20$
  - This means we are willing to make a Type II error 20% of the time (i.e., 80% power).
- $\alpha =.05$
  - This means we are willing to make a Type I error only 5% of the time (i.e., significance < .05.
- $1-\beta = .80$ (should be .9)

Power

What does it mean if we say: "we compare retrieval practice to re-reading with power = .75"

- If retrieval practice is actually beneficial, there is a 75% chance we'll get a significant result when we do this study MANY MANY TIMES
- We compare bilinguals to monolinguals on a test of non-verbal cognition with power = .35
- If there is a difference between monolinguals & bilinguals, there is a 35% chance we'll get p < .05 IF WE DO THIS MANY MANY TIMES

Why do we care?

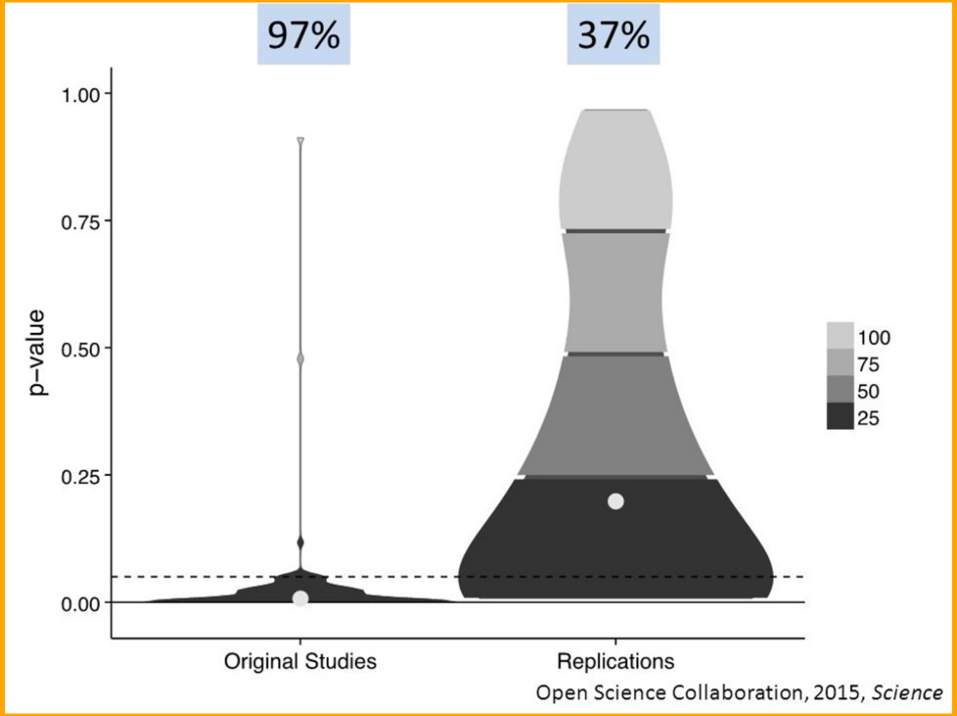
- Efficient use of resources
  - Power analyses tell us if our planned sample size (n) is:
  - Large enough to be able to find what we’re looking for
    - Not too large that we’re collecting more data than necessary
- This is about good use of our resources!
  - Societal resources: money, participant hours
  - Your resources: Time!!

Why do we care?

- Avoid p-hacking (Simmons et al., 2011)
    - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig
    - Power analysis decides sample in advance

Why do we care?

- Understand non-replication (Open Science Collaboration, 2015)
    - Even if an effect exists in the population, we'd expect some non-significant results
        - Power is almost never 100%
        - In fact, many common designs in psychology have low power (Etz &



Open Science Collaboration, 2015, *Science*

Why do we care?

- Understand null results
    - Non-significant result, by itself, doesn’t prove an effect doesn’t exist
    - With high power, null result is more informative
        - E.g., null effect of cloaks on behavior 20% power
            - Maybe cloaks work & we just couldn’t detect the effect?
            - But: null effect of cloaks on behavior with power of 90%
            - Makes me more sure

Why do we care?

