MA256 Lesson 6 - Generalization (2.3, 2.4)

Warm up (quiz)

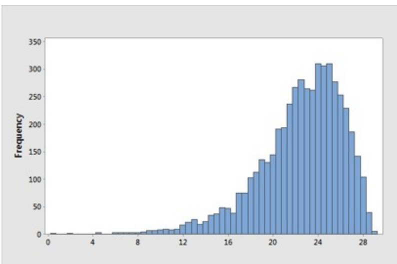
1. What are the 3 S's in the 3S Strategy?

Statistic, Simulate, Strength of Evidence

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

1-Shape, 2-Center, 3-Variability, & 4-Unusual Observations. (See P-2, p. 10)

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



This distribution is slightly left skewed. For skewed distributions, the mean is pulled in the direction of the skew, so the median is better for defining the Center.

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

Calculate the P-value or Z-statistic (standardized statistic). Simulate with One Proportion Applet or Theory-Based method.

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

Calculate the P-value or t-statistic (standardized statistic). Simulate with One Mean Applet or Theory-Based method.

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)	$\geq$ 10 success and $\geq$ 10 failures in sample.	Calculate standard deviation of $\hat{p}$: $s = \sqrt{\pi(1-\pi)/n}$	flipping a coin
Sampling from a Population (Categorical Variables)	$\geq$ 10 success and $\geq$ 10 failures in sample AND population is $\geq 20\times$ sample size.	Calculate standard deviation of $\hat{p}$: $s = \sqrt{\pi(1-\pi)/n}$	boxing uniform color example (Expl. 1.4)
Sampling from a Population (Quantitative Variables)	Symmetric Distribution -OR- $\geq$ 20 Observations AND Sample Distribution is not Strongly Skewed	Calculate t-statistic on a Population Mean: $t = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$	# of sick-call Soldiers in your BN

Bypass simulation and jump straight to strength of evidence using intuitive π or μ_0 and a formula for s.

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.

```
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(ACFT)
```

```
## # A tibble: 6 x 24
##   MDL_Raw MDL_Score SPR_Raw SPR_score HRPU_Raw HPRU_sc~1 SDC_Raw SDC_s~2 LTK_Raw
##     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <time>    <dbl>   <dbl>
## 1     260        82     9.7        81       22        66 01:30       100      11
## 2     300        90    12.1        97       28        69 01:28       100       2
## 3     340       100    10.3        85       34        74 01:40        98      11
## 4     240        78    10          83       23        66 01:49        91      11
## 5     340       100    12.6       100       42        82 01:29       100      18
## 6     200        70     9.4        79       27        68 01:55        85      10
## # ... with 15 more variables: LTK_score <dbl>, X2MR_Finish <time>,
## #   ‘2MR_score‘ <dbl>, APFT_pushups <dbl>, APFT_situps <dbl>,
## #   APFT_runtime_sec <dbl>, APFT_score <dbl>, IOCT_Score <dbl>,
## #   IOCT_time <dbl>, age <dbl>, hgt_in <dbl>, wt_lb <dbl>, BMI <dbl>,
## #   Sex <chr>, ACFT_score <dbl>, and abbreviated variable names 1: HPRU_score,
## #   2: SDC_score
```

```
colnames(ACFT)
```

```
##  [1] "MDL_Raw"          "MDL_Score"        "SPR_Raw"          "SPR_score"
##  [5] "HRPU_Raw"         "HPRU_score"       "SDC_Raw"          "SDC_score"
##  [9] "LTK_Raw"          "LTK_score"        "X2MR_Finish"      "2MR_score"
## [13] "APFT_pushups"     "APFT_situps"      "APFT_runtime_sec" "APFT_score"
## [17] "IOCT_Score"       "IOCT_time"        "age"              "hgt_in"
## [21] "wt_lb"            "BMI"              "Sex"              "ACFT_score"
```

1. What are the observational units in this study?

Each individual cadet

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

Some variation, but a categorical could be "Sex", quantitative could include age, IOCT_Score, AFCT_score, etc.

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?

Is the Corps of Cadets average APFT score equal to 270 or not?

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

The population of interest is the Corps of Cadets

Step 2: Design a Study and Collect Data

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

$H_0 : \mu = 270$. The true mean APFT score for the Corps of Cadets is 270.
$H_a : \mu \neq 270$. The true mean APFT score for the Corps of Cadets is not 270.

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

Our variable of interest is APFT_score, which is a quantitative variable.

