MA206, Lesson 7 - Generalization

What is **generalization**?

How **broadly** the conclusions apply. It is to which, if any, larger group of individuals our results may apply to outside of the assessed sample.

Define:

Population:

Indicated by N

The population is the entire set of observational units of a certain group

Sample:

Indicated by n

Some smaller subset of a larger population

Convenience Sample:

A nonrandom sample of a population which often tends to oversample a certain group within the population while undersampling others.

Biased Sampling Method:

A sampling method which tends to consistently overestimate or underestimate the population parameter of interest.

Simple Random Sample:

A random sample which ensures each observational unit has the same chance of being selected in the sample. Typically done through a Random Number Generator, pulling names from a hat, rolling dice, etc.

Does a larger sample size fix sampling bias?

No. Larger sample size will not fix a biased sampling method.

What are some Nonsampling concerns which might also introduce bias into our data?

Some examples might include the wording of a question, lack of anonymity, extreme phrasing, uncalibrated scales, time of day, intimidation, etc.

1) Suppose you have a massive dessert bowl containing 40% red skittles and 60% green skittles. You take many, many random samples of 25 skittles and each time note the proportion that are red. From this, you create a distribution of all your sample proportions of red skittles.

a) What is the expected mean of your distribution of sample proportions?

0.4, the true proportion of red skittles (40%)

b) What is the expected standard deviation of your distribution?

$\sqrt{\frac{0.4*(1-0.4)}{25}} = 0.0979796$

```
pi <- 0.40
n <- 25

sd <- sqrt((pi * (1 - pi))/n)
sd

## [1] 0.09797959
```

2) Suppose the leadership at Arvin Gym wants to get a sense of how many cadets actually want to reopen the weight room on the third floor. They know that sending a survey out to the entire Corps is destined to fail, and so have come up with four courses of action.

Which course of action below should be used? Justify your answer.

a) Send a survey to the Football Team to gather their opinion.

b) Have the front desk ask everybody who comes into Arvin.

c) Compile a list of names of all the Cadets on Corps Squad teams, randomly select a sample of those names using a random number generator, and survey those cadets.

d) Compile a list of names of all Cadets, Staff, and Faculty at USMA, randomly select a sample of those names using a random number generator, and survey those individuals.

e) Compile a list of names of all Cadets, randomly select a sample of those names using a random number generator, and survey those individuals.

Compile a list of names of all Cadets, randomly select a sample of those names using a random number generator, and survey those individuals. This simple random sampling method ensures all Cadets (the population of interest) are equally likely to be chosen, and thus can be generalized to the Corps of Cadets.

3) As part of the General Social Survey (GSS) in 2018, a random sample of U.S. adults were asked whether they have ever been told by a health professional that they have depression. In the sample of 1,414 people that received this question, 271 of them said that they have been told that they did have depression.

a) Suppose in the population of all U.S. adults, 20% have been told by a health professional that they had depression. What should be the mean and standard deviation be if we were to sample from this population many times?

From CLT, Mean = 0.2 (the true population parameter)

SD = $\sqrt{\frac{0.2\times(1-0.2)}{1414}} = 0.0106374$

```
pi <- 0.20
n <- 1414

sd <- sqrt((pi * (1 - pi))/n)
sd

## [1] 0.0106374
```

b) How many standard deviations below the mean of the distribution described in part (a) is the sample proportion from the GSS?

$\hat{p} = \frac{271}{1414} = 0.19165$

$z = \frac{0.19165-0.2}{0.010637395} = -0.7845$

The sample proportion from GSS is 0.7845 standard deviations below the mean.

```
phat <- 271/n
phat

## [1] 0.1916549

z <- (phat - pi)/sd
z

## [1] -0.7845079
```

c) Based on your answer from part (b), is it very unlikely that a random sample of 1,414 U.S. adults would only find 271 of them that would say that they have been told they had depression? Explain.

No, it is not unlikely. With only -0.7845 standard deviations from the mean, this is weak evidence against the null hypothesis that the true proportion of Americans who have ever been told they've had depression is 20%.

d) Using your work, calculate a p-value for an alternative hypothesis that the true proportion of adults that have been told by a health professional that they had depression is not 20% .

z = -0.7845 gives a p-value of 0.433.

```
pvalue <- 2*(1-pnorm(abs(z)))
pvalue

## [1] 0.4327422
```

4) A survey was conducted on 56 West Point cadets in MAJ McD's AY23-2 MA206 class about their preferences for original Starburst flavors between Pink, Orange, and Yellow. The results are compiled in the Starburst.csv file on Teams. You may use the course guide as a reference to read in this file. We want to validate the claim that, if given a choice between Pink, Orange, and Yellow from the original starburst colors, cadets think that Yellow starburst are the worst.

a) Write the null and alternate hypotheses using symbols and describe the parameter of interest in words.

