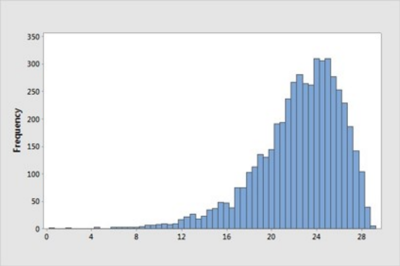
MA256 Lesson 6 - Generalization (2.3, 2.4)

Warm up (quiz)

1. What are the 3 S's in the 3S Strategy?

2. What are the 4 elements used to describe a distribution?

3. Describe the shape of the following distribution and determine whether the mean or median better represents the Center.



4. What are two ways to measure the strength of evidence for a single categorical variable?

5. What are the two ways to measure strength of evidence for a single quantitative variable?

6. What is the point of Central Limit Theorem (CLT)? Build the table below showing validity conditions by scenario & key formula.

Applications	Validity Conditions	Formula	Example
Endless Random Process with Fixed Probability of Success (Categorical Vars)			
Sampling from a Population (Categorical Variables)			
Sampling from a Population (Quantitative Variables)			

ACFT, IOCT, and APFT Dataset

Load the ACFT Dataset posted on Microsoft Teams. This constitutes data collected on or about 2019 (pre-Covid) from randomly selected cadets in 1st Regiment on physical fitness and includes APFT, IOCT, and ACFT performance. Researchers were curious to know how cadets performed physically on various tests, specifically if they were above average. We will conduct analysis on all three tests.

```r
library(tidyverse)
# ACFT <- read_csv("ACFT1.csv")
ACFT <- read_csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/ACFT1.csv")
```

We can begin exploring our available data to see what variables are captured.

```
# head(XXXXX)
# colnames(XXXXX)
```

1. What are the observational units in this study?

2. What are some categorical and quantitative variables of interest? List and classify some of interest.

APFT

The APFT is scored with 3 events on a 100 point scale, where 60 is passing in each of the three events (pushup, situp, and 2 mile run). If a Soldier scores at least 100 points on each event, they can push past 300 to an "extended scale," but if any event falls beneath the 100 point threshold then the "additional" points are lost. It has been suggested that the Corps of Cadets had an average APFT score of 270, and we want to assess if this is true.

Step 1: Ask a Research Question

A.3) What is the research question?

A.4) Based on these questions, what is the population of interest?

Step 2: Design a Study and Collect Data

A.5) What is our null and alternate hypothesis, stated in both words and symbols?

A.6) What is our variable of interest? Is it categorical or quantitative?

A.7) Based on the classification in **A.6**, what theory-based test will we conduct? (hint: z or t)

Step 3: Explore the Data

A.8) Use R to create a histogram of the APFT scores and describe the shape, center, variability, and any unusual observations. (Check the Tidyverse Tutorial for reference.)

```
# ACFT %>% ggplot(aes(x = XXXXX)) + XXXXX +
#   labs(x = "Overall Score", y="Count", title="Cadet APFT Scores")
```

A.9) Calculate the mean, median, standard deviation, and sample size of the ACFT scores. Include the proper notation as appropriate. (Check the Tidyverse Tutorial for reference.)

```
# ACFT %>% summarise(mean = XXXXX,
#                     median = XXXXX,
#                     s = XXXXX,
#                     n = XXXXX)
```

Step 4: Draw Inferences Beyond the Data

A.10) Have we met our validity conditions to use theoretical methods? Why or why not?

A.11) Assume we met our validity conditions. Using theory, calculate the appropriate standardized statistic.

```
# stat <- XXXXX
# stat
```

A.12) Calculate the appropriate p-value using the theoretical method and standardized statistic from **11**.

```
# XXXXX
```

A.13) Comment on the strength of evidence as it applies to our null and alternate hypotheses.

Step 4: REDUX (using the median)

A.14) After looking at your figure and summary statistics above (A.8 and A.9), does it make sense that the mean is higher than the median APFT score? Explain why the mean and median have this relationship.

A.15) The largest APFT score in the data was 375 points. What would happen to the mean if the score was incorrectly recorded as 475? What would happen to the median?

A.16) For this dataset, the difference between the mean and median is not large, but we still might consider the median a more appropriate statistic to focus on, as it would represent a *typical* score, rather than the mean score which is being pulled to the right by a few large percentages. The problem is, all the theory you saw in Sections 2.2 and 2.3 only applies to means. So how can we get a sense for how much sample-to-sample variation there is in the sample medians? How does this work?

