Effect, Power, ANOVA, and NHST Review

Presented by David John Baker
August 2019

FLATIRON SCHOOL

Power, Effects, ANOVA, Review

- Power tells us how likely are we to find an effect if there is one
- In practice, this means allocating the appropriate amount of resources (time and money!)
- Power relates to the other concepts this week and we will dive into all those relationships while reviewing others this week
- Introduce ANOVA and problems of multiple comparison
- Review Terms Together

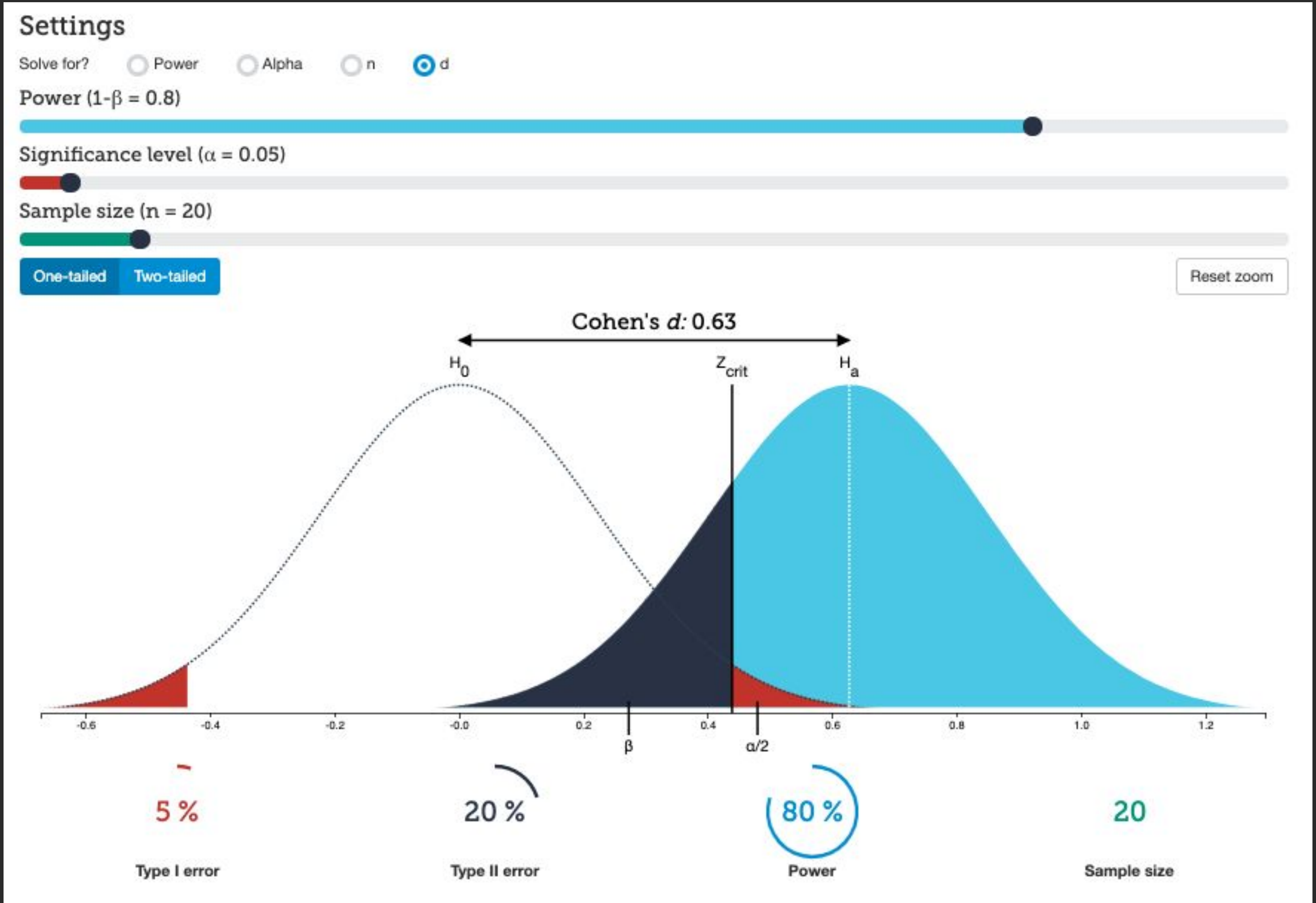
Types of Errors

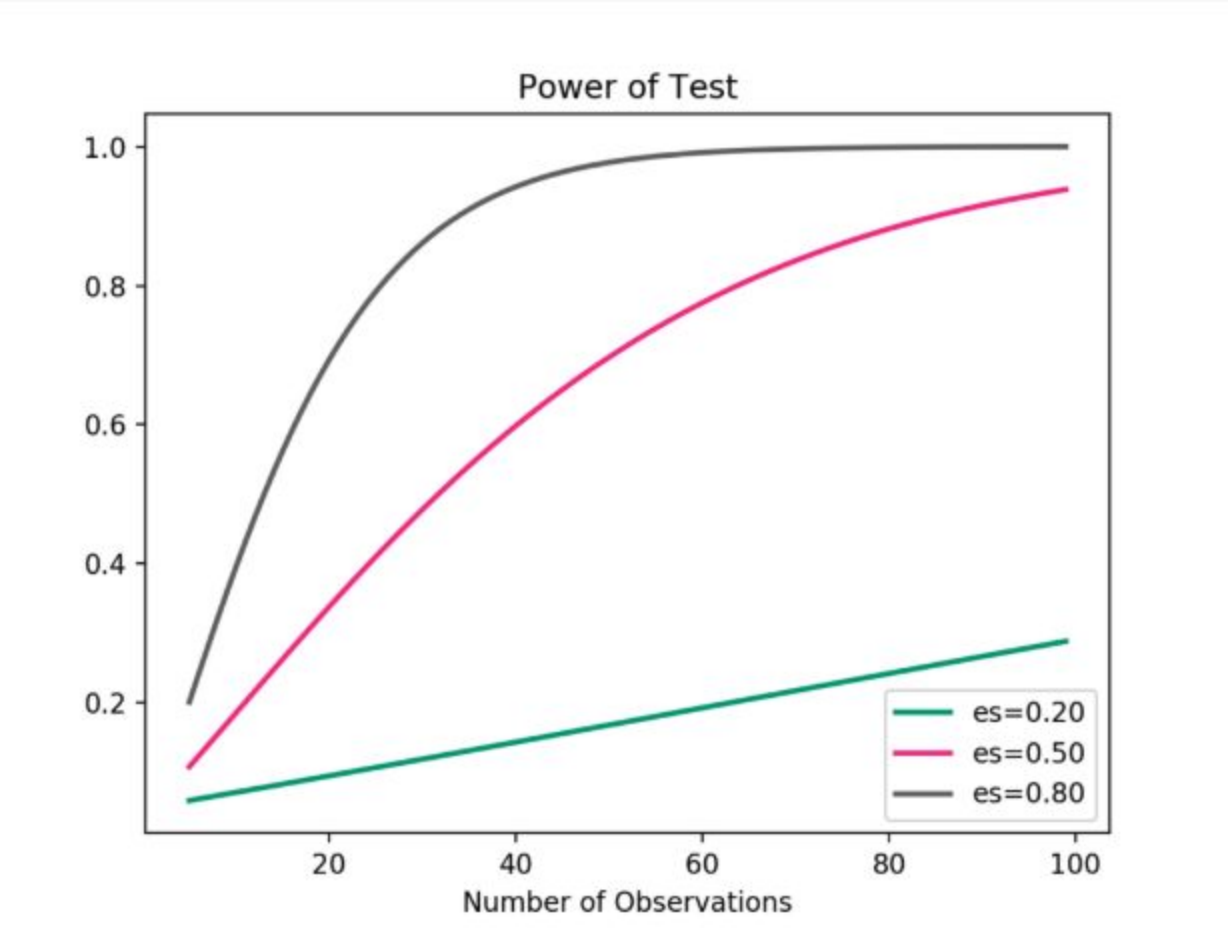
	H0 True	H1 True
Significant Finding	False Positive	True Positive
Non-Significant Finding	True Negative	False Negative

Review (Interview Questions)