- Granting agencies want them now
    - Don’t want to fund a study with low probability of showing anything
        - e.g., Our theory predicts greater activity in Broca’s area in condition A than condition B. But our experiment has only a 16% probability of detecting the difference. Not good!

Why do we care?

- Scientific accuracy!
    - If there is an effect, we want to know about it!

Power Analysis

- Power analysis: Do we have the power to detect the effect we’re interested in? If not, what is it going to take?

- Depends on:

  1. Sample size
  2. Effect size (e.g., d)
  3. Statistical significance criteria (α)
  4. Variability

Power to Estimate Sample Size

- Calculate required sample size given a) effect size (e.g., d) b) significance level (α), c) desired power.

- How do we determine the effect size?

1. Smallest effect size of interest (SESOI) (e.g., d = 0.5) (Lakens)

2. A priori

    - Use literature to estimate effect size (set α, desired power)

3. Pilot data

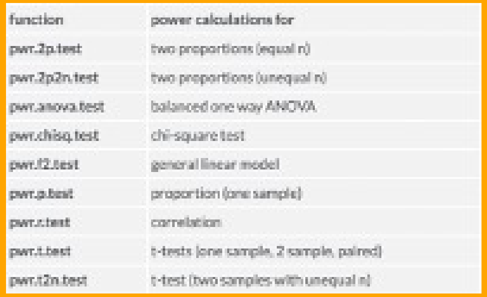
    - Estimate effect size using pilot data (set α, desired power)

For SESOI, you consider the smallest effect size you care about. For both a priori and pilot data power analyses, you need to get an estimate of the effect size (e.g., d)

Determining Power

R Packages

- `pwr` package



function	power calculations for
pwr.2p.test	two proportions (equal n)
pwr.2p2n.test	two proportions (unequal n)
pwr.anova.test	balanced one way ANOVA
pwr.chisq.test	chi-square test
pwr.f2.test	general linear model
pwr.p.test	proportion (one sample)
pwr.r.test	correlation
pwr.t.test	t-tests (one sample, 2 sample, paired)
pwr.t2n.test	t-test (two samples with unequal n)

- Enter three of the four parameter options above (sample size, effect size, statistical significance, and power) and the package will calculate the fourth parameter.

Determining Power

R Packages: Power

- Superpower
  - https://aaroncaldwell.us/SuperpowerBook/

```r
#install.packages("Superpower")
library(Superpower)
```

- Mixedpower(simulations for more complicated models)

```r
# if (!require("devtools")) {
 #   install.packages("devtools", dependencies = TRUE)}

 #  devtools::install_github("DejanDraschkow/mixedpower")
```

A Priori Power

From last lecture: A math test was given to 17 year old students in 1978 and again to another 17 year old students in 1992. From that data we have an estimate of the following parameters

- Group 1: X1 = 300.4, s1 = 34.9, n = 300
- Group 2: X2 = 306.7, s2 = 30.1, n = 350

We want to conduct a similar experiment and estimate how many people we should collect to achieve a desired power of 80%

- What is the effect size?