A.17) Is it plausible that this sample came from a population with a population median of 270 points? Use the bootstrap to answer. Do we see a difference between the bootstrapped mean and median?

```
# set.seed(256)
# M <- XXXXX
# apft.scores <- XXXXX
# n <- XXXXX
#
# RES <- data.frame(res.med = rep(NA, M),
#                   res.mean = rep(NA, M))
#
# for(i in seq(1:M)){
#   x <- sample(XXXXX, XXXXX, replace = TRUE)
#   RES$res.med[i] <- XXXXX
#   RES$res.mean[i] <- XXXXX
# }
# # calculate the difference between the mean (under null hypothesis) and the observed median (your data)
# my.high.median <- XXXXX
# my.shift <- XXXXX
# my.low.median <- XXXXX
#
# RES %>% ggplot(aes(x=res.mean-my.shift)) + geom_histogram()
# RES %>% ggplot(aes(x=res.med-my.shift)) +
#   geom_histogram() +
#   geom_vline(xintercept= my.high.median, linetype="dashed", color = "red") +
#   geom_vline(xintercept= my.low.median, linetype="dashed", color = "red")
# ( sum(RES$res.med-my.shift > my.high.median) +
#     sum(RES$res.med-my.shift < my.low.median)) / M
```

Step 5: Formulate Conclusions

A.18) Do you feel comfortable generalizing the results of your analysis to all Army cadets (USMA, ROTC, G2G)? Explain your reasoning.

A.19) Do you feel comfortable generalizing the results of your analysis to the Corps of Cadets? Explain your reasoning.

A.20) What population could you generalize your results to? Explain your reasoning.

A.21) How confident are you to say that we have proven the alternate hypothesis?

Step 6: Look Back and Ahead

A.22) Suggest how you might redesign the experiment to allow you to draw a more broad conclusion.

A.23) Suppose your research question had an alternate hypothesis for a one-sided test. That is, do USMA cadets, on average, perform better than 270? Report your new alternate hypothesis, p-value and the significance of your findings. Is this surprising?

```
# XXXXX
```

A.24) Suppose you wanted to assess the data for only males or only females. Repeat your original ($\neq$) }\analysis with the proper subsets and report on your findings.

```
# ACFT %>% group_by(Sex) %>% summarise(mn = XXXXX, sd = XXXXX, n=XXXXX)
# t.m <- XXXXX; t.m
# t.f <- XXXXX; t.f
# # (males, females)
# c(1-pt(t.m,234), c(1-pt(t.f,59)))
```

Heavy Backpacks

Carrying a heavy backpack can be a source of “chronic, low-level trauma,’’ and can cause long-lasting shoulder, neck, and back pain. The American Academy of Orthopedic Surgeons recommends that a backpack not weigh more than 10 to 15% of the wearer’s body weight. As the problem of over heavy backpacks has received more attention recently among primary and secondary school students, a student group decided to investigate the issue with students at their own university (approximately 20,000 enrolled students). Twenty-five students were sampled from each of four locations across the campus, during the course of four days, and four different times of the day (early morning, lunch time, evening, late afternoon). Scales were used to measure backpack weight to the nearest pound, and students were asked to report their body weight on a survey. One researcher was the designated talker and introduced the survey to the participants (survey questions included major, year in school, whether have back problems, and gender). Another person in the group was in charge of reading the scale, and another in charge of gathering the surveys.

B.1) Based on the study description, phrase a research question that you would be interested in testing with these students’ data. Be sure to identify the population and parameter of interest. What symbol could you use for this parameter?

B.2) The data in `backpack.xls` contain several variables of interest, including backpack weight. But the recommendation was stated in terms of “percentage of person’s weight,” so we first convert this variable into the percentage of body weight for each individual in the dataset.

```
# backpack <- read_csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Backpack.csv")
#
# backpack$perc <- backpack$BackpackWt / backpack$BodyWt * 100
# # backpack <- backpack %>% mutate(perc = backpack$BackpackWt / backpack$BodyWt * 100)
```

B.3) Describe the distribution of the percentages for this sample of 100 students. Remember to focus on shape, center, and variability, and any unusual observations or outliers.

```
# backpack %>% ggplot(aes(x=perc)) + geom_histogram(bins=10)
# backpack %>% summarise(mn = mean(perc),
#                        sd = sd(perc))
```

B.4) Predict how the median percentage will compare to the mean percentage, keeping in mind that the median will divide the dataset in half, 50 students below the median and 50 students above the median.

```
# median(backpack$perc)
```

B.5) The largest percentage in this dataset is 18.1%. Suppose this had been incorrectly recorded as 181%! How will the mean change? How will the median change?

Is it plausible that this sample came from a population with a population median of 10%?

B.6) State appropriate null and alternative hypotheses, in words, about the population median for this research question. Assume a one-sided hypothesis test.

The code below provides a bootstrap simulation to estimate the median percentage of backpack weight.

```
# set.seed(256)
#
# M <- 1000
# RES <- data.frame(res.med = rep(NA, M),
#                   res.mean = rep(NA, M))
# n <- 100
#
# for(i in seq(1:M)){
#   x <- sample(backpack$perc, n, replace = TRUE)
#   RES$res.med[i] <- median(x)
#   RES$res.mean[i] <- mean(x)}
#
# my.median <- 7.143  #obs. median
# my.shift <- 10 - my.median
#
# RES %>% ggplot(aes(x=res.med + my.shift)) +
#   geom_histogram() +
#   geom_vline(xintercept= my.median, linetype="dashed", color = "red")
```

B.7) Estimate the p-value using the results of the bootstrap simulation.

```
# estimated p-value
# sum(RES$res.med + my.shift <= my.median) / M
```

B.8) Write a paragraph summarizing your conclusions from this study. Be sure to discuss both what you learned from the sample and what you believe to be plausible about the population.

B.9) Return to your critique of the study design. If you were to carry out such a study on your own campus, discuss what you would do differently and why.