MA206, Lesson 14 - Two Groups - Two Proportions

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

What is the parameter of interest when comparing two proportions?
The long-term difference between two variables (a process); The true difference between two groups (populations).
$\pi_1 - \pi_2$

What is the statistic of interest used to infer about the parameter?
The difference between the two observed statistics for each of the two groups.
$\hat{p}_1 - \hat{p}_2$

What do we use to compare two proportions?
The difference of the proportions, $\hat{p}_1 - \hat{p}_2$
Alternatively, there is also the Relative Risk $\frac{\hat{p}_1}{\hat{p}_2}$

What is the null hypothesis if we assume no association between two groups?
$H_0 : \pi_1 - \pi_2 = 0$

How do we find the standardized statistic for 2 proportions?
z = $\frac{observed - null}{SD(null)} = \frac{(\hat{p}_1 - \hat{p}_2) - (\pi_1 - \pi_2)}{SD(null)}$
So, for $H_0 : \pi_1 - \pi_2 = 0$, we have that $z = \frac{(\hat{p}_1 - \hat{p}_2)}{SD(null)}$

How do we calculate the confidence interval for the difference in two proportions?
observed $\pm$ M $\times$ SE = $(\hat{p}_1 - \hat{p}_2) \pm M \times SE$

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?

$H_0 : \pi_A - \pi_B = 0$. The true proportion of success for both groups A and B are the same.
$H_a : \pi_A - \pi_B > 0$. The true proportion of success for Group A is larger than Group B.

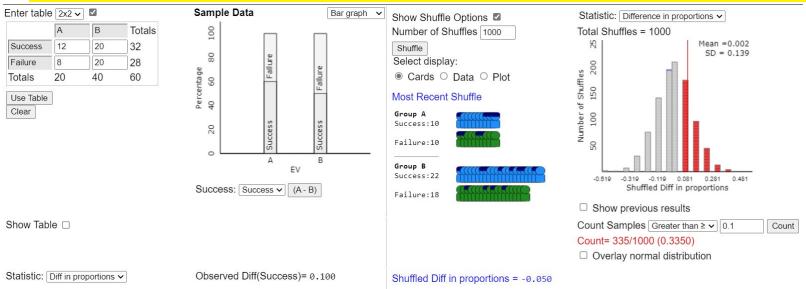
b) List the values for $\hat{p}_1$, $\hat{p}_2$, and the difference of proportions.

$\hat{p}_1(A) = \frac{12}{20} = 0.60$.
$\hat{p}_2(B) = \frac{20}{40} = 0.50$.
$\hat{p}_1 - \hat{p}_2 = 0.6 - 0.5 = 0.10$.

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

No, as this was an observational study and not a randomized experiment, we have not controlled for possible confounding variables.

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



Answers may vary some. SD = 0.139, p-value = 0.335. There is weak evidence against the null hypothesis that the two groups have the same proportion of success in the long run.

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

Answers may vary some. $z = \frac{0.1-0}{0.139} = 0.7194$. Our results are only 0.72 Standard Deviations away from the hypothesized mean value, which is weak to no evidence against the null hypothesis.

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

Answers may vary some. $CI = 0.1 \pm 2 * 0.139 = (-0.178, 0.378)$. Note that this interval has both negative and positive values and that our null hypothesis, 0, is within the interval. We would conclude that 0 is a feasible value, which is consistent with our previous findings.

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.

The observational units are each of the 4,167 births from the local hospitals.

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

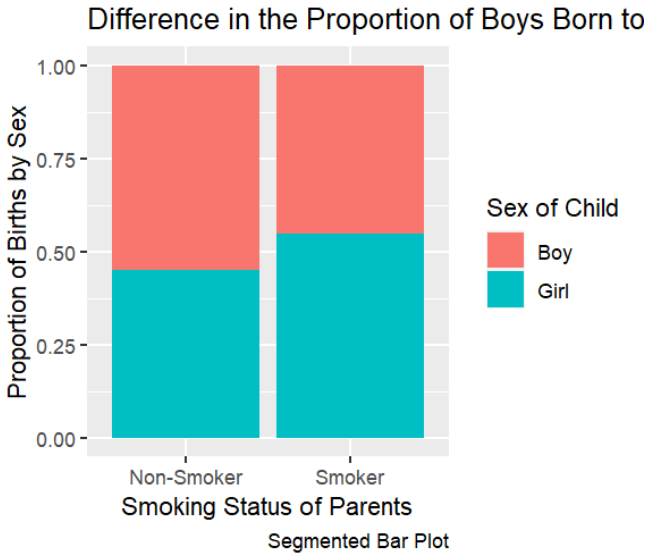
Whether both parents smoke or not - Categorical.
If the baby was a boy or girl - Categorical.

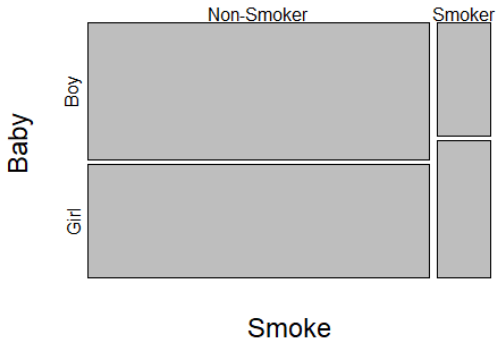
c) Which variable would you regard as explanatory and which is response? Draw the causal diagram.