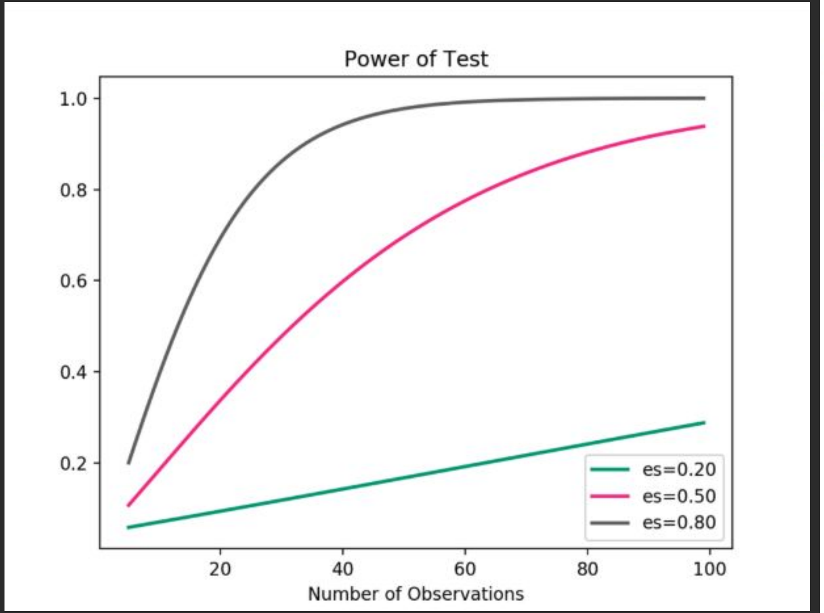
- What is hypothesis testing and how is it used in research?
- Why do we set up a null hypothesis and an alternative hypothesis?
- What are Type I errors?
- What are Type II errors?
- What can be done in order to lower your error rates?
- Describe the differences between frequentist and Bayesian schools of thought on data and its relation to the world.

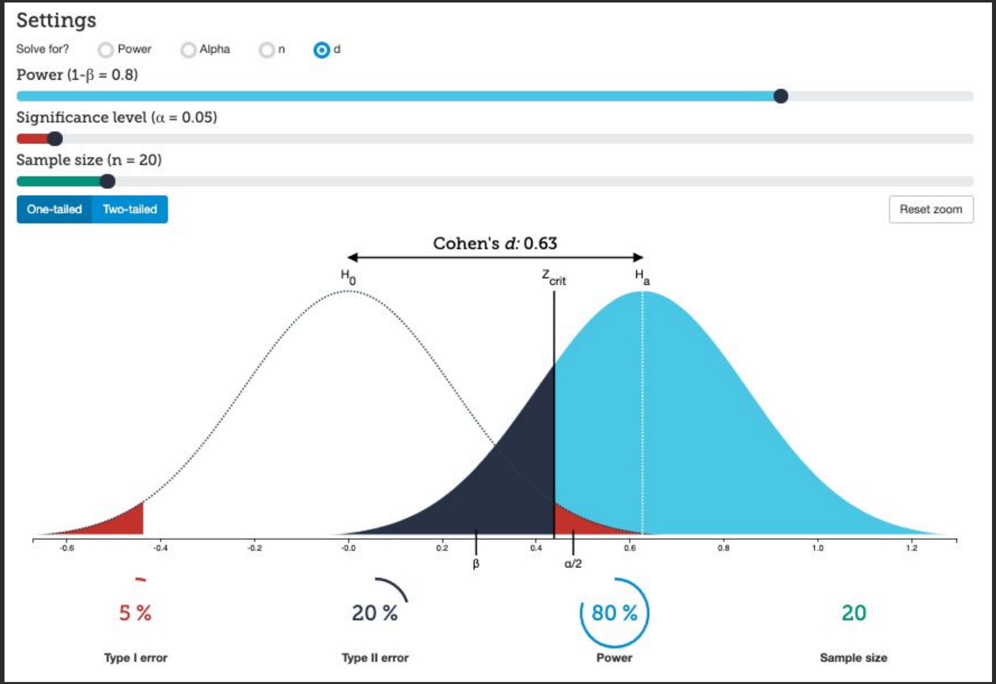
https://rpsychologist.com/d3/NHST/


Settings
Solve for?
Power
Alpha
n
d
Power (1-β = 0.8)
Significance level (α = 0.05)
Sample size (n = 20)
One-tailed
Two-tailed
Reset zoom
Cohen's d: 0.63
H0
Zcrit
Ha
-0.6
-0.4
-0.2
-0.0
0.2
0.4
0.6
0.8
1.0
1.2
β
α/2
5 %
Type I error
20 %
Type II error
80 %
Power
20
Sample size



Power of Test
1.0
0.8
0.6
0.4
0.2
20
40
60
80
100
Number of Observations
es=0.20
es=0.50
es=0.80


Power of Test
1.0
0.8
0.6
0.4
0.2
20
40
60
80
100
Number of Observations
es=0.20
es=0.50
es=0.80


Settings
Solve for?
Power
Alpha
n
d
Power (1-β = 0.8)
Significance level (α = 0.05)
Sample size (n = 20)
One-tailed
Two-tailed
Reset zoom
Cohen's d: 0.63
Z crit
β
α/2
5 %
Type I error
20 %
Type II error
80 %
Power
20
Sample size

//

Special Relationships

- There is a special relationship between effect size, sample size, alpha, and beta. If you know three, you can always calculate the fourth.
- In groups, come up with a situation in which you have three and would need to calculate the fourth. Be prepared to justify your answers. You should have four scenarios.