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

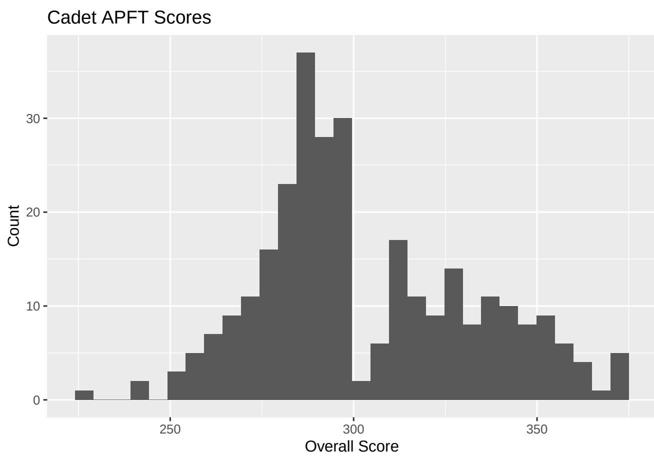
A quantitative variable will conduct a single-mean test with the t statistic.

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 = APFT_score)) +
  geom_histogram() +
  labs(x = "Overall Score", y="Count", title="Cadet APFT Scores")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



The shape appears to have two peaks split at the 300 cutoff point, which makes sense due to the grading rules. It has a mean at 304 and median at 296, indicating some right skew. Our standard deviation is 29.1, with values ranging from about 225 up to 275.

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 = mean(APFT_score),
                   median = median(APFT_score),
                   s = sd(APFT_score),
                   n = n())
```

```
## # A tibble: 1 x 4
##    mean median     s     n
##   <dbl>  <dbl> <dbl> <int>
## 1  304.    296  29.1   293
```

From the code above, we find that mean = $\mu = 303.5$
median = 296
standard deviation = $s = 29.1$
sample size = $n = 293$

Step 4: Draw Inferences Beyond the Data

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

Yes, for quantitative data we have more than 20 observations and the data is not strongly skewed.

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

```r
t <- (304-270) / (29.1 / sqrt(293))
t
```

```
## [1] 19.99953
```

$t = \frac{304-270}{\frac{29.1}{\sqrt{293}}} = 19.99953$

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

```r
2*(1 - pt(abs(t), 292))
```

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

$2 * (1 - pt(abs(19.99953), 292)) = 0$

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

With a p-value of (computationally) 0, we have very strong evidence that the true mean of corps of cadets APFT scores is not equal to 270.

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.

Yes, it makes sense since the distribution is slightly right skewed. The larger values in the right tail will "pull up" the mean.

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?

The mean would increase, while the median would remain the same.

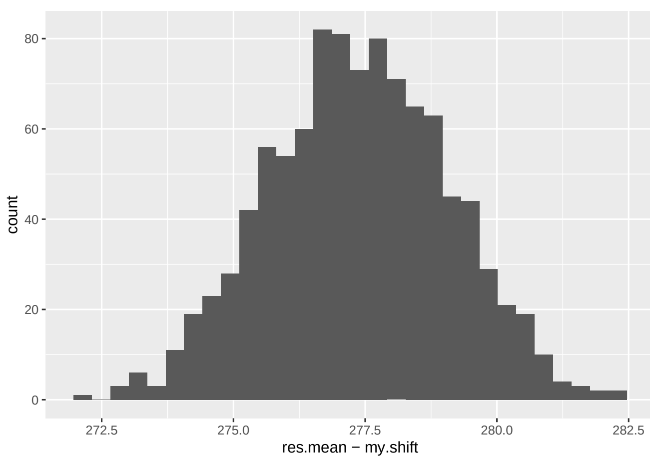
A.16) For this dataset, the difference between the mean and median is not large, but we still might consider the median a more appropriate statistic to focus on, as it would represent a *typical* score, rather than the mean score which is being pulled to the right by a few large percentages. The problem 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?

Yes... bootstrap!
Resampling works as such: given a set X = {250, 275, 300}, one sample could be {300, 275, 275}; another {250, 300, 300}; and another {250, 250, 250}; .... etc.
Using the bootstrap will allow us to estimate the standard deviation/variance of the data (that may be non-normal).

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?

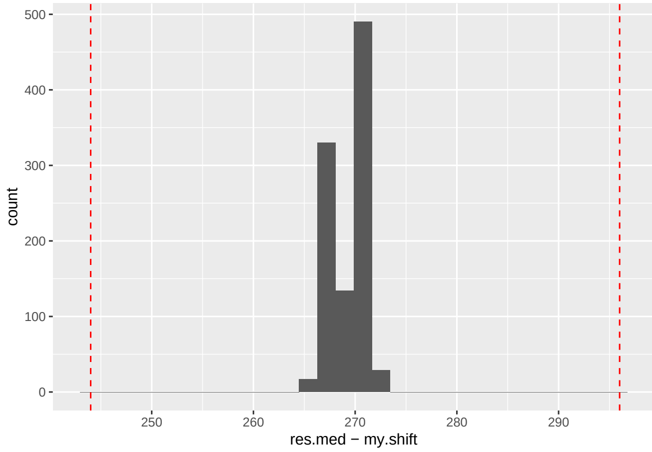
```r
set.seed(256)
M <- 1000
apft.scores <- ACFT$APFT_score
n <- length(ACFT$APFT_score)