- Calculate power

```r
library(pwr)
d2abs = .2
pwr.t.test()
```

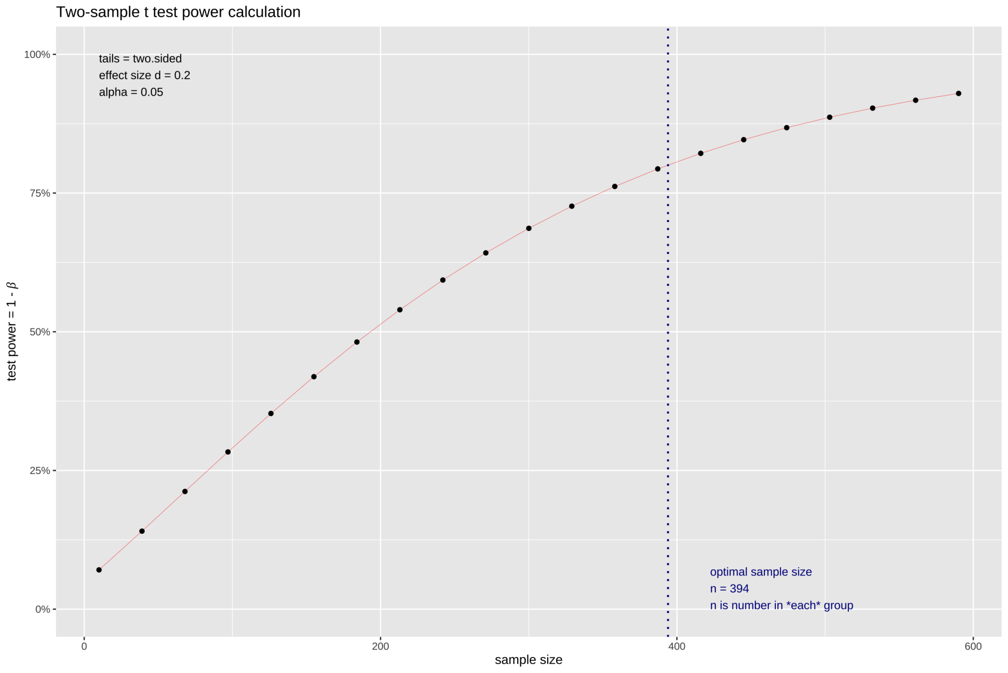
Power Curves

- Visualization showing power as a function sample size

```
d2abs=.2

p.out <- pwr.t.test(d = d2abs, power = 0.80, sig.level = 0.05,
type = "two.sample", alternative = "two.sided")

plot(p.out)
```



- From last lecture: A cookie company claims that there are 15 chocolate chips per cookie, but you aren't convinced. You take 10 cookies and count the number of chocolate chips in each cookie. Here is what the data looks like: [13,14,15,17,18,19,21,20,19,20]. Are the number of chocolate chips significantly different from 15 (reminder: yes, p = 0.015)?

How many cookies do you need to get 90% power?

```r
data1 = c(13,14,15,17,18,19,21,20,19,20)

r=t.test(data1, mu=15)

d1=effectsize::cohens_d(r, data = data1, mu=15)
```

```r
pwr.t.test(d = d1$Cohens_d, power = 0.90, sig.level = 0.05, type = "one.sample", alternative = "two.sid
```

```
## [1] 13.87667
```

Power Questions

- What would our power be if I could only collect 3 cookies?
- Can calculate current power by slightly adjusting the function

```r
data1 = c(13,14,15,17,18,19,21,20,19,20)

pwr.t.test(n = , d = , power = , sig.level = 0.05,
type = "one.sample", alternative = "two.sided")$power
```

Power Question

- Using the cookie data, how many cookies would we need to achieve 90% power?

```r
library(pwr)
pwr.t.test(n =NULL, d =, #effect size
           sig.level = .05, #alpha
           power = , #power
           type = "one.sample", #independent
           alternative = "two.sided") #two tailed test
```

Simulation

Remember the definition of power?

- The probability of observing a significant effect in our sample if the effect truly exists in the population
- What if we knew for a fact that the effect existed in a particular population?
  - Then, a measure of power is how often we get a significant result in a sample (of our intended n)
- Solution: We create (“simulate”) the data.

