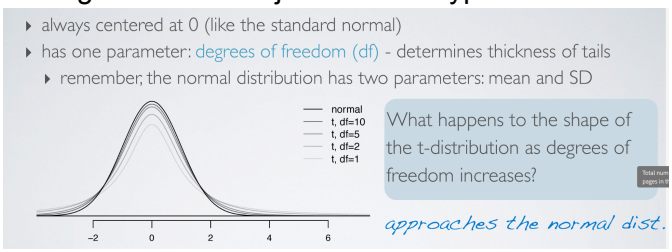
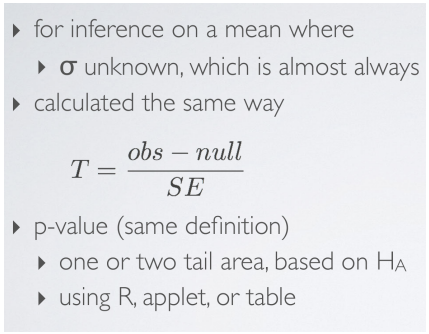
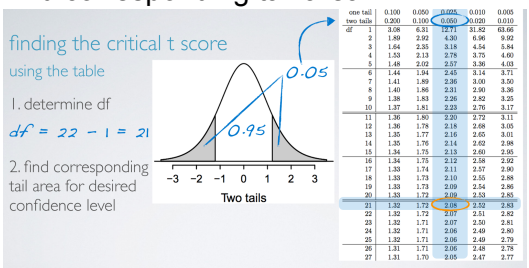
T-Distribution and Comparing Two Means	
T-Distribution	When population data is unknown, use the t-distribution to address the uncertainty of the standard error estimates. - Has thicker tails to account for less reliable estimation, making it harder to reject the null hypothesis - 
T-Statistic Calculation	Similar to a sample statistic only we reference the t-table and calculate a t-value than a z-table/z-value 
T-Stat Notations:	Critical T-Score 1. Using t-table: a. Determine df = n-1 b. Find corresponding tail area  2. Using R a. `> qt(0.025, df = 21)` `[1] -2.079614` Estimating the T-Statistic Mean (margin of error)

	$\bar{x} \pm t^\star_{df} SE_{\bar{x}}$ $\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$ $\bar{x} \pm t^\star_{n-1} \frac{s}{\sqrt{n}}$ **Degrees of freedom for t statistic for inference on one sample mean** $df = n - 1$
Inference for ONE mean (using t-stat)	**sample:** 44 patients: 22 men and 22 women **study design:** - randomized into two groups: (1) play solitaire while eating - "win as many games as possible" (2) eat lunch without distractions - both groups provided same amount of lunch - offered biscuits to snack on after lunch | *biscuit intake* | $\bar{x}$ | s | n | | **solitaire** | 52.1 g | 45.1 g | 22 | | **no distraction** | 27.1 g | 26.4 g | 22 | Estimate the average after-lunch snack consumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval. Find the Confidence Interval of 95% Step 1: Identify the given items - x(distracted average) = 52.1 - n(sample size) = 22 - sd = 45.1 Step 2: Find the t-score - t* = degree of freedom via the table - df = n-1 = 22-1 = 22 - t*21 = 2.08 (via t-table) Step 3: Find the Confidence Interval $\bar{x} \pm t^* SE = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$ $= 52.1 \pm 2.08 \times 9.62$ $= 52.1 \pm 20 = (32.1, 72.1)$
Pointed estimate using t-score Example: (One Mean)	Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size? Step 1: identify the given factors - x(distracted average) = 52.1 - n(sample size) = 22 - sd = 45.1 - SE = sd/sqrt(n) = 9.62 Step 2: Define hypothesis

- U = amount of snacks consumed by distracted eaters
- H0: u=30
- HA: u != 30

Step 3: Find t-score

- Tscore = (observe value - null)/SE
- $T = \frac{52.1 - 30}{9.62} = 2.3$

Step 4: Find df

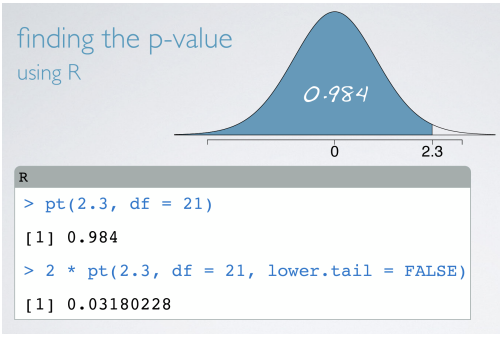
- $df = 22 - 1 = 21$

Step 4: Find p-value

- Locate on the t-table, 2.3 is between 0.05 and 0.02
- So the percentage is 2%<pval<5%
-

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77

- Locate on R. since the left curve is 98% and we want the the extremes, do (1-98%) * 2
-



Step 5: Interpret the result

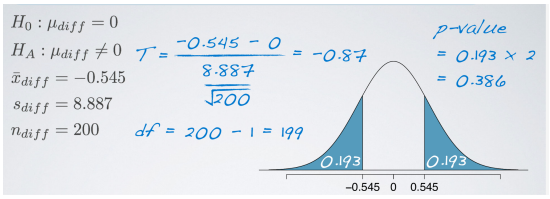
- 0.03 < 0.05
- Reject H0

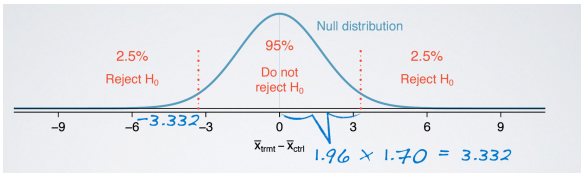
Note: The data is naturally right skewed because you can't eat less than 0g of food but can eat unlimited more grams of food.