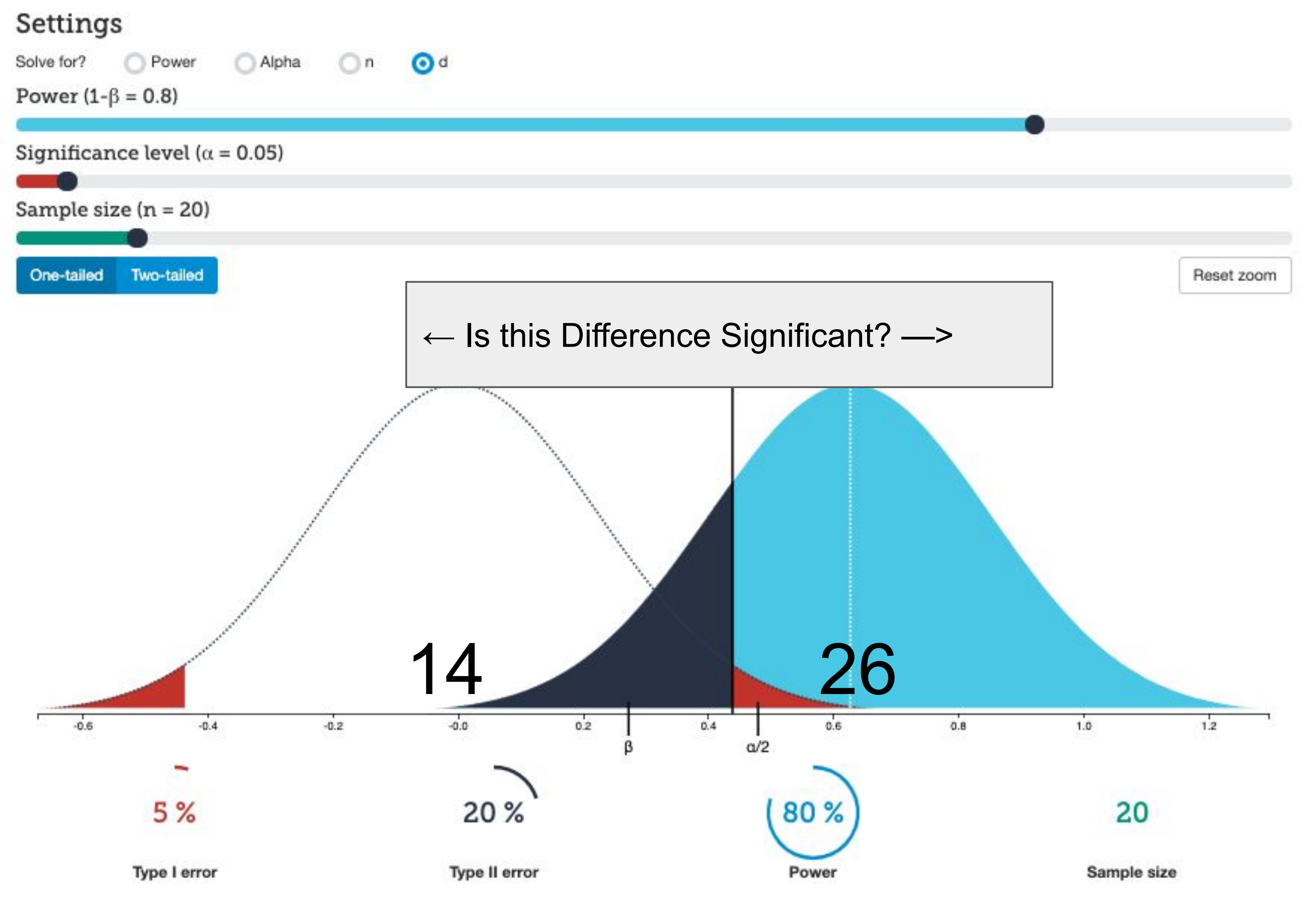
Discussion Questions

- Describe the relationship between the power of a test and the size of its effect
- What affects effect size? ***

Calculating Effect Size

Cohen's d = (Mean 1 - Mean 2) / SD*

$*SD_{pooled} = \sqrt{((SD_1^2 + SD_2^2)/2)}$


Settings
Solve for?
Power
Alpha
n
d
Power (1-β = 0.8)
Significance level (α = 0.05)
Sample size (n = 20)
One-tailed
Two-tailed
Reset zoom
← Is this Difference Significant? —>
14
26
-0.6
-0.4
-0.2
-0.0
0.2
0.4
0.6
0.8
1.0
1.2
β
α/2
5 %
Type I error
20 %
Type II error
80 %
Power
20
Sample size

Calculating Effect Size

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

c. **Step three**: Select the sample and collect your data.

d. **Step four**: Locate the region of rejection and the critical value(s) of your test statistic. Again, directionality is important to consider.

