MA206, Lesson 15 - Two Groups - Two Proportions

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

Review: What is the parameter of interest when comparing two proportions?
The difference between those two groups' long run proportions or true difference in population proportions.
$\pi_1 - \pi_2$

The true association between two categorical variables.

Review: 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 is $\hat{p}_{Total}$?
The proportion for the sample without regard for different groups.

$\frac{Count\ of\ Successes_1\ +\ Count\ of\ Successes_2}{n_1+n_2}$

How do we find the standardized statistic for 2 proportions using theory?

z $= \frac{observed\ -\ null}{SD(null)} = \frac{(\hat{p}_1-\hat{p}_2)-(\pi_1-\pi_2)}{SD(null)}$

$SD(null) = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$

So, z $= \frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$

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

observed $\pm$ M $\times$ SE $= (\hat{p}_1 - \hat{p}_2) \pm M \times SE$

$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

So, Confidence Interval $= (\hat{p}_1 - \hat{p}_2) \pm qnorm(1 - \frac{\alpha}{2}) \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veterans Administration Hospital in Northampton, Massachusetts. She became one of the nurses that others looked up to as an example of skill as she was particularly good in a crisis. When a victim went into cardiac arrest, for example, she was often the first to notice that something was wrong and calmly apply adrenaline to restart the patient's heart. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. An investigation was launched of 1,641 eight-hour shifts during her time in the ICU to see if there was indeed an increase in death rate while she was on shift. For each shift, two variables were recorded: Did Gilbert work on the shift, and did at least one patient die during this shift.

a) Identify the observational units in this study.

The observational units are each of the 1,641 eight-hour hospital shifts

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

Whether Gilbert worked on shift or not - Categorical.
If at least one patient died during the shift - Categorical.

c) Which variable would you classify as explanatory and which is response? Draw the Causal Diagram.

If Gilbert worked is the explanatory variable.
If at least one person died is the response variable.
$Gilbert\ Worked \rightarrow Patient\ Died$

d) In words and symbols, write the null and alternate hypotheses.

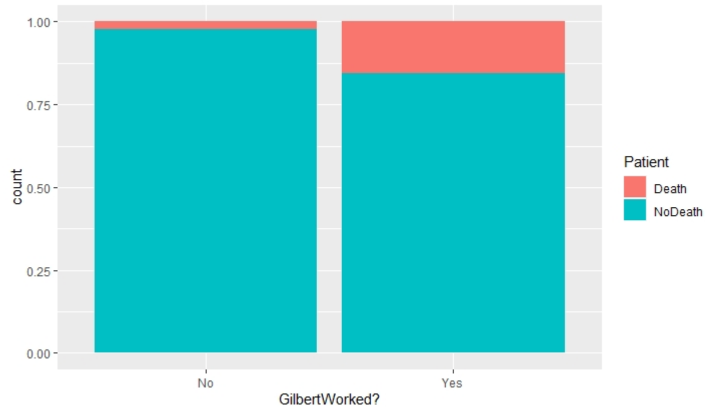
$H_0 : \pi_1 - \pi_2 = 0$. There is no difference in the long run proportion of shifts which had deaths where Gilbert worked and shifts which had deaths where Gilbert did not work.
That is, there is no association between Gilbert working on shift and the proportion of shifts with at least one death.

$H_a : \pi_1 - \pi_2 > 0$. The long-run difference in proportions between shifts with at least one death that Gilbert worked and shifts with at least one death that Gilbert did not work is greater than zero.
That is, there is a positive association between Gilbert working on shift and there being at least one death on the shift.

The data from this study is available and can be imported using R by the code below. Copy and run this code to begin analysis and provide evidence for or against Gilbert's case at trial.

```
Nurse <- read_delim("http://www.isi-stats.com/isi/data/chap5/Gilbert.txt", delim="\t")
```

e) Generate a Segmented Bar Graph and a Mosaic Plot for the Gilbert data. Interpret the results.



The proportion of shifts with patient deaths seems to be much higher for shifts that Gilbert worked than it does for days that Gilbert did not work, as evidenced by the sharp contrast in where the fill changes.

.

We see that Gilbert worked less than 1/4 of the total shifts, evidenced by the vertical divide. The proportion of deaths on her shift, however, are much higher than the proportion of deaths on shifts she did not work.

f) Do we meet the validity conditions required to use theoretical methods?

Yes, we have at least 10 observations in each category. (34, 34, 1350, and 217).

	Worked	**Did Not Work**	**Total**
Death	40	34	74
No Death	217	1350	1567
Total	257	1384	1641

h) What is our statistic (difference of proportions)?

The difference in proportions is $0.1556 - 0.0246 = 0.1310$

i) Calculate the standard deviation and standard error.

