Lab 10

Taylor Blair

Math 141, Week 11

Due: Before your Week 12 lab meeting

Goals of this lab

1. Practice using the normal distribution to find probabilities and compute quantiles.
2. Derive the formulae for sample size calculations related to a single mean or a single proportion.
3. Practice using theoretical probability distributions in statistical inference problems.

Problems

```
# Insert libraries here
library(tidyverse)
library(infer)
```

Problem 1

Let $Z \sim N(\mu = 0, \sigma = 1)$. For this problem, we want you to practice computing probabilities using `pnorm()` and finding quantiles/percentiles using `qnorm()`.

a. Find $P(Z < -0.5)$.

```
pnorm(-0.5, mean = 0, sd = 1)
```

```
## [1] 0.3085375
```

b. Find $P(Z > 1.96)$.

```
1-pnorm(1.96, mean = 0, sd = 1)
```

```
## [1] 0.0249979
```

c. Find the 95th percentile.

```
qnorm(0.95, mean = 0, sd=1)
```

```
## [1] 1.644854
```

d. Suppose we want to construct a 95% confidence interval. Find the z^* value such that 95% of the standard Normal is between $-z^*$ and z^*.

```
cat("Lower: ", qnorm(0.025, mean = 0, sd=1))
```

```
## Lower:  -1.959964
```

```
cat("Upper: ", qnorm(0.975, mean = 0, sd=1))
```

```
## Upper:  1.959964
```

e. Repeat d but this time find the z^* for an 80% confidence interval.

```
ci_val <-0.8
cat("Lower: ", qnorm((1-ci_val)/2, mean = 0, sd=1))
```

```
## Lower:  -1.281552
```

```
cat("Upper: ", qnorm(1-(1-ci_val)/2, mean = 0, sd=1))
```

```
## Upper:  1.281552
```

f. Suppose we needed a critical value $t*$ from a $t(df = 29)$ for an 80% confidence interval. Find $t*$. Why is $t*$ larger than the z^* you found in e?

Problem 2

Often before we conduct a study we want to determine how **large our sample** should be to achieve a certain level of precision in our confidence interval. This determination is called a **sample size calculation**. To determine the sample size, we want bound our Margin of Error by B (ME $\leq$ B) and solve for n.

a. Derive the sample size formula for

- a CI for p, the population proportion.
- a CI for μ, the population mean.

In your solution give

$$n \geq \frac{fill - in}{fill - in}$$

for both cases.

- Population prop
  - $z * \sqrt{\frac{p(1-p)}{n}} \leq B$
  - $\sqrt{\frac{p(1-p)}{n}} \leq \frac{B}{z}$
  - $\frac{p(1-p)}{n} \leq \frac{B^2}{z^2}$
  - $\frac{1}{n} \leq \frac{B^2}{z^2 p(1-p)}$
  - $n \geq \frac{z^2 p(1-p)}{B^2}$
- Population mean
  - It's 11:22 PM. I am eyeballing this based on the above
  - $\frac{\bar{x}-\mu_o}{s/\sqrt{n}} \leq B$
  - $n \geq (\frac{B}{s\bar{x}-\mu_o})^2$

Note: Unfortunately, BOTH of our sample size formulae rely on statistics from the study. But remember that the goal is to determine the sample size BEFORE collecting data. To remedy this issue, researchers will usually either:

- Estimate the unknowns from previous studies.
- Conduct a pilot study on a small group of cases and use this to estimate unknowns.

---

b. Suppose we want to estimate the proportion of Reedies from Oregon with 90% confidence and want to bound the Margin of Error by 0.04. From Reed's admissions page, it says 17% of Reedies are from the Northwest so let's estimate that 10% of Reedies are from Oregon. How many students should we sample? Why should we ALWAYS round up instead of down with our sample size calculations?

```
ceiling((0.9^2*0.1*(1-0.1))/(0.04^2))
```

```
## [1] 46
```

- We want to round up because we can't have part of a person.

---

c. Suppose we actually want to estimate the proportion of Reedies from Oregon AND the proportion from California, AND the proportion from Iowa, ... We still want 90% confidence and want to bound the Margin of Error for all of the CIs by 0.04. Why should we then estimate $\hat{p}$ with 0.5 in the sample size calculation?

```
ceiling((0.9^2*0.5*(1-0.5))/(0.04^2))
```

```
## [1] 127
```

- The proportion has increased, by how much is unkown. Because we are counting two states outside of the northwest the p hat will inctrase an unkown amount.