Cohen's d = (Mean 1 - Mean 2) / SD*

Cohen's d = (26 - 14) / 4 = 3

$*SD_{pooled} = \sqrt{((SD_1^2 + SD_2^2)/2)}$

https://rpsychologist.com/d3/cohend/

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

What other factors might contribute to a measure of effect size?

Where else have we seen an an effect size?

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

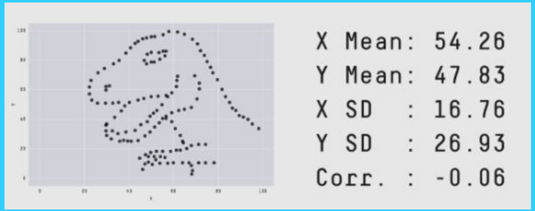
What other factors might contribute to a measure of effect size?

Correlation coefficients are effect sizes!!!

$$r_P = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

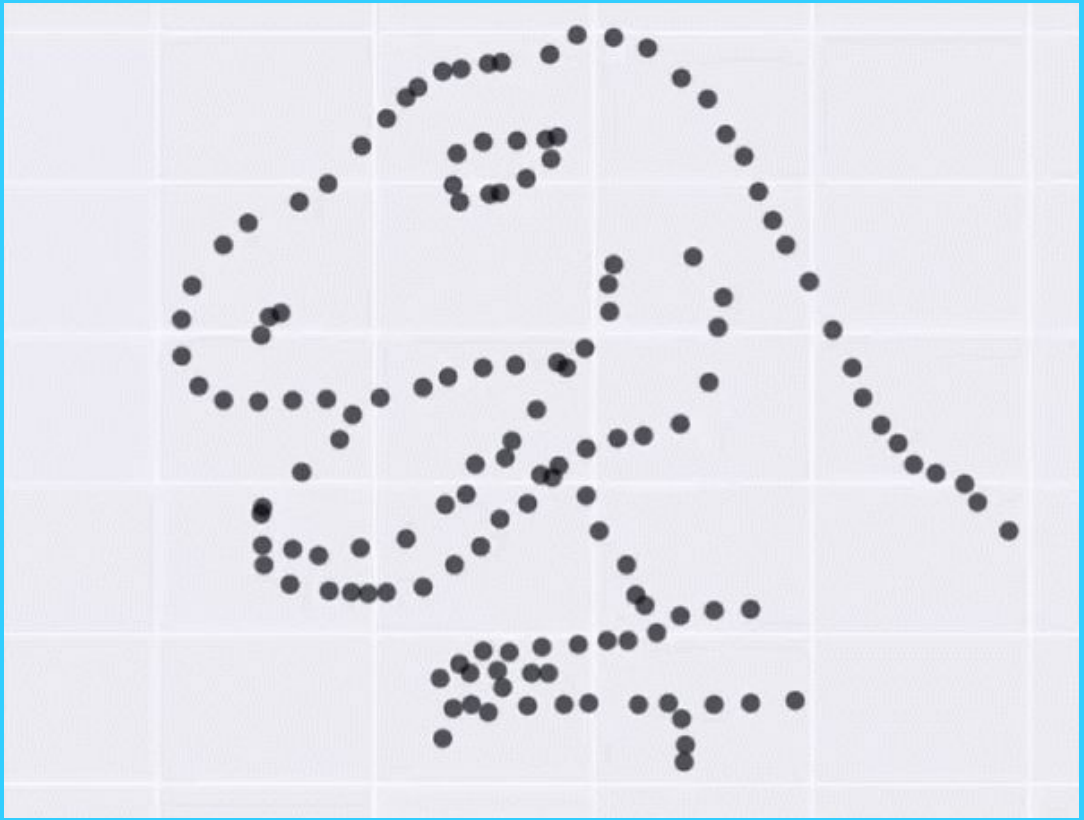
$$\beta_1 = r_P \frac{\sigma_y}{\sigma_x}$$

$$\beta_0 = \bar{y_1} - \beta_1 \bar{x}$$



Correlation coefficients are effect sizes!!!