Inference for Comparing Two independent Means (Two Independent Mean)	Estimating the difference between two independent means point estimate ± margin of error $(\bar{x}_1 - \bar{x}_2) \pm t^\star_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$ SE and df are calculated differently now. **Standard error of difference between two independent means:** $SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ **DF for t statistic for inference on difference of two means** $df = min(n_1 - 1, n_2 - 1)$ Conditions: **Conditions for inference for comparing two independent means:** 1. ***Independence:*** ✓ **within groups:** sampled observations must be independent ‣ random sample/assignment ‣ if sampling without replacement, n < 10% of population ✓ **between groups:** the two groups must be independent of each other (non-paired) 2. ***Sample size/skew:*** The more skew in the population distributions, the higher the sample size needed.
Point estimate for difference of two means Example:	Estimate the difference between the average post-meal snack consumption between those who eat with and without distractions. | biscuit intake	$\bar{x}$	s	n
---	---	---	---
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22	 Step 1: Find the Confidence Interval $(\bar{X}_{wd} - \bar{X}_{wod}) \pm t^*_{df} \; SE = (52.1 - 27.1) \pm 2.08 \times \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}$ $= 25 \pm 2.08 \times 11.14$ $= 25 \pm 23.17$ $= (1.83, 48.17)$ Step 2: Interpret the Data - Says that those who eat distracted consume about 1.83 to 48.17 more grams than no distracted eating
Point estimate for difference of two means Example:	Do these data provide convincing evidence of a difference between the average post-meal snack consumption between those who eat with and without distractions? | biscuit intake	$\bar{x}$	s	n
---	---	---	---
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22	

	Step 1: Establish your Hypothesis $H_0: \mu_{wd} - \mu_{wod} = 0$ $H_A: \mu_{wd} - \mu_{wod} \neq 0$ Step 2: Find t-score - Find df = n-1 = 21 - Find t-score - $T_{21} = \frac{25 - 0}{11.14} = 2.24$ Step 3: Find p-value - T-score of 2.24 is appx 0.98 - P-value = 2*(1-0.98) = 0.04 Step 4: Interpret - 0.04 < 0.05, reject H0 - It does provide evidence of a difference
Inference for comparing two PAIRED means (not independent)	Paired Data can happen when we take repeated measures or the nature of the data. For example, reading scores are likely not independent from writing scores. What do we do? Create a new variable that takes the difference of the two paired data and perform all the sample statistics steps on that difference variable.
	Step 1: Create a difference variable Reminder that the udiff is the population and what we define for the hypothesis while xdiff is the sample or point estimate parameter of interest: Average difference between the reading and writing scores of **all** high school students. μ_{diff} point estimate: Average difference between the reading and writing scores of **sampled** high school students. $\bar{x}_{diff}$

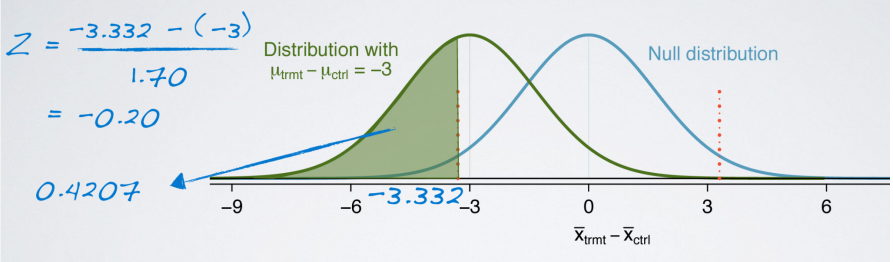
	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
…	…	…	…	…
200	137	63	65	-2

	Step 2: Find the average, standard deviation, and n of the difference column $\bar{x}_{diff} = -0.545$ $s_{diff} = 8.887$ $n_{diff} = 200$ Step 3: Formulate hypothesis $H_0 : \mu_{diff} = 0$ There is no difference between the average reading and writing scores. $H_A : \mu_{diff} \neq 0$ There is a difference between the average reading and writing scores. - Step 4: Find the T-score - Find df = 200-1 = 199 - Find t-score = (below) Step 5: Find p-value - Tscore results in 0.193 (we want the extremes) - P-value = 0.386 
Power	The probability of correctly rejecting H0 and is denoted as 1-B  Increasing sample size can decrease both alpha and beta low. So we can do this calculation to determine the ideal sample size the increase the power of a test.

Power Calculation Example:	Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication, and half of the subjects are given the new drug (treatment) and the other half continue to take their current medication through generic-looking pills to ensure blinding (control). What are the hypotheses for a two-sided hypothesis test in this context? Step 1: Formulate the Hypothesis - $H_0: \mu_{trmt} - \mu_{ctrl} = 0$ $H_A: \mu_{trmt} - \mu_{ctrl} \neq 0$ Suppose researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric. If we had 100 patients per group, what would be the approximate standard error for difference in sample means of the treatment and control groups? Step 2: Find Standard Error - $SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$ For what values of the difference between the observed averages of blood pressure in treatment and control groups (effect size) would we reject the null hypothesis at the 5% significance level? Step 3: Find confidence interval - Since the null=0, the t-statistic distribution would be centered at 0 and we just need to find the margin of error which is t*SE -  Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. What is the power of the test that can detect this effect? Step 4: Find the power of the test - Aka find the probability that we accept the HA hypothesis - Step 1: Question is asking for a difference that is 3mg or

larger. So we can first re-center the distribution to be at -3 (or +3, doesn't matter)