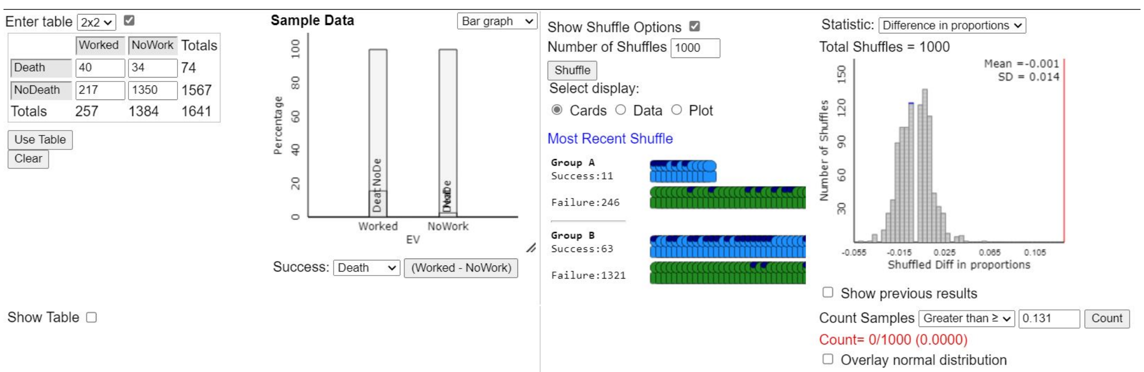
$$SD = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})} = \sqrt{0.045*(1-0.045)*(\frac{1}{257}+\frac{1}{1384})} = 0.01409$$
$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.156(1-0.156)}{257}+\frac{0.025(1-0.025)}{1384}} = 0.02299$$

j) Report your standardized statistic and p-value. Interpret the results.

$$\text{z} = \frac{observed - null}{SD(null)} = \frac{0.1310}{0.01409} = 9.2995.$$

For a greater than test, p-value = 1-pnorm(z) = 0
With a p-value of essentially 0, there is very strong evidence that the difference in death rates between shifts Gilbert worked and shifts Gilbert didn't work did not occur by random chance alone. Alternatively worded, we have very strong evidence against the null that there is no association between the two.

k) Using the applet, visualize this scenario. Do these results line up with the theoretical analysis?



With a SD(Null) of 0.014 and a p-value of computationally 0, there is very strong evidence that the difference in rates was not by chance alone. This is in line with theoretical computations.

l) Calculate the 90% confidence interval and interpret your results.

$$(\hat{p}_1-\hat{p}_2) \pm qnorm(1-\frac{\alpha}{2}) \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.1310 \pm qnorm(1-\frac{0.1}{2}) \times 0.02299 = (0.0933, 0.1689)$$

We are 90% confident that the true difference in proportions of shifts with at least one death between shifts Gilbert worked and shifts she did not work is between 0.0933 and 0.1689. As this interval is strictly positive, our evidence indicates there is a positive association.

j) Based on these results, do we have reasonable suspicion that the differences in death rates is not due to random chance?
Yes, we can surmise that the difference in rates is likely not due to chance alone. However, that does not give insight into what the actual reason might be.

k) Based on these results, have we proven that Kristen Gilbert caused the deaths of her patients?
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 between shifts Gilbert did work and shifts she did not work did no occur by random chance. That does not mean we can infer as to what that reason is.

l) If you were on the prosecution for Kristen Gilbert, what might you say to sway the jury towards a guilty verdict? You may refer to your calculations and Figures.

You can say that it is highly unlikely to have happened by random chance, pointing to an association. You must be careful not to overstretch and state that with this many excess deaths, then the chance is computationally zero that Gilbert is innocent. This is known as the **Prosecutor's Fallacy** and cannot be inferred from our analysis.

m) If you were on the defense for Kristen Gilbert, what might you say to sway a jury against this result? You may refer to your calculations and Figures.

You might point to potential confounding variables which explain the association and that, as it is an observational study, it cannot infer cause and effect. For example, perhaps Kristen only worked holiday weekends and holiday weekends tend to have a higher death rate than other days. There could be another unknown variable explaining the association.

Gilbert was eventually arrested and charged with these murders. Part of the evidence presented against Gilbert to bring her to trial was a statistical analysis of these eight-hour shifts during the time Gilbert worked in the ICU. She is currently serving four consecutive life sentences, plus 20 years, without the possibility of parole for the death of 4 hospitalized veterans, though coworkers suspect she was responsible for 80 or more deaths and more than 300 medical emergencies. She tried to appeal her case, but withdrew after the supreme court ruled that it would have allowed prosecutors to pursue the death penalty in the event of a retrial. You can read more about it in the "Murder Nurse What Happened" document on Teams.

2) Use your own project dataset to explore the relationship between two categorical variables. If your categorical variables have more than one category, you may wish to transform them into binary variables such that one category is treated as a "success" and the other is treated as a "failure." Alternatively, you can manually do the math from your resulting tables.

If your dataset does not have two different categorical variables, you may convert your response variable to a categorical variable to look for an association. For example code, see below and adjust for your variables. The code below takes a numeric response (Score) and converts them to categories based on value (A, B, C, D, or F).

```
Data2 <- Data1 %>%
  mutate(ResponseCat = case_when(
    Score <= 65 ~ "F",
    Score < 70 ~ "D",
    Score < 80 ~ "C",
    Score < 90 ~ "B",
    TRUE ~ "A"
  )
)
```

a) Provide a brief description of your dataset. What are the observational units?

b) Describe the two categorical variables you are comparing. What is your null and alternate hypothesis about their association?

c) Generate a two-way table of your data. Do you meet the validity conditions to calculate using theory?

d) Generate a segmented bar chart. Describe the shape and its implications.

e) Generate a mosaic plot. Describe the shape and its implications. Does this provide more information than the segmented bar chart previously?

f) Using your data and hypothesis above, calculate your observed statistic, standard deviation, standardized statistic, and p-value.

g) Calculate a 99% confidence interval. Interpret your results.