If both parents smoked the explanatory variable and sex of the baby is the response.

d) Generate a Segmented Bar Graph and a Mosaic Plot for the study results.

Comment on what the Figures are telling you.





The figures tell us that there are more families where both parents do not smoke than where both parents do. Additionally, parents who both smoke have a slightly lower proportion of children that are boys in our sample.

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

$H_0 : \pi_1 - \pi_2 = 0$. The null hypothesis is that there is no difference between the proportion of babies born as boys to parents who both smoke and the proportion of babies born as boys to parents who do not both smoke.

$H_a : \pi_1 - \pi_2 \neq 0$. The alternate hypothesis is that there is a difference between the proportion of babies born as boys to parents who both smoke and the proportion of babies born as boys to parents who do not both smoke.

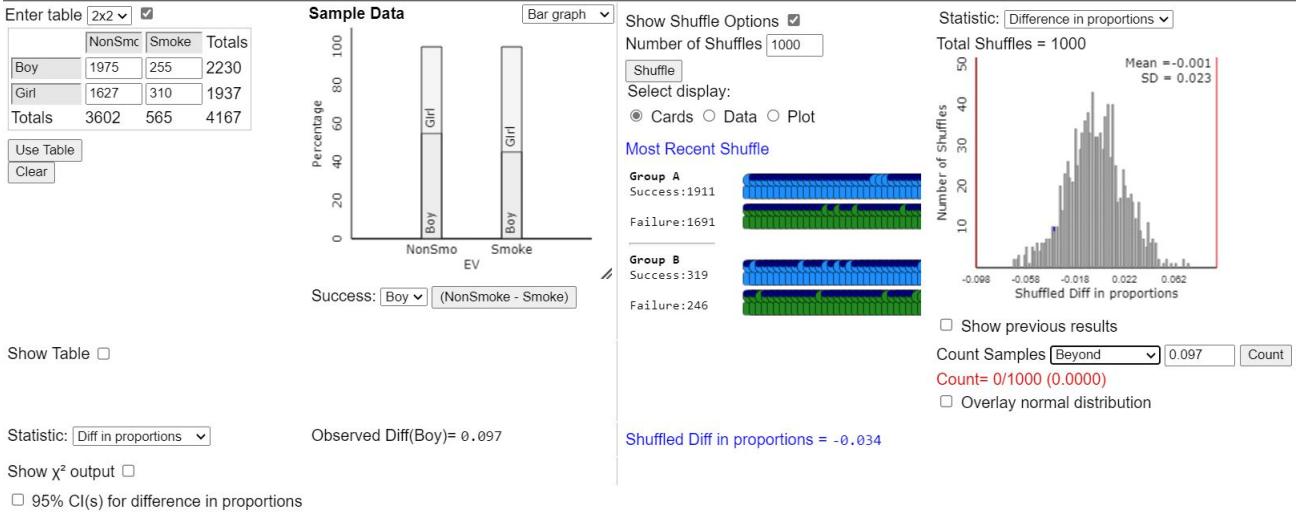
f) Using the data given, fill in the table below.

	Non-Smoker	**Smoker**	**Total**
Boys	1975	255	2230
Girls	1627	310	1937
Total	3602	565	4167

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

$\hat{p}_1 = \frac{1975}{3602} = 0.5483$. The sample proportion of boys born to non-smoking parents is 0.5483.
$\hat{p}_2 = \frac{255}{565} = 0.4513$. The sample proportion of boys born to smoking parents is 0.4513.
The difference in proportions is $0.5483 - 0.4513 = 0.09698$

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



Answers may vary, but should be around mean of 0, SD of 0.023.

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

Answers may vary, but should be very small = after 200000 simulations, had 0 as extreme. The probability of observing a difference of proportions at least as extreme as our observed result, assuming the null hypothesis is true, is computationally 0. This is very strong evidence against the null hypothesis.

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

$z = \frac{0.5483-0.4513-(0)}{0.023} = 4.216$. That our observed value is more than 4 standard deviations away from the null, we would say we have very strong evidence against the null hypothesis that there is no difference between birth rates of boys for parents who smoke compared to parents who do not smoke.

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.

$CI = observed \pm M \times SE = 0.09698 \pm 1.95996 \times 0.023 = (0.0519, 0.1421)$.

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

No, we haven't proven anything. As this was an observational study and not an experiment, we cannot infer causation. We have simply shown that there is very strong evidence that the differences in proportions of baby boy births between parents who smoke and parents who do not smoke did not occur by chance alone.

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?

$\hat{p}_A = \frac{800}{1000} = 0.8$
$\hat{p}_B = \frac{900}{1000} = 0.9$
Hospital B saved a 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).

Fair: $\hat{p}_A = \frac{590}{600} = 0.983$
$\hat{p}_B = \frac{870}{900} = 0.967$
Hospital A saved a higher proportion of fair condition patients.

Fair: $\hat{p}_A = \frac{210}{400} = 0.525$
$\hat{p}_B = \frac{30}{100} = 0.3$
Hospital A saved a higher proportion of poor condition patients.

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

Yes, because whether a patient had a fair or poor condition changes the behavior of both which hospital they went to and if they survived or not. Therefore, Condition is associated with both our explanatory variable and our response variable, so it is a confounding variable.

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

I would rather go to Hospital A, as it has a higher survivability rate when controlled for condition of its patients for both Fair and Poor conditions.

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/