Monte Carlo Simulations

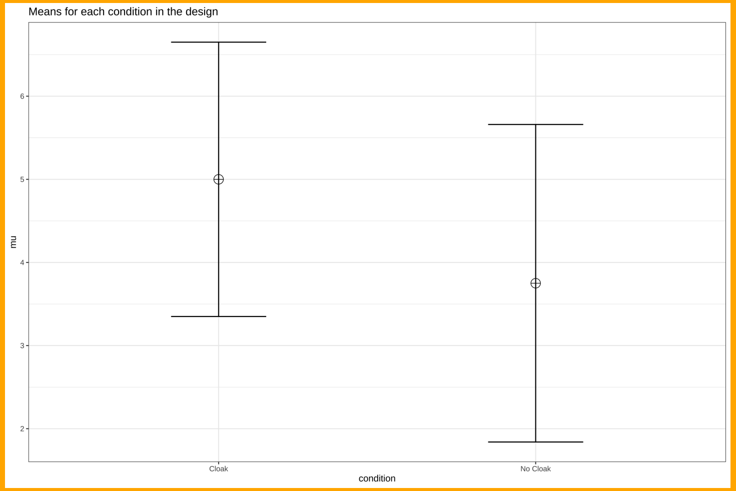
1. Set population parameters

   from already conducted studies or pilot data

Cloak Data

```r
design <- "2b"
n <- 12
mu <- c(5, 3.75)
sd <- c(1.65, 1.91)
label_list = list("condition" = c("Cloak", "No Cloak")) #

design_result <- ANOVA_design(design = design,
                              n = n,
                              mu = mu,
                              sd = sd,
                              label_list = label_list)
```



Monte Carlo Simulations

1. Set population parameters
2. Create a random sample from these data
3. Do this multiple times
4. Calculate how many times you get a significant result
    - E.g., 5 out 10 times (50% power)

```r
nsims=1000 # number of times we do this

power_result_vig_2 <- ANOVA_power(design_result,
                                  nsims = nsims,
                                  seed = 1234)
```

```
## Power and Effect sizes for ANOVA tests
##                 power effect_size
## anova_condition  38.2      0.1461
##
## Power and Effect sizes for pairwise comparisons (t-tests)
##                                        power effect_size
## p_condition_Cloak_condition_No Cloak  38.2     -0.7335
```

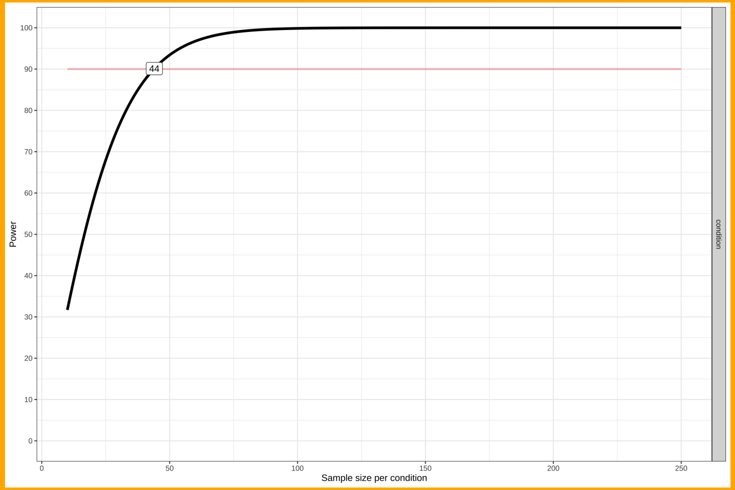
```r
#Note we do not specify any correlation in the ANOVA_design function (default r = 0), nor do we specify

knitr::kable(confint(power_result_vig_2, level = .98))
```

	power	lower.ci	upper.ci
anova_condition	38.2	35.23915	41.25117

Power Curve

```
plot_power(design_result, min_n = 10, max_n = 250)
```



```
## Achieved Power and Sample Size for ANOVA-level effects
##    variable                  label  n achieved_power desired_power
## 1 condition Desired Power Achieved 44          90.12            90
```

Dependent: Power

- The test type will affect power, as our independent results suggested we needed 44 or more people

```
pwr.t.test(n = 12,
           d = 1.10,
           sig.level = .05,
           power = NULL,
           type = "paired",
           alternative = "two.sided")
```

```
##
##      Paired t test power calculation
##
##               n = 12
##               d = 1.1
##       sig.level = 0.05
##           power = 0.9333861
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```r
# if (!require("devtools")) {
 #   install.packages("devtools", dependencies = TRUE)}

 #  devtools::install_github("DejanDraschkow/mixedpower")

library(mixedpower)

longdata$id<-rep(1:12, length(longdata))

d_reg<-lme4::lmer(Mischief~Cloak + (1|id),  data=longdata)

d_reg_mixed <- mixedpower(d_reg, data=longdata, fixed_effects = c("Cloak"), simvar = "id", steps = c(20
```

Posthoc Power

- Sometimes reviewers will ask you to conduct a post-experiment power calculation in order to interpret non-significant findings
    - Do not do this!

