MA256 Lesson 11 - Two Proportions (5.1-5.3)

Review for Single Proportion:

Hypotheses (in symbols):

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Two proportions:

Hypotheses (in symbols):

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Validity conditions:

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(XXXXX)
# population <- c(272.69, 287.97)
# XXXX /XXXXX
```

Are we comparing apples to apples? Why or why not? What can we do to compare the two years? Given the following table, calculate 1) the difference in the propostions of success for Group A compared to Group B ($\hat{P}_A - \hat{P}_B$); and 2) the relative risk for success in Group A compared to Group B ($\hat{P}_A / \hat{P}_B$).

	Group A	Group B	Total
Success	10	20	30
Failure	10	20	30
Total	20	40	60

```
# phat.A <- XXXX / XXXX; phat.A
# phat.B <- XXXX / XXXX; phat.B
# XXXXX # difference?
# XXXXX # relative risk?
```

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.

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

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

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

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

# table(XXXXX)
```

6) Create a stacked barchart to help visualize the difference in proportions. What do you see here? Hint: Use the MA256 Course guide for help.

```
# nurse %>%
#
```

7) Create a mosaic plot to help visualize the difference in proportions. Explain what you see. Hint: Use the MA256 Course guide for help.

```
#
```

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 Course guide for help.

```
#
```

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.

```
#   # hint: use the tabyl command + some other stuff
```

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?

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

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

```
 # ## GilbertWorked. Death NoDeath Total
 # ##            No    34    1350  1384
 # ##            Yes    40     217   257
 # ##          Total    74    1567  1641
#
# nG <- XXXX
# nNG <- XXXX
# n <- XXXX
# pG <- XXXX
# pNG <- XXXX
# pi <- XXXX
#
# zstat <- XXXX
#
# pval <- XXXX
# c(zstat, pval)
```

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?

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 <- XXXX
# alpha <- XXXX # 1-0.95
# multiplier <- XXXX
# SE <- sqrt(XXXX)
# c(XXXX)  # CI
```

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

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$pat.shuff <- sample(XXXX)
#   Gilworked <- nurse2$pat.shuff[nurse2$GilbertWorked. == XXXX]
#   numerator <- sum(Gilworked == XXXX)
#   RES$res[rep] <- numerator / length(Gilworked)
# }
#
# RES %>% ggplot(aes(x=res)) +
#   geom_histogram() +
#   geom_vline(xintercept = myphat, color="red")
#
# # estimate p-value
# sum(RES$res >= myphat) / M
```

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