---

d. Let's determine the sample size needed to estimate μ with 95% confidence when we approximate t^* with z^* and $s = 10$. Find the needed sample size for a margin of error within 1 unit, 2 units, 5 units. Comment on the relationship between the sample size and the desired margin of error.

```
ceiling((0.95^2*0.5*(1-0.5))/(0.04^2))
```

```
## [1] 142
```

- Multiply by 1, 2, 5 respectively

---

e. With a margin of error of 2 units, determine the sample size needed to estimate μ when we approximate t^* with z^* and $s = 10$ at a confidence level of 90%, 95%, and 99%. Comment on the relationship between the sample size and the confidence level.

```
cat("90%", ceiling((0.90^2*0.5*(1-0.5))/(0.04^2)))
```

```
## 90% 127
```

```
cat("95%", ceiling((0.95^2*0.5*(1-0.5))/(0.04^2)))
```

```
## 95% 142
```

```
cat("99%", ceiling((0.99^2*0.5*(1-0.5))/(0.04^2)))
```

```
## 99% 154
```

- Grows

Problem 3

The TV show Mythbusters wanted to determine if yawns really are contagious. Therefore, they recruited 50 people at a local flea market to participate in a study. For each subject, the attendee would take the subject to a small room with a hidden camera. The attendee would either yawn (treatment group) or not yawn (control group) as they led the subject into the room. Then as the subject sat in the room, the researchers observed whether or not they yawned. Mythbusters concluded that yawns are contagious. Let's see what conclusions we want to draw from this data.

```
mythbusters <- read_csv("/home/courses/math141f20/Data/mythbusters.csv")
```

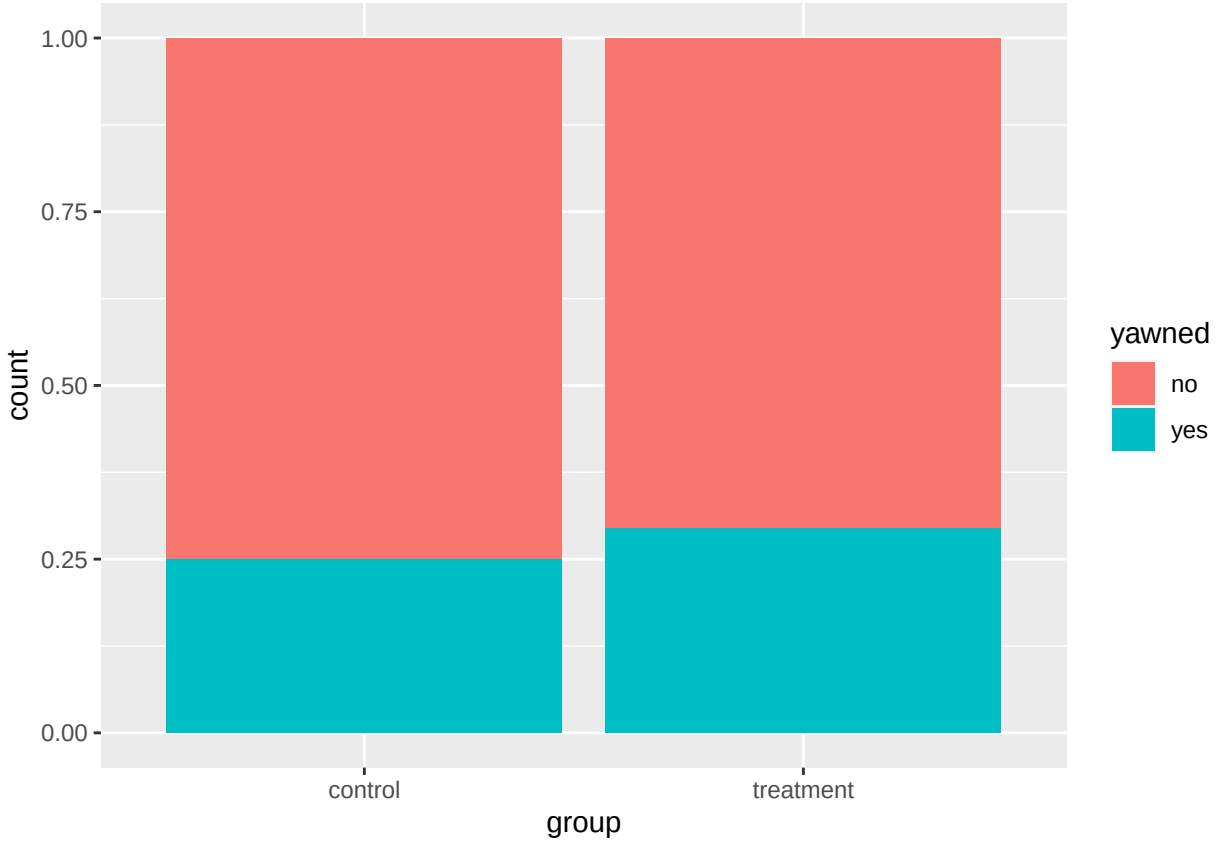
a. Identify the explanatory variable and the response variable.