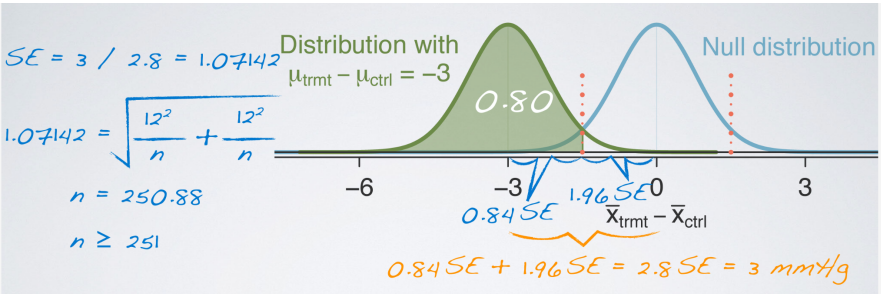
- Step 2: since question is asking for 3 or larger, it would be to the left of the curve. For this to keep in the 95% confidence, the area would be to the left of the -3.332 confidence interval.
- Step 3: Find the area under the curve via z-score. Power is 0.4207



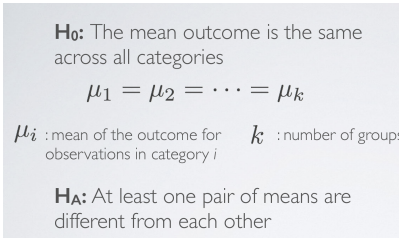
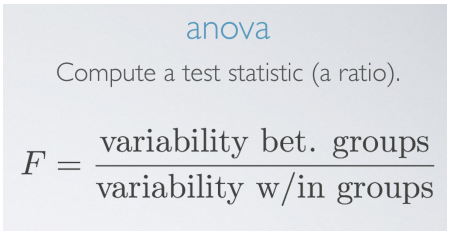
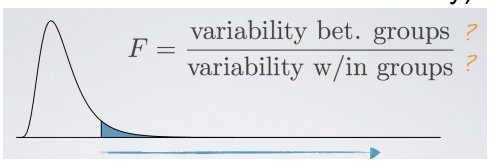
What sample size will lead to a power of 80% for this test?

Step 6: Find the sample size

- A z-score of 0.84 yields 80%
- Z-score = (observed value - null)/SE
- (observed value-null) = distance from the center
- Therefore….distance from -3 = z-score*SE = 0.84
- Then leverage SE to find the sample size



- We know that the distance between the null center and the rejection region is 1.96SE. That was predetermined when we wanted the region to be 95% confident. We are calibrating the green region to be 80%

ANOVA and Bootstrapping	
ANOVA notations: F-distribution	Comparing more than two means Hypothesis Construction: -  Test Statistics: -  Distribution Curve: - Alway straight skewed and positive (because it's a ratio between two measures of variability) -  - To reject H0 will need a large f-statistic. Large f-statistic will require bigger variability between groups than within groups
ANOVA interpretation (do not need to calculate by hand)	Conceptually, ANOVA provides a way to compare multiple means by interpreting variability in the different data. Total Variability is split into variability **between** groups and variability **within** groups. Ex) In determining if there is a relationship between vocabulary score and social class, there is variability attributed to class (working, middle, high) and variability attributed to other factors.

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

The table shows Group variability and Total variability. Error is the difference of Total and Group

ANOVA SST (sum of squares total)

Sum of Squares Total (***Total: 3106.36***)

- Measures the total variability in the response variable
- Calculation is similar to variance, just not divided by sample size

Sum of squares total (SST):

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

y_i : value of the response variable for each observation

$\bar{y}$: grand mean of the response variable

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
...	...	...
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

$SST = (6-6.14)^2 + (9-6.14)^2 + (6-6.14)^2 + \cdots + (9-6.14)^2 = 3106.36$

ANOVA SSG (sum of squares group)

Sum of Squares Group (***Group: 236.56)***

- Measures variability between groups
- Squared deviation of group means from overall mean.
- Weighted by sample size of each group

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

n_j : number of observations in group j

$\bar{y}_j$: mean of the response variable for group j

$\bar{y}$: grand mean of the response variable

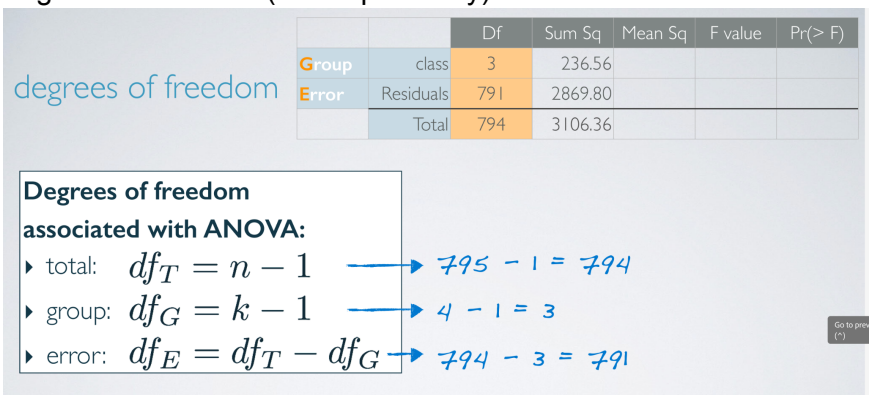
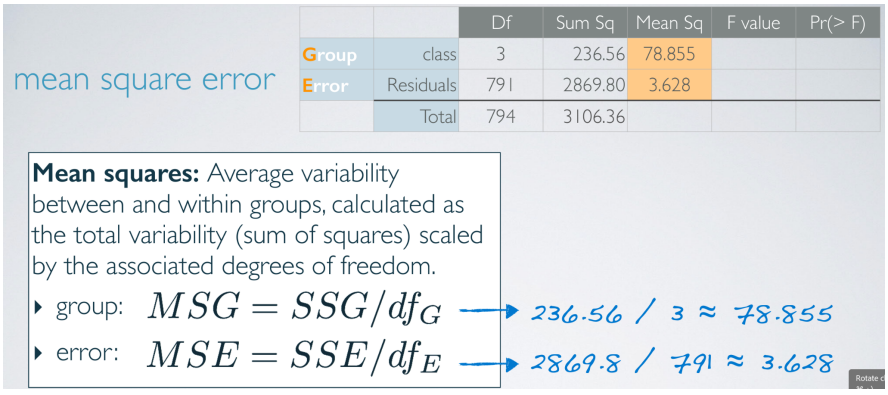
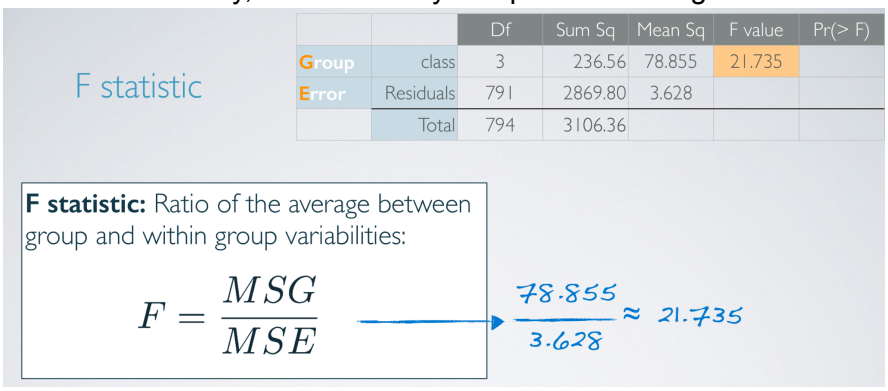
	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

