MA206, Lesson 14 - Two Groups - Two Proportions

What types of variables do we compare when testing two proportions?

What is the parameter of interest when comparing two proportions?

What is the statistic of interest used to infer about the parameter?

What do we use to compare two proportions?

What is the null hypothesis if we assume no association between two groups?

How do we find the standardized statistic for 2 proportions?
z =

How do we calculate the confidence interval for the difference in two proportions?

1) Suppose an observational study between two groups yielded the results in the chart below. The researcher wishes to know if people from Group A are more likely to succeed than the people in Group B.

	Group A	**Group B**	**Total**
Success	12	20	32
Failures	8	20	28
Total	20	40	60

a) Using symbols and words, what is the null and alternate hypothesis?

b) List the values for $\hat{p}_1$, $\hat{p}_2$, and the difference of proportions.

c) Can we infer that there is a cause and effect relationship between assigned group and performance?

d) Through simulation, report the standard deviation and p-value. Comment on the strength of evidence.

e) Using the SD found above, calculate the standardized statistic. Interpret the results.

f) Using the SD found above to estimate SE, estimate the 95% confidence interval using a multiplier = 2.

2) An area of research that generates a lot of media coverage is examining how parents' behavior may be associated with the sex of their children. One 2002 study conducted by Fukuda et al wanted to see if there was any difference in the proportion of babies born as boys to parents who both smoked when compared to parents who do not smoke. They gathered a random sample of birth information from the local hospitals. Their results are in the file **smokers.csv** where an entry of "Smoker" means both parents smoke and "Non-Smoker" means neither parent smokes.

a) Identify the observational units in this study.

b) List the applicable variables and classify them as categorical or quantitative.

c) Which variable would you regard as explanatory and which is response? Draw the causal diagram.

d) Generate a Segmented Bar Graph and a Mosaic Plot for the study results. Comment on what the Figures are telling you.

e) In words and symbols, what is the null and alternate hypotheses?

f) Using the data given, fill in the table below.

	Non-Smoker	**Smoker**	**Total**
Boys			
Girls			
Total			

g) In words and symbols, what are our observed proportions? What is our difference of proportions?

h) Using the applet, report the simulated difference in proportions and the standard deviation.

i) Report the p-value given by the simulation and the research hypotheses. Interpret the result.

j) Using the Standard Deviation from your simulation, calculate the standardized statistic. What would we say about this strength of evidence?

k) Using the Standard Deviation from your simulation above, calculate the 95% Confidence Interval using $M = qnorm(1 - \frac{\alpha}{2})$ and estimating SE with SE = SD.

l) Based on these results, have we proven that smoking causes boys to be born at a lower rate?

3) Simpson's Paradox: The following two-way table classifies hypothetical hospital patients with a certain disease, according to the hospital that treated them and whether they survived or died:

	Hospital A	**Hospital B**
Survived	800	900
Died	200	100
Total	1,000	1,000

a) Calculate the proportion of Hospital A's patients who survived and the proportion of Hospital B's patients who survived. Which hospital saved the higher proportion of patients?

Suppose we further classify each patient according to a third variable: whether they were in fair condition or poor condition prior to treatment. We obtain the following two-way tables:

Fair condition:

	Hospital A	**Hospital B**
Survived	590	870
Died	10	30
Total	600	900

Poor condition:

	Hospital A	**Hospital B**
Survived	210	30
Died	190	70
Total	400	100

b) Calculate the proportions for Hospital A's and Hospital B's patients that survived given each condition (Fair and Poor).

c) Would you consider Condition to be a confounding variable? Explain.

d) Which hospital would you rather go to? Why?

This change in association is borne out of the skewness of categories (Note how 90% of Hospital B patients are in Fair Condition) is known as **Simpson's Paradox**. Specifically, Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations. There are many instances of the paradox, including in

epidemiology and in studies of discrimination, where understanding the paradox is essential for drawing the correct conclusions from the data.[1]

[1]Stanford Encyclopedia, https://plato.stanford.edu/entries/paradox-simpson/