- **Explanatory**
  - `group`
  - Control or treatment
- **Response**
  - `yawned`
  - Yes or no

---

b. Produce a graph that provides the proportion who did and did not yawn for each of the explanatory groups. Draw some initial conclusions about the research question from the graph.

```
ggplot(mythbusters, aes(group, fill=yawned))+
  geom_bar(position="fill")
```




```
cat("There are ", sum(mythbusters$group=="control"), "indiviuals in the control")
```

```
## There are  16 indiviuals in the control
```

```
cat("There are ", sum(mythbusters$group=="treatment"), "indiviuals in the treatment")
```

```
## There are  34 indiviuals in the treatment
```

- More individuals yawned than did not yawn
- The there is a slightly larger proportion of yawners in the treatment

---

c. State the null and alternative hypotheses in terms of conjectures and in terms of the population parameter.

- Null
  - There is no difference in proportions between the two groups.
  - $P_t - P_c = 0$
- Alternative
  - There is a higher proportion of yawners in the treatment group
  - $P_t > P_c$
- Alpha
  - $\alpha = 0.05$

---

d. Using R as a calculator, Compute the appropriate Z-score test statistic. Based on the test statistic

alone (without calculating the p-value), do you think the sample results will be likely or unlikely under the null hypothesis? Justify your answer.

```
P_t <- sum(mythbusters$group[mythbusters$yawned=="yes"]=="treatment")/sum(mythbusters$group=="treatment

P_c <- sum(mythbusters$group[mythbusters$yawned=="yes"]=="control")/sum(mythbusters$group=="control")


(P_t-P_c)/sqrt((P_t*(1-P_t))/sum(mythbusters$group=="treatment")+(P_c*(1-P_c))/sum(mythbusters$group=="
```

```
## [1] 0.3304438
```

- Likely under null

---

e. The CLT sample size assumption here is that there are at least 10 successes and failures in each explanatory group. Is that condition met? Justify your answer.

- control
  - yes
    - 4
  - no
    - 12
- treatment
  - yes
    - 10
  - no
    - 24

NO, missing 10 yawners in the control

---

f. What distribution does the test statistic follow when the sample size is large?