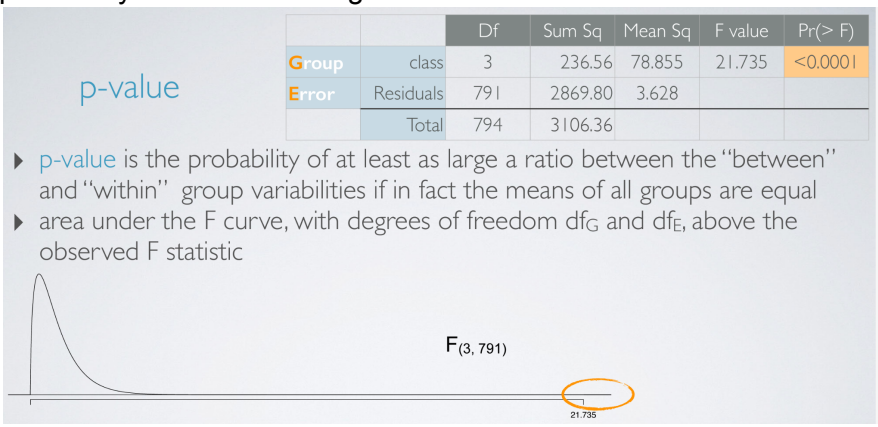
$SSG = (41 \times (5.07 - 6.14)^2) + (407 \times (5.75 - 6.14)^2) + (331 \times (6.76 - 6.14)^2) + (16 \times (6.19 - 6.14)^2) \approx 236.56$

ANOVA SSE (Sum of Squares Error)

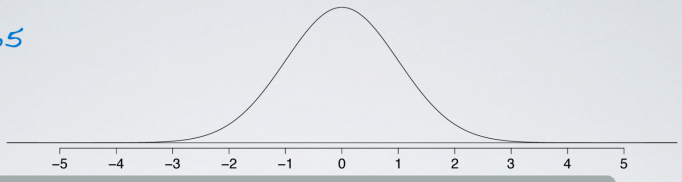
Sum of Squares Error (***Error: 2869.8***)

- measures variability within groups (the variability that is NOT in group)
- Thought of as the variability unexplained by group variable due to other reasons.

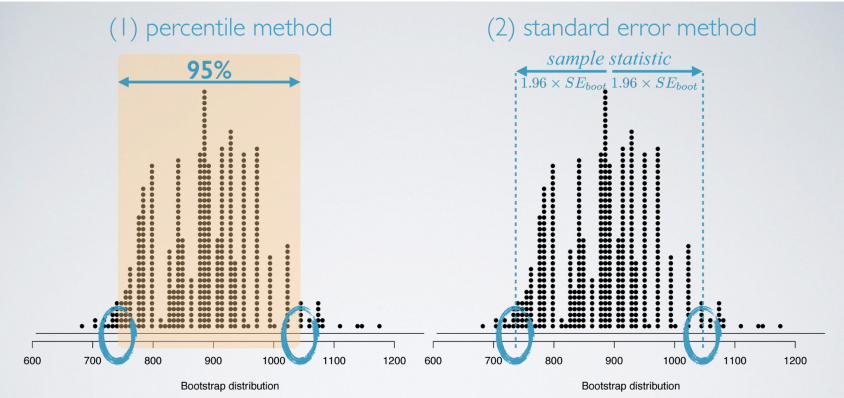
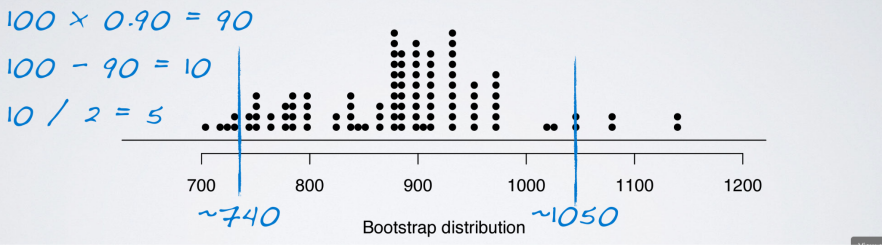
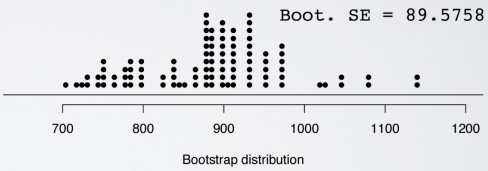
	- **Sum of squares error (SSE):** $SSE = SST - SSG$
ANOVA Df	Degrees of freedom (self explanatory) 
ANOVA Mean Square Error	Average variability between and within groups Sum of Squares divided by degrees of freedom 
ANOVA F-Value	A ratio of variability, so it will always be positive and right skewed 
ANOVA P-value	Given that the means of all groups are equal to each other, the