RES <- data.frame(res.med = rep(NA, M),
                  res.mean = rep(NA, M))

for(i in seq(1:M)){
  x <- sample(apft.scores, n, replace = TRUE)
  RES$res.med[i] <- median(x)
  RES$res.mean[i] <- mean(x)
}
# calculate the difference between the mean (under null hypothesis) and the observed median (your data)
my.shift <- 296 - 270
my.high.median <- 296
my.low.median <- 270 - (my.high.median - 270)

RES %>% ggplot(aes(x=res.mean-my.shift)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
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")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
( sum(RES$res.med-my.shift > my.high.median) +
    sum(RES$res.med-my.shift < my.low.median)) / M
```

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

It does not look to be very plausible that the median APFT score is 270. Our estimated p-value is 0.

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.

No, those observational units did not have an equal chance of being selected and so we cannot generalize these results.

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

No, observational units from different regiments did not have an equal chance of being selected and so we cannot generalize these results to populations outside of 1st REG.

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

We can generalize these results to the 1st Regiment population of cadets, as those cadets had an equal probability of being sampled for this study.

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

We never say we proved or disproved a hypothesis, only that we have very strong evidence against the null 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.

To generalize to the USMA Corps of Cadets, we would randomly sample from the entire population of the corps of cadets instead of just one regiment. To generalize to all cadets, we would need to be able to randomly sample from the Corps of Cadets, ROTC, and G2G populations with equal probability per cadet.

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?

```
1 - pt(t, 292)
```

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

$H_a : \mu > 270; \quad 1 - pt(19.99953, 292)) = 0$
This is not surprising, it should be half the two-sided test, and $\frac{1}{2}$ of 0 is 0.

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 = mean(APFT_score), sd = sd(APFT_score), n=n())
```

```
## # A tibble: 2 x 4
##   Sex       mn    sd     n
##   <chr>  <dbl> <dbl> <int>
## 1 Female  313.  29.4    59
## 2 Male    301.  28.6   234
```

```
t.m <- (301-270) / (28.6/ sqrt(234)); t.m
```

```
## [1] 16.58073
```

```
t.f <- (312-270) / (29.4 / sqrt(59)); t.f
```

```
## [1] 10.97307
```

```
# (males, females)
c(1-pt(t.m,234), c(1-pt(t.f,59)))
```

```
## [1] 0.000000e+00 3.330669e-16
```

MALES: $t = \frac{301-270}{\frac{28.6}{\sqrt{234}}} = 16.58073$, p-value ≈ 0 \ FEMALES: $t = \frac{313-270}{\frac{29.4}{\sqrt{59}}} = 11.23433$, p-value $= 2.22\text{e}^{-16}$

ACFT

The ACFT is scored with 6 events on a 100 point scale, where 60 is passing in each of the six events. If the arbitrary `average'' score is 80 in each event (480 points), is the Corps of Cadets`average'' at the ACFT?

Step 1: Ask a Research Question

C.1) What is the research question?

Is the Corps of Cadets average ACFT score equal to 480 or not?

C.2) Based on these questions, what is the population of interest?

The population of interest is the Corps of Cadets

Step 2: Design a Study and Collect Data

C.3) What is our null and alternate hypothesis, stated in both words and symbols?

$H_0 : \mu = 480$. The true mean ACFT score for the Corps of Cadets is 480.
$H_a : \mu \neq 480$. The true mean ACFT score for the Corps of Cadets is not 480.

C.4) What is our variable of interest? Is it categorical or quantitative?

Our variable of interest is ACFT_score, which is a quantitative variable.