But make sure to always look at your data



https://www.autodeskresearch.com/publications/samestats https://rpsychologist.com/d3/correlation/

```
#Load pwr package to easily calculate the statistical power
if(!require(pwr)){install.packages('pwr')}
library(pwr)
#Disable scientific notation (1.05e10)
options(scipen=999)
#Set number of simulations
nSims <- 100000 #number of simulated experiments

M<-106 #Mean IQ score in the sample (will be compared with 100 in a one-sample t-test)
n<-26 #set sample size
SD<-15 #SD of the simulated data
#With a mean difference of 6, and SD of 15, and a sample size of 26, the test has 50% power)

p <-numeric(nSims) #set up empty variable to store all simulated p-values
bars<-20
#Run simulation
for(i in 1:nSims){ #for each simulated experiment
  x<-rnorm(n = n, mean = M, sd = SD) #Simulate data with specified mean, standard deviation, and sample size
  z<-t.test(x, mu=100) #perform the t-test against mu (set to value you want to test against)
  p[i]<-z$p.value #get the p-value and store it
}

#Check power by summing significant p-values and dividing by number of simulations
(sum(p < 0.05)/nSims) #power
#Calculate power formally by power analysis
power<-pwr.t.test(d=(M-100)/SD, n=n,sig.level=0.05,type="one.sample",alternative="two.sided")$power #determines M when power > 0. When power = 0, will set  M = 100.

#Plot figure
#png(file="P-valueDist.png",width=4000,height=3000, , units = "px", res = 500)
op <- par(mar = c(5,7,4,4)) #change white-space around graph
hist(p, breaks=bars, xlab="P-values", ylab="number of p-values\n", axes=FALSE,
     main=paste("P-value Distribution with",round(power*100, digits=1),"% Power"),
     col="grey", xlim=c(0,1),  ylim=c(0, nSims))
axis(side=1, at=seq(0,1, 0.1), labels=seq(0,1,0.1))
axis(side=2, at=seq(0,nSims, nSims/4), labels=seq(0,nSims, nSims/4), las=2)
abline(h=nSims/bars, col = "red", lty=3)
#dev.off()

#Š Daniel Lakens, 2016.
# This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. https://creativecommons.org/licenses/by-nc-sa/4.0/
```

ANOVA

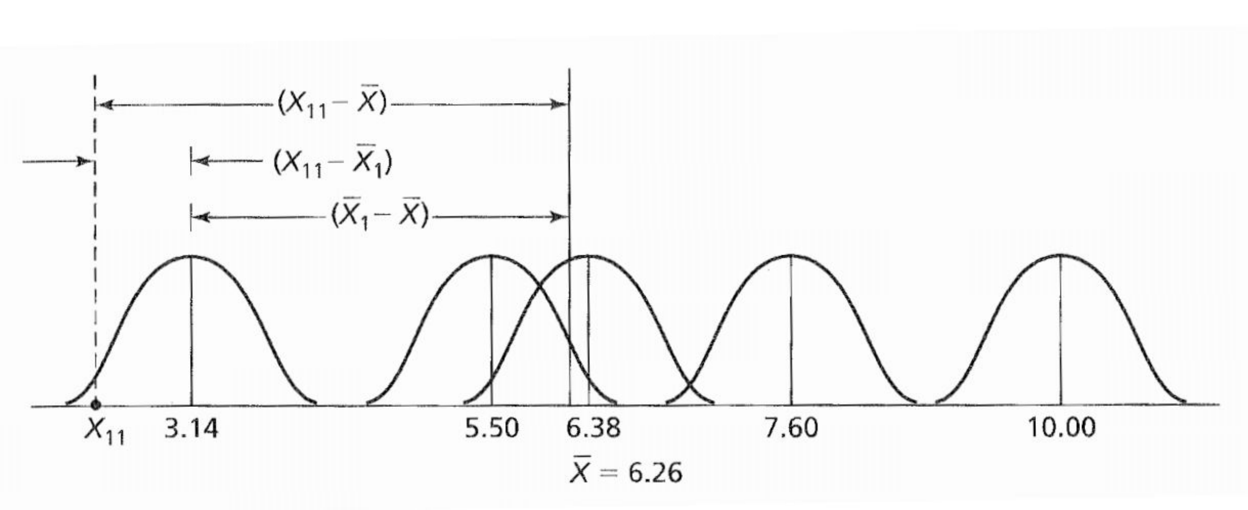
- For the past two lectures, we have been focusing on differences between 2 distributions, quantifying our uncertainty, but we can't just run tons of t-tests our whole lives
- Knowing what we know about Type I error rates, what problem would there be in running t-test after t-test after t-test?