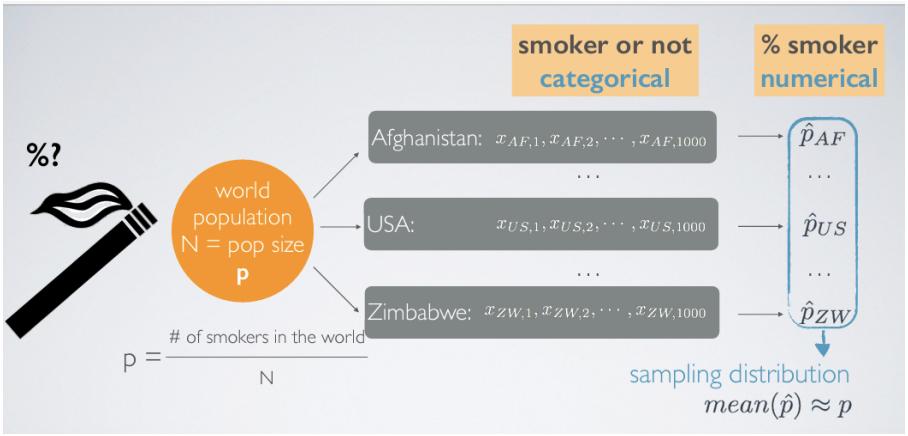
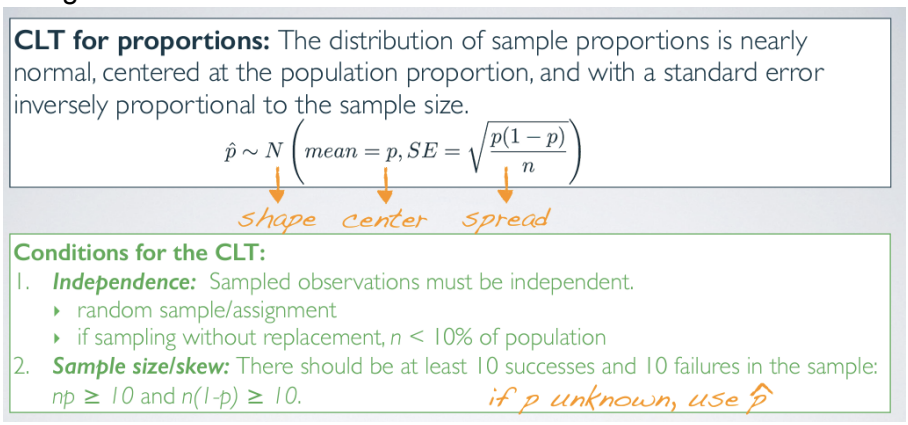
	probability of at least as large a F-value ratio is close to 0%.  P-value is small (< 0.05), reject H0 - Data provides convincing evidence that at least one pair of population means are different from each other P-value is large (> 0.05), fail to reject H0 - High probability that the same mean exists given that all the means are already equal to each other. - The observed differences are then attributable to sampling variability or chance.
ANOVA Conditions	**Conditions for ANOVA** 1. ***Independence:*** ✓ **within groups:** sampled observations must be independent ✓ **between groups:** the groups must be independent of each other (non-paired) 2. ***Approximate normality:*** distributions should be nearly normal within each group 3. ***Equal variance:*** groups should have roughly equal variability Where within group Independence is true if - Random sample - Each n < 10% of population Where between group independence is true if - carefully considered
Modified Significance Level/ for Multiple Comparisons	Multiple Comparisons can lead to higher rates of Type 1 error. Therefore we can adjust the significance level by **Bonferroni correction**. **Bonferroni correction:** $\alpha^\star = \alpha/K \quad K$: number of comparisons, $K = \frac{k(k-1)}{2}$ Pairwise Comparisons (two items). - To have constant variance SE, and df so we can compare

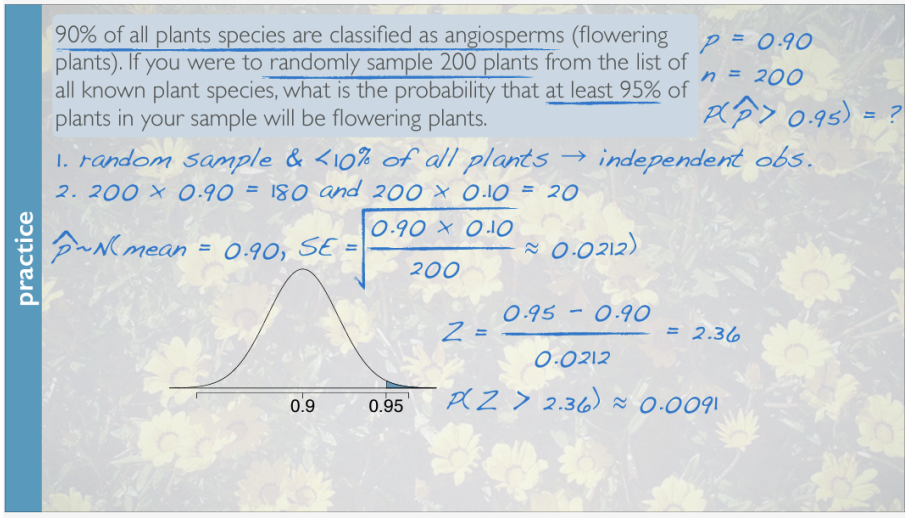
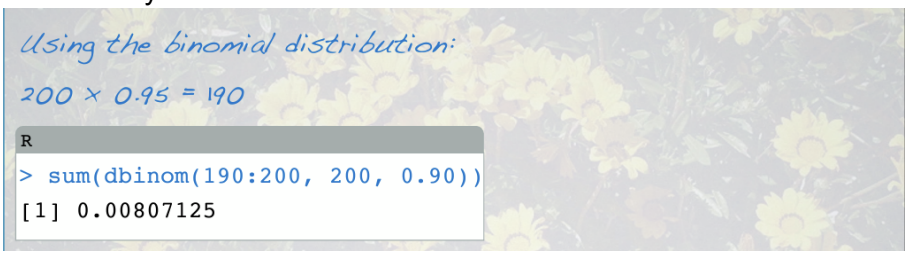
	the p-values from each test to the modified significance level **Standard error for multiple pairwise comparisons:** $SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$ indep. groups test: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ **Degrees of freedom for multiple pairwise comparisons:** $df = df_E$ $df = min(n_1 - 1, n_2 - 1)$