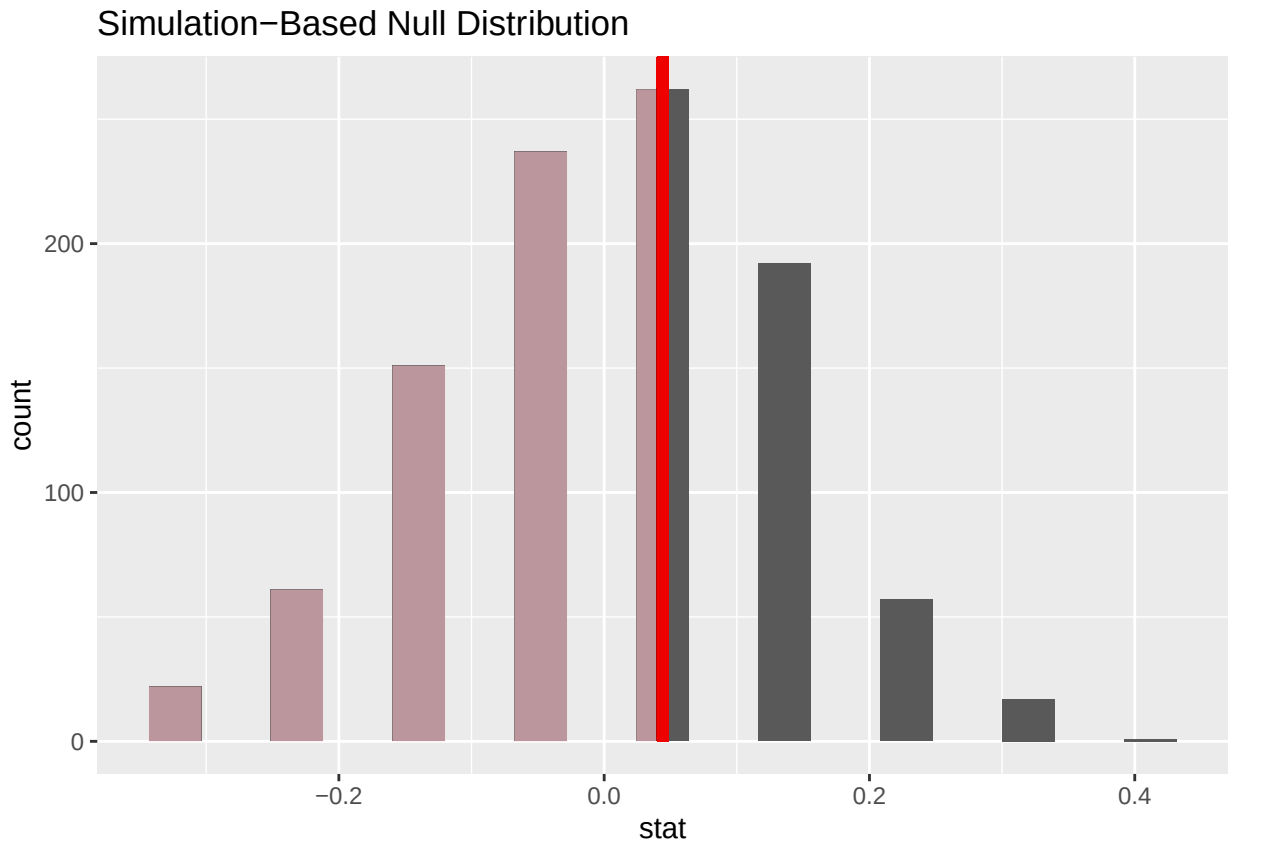
- Normal
- Centered at 0

---

g. Find the p-value two ways:
   - Using a simulated null distribution
   - Using probability function that approximates the null distribution (even if the CLT assumption isn't met)

```
# Using infer

i_reject <- P_t-P_c #someone please get the joke


your_reality <- mythbusters %>%
  specify(explanatory = group, response = yawned, success = "yes") %>%
  hypothesise(null = "independence") %>%
  generate(reps=1000, type ="permute") %>%
  calculate("diff in props", order = c("treatment", "control"))
```

```r
and_substitute <- your_reality %>%
  get_p_value(i_reject, "left")

your_reality %>%
  visualise() +
  shade_p_value(i_reject, "left")
```




```r
# Approximating with probability function
```

h. Compare the p-values. Will your conclusion to the question "Is yawning contagious?" change depending on which p-value you use?

- Yes, the simulated p value is 0.733 which is so closely centered around 0.5 and fluctuates so often that most alpha will not fit.

---

i. Did you just show that yawning is not contagious? Justify your answer.

- No, failed to reject the null hypothesis.

Problem 4

Does mindset matter? Does it positively impact weight loss? A 2007 study sought to understand the impact of mindset on weight. They recruited 75 female maids working at different hotels to participate in the study and 41 randomly chosen maids were told that their work satisfies the Surgeon General's recommendation for an active lifestyle (which is true), giving the maids examples on how their work qualifies as good exercise. The other 34 maids were told nothing. Each maid's weight was measured at the start of the study (`Wt`) and four weeks later (`Wt2`). For the variable, `Cond`: 1 = informed, 0 = not informed.

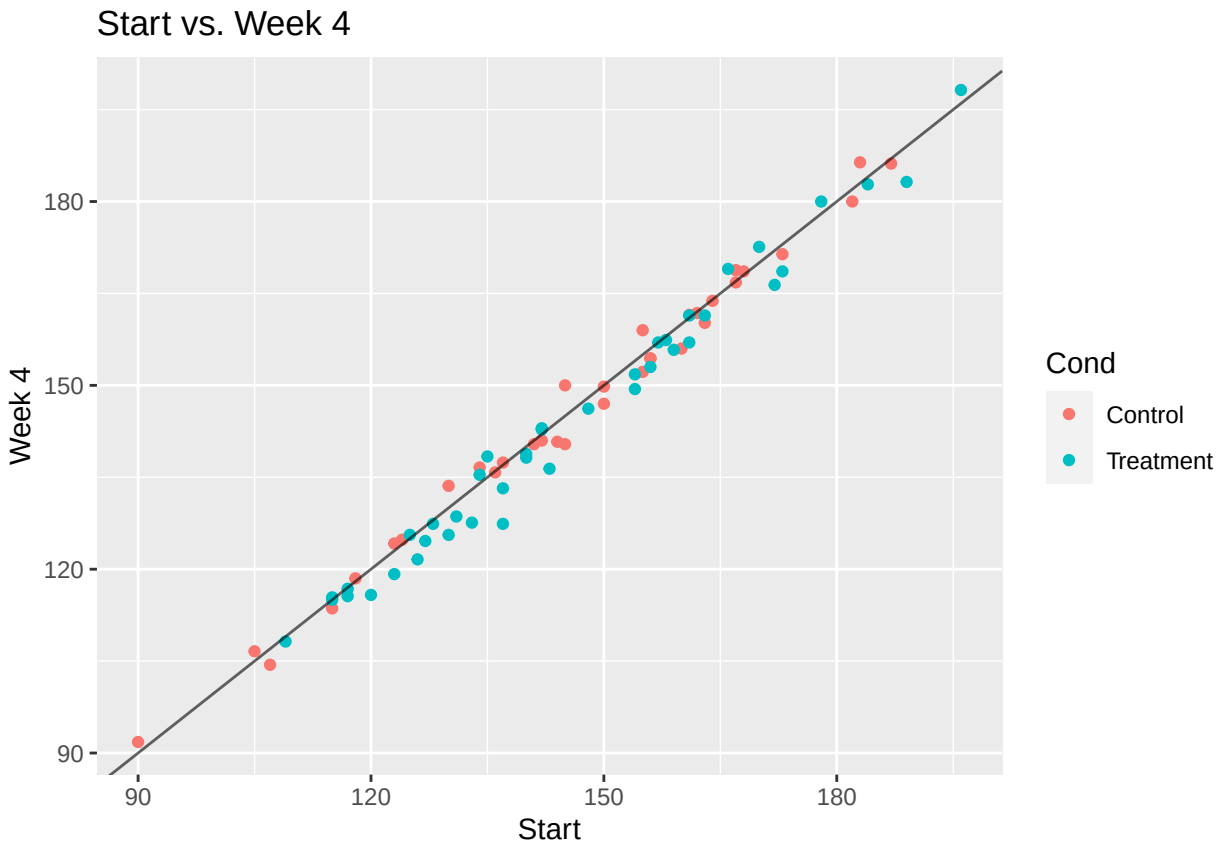
```
MindsetMatters <- read_csv("/home/courses/math141f20/Data/MindsetMatters.csv")
```