$H_0 : \pi = \frac{1}{3}$
$H_a : \pi > \frac{1}{3}$
The parameter π is the true long-run proportion of cadets who chose yellow as their least favorite flavor.

```
Results <- read_csv("Starburst.csv")
head(Results)

## # A tibble: 6 x 2
##   Best  Worst
##   <chr> <chr>
## 1 Pink  Pink
## 2 Pink  Pink
## 3 Pink  Pink
## 4 Pink  Pink
## 5 Pink  Orange
## 6 Pink  Orange
```

b) List your observed statistic, sample size, standardized statistic, and p-value (using appropriate methods).

$\hat{p} = \frac{17}{56} = 0.3035714$
$n = 56$
$z = \frac{0.3035714 - 0.333333}{\sqrt{\frac{0.3333 \times (1 - 0.3333)}{56}}} = -0.4724556$
$p - value = 1 - pnorm(z)[greaterthantest] = 0.6816992$

```
Results %>%
  tabyl(Worst) %>%
  adorn_totals()
```

```
n <- 56
phat <- 17/n
pi <- (1/3)

sd <- sqrt((pi * (1-pi))/n)

z <- (phat - pi)/sd

pvalue <- 1-pnorm(z)

phat

## [1] 0.3035714

sd

## [1] 0.06299408
```

```
z
```

```
## [1] -0.4724556
```

```
pvalue
```

```
## [1] 0.6816992
```

c) Interpret the results of your analysis. Ensure you include your calculated p-value.

With a p-value of 0.6816992, we conclude that we have very weak to no evidence against the null hypothesis and therefore cannot support the claim that cadets believe yellow is the worst flavor.

d) Do you feel comfortable generalizing these results to the entire Corps of Cadets?

No, the samples were limited to MAJ McD's AY23-2 MA206 class, which is comprised mostly Yearlings excited about statistics. As not every cadet had an equal chance to be selected, there may be sampling bias and we cannot generalize.

e) To what population would you feel comfortable generalizing these results to?

As this was a convenience sample and not a random sample, I do not feel comfortable generalizing these results to any outside population.

5) According to the National Coffee Drinking Study from the National Coffee Association, 40% of 18- to 24-year-olds in the United States regularly drink coffee every day. Suppose this number is accurate. Researchers are interested in testing if the daily coffee-drinking habits of cadets at West Point differ from the national average. To test this, they sampled 140 cadets in Grant Hall after lunch. Of those, 73 indicated they regularly drank coffee every day.

a) Write the null and alternate hypotheses, in words and symbols.

$H_0 : \pi = 0.4$. The true proportion of West Point cadets who drink coffee every day is 40%.
$H_a : \pi \neq 0.4$. The true proportion of West Point cadets who drink coffee every day is not equal to 40%.

b) Do we meet validity conditions to use theoretical methods in this analysis?

Yes, because there are 73 "successes" and 67 "failures", both greater than 10.

c) Using simulation, report the standardized statistic and p-value.

Answers may vary. $z = \frac{stat - mean(null)}{sd(null)} = \frac{0.5214 - 0.4}{0.043} = 2.82392$
p-value $= 0.005$

d) Using theoretical methods, report the standardized statistic and p-value.

$z = \frac{stat - mean(null)}{sd(null)} = \frac{0.5214 - 0.4}{\sqrt{\frac{0.4 \times (1 - 0.4)}{140}}} = \frac{0.1214}{0.0414} = 2.9328$
p-value $= 2 * (1 - pnorm(abs(z))) = 0.00336$

```
n <- 140
phat <- 73/n
pi <- 0.40
sd <- sqrt((pi * (1 - pi))/n)
z <- (phat - pi)/sd; z

## [1] 2.932779

pvalue <- 2*(1-pnorm(abs(z))); pvalue

## [1] 0.003359433
```

e) In words, summarize your findings. Use your findings from **c)** or **d)** to justify your conclusion, using the appropriate based on validity conditions.

With a theoretical p-value of 0.00336, we have very strong evidence against the null hypothesis that West Point cadets have the same coffee consumption as the National average.

f) Comment on the generalize-ability of these results.

This was a convenience sample from Grant, so the results cannot be generalized. Even if the sample was a random sample, as best the results could be generalized to cadets in Grant Hall after lunch, not the rest of the Corps of Cadets who do not frequent Grant Hall.