Example of Multiple Comparison	Is there a difference between the average vocabulary scores between middle and lower class Americans?
	
	|
		Df	Sum Sq	Mean Sq	F value	Pr(> F) |
						|---|---|---|---|---|---|
						| class | 3 | 236.56 | 78.855 | 21.735 | <0.0001 |
						| Residuals | 791 | 2869.80 | 3.628 | | |
						| Total | 794 | 3106.36 | | | |
						
						H_0: $\mu_{middle} - \mu_{lower} = 0$
						H_A: $\mu_{middle} - \mu_{lower} \neq 0$
						
						|
							n	mean |
								|---|---|---|
								| lower class | 41 | 5.07 |
								| middle class | 331 | 6.76 |
								
								$T = \frac{(\bar{X}_{middle} - \bar{X}_{lower}) - 0}{\sqrt{\frac{MSE}{n_{middle}} + \frac{MSE}{n_{lower}}}} = \frac{(6.76 - 5.07)}{\sqrt{\frac{3.628}{331} + \frac{3.628}{41}}} = \frac{1.69}{0.315} = 5.365$
								$df = 791$
								
								Step 1: Formulate the hypothesis
								H0 = no difference between lower and middle class
								HA = there is a difference between lower and middle class
								
								Step 2: Create constant SE, Df, and T for comparison
								- Find the standard error for multiple pairwise comparison
								- Find the pointed estimate for difference in mean
								- Find the degrees of freedom equal to the “error”
								Step 3 :Find the p-value for multiple comparison
								- Using R or locate on a t-table
								
								$T = 5.365$
								$df = 791$
								
								 
								
								R
								`> 2 * pt(5.365, df = 791, lower.tail = FALSE)`
								`[1] 1.063895e-07`
								
								$\alpha^* = 0.0083$
								p-value $< \alpha^*$ → Reject H_0

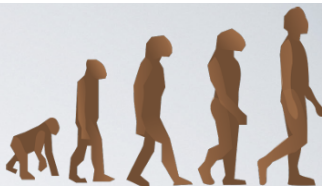
Bootstrapping	Simulation based inference. The “impossible task” is estimating a population parameter using data from only the given sample.

	Bootstrap Process: - Step 1: Take a bootstrap sample (random sample taken with replacement from the original sample) - Step 2: calculate the bootstrap statistics - Step 3: repeat 1 and 2 many times to create a distribution 
Bootstrap, Percentile Method	The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the percentile method.  - Take the sample and find the 90th and 10th percentile which is the 5th and 95th number
Bootstrap, SE method	The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the standard error method. sample median ± t^* SE_{boot} = = 887 ± 1.66 × 89.5758 ≈ (738, 1036) 

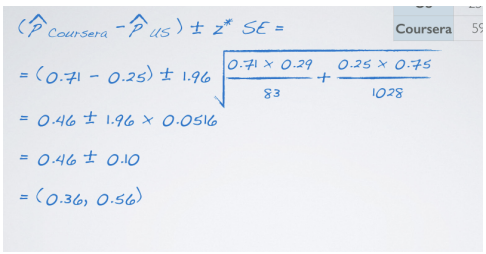
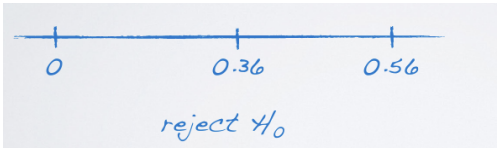
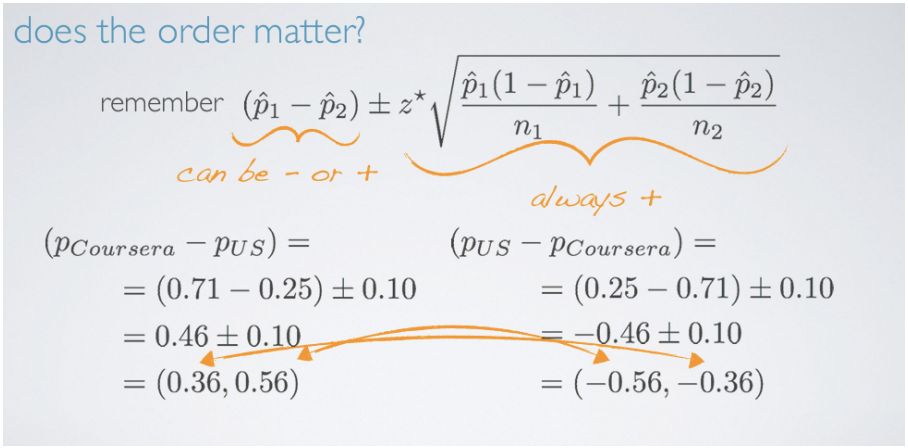
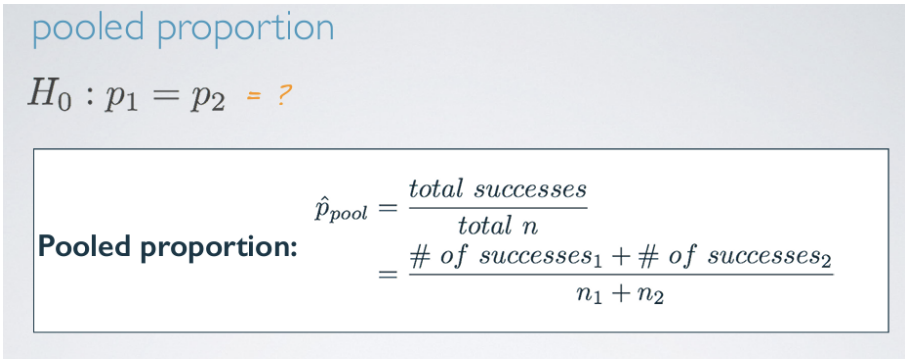
Inference for Proportions (Categorical Variable)	
Sampling Variability and CLT for Proportions. Categorical Variable	When you have different groups in the data and want to take a proportional approach. Step 1: categorize the sample distribution intro groups (ex: by country) Step 2: take the sampling distribution of each (ex: average of smokers per country)  Step 3: Central Limit Theorem for Proportions. SE calculation changes but conditions remain similar to before.  Step 4: Find the z score after obtaining the mean and SE Step 5: Find the p-value (probability)