We want you to conduct a hypothesis test to answer the question "Does mindset matter?" and to estimate the effect of mindset. For your inference procedures, use an approach that approximates the distributions with a probability function (i.e. Don't generate a null distribution or bootstrap distribution).

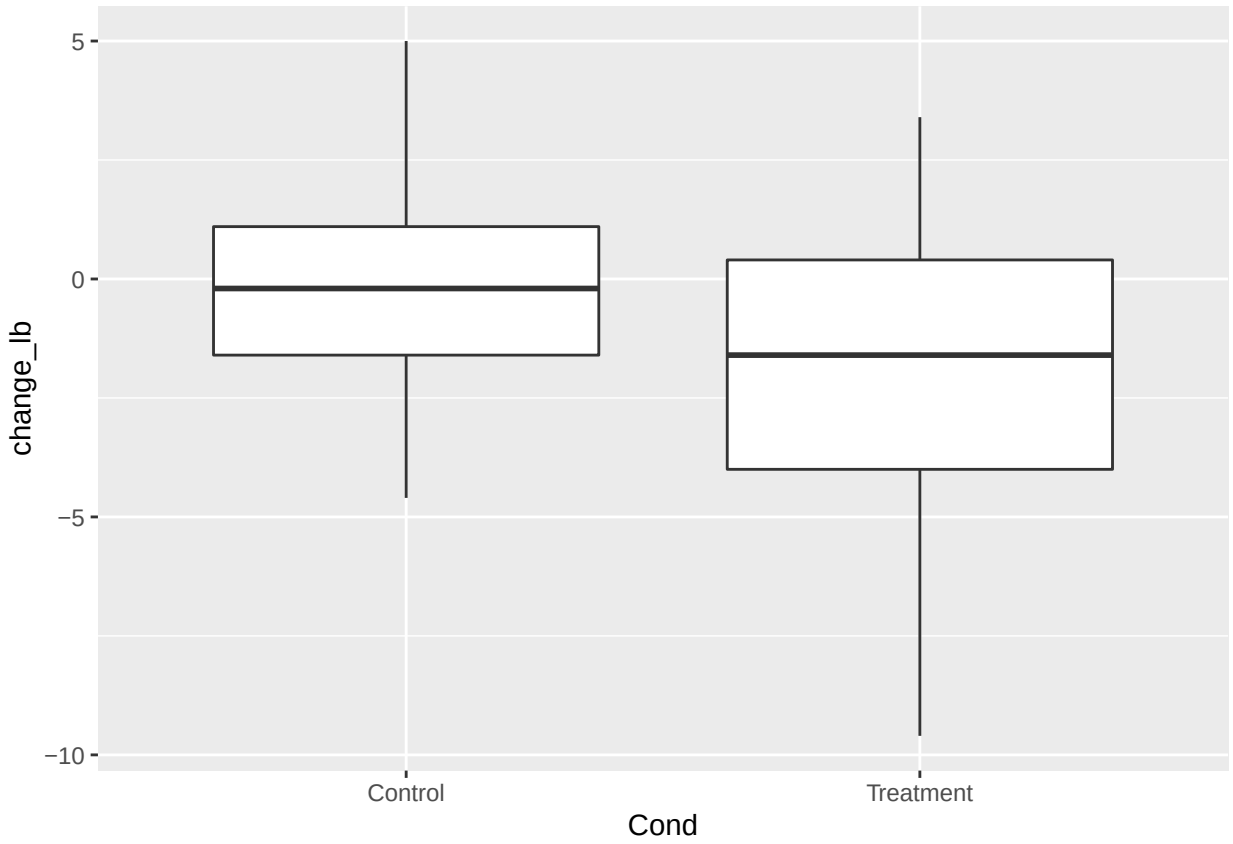
Make sure to include:

- Any useful data wrangling. (Hint: You will need to create the response variable!)
- Useful exploratory analyses.
- The null and alternative hypotheses, the observed test statistic, and the p-value.
- Interpret the p-value in the context of the problem.
- A confidence interval estimate of the mindset effect.
- Some conclusions about the conjecture.