ANOVA II

- Performing multiple t-tests increases our Type I error rate!!
- In order to not fool ourselves in the long run, we need to set up a way to protect ourselves against Type I error rates
- Experimentwise Type I Error Rate
  - $1-(1-\text{alpha})^c$
  - Where c = number of independent t tests
  - For example, if I had three groups, my experiment type I error rate would be $1 - (1-.05)^3 = .142$

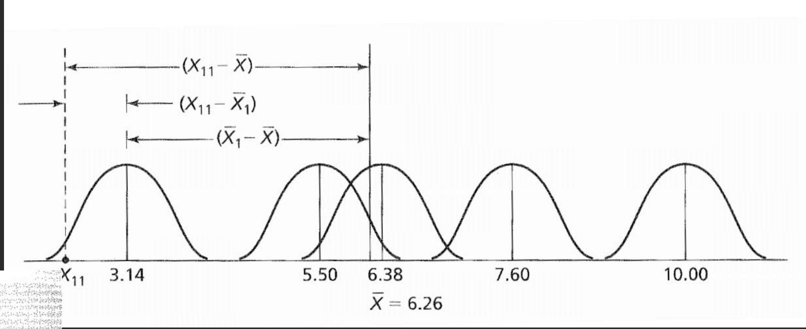
ANOVA III

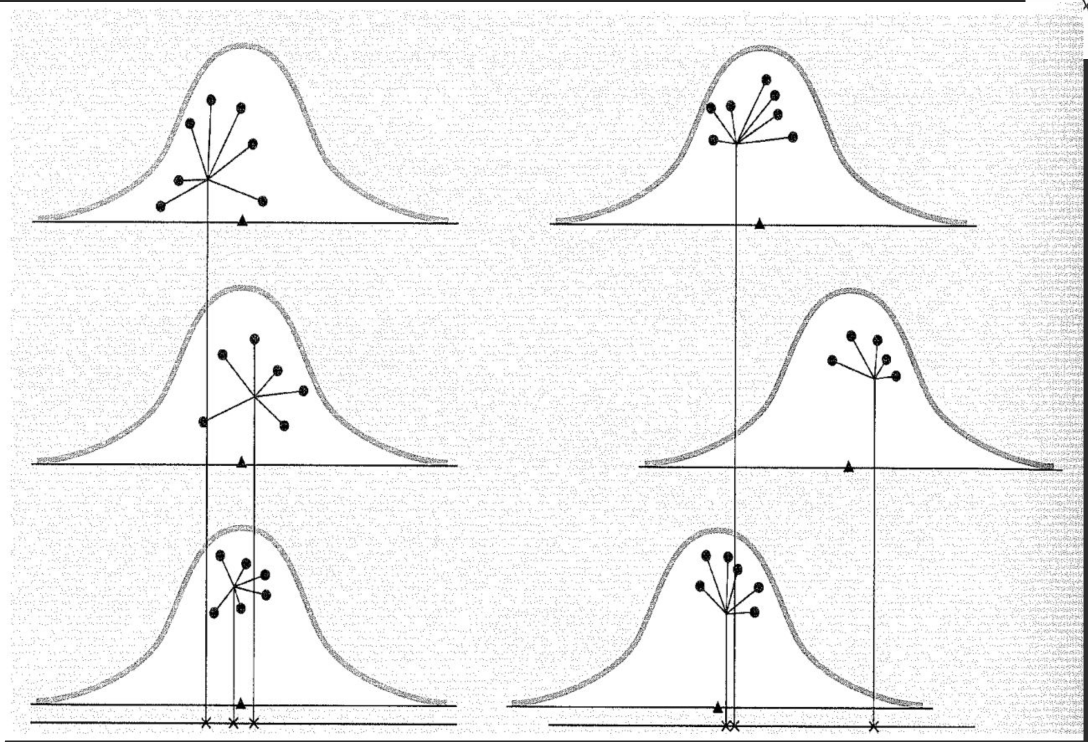
- ANOVA or ANalysis Of VAriance or F test attempts to fix this problem



//

ANOVA IV


$(X_{11} - \bar{X})$
$(X_{11} - \bar{X}_1)$
$(\bar{X}_1 - \bar{X})$
X_{11}
3.14
5.50
6.38
7.60
10.00
$\bar{X} = 6.26$



Scenario: Does the ethnicity of a defendant affect the likelihood that he is judged guilty? People were given transcripts of a trial and asked to judge the likelihood that a defendant was guilty, on a 0 – 10 scale. The transcript was identical, but across 3 conditions, the reported ethnicity of the defendant varied. The results were as follows (study based on Stephen, 1975):