CLT for Proportions Example: Categorical Variable	 Step 1: identify the population data p=0.9 n=200 Step 2: find SE Step 3: Find z-score Step 4: Find probability Alternatively can use binomial distribution 
Confidence Interval for Proportions: Categorical Variable	General idea remains the same, just the SE calculation has now changed **Standard error for a proportion, for calculating a confidence interval:** $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Confidence Interval for Proportions Example: Categorical Variable	The GSS found that 571 out of 670 (~85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

	Step 1: Identify the conditions 1. independence: 670 < 10% of Americans, and GSS samples randomly Whether one American in the sample has good intuition about experimental design is independent of another. 2. sample size / skew: 571 successes, 670 − 571 = 99 failures Since the success-failure condition is met, we can assume that the sampling distribution of the proportion is nearly normal. Step 2: Conditions are met, calculate $\hat{p} \pm z^* SE = 0.85 \pm 1.96\sqrt{\frac{0.85 \times 0.15}{670}}$ $= 0.85 \pm 1.96 \times 0.0138$ $= 0.85 \pm 0.027$ $= (0.823, 0.877)$ Step 3: Interpret We are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design.
Confidence Interval for Proportions Example (find sample size)	The margin of error for the previous confidence interval was 2.7%. If, for a new confidence interval based on a new sample, we wanted to reduce the margin of error to 1% while keeping the confidence level the same, at least how many respondents should we sample? $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $ME = 0.01 = 1.96\sqrt{\frac{0.85 \times 0.15}{n}}$ $0.01^2 = \frac{1.96^2 \times 0.85 \times 0.15}{n}$ $n = \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} = 4898.04 \rightarrow$ at least 4899 If you don't know the p(1-p), go with 50-50 for a conservative estimate

Hypothesis Test for a Single Categorical Variable Proportion	**Hypothesis testing for a single proportion:** 1. Set the hypotheses: $H_0 : p = null\ value$ $H_A : p <\ or\ >\ or\ \neq\ null\ value$ 2. Calculate the point estimate: $\hat{p}$ 3. Check conditions: 1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population) 2. ***Sample size/skew:*** $np \geq 10$ and $n(1-p) \geq 10$ 4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$ 5. Make a decision, and interpret it in context of the research question: ▸ If p-value < α, reject H_0; the data provide convincing evidence for H_A. ▸ If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.
Hypothesis Test for a Single Categorical Variable Proportion Example:	A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?  Step 1: Define the hypothesis - $H_0: p = 0.5$ $H_A: p > 0.5$ $\hat{p} = 0.6$ $n = 1983$ Step 2: Check the Conditions - 1. Independence: 1983 < 10% of Americans & random sample Whether one American in the sample believes in evolution is independent of another. 2. Sample size / skew: $1983 \times 0.5 = 991.5 > 10$ S-F condition met → nearly normal sampling distribution Step 3: Calculate CLT for a Proportion - $\hat{p} \sim N(mean = 0.5, SE = \sqrt{\frac{0.5 \times 0.5}{1983}} \approx 0.0112)$ - P-hat is the proportion - CLT states the mean of p-hat is centered at the population mean Step 4: Find the test statistic (aka z-score) - $Z = \frac{0.6 - 0.5}{0.0112} \approx 8.92$ Step 5: Find the probability (p-value) - p-value = $P(Z > 8.92)$ = almost 0 → reject H_0
Estimating the Difference between	This is the same as we are doing before, only this time, just know that we are taking the difference between the proportions of sample