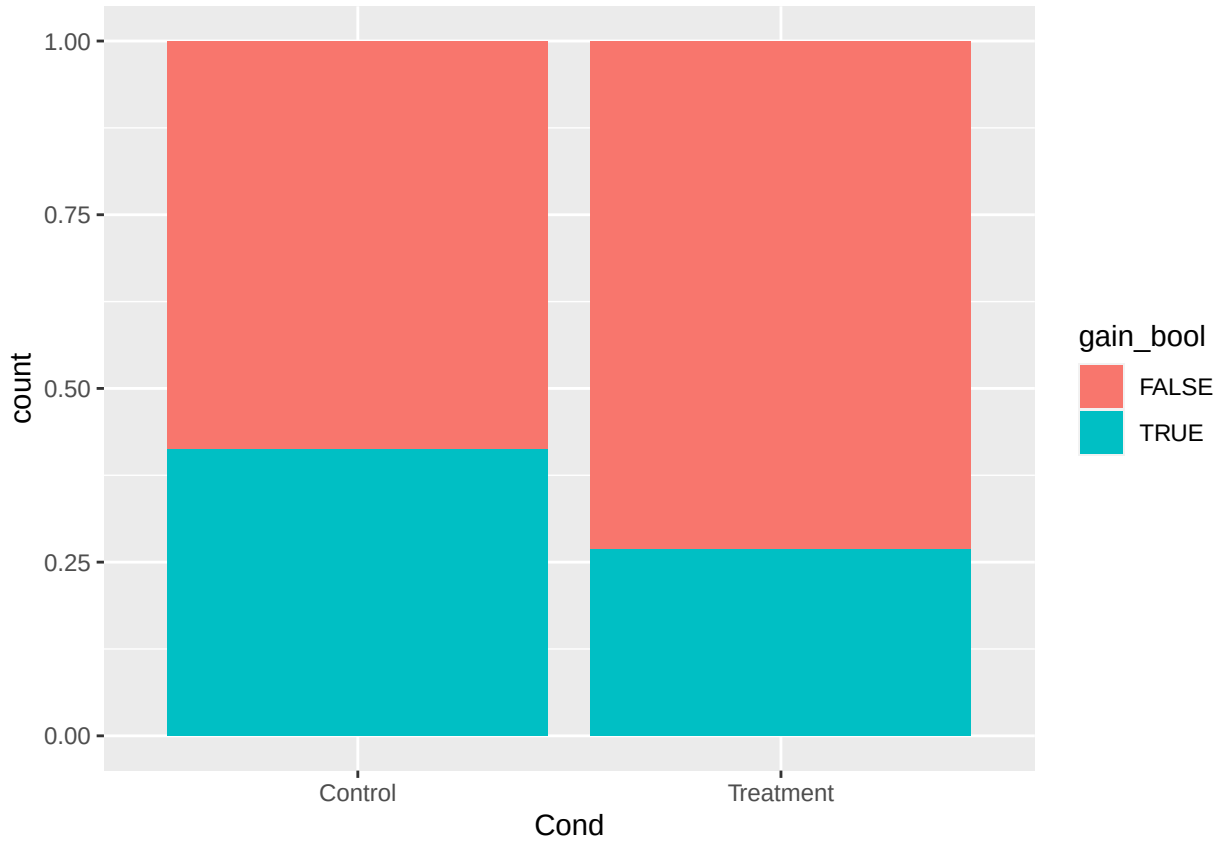
```
MindsetMatters <- MindsetMatters %>%
  mutate(change_lb=Wt2-Wt,
         gain_bool = change_lb>0) %>%
  transform(Cond = as.character(Cond))

MindsetMatters$Cond[MindsetMatters$Cond=="0"] <- "Control"

MindsetMatters$Cond[MindsetMatters$Cond=="1"] <- "Treatment"

ggplot(MindsetMatters, aes(Wt, Wt2, color=Cond)) +
  geom_point() +
  geom_abline(alpha=0.6) +
  labs(title = "Start vs. Week 4") +
  xlab("Start") +
  ylab("Week 4")
```