C.5) Based on the classification in **6**, what theory-based test will we conduct? (hint: z or t)

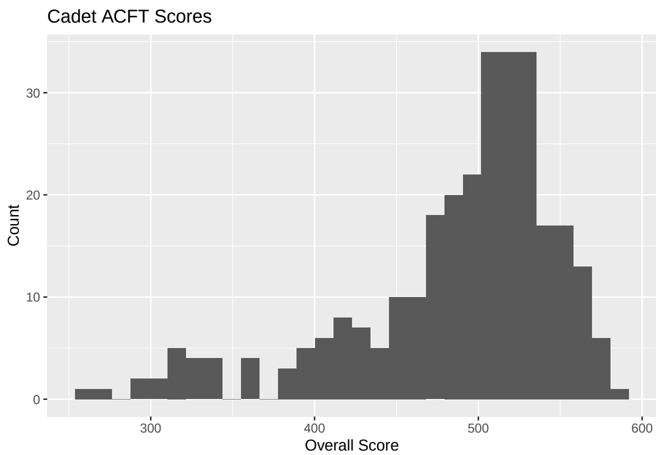
A quantitative variable will conduct a single-mean test with the t statistic.

Step 3: Explore the Data

C.6) Use R to create a histogram of the ACFT scores and describe the shape, center, variability, and any unusual observations.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The shape is left skewed with a peak at about 500, centered between 450 and 500. Values range from nearly 600 down to 250 with no obvious outliers.

C.7) Calculate the mean, median, standard deviation, and sample size of the ACFT scores. Include the proper notation as appropriate.

```
ACFT %>% summarize(median = median(ACFT_score),
                   mean = mean(ACFT_score),
                   s = sd(ACFT_score),
                   n = n() )
```

```
## # A tibble: 1 x 4
##   median  mean     s     n
##    <dbl> <dbl> <dbl> <int>
## 1    505  487.  64.7   293
```

From the code above, we find that mean $= \mu = 487$
median $= 505$
standard deviation $= s = 64.7$
sample size $= n = 293$

Step 4: Draw Inferences Beyond the Data

C.8) Have we met our validity conditions to use theoretical methods? Why or why not?

Yes, for quantitative data we have more than 20 observations and the data is not **strongly** skewed.

C.9) Assume we met our validity conditions. Using theory, calculate the appropriate standardized statistic.

```
t <- (487 - 480) / (64.7 / sqrt(293))
t
```

```
## [1] 1.851943
```

$t = \frac{487-480}{\frac{64.7}{\sqrt{293}}} = 1.851943$

C.10) Calculate the appropriate p-value using the theoretical method and standardized statistic from **11**.

```
2*(1 - pt(abs(1.851943), 292))
```

```
## [1] 0.06504327
```

$2 * (1 - pt(abs(1.851943), 292)) = 0.06504327$

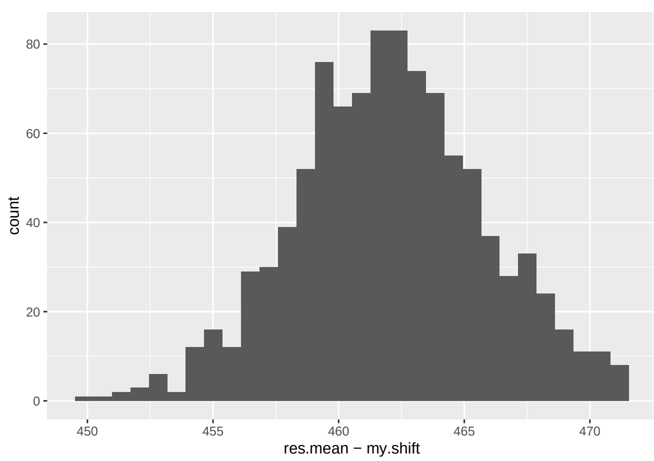
C.11) Comment on the strength of evidence as it applies to our null and alternate hypotheses.

With a p-value of 0.065, we have moderate evidence that the true mean of corps of cadets ACFT scores is not equal to 480.

C.12) Repeat the analysis but use the median as the statistic of interest. Note: You will have to use the bootstrap method that you used with the APFT data.