White	African-American	Hispanic
6	10	10
7	9	6
2	4	5
3	10	5
5	10	2
0	3	10

Step one: null and alternative hypotheses

Null	Alternative
$\mu_W = \mu_{AA} = \mu_H$	$\mu_i \neq \mu_k$, for some groups i and k
	$\mu_W < \mu_{AA} = \mu_H$
	$\mu_W = \mu_{AA} \neq \mu_H$
	$\mu_W = \mu_H \neq \mu_{AA}$

Step two: select the test and significance level

F test, $\alpha = .05$

Step three: select samples and collect data

Step four: locate region of rejection (i.e., critical value)

F table with *df* in numerator = $K - 1$, *df* in denominator = $N - K$

$$F_{2,15;\alpha=.05} = 3.68$$

Step five: calculate the test statistic

	White	A-A	Hispanic	
	6	10	10	
	7	9	6	
	2	4	5	
	3	8	5	
	5	10	2	
	0	3	8	
n_k	6	6	6	N = 18
s^2	6.97	9.47	7.60	
$\bar{X}$	3.83	7.33	6.00	5.72

when group n is NOT equal:

$$MS_W = \frac{\sum (n_i - 1)s_i^2}{\mathrm{df}_W}$$

$$MS_W = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + \cdots + (n_k - 1)s_k^2}{N_T - k}$$

$$MS_W = \frac{\sum s^2}{k} = \frac{6.97+9.47+7.60}{3} = 8.01, \text{ when group } n \text{ is equal}$$

$$MS_{\mathrm{B}} = \frac{\sum n_i(\bar{X}_i - \bar{X}_G)^2}{\mathrm{df_B}} = \frac{6(3.83-5.72)^2 + 6(7.33-5.72)^2 + 6(6.00-5.72)^2}{3-1}$$

$$MS_{\mathrm{B}} = \frac{21.43 + 15.55 + 0.47}{2} = 18.73$$

$$F = \frac{MS_B}{MS_W} = \frac{18.73}{8.01} = 2.34$$

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

e. **Step five**: Compute the appropriate test statistic. σ is known, so we use the z test.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{14}} = 1.07 \qquad z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{23.36 - 26}{4/\sqrt{14}}$$

$$z = \frac{-2.64}{1.07} = -2.47$$

$$MS_{\mathrm{B}} = \frac{\sum n_i (\bar{X}_i - \bar{X}_G)^2}{\mathrm{df}_{\mathrm{B}}} = \frac{6(3.83-5.72)^2 + 6(7.33-5.72)^2 + 6(6.00-5.72)^2}{3-1}$$

$$MS_{\mathrm{B}} = \frac{21.43 + 15.55 + 0.47}{2} = 18.73$$

$$F = \frac{MS_B}{MS_W} = \frac{18.73}{8.01} = 2.34$$

Scenario: On a standardized anagram task, $\mu = 26$ anagrams solved with a $\sigma = 4$. A researcher tests whether the arousal from anxiety is distracting and will decrease performance. A sample of $n = 14$ anxiety patients is tested on the task. There average performance is 23.36 anagrams.

e. **Step five**: Compute the appropriate test statistic. σ is known, so we use the z test.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{14}} = 1.07 \qquad z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{23.36 - 26}{4/\sqrt{14}}$$

$$z = \frac{-2.64}{1.07} = -2.47$$

$$MS_{\mathrm{B}} = \frac{\sum n_i (\bar{X}_i - \bar{X}_G)^2}{\mathrm{df}_{\mathrm{B}}} = \frac{6(3.83-5.72)^2 + 6(7.33-5.72)^2 + 6(6.00-5.72)^2}{3-1}$$

$$MS_{\mathrm{B}} = \frac{21.43 + 15.55 + 0.47}{2} = 18.73$$

$$F = \frac{MS_B}{MS_W} = \frac{18.73}{8.01} = 2.34$$

//

ANOVA summary table

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	37.45	2	18.73	2.34	0.130761	3.682317
Within Groups	120.17	15	8.01			
Total	157.61	17				

Review Questions

- What is statistical power?
- How does it relate to effect size?
- What are some common effect size measures?
- What contributes to effect size?
- What parameters would change when you manipulate…
    - Sample size?
    - Alpha Levels?
    - Beta Levels?

Review Questions II — True, False, and Why?

- A significant p value indicates the degree of evidence for the alternative hypothesis
- Collecting larger samples will result in larger effect sizes
- The null hypothesis and the alternative hypothesis should be probabilistically mutually exclusive
- Confidence intervals capture the range where the point estimate is most likely to occur