MA256 Lesson 9 - Two Proportions (5.1-5.3)

U.S. Crimes

In 1999, there were 11.61M total crimes in the US compared to 11.88M in 2002; only 3 years later. What if I claimed the US was more dangerous in 2002 because there were almost 271,000 more crimes committed?

```
crimes <- c(11.61, 11.88)  # 1999, 2002
diff(crimes)
```

```
## [1] 0.27
```

Are we comparing apples to apples? Why or why not?

No! We just learned that we can't just compare the numbers, we need to use proportions so we can compare apples to apples. How can I get a proportion so I can compare apples to apples?

What can we do to compare the two years?

Population probably changed from 1999 to 2002. I should divide the number of crimes by the US population for each respective year. The proportion of crime actually went down from 4.26% in 1999 to 4.13% in 2002 so the country is actually safer than it was!

```
population <- c(272.69, 287.97)
crimes/population
```

```
## [1] 0.04257582 0.04125430
```

Review for Single Proportion:

Hypotheses (in symbols):

$H_0 : \pi = 0$

$H_a : \pi \neq 0$

Strength of Evidence: Calculate z statistic: $z = \frac{\hat{p}-\pi_0}{\sqrt{\pi(1-\pi)/n}}$

reject H_0 for z "more extreme" using the guidlines for appropriate significant level.

Confidence Interval: $\hat{p} \pm z^*_{(1-\alpha/2)} \times \sqrt{(\hat{p}(1-\hat{p})/n)}$

reject H_0 if π_0 is NOT in CI.

Two proportions:

Hypotheses (in symbols):

$H_0 : \pi_{treatment} - \pi_{control} = 0 \quad OR \quad \frac{\pi_{treatment}}{\pi_{control}} = 1$

$H_a : \pi_{treatment} - \pi_{control} \neq 0 \quad OR \quad \frac{\pi_{treatment}}{\pi_{control}} \neq 1$

Strength of Evidence: Calculate z statistic: $z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, where $\hat{p}$ is the total proportion of successes

Confidence Interval: $\hat{p}_2 - \hat{p}_1 \pm z^*_{(1-\alpha/2)} \times \sqrt{(\hat{p}_1(1-\hat{p}_1)/n_1) + (\hat{p}_2(1-\hat{p}_2)/n_2)}$

Validity conditions: Must have at least 10 observations in each category of the response variable (≥ 10 for each cell in a two-way table)

Nurse Gilbert

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, MA. 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.

Gilbert was eventually arrested and charged with these murders. Part of the evidence presented against Gilbert at her murder trial was a statistical analysis of 1,641 eight-hour shifts during the time Gilbert worked in the ICU. For each of these shifts, researchers recorded two variables: whether Gilbert worked on the shift and whether at least one patient died during the shift.

1) Identify the observational units in this study.

The eight-hour shifts ($n = 1641$)

2) What are the variables in this study? Classify them.

Whether Gilbert worked on the shift, categorical (binary)
Whether at least one patient died on the shift, categorical (binary)

3) Which variable would you regard as explanatory and which as response?

Explanatory: Whether Gilbert worked on the shift
Response: Whether at least one patient died on the shift

4) Is this an observational study or an experiment? Explain.

Observational study

5) Read in the table as a data frame. Use the `table()` command to count the number of instances of death. Why doesn't this information provide any clues about whether there is an association between Gilbert working on a shift and a patient dying on a shift? What other information do we need to know?

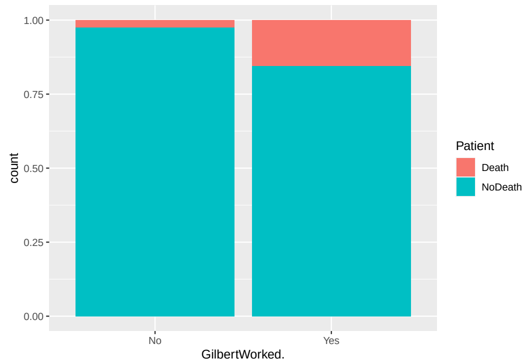
```
library(tidyverse)
library(janitor)
library(ggmosaic)
library(ggthemes)
nurse <- read.table("http://www.isi-stats.com/isi/data/chap5/Gilbert.txt",
                    header=TRUE, stringsAsFactors = TRUE)

table(nurse)
```

```
##                Patient
## GilbertWorked. Death NoDeath
##             No    34    1350
##            Yes    40     217
```

We can see that there were 40 deaths when Nurse Gilbert was on a shift and 34 deaths otherwise. We would need to know the total number of shifts Gilbert worked/didn't work. Then we could compare the proportions.