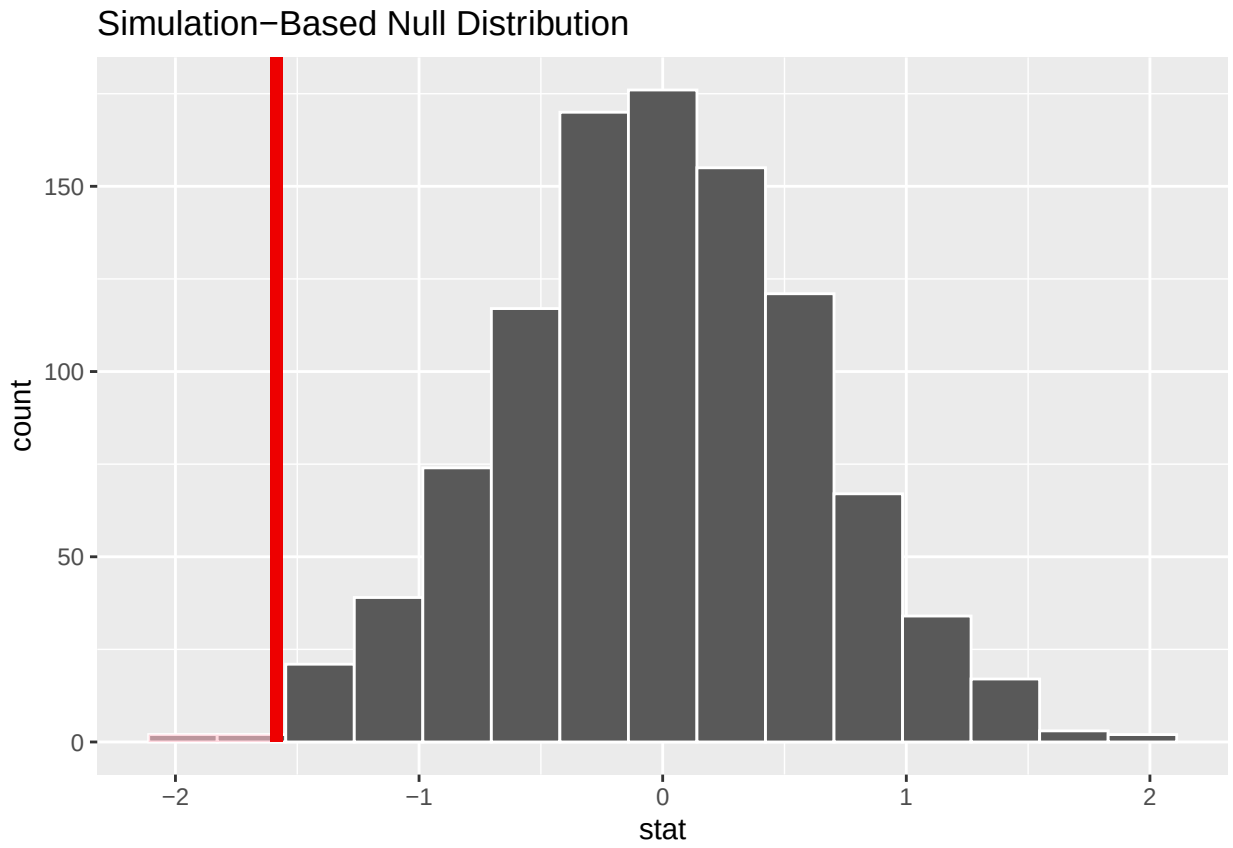

```
ggplot(MindsetMatters, aes(Cond, change_lb)) +
  geom_boxplot()
```