Two Proportions (Confidence Interval)	data vs the full data. The calculations will be on p-hat. **parameter of interest**: Difference between the proportions of **all** Coursera students and **all** Americans who believe there should be a ban on possession of handguns. $p_{Coursera} - p_{US}$ **point estimate**: Difference between the proportions of **sampled** Coursera students and **sampled** Americans who believe there should be a ban on possession of handguns. $\hat{p}_{Coursera} - \hat{p}_{US}$ Note that we are now calculating SE differently again. point estimate ± margin of error $(\hat{p}_1 - \hat{p}_2) \pm z^\star SE_{(\hat{p}_1 - \hat{p}_2)}$ **Standard error for difference between two proportions, for calculating a confidence interval:** $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ Conditions remain the same as before. **Conditions for inference for comparing two independent proportions:** 1. ***Independence:*** ✓ **within groups:** sampled observations must be independent within each group ‣ random sample/assignment ‣ if sampling without replacement, n < 10% of population ✓ **between groups:** the two groups must be independent of each other (non-paired) 2. ***Sample size/skew:*** Each sample should meet the success-failure condition: ✓ $n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$ ✓ $n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$
Estimating the Difference between Two Proportions Example: (Confidence Interval)	Using a 95% confidence interval, estimate how Coursera students and the American public at large compare with respect to their views on laws banning possession of handguns. |	suc.	n	$\hat{p}$ | | US | 257 | 1028 | 0.25 | | Coursera | 59 | 83 | 0.71 | Step 1: Identify the hypothesis - $H_0: p_{Coursera} - p_{US} = 0$ - HA: Pcousera - Pus != 0 Step 2: Check the conditions 1. Independent - random sample and 10% met for both 2. Sample size fits the # success/#failure minimum requirements Step 3: Calculate the Confidence Interval using the formula

	 - Step 4: Interpret the results  - - Difference does not equal 0, reject null hypothesis - So there is significant level of difference NOTE: Order doesn't matter 
Hypothesis Test for Two Proportions	What is different is.. - we use a pooled proportion for the SE calculation - Observed value for the z-score calculation is the difference of p-hat between the two proportions 

	observed / confidence interval vs. expected / hypothesis test success-failure condition: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, $n_2(1-\hat{p}_2) \geq 10$ (observed); $n_1\hat{p}_{pool} \geq 10$, $n_1(1-\hat{p}_{pool}) \geq 10$, $n_2\hat{p}_{pool} \geq 10$, $n_2(1-\hat{p}_{pool}) \geq 10$ (expected) standard error: $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ (observed); $SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$ (expected)
Hypothesis Test for Two Proportions Example:	Conduct a hypothesis test, at 5% significance level, evaluating if males and females are equally likely to answer "Yes" to the question about whether any of their children have ever been the victim of bullying. Step 1: Identify the hypothesis Step 2: Find the P-hat pool Step 3: CLT to frame the distribution. - Where is the distribution centered - What is the SE (like SD) - What is the point estimate we are trying to calculate the proportion of?

	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

$H_0: p_{male} - p_{female} = 0$ $H_A: p_{male} - p_{female} \neq 0$

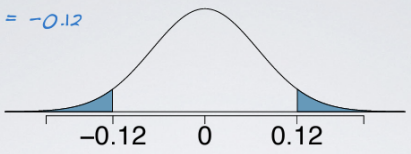
Calculate the estimated pooled proportion of males and females who said that at least one of their children has been a victim of bullying.

$\hat{p}_{pool} = \frac{34 + 61}{90 + 122} \approx 0.45$

	Male	Female
Yes	34	61
No	52	61
Not sure	4	0
Total	90	122
$\hat{p}$	0.38	0.50

$(\hat{p}_{male} - \hat{p}_{female}) \sim N(mean = 0, SE = \sqrt{\frac{0.45 \times 0.55}{90} + \frac{0.45 \times 0.55}{122}} \approx 0.0691)$

point estimate $= \hat{p}_{male} - \hat{p}_{female} = 0.38 - 0.50 = -0.12$

	- point estimate = -0.12 null value = 0 SE = 0.0691  Step 4: Find the test statistic - $Z = \frac{-0.12 - 0}{0.0691} \approx -1.74$ Step 5: Find the p-value - $p\text{-value} = P(\lvert Z \rvert > 1.74) \approx 0.08$

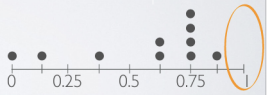
Simulation Based Inference for Proportions	
Inference via Simulation	- Devise a simulation scheme that assumes the null hypothesis is true and repeat the simulation many times and record the relevant sample statistic. - Then calculate the proportion of the simulations that yield a result favorable to the alternative hypothesis. - Ultimate goal is to p-value
Simulation Example:	Paul the Octopus predicted 8 World Cup games, and predicted them all correctly. Does this provide convincing evidence that Paul actually has psychic powers, i.e. that he does better than just randomly guessing? $H_0: p = 0.5$ $H_A: p > 0.5$ $n = 8$ $\hat{p} = 1$ Each guess he has a 50/50 chance of guessing it right/wrong. We can simulate using a coin toss with head=correct. tail=wrong. There are 8 guesses in one simulation $\hat{p}_{sim}$ Then we can repeat the simulation several times. Then we can calculate the percentage of simulations where the proportion of heads is at least as extreme as the observed proportion $\hat{p}_{sim,1}, \hat{p}_{sim,2}, \cdots, \hat{p}_{sim,N}$ Our null hypothesis is that the true rate of success is 50%.

Alternative hypothesis states that over 50% of simulation yields heads.

$$H_0 : p = 0.5$$
$$H_A : p > 0.5$$

Our p-value is the proportion of times he got over 100% (p-hat) correct given that 50% chance of success is true. In our simulation result, there is nothing for 1+ so p-value is 0.

$$\text{p-value} = P(\hat{p}_{sim} \geq 1 \mid p = 0.5) = 0$$



Can also do this in R

- Paul = the data set that we are working with where we define the number of success and number of failures
- Inference () = performs the hypothesis testing

```
R
> source("http://bit.ly/dasi_inference")
> paul = factor(c(rep("yes", 8), rep("no", 0)), levels = c("yes","no"))
> inference(paul, est = "proportion", type = "ht", method = "simulation",
success = "yes", null = 0.5, alternative = "greater")
Single proportion -- success: yes
Summary statistics: p_hat = 1 ;  n = 8
H0: p = 0.5
HA: p > 0.5
p-value =  0.0037
```