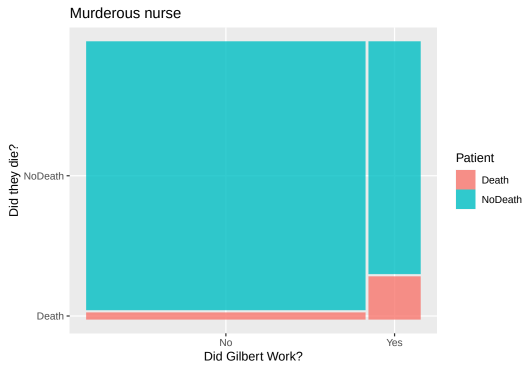
6) Create a stacked barchart to help visualize the difference in proportions. What do you see here?

```
nurse %>%
  ggplot(aes(x=GilbertWorked., fill = Patient)) +
  geom_bar(position = position_fill())
```



7) Create a mosaic plot to help visualize the difference in proportions. Explain what you see.

```
nurse %>%
  ggplot()+
  geom_mosaic(aes(x = product(Patient, GilbertWorked.), fill = Patient)) +
  labs(y="Did they die?", x="Did Gilbert Work?", title = "Murderous nurse")
```



8) Calculate the number Deaths/No Deaths for the shifts where Gilbert worked/didn't work. Calculate the percentage of deaths for shifts where Gilbert worked/didn't work. Hint: Use the MA256 AY23-2 Course guide for help.

```
nurse%>%
  count(GilbertWorked.,Patient) %>%
  pivot_wider(names_from = GilbertWorked., values_from = n) %>%
  adorn_totals(c("row", "col"))
```

```
##  Patient   No Yes Total
##    Death   34  40    74
##  NoDeath 1350 217  1567
##    Total 1384 257  1641
```

9) Calculate the conditional proportion of deaths for shifts where Gilbert worked/didn't work. Do the conditional proportions and graph appear to provide evidence that at least one patient was more likely to die on a Gilbert shift than on a non-Gilbert shift? In other words, does there appear to be a tendency for Gilbert shifts to have at least one death more often than non-Gilbert shifts? Explain.

```
nurse%>%
  count(GilbertWorked.,Patient) %>%
  pivot_wider(names_from = GilbertWorked., values_from = n) %>%
  adorn_percentages(denominator = "col")
```

```
##  Patient         No      Yes
##    Death 0.02456647 0.155642
##  NoDeath 0.97543353 0.844358
```

We can see that the conditional probability for shifts that Gilbert did/didn't work is 15.6% and 2.46%. Yes, there appears to be evidence that patients died more often when Gilbert was working.

10) Without doing any further analysis, do you consider the difference between the conditional proportions to be striking and/or worth reporting to a jury?

11) There are two long-run probabilities (parameters) in this study. What are they?

$\pi_G =$ the long-run probability at least one patient died and Gilbert was working.
$\pi_{NG} =$ the long-run probability at least one patient died and Gilbert was not working.

12) Write the null and alternative hypotheses in terms of the two long-run proportions in words and using appropriate symbols.

Null: The probability at least one patient dies is the same when Gilbert works as when Gilbert does not work; $\pi_G = \pi_{NG}$ OR $\pi_G - \pi_{NG} = 0$
Alt: The probability at least one patient dies is NOT the same when Gilbert works as when Gilbert does not work; $\pi_G \neq \pi_{NG}$ OR $\pi_G - \pi_{NG} \neq 0$

13) What single number (statistic) can we compute to summarize the data? Calculate the statistic and the p-value.

```
 #  Patient   No Yes Total
 #    Death   34  40    74
 # NoDeath 1350 217  1567
 #    Total 1384 257  1641

nG <- 257
nNG <- 1348
n <- 1641
pG <- 40 / nG
pNG <- 34 / nNG
pi <- (40+34) / n

zstat <- (pG - pNG) / sqrt(pi*(1-pi)*(1/nG + 1/nNG))

pval <- 2*(1-pnorm(abs(zstat)))
c(zstat, pval)
```

```
## [1] 9.233676 0.000000
```

14) Do we meet the validity conditions for a two-proportion z-test? If we didn't meet the validity conditions, what would we have to do in order to find the p-value?

At least 10 values in each of the four cells of the two way table, yes we met them. If we didn't meet the validity conditions, would have to simulate.