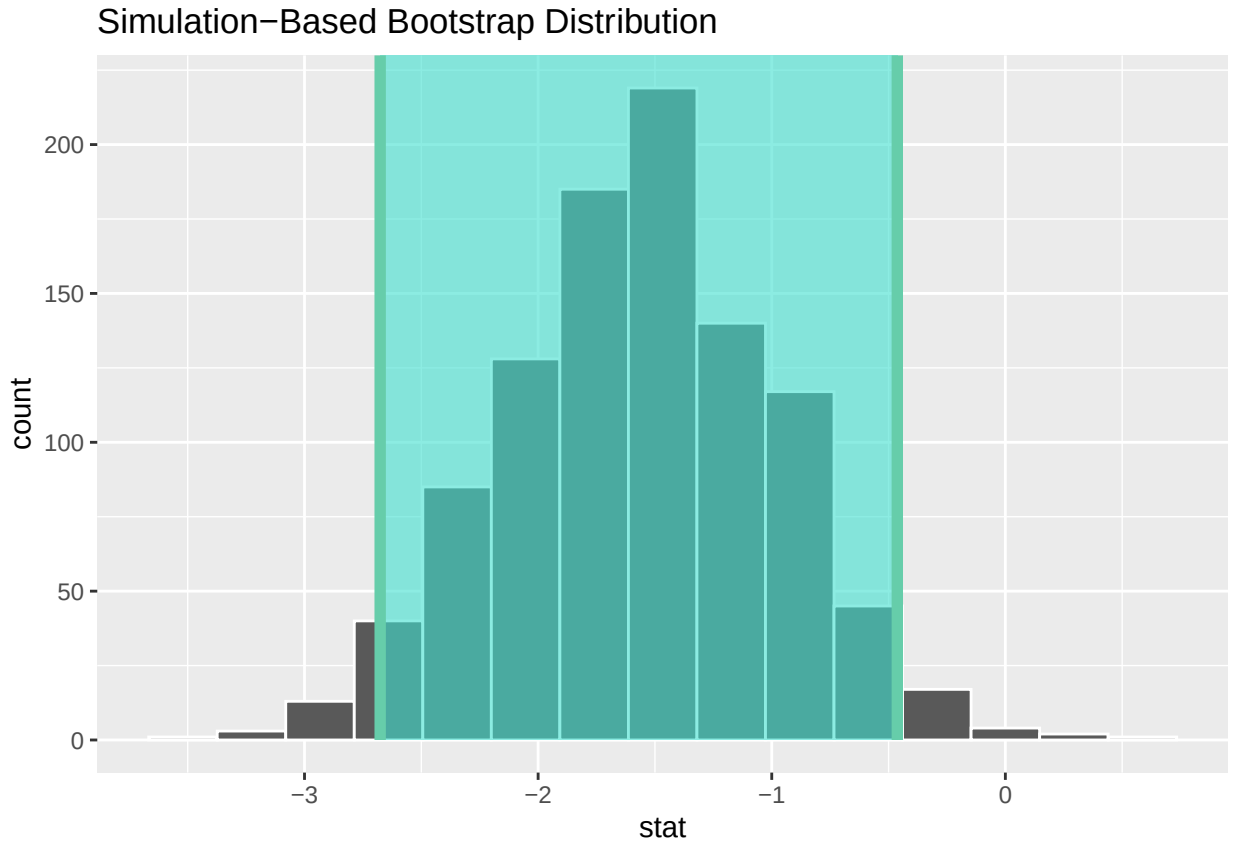

```
ggplot(MindsetMatters, aes(Cond, fill=gain_bool)) +
  geom_bar(position = "fill")
```




```
control_bar <- mean(MindsetMatters$change_lb[MindsetMatters$Cond=="Control"])
treatment_bar <- mean(MindsetMatters$change_lb[MindsetMatters$Cond=="Treatment"])
test_stat <- treatment_bar-control_bar

null <- MindsetMatters %>%
  specify(explanatory = Cond, response = change_lb) %>%
  hypothesise(null = "independence") %>%
  generate(reps=1000, type ="permute") %>%
  calculate("diff in means", order = c("Treatment", "Control"))

p_val <- null %>%
  get_p_value(obs_stat = test_stat, direction = "left")

null %>%
  visualise() +
  shade_p_value(obs_stat = test_stat, direction = "left")
```




```
boot <- MindsetMatters %>%
  specify(explanatory = Cond, response = change_lb) %>%
  generate(reps=1000, type ="bootstrap") %>%
  calculate("diff in means", order = c("Treatment", "Control"))

ci <- boot %>%
  get_ci(level = 0.95)

boot %>%
  visualise() +
  shade_ci(ci)
```




```
MindsetMatters %>%
  group_by(Cond, gain_bool) %>%
  summarize(count = n())
```

```
## # A tibble: 4 x 3
## # Groups:   Cond [2]
##   Cond      gain_bool count
##   <chr>     <lgl>     <int>
## 1 Control   FALSE        20
## 2 Control   TRUE         14
## 3 Treatment FALSE        30
## 4 Treatment TRUE         11
```

- Null hypothesis is there is no difference in mean change between groups, $\alpha = 0.05$
  - Null hypothesis can be rejected as the simulated p-value is 0.003 which is less than 0.05
- Alternative hypothesis is that individuals in the treatment group lost more weight than those in the control group
  - Points in the treatmant group were under the line at a greater rate than the control indicating they gained less weight (*see figure 1*)
  - The box plot shows that the individuals in the treatment on average lost more weight than those in the control group. (*figure 2*)
  - Higher proportion of individuals lost weight in the treatment group than in the control (*Figure 3*)
  - Lastly the bootstrap had a 95% confidence interval from -2.6762003 to -0.462546
- *Other things to note*
  - There are at least ten in each subgroup of `Cond` and `gain_bool`
  - Individuals in the control gained -0.2 on average compared to -1.7853659 (*Negative numbers signify*

losing weight)