1. Sample effect size not rep of population size

2. Adds nothing over a *p*-value
    - p > .05 = low power(Duh!)

```r
set.seed(11)
x1 <- rnorm(10, mean = 10, sd = 1)
x2 <- rnorm(10, mean = 10.1, sd = 1)

#run t test
# get effect size
```

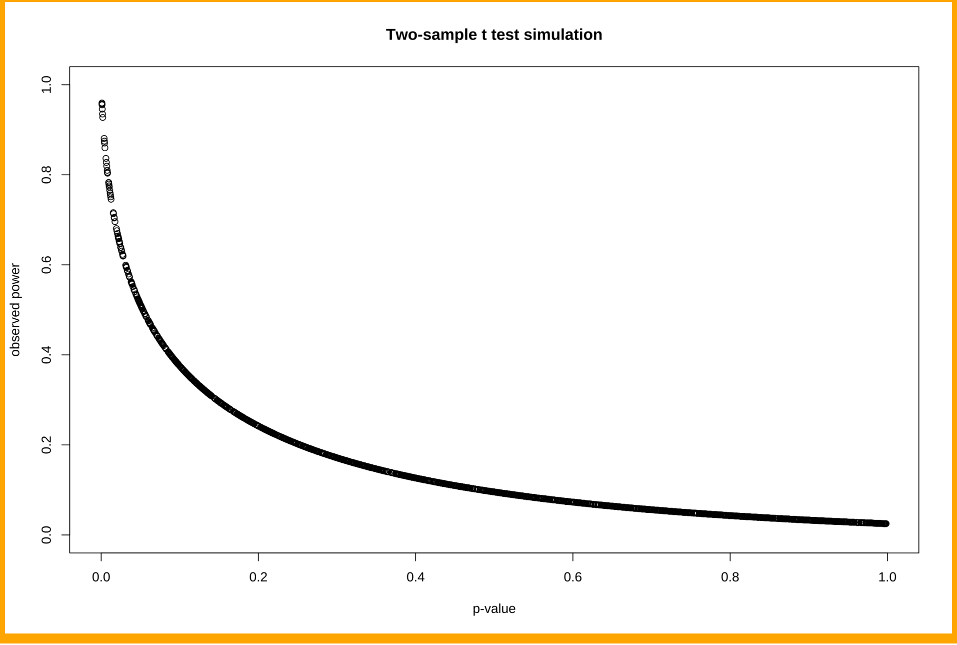
Posthoc Power

```
#calculate posthoc power from this
```

Posthoc Power

```r
sim_out <- replicate(n = 2000, expr = {
  x1 <- rnorm(10, mean = 10, sd = 1)
  x2 <- rnorm(10, mean = 10.1, sd = 1)
  ttest <- t.test(x1, x2, var.equal = TRUE)
  pwr <- power.t.test(delta = diff(ttest$estimat
                sd = sqrt((var(x1) + var(x2))/2),
                sig.level = 0.05,
                n = 10)
  c(pvalue = ttest$p.value, obs_power = pwr$powe
})
```

```r
plot(t(sim_out),
     xlim = c(0,1), ylim = c(0,1),
     xlab = "p-value", ylab = "observed power",
     main = "Two-sample t test simulation")
```



The Minimal Detectable Effect Size

- Report the smallest effect size that could be detected in your study for this particular sample size
- We do not live in a perfect world
    - Sometimes we cannot collect all the data we need