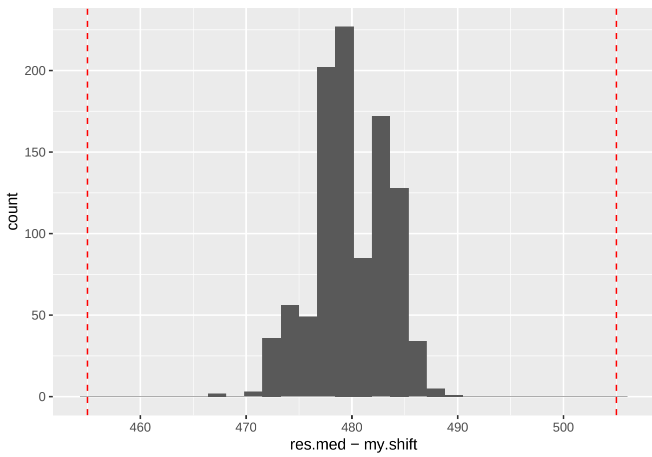
```
set.seed(256)
M <- 1000
acft.scores <- ACFT$ACFT_score
n <- length(ACFT$ACFT_score)

RES <- data.frame(res.med = rep(NA, M),
                  res.mean = rep(NA, M))

for(i in seq(1:M)){
  x <- sample(acft.scores, n, replace = TRUE)
  RES$res.med[i] <- median(x)
  RES$res.mean[i] <- mean(x)
}

my.shift <- 505 - 480
my.high.median <- 505
my.low.median <- 480 - (my.high.median - 480)

RES %>% ggplot(aes(x=res.mean-my.shift)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
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")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
( sum(RES$res.med-my.shift > my.high.median) +
    sum(RES$res.med-my.shift < my.low.median)) / M
```

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

Step 5: Formulate Conclusions

C.13) Do you feel comfortable generalizing the results of your analysis to all Army cadets (USMA, ROTC, G2G)? Explain your reasoning.

No, those observational units did not have an equal chance of being selected and so we cannot generalize these results.

C.14) Do you feel comfortable generalizing the results of your analysis to the Corps of Cadets? Explain your reasoning.

No, observational units from different regiments did not have an equal chance of being selected and so we cannot generalize these results to populations outside of 1st REG.

C.15) What population could you generalize your results to? Explain your reasoning.

We can generalize these results to the 1st Regiment population of cadets, as those cadets had an equal probability of being sampled for this study.

C.16) How confident are you to say that we have proven the alternate hypothesis?

We never say we proved or disproved a hypothesis, only that we have moderate evidence against the null hypothesis.

Step 6: Look Back and Ahead

C.17) Suggest how you might redesign the experiment to allow you to draw a more broad conclusion.

To generalize to the corps of cadets, we would randomly sample from the entire population of the corps of cadets instead of just one regiment. To generalize to all cadets, we would need to be able to randomly sample from the Corps of Cadets, ROTC, and G2G populations with equal probability per cadet.

C.18) Suppose your research question had an alternate hypothesis for a one-sided test. Report your new new alternate hypothesis, p-value and the significance of your findings. Is this surprising?

```
1 - pt(1.851943, 292)
```

```
## [1] 0.03252164
```

$H_a : \mu > 480; \quad 1 - pt(1.851943, 292)) = 0.03252164$
This is not surprising, it should be half the two-sided test, and $\frac{1}{2}$ of 0.065 is 0.0325

C.19) 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 = mean(ACFT_score), sd = sd(ACFT_score), n=n())
```

```
## # A tibble: 2 x 4
##   Sex       mn    sd     n
##   <chr>  <dbl> <dbl> <int>
## 1 Female  386.  58.9    59
## 2 Male    512.  33.9   234
```

```
t.m <- (512-480) / (33.9 / sqrt(234)); t.m
```

```
## [1] 14.4397
```

```
t.f <- (385-480) / (58.938 / sqrt(59)); t.f
```

```
## [1] -12.38096
```

```
# (males, females)
c(1-pt(t.m,234), c(pt(t.f,59)))
```

```
## [1] 0.000000e+00 2.384957e-18
```

MALES: $t = \frac{512-480}{\frac{33.9}{\sqrt{234}}} = 14.4397$, p-value ≈ 0

FEMALES: $t = \frac{386-480}{\frac{58.9}{\sqrt{59}}} =$ -12.25853, p-value ≈ 0

C.20) What do the signs of the t statistic tell you that the p-value alone does not in this case?

Here, it tells us two things. One, the evidence is strong that the male mean is larger than the null hypothesis (positive t) while the evidence is strong that the female mean is smaller than the null hypothesis (negative t). Furthermore, despite both p-values being computationally 0, we can see that the distance of the male scores is farther away, in terms of standard deviations, than the female scores.