15) What is the 95% confidence interval? What does this confidence interval mean? . Based on the confidence interval, how likely is random chance that people were dying when Nurse Gilbert was on shift? Why?

```
stat <- (pG - pNG)
alpha <- 0.05 # 1-0.95
multiplier <- qnorm(1 - alpha/2)
SE <- sqrt(pG*(1-pG)/nG + pNG*(1-pNG)/nNG)
c(stat - multiplier*SE, stat + multiplier*SE)
```

```
## [1] 0.0853151 0.1755238
```

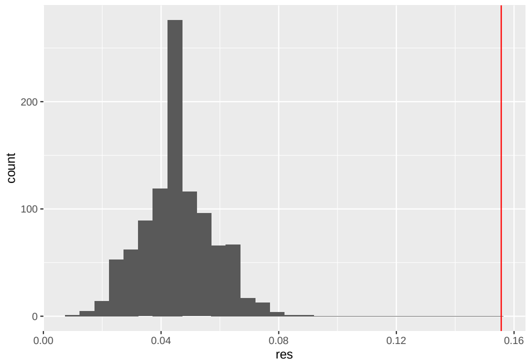
[0.853, 0.1755]
I am 95% confident, that when Nurse Gilbert is on shift, the probability of a patient dying is between 0.0853 and 0.1755 higher than when Gilbert is not on shift.
It is not very likely since the CI does not contain 0.

16) How to simulate: If the null hypothesis is true, what does that mean about the labels 'Death' and 'NoDeath'?

If the null hypothesis is true, the lables 'Death' and 'NoDeath' would be arbitrary.

17) Operating under the null hypothesis, we could simply shuffule the labels 'Death' and 'NoDeath'. Calculate the number of instances where the proportion of deaths where Gilbert is working is greater than the observed proportion of deaths. (Note: our statistic is different than above.)

```
set.seed(256)

myphat <- pG

M <- 1000
nurse2 <- nurse
RES <- data.frame(res = rep(NA, M))

for(rep in 1:M){
  nurse2$pat.shuff <- sample(nurse$Patient)
  Gilworked <- nurse2$pat.shuff[nurse2$GilbertWorked. == "Yes"]
  numerator <- sum(Gilworked == "Death")
  RES$res[rep] <- numerator / length(Gilworked)
}

RES %>% ggplot(aes(x=res)) +
  geom_histogram() +
  geom_vline(xintercept = myphat, color="red")
```




```r
# estimate p-value
sum(RES$res >= myphat) / M
```

```
## [1] 0
```

Nurse Gilbert Follow-up

First, the grand jury elected to send her case to full jury, based at least in part on the statistical testimony.

The judge to rule whether or not the jury should be allowed to hear the statistical evidence. Why is this an issue?

There is a clearly and association between Nurse Gilbert being on the floor and patients dying, but can we make the claim she caused their deaths?

No! Association is not causation! A pvalue doesn't say "Nurse Gilbert is the reason for excess deaths" all it says is "whatever the explanation may be, you can be quite confident that its not mere chance variation". The issue here stems from the difference between an observational study and a randomized experiment!

Additionally, another factor in not using the statistical analysis in the jury trial was the Prosecutor's Fallacy.

"Suppose Gilbert is not guilty, and that the deaths behave in a chancelike way, like a coin toss. Then the probability is less than 1 out of 100 million that you would see so many excess deaths on Gilbert's shifts. (Correct)

It's a quick jump to the following shorter version. "If Gilbert is innocent, then it would be almost impossible to get so many excess deaths (also correct).

And then, "with this many excess deaths, the chance is less than one in 100 million that Gilbert is innocent (NOT VALID).

Because of this the defense claimed that the statistical evidence will be misinterpreted by the jury in a way that favored the prosecution.

This is why we talk about DATA LITERACY!!!!!

She induced cardiac arrest in patients by injecting their intravenous therapy bags with massive doses of epinephrine, an untraceable heart stimulant. She would then respond to the coded emergency, often resuscitating the patients herself. Although it is believed that she may have been responsible for 350 or more deaths

Although other nurses noticed a high number of deaths on Gilbert's watch, they passed it off and jokingly called her "The angel of Death." In 1996, however, three nurses reported their concern about an increase in cardiac arrest deaths and a decrease in the supply of epinephrine, and an investigation ensued. Gilbert telephoned in a bomb threat to attempt to derail the investigation

Gilbert forced an untrained colleague to use cardiac defibrillation paddles on a patient during a medical emergency on Nov. 17, 1995, by refusing to use the equipment herself