Notes and Exercises Chapter 20 – Inference for Comparing 2 Independent Means
Math 217 **Name** _____
Saidi **Date** _____

20.1 Randomization test for the difference in means

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, they shuffled the exams together to ensure each student received a random version. Anticipating complaints from students who took Version B, they would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

The classdata data can be found in the **openintro** R package.

20.1.1 Observed data

Summary statistics for how students performed on these two exams are shown in Table 20.1 and plotted in Figure 20.1.

Table 20.1: Summary statistics of scores for each exam version.

Group	n	Mean	SD	Min	Max
A	58	75.1	13.9	44	100
B	55	72.0	13.8	38	100

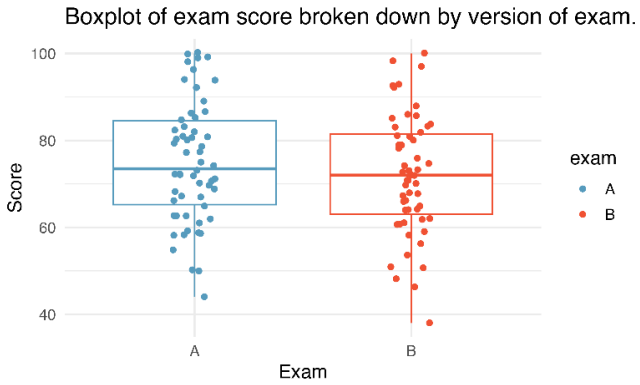


Figure 20.1: Exam scores for students given one of three different exams.

Construct hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 3.1$, is likely to have happened due to chance, if the null hypothesis is true. We will later evaluate these hypotheses using α=0.01.

20.1.2 Variability of the statistic

In the exam example, the null hypothesis is that exam A and exam B are equally difficult, so the average scores across the two tests should be the same. If the exams were equally difficult, *due to natural variability*, we would sometimes expect students to do slightly better on exam A ($\bar{x}_A > \bar{x}_B$) and sometimes expect students to do slightly better on exam B ($\bar{x}_B > \bar{x}_A$). The question at hand is: does $\bar{x}_A - \bar{x}_B = 3.1$ indicate that exam A is easier than exam B.

Figure 20.2 shows the process of randomizing the exam to the observed exam scores. If the null hypothesis is true, then the score on each exam should represent the true student ability on that material. It shouldn't matter whether they were given exam A or exam B. By reallocating which student got which exam, we are able to understand how the difference in

average exam scores changes due only to natural variability. There is only one iteration of the randomization process in Figure 20.2, leading to one simulated difference in average scores.

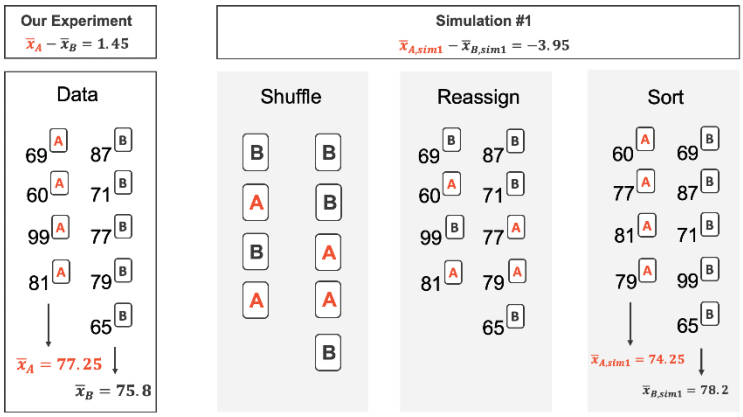


Figure 20.2: The version of the test (A or B) is randomly allocated to the test scores, under the null assumption that the tests are equally difficult.

Building on Figure 20.2, Figure 20.3 shows the values of the simulated statistics $\bar{x}_{1\_sim} - \bar{x}_{2\_sim}$ over 1,000 random simulations. We see that, just by chance, the difference in scores can range anywhere from -10 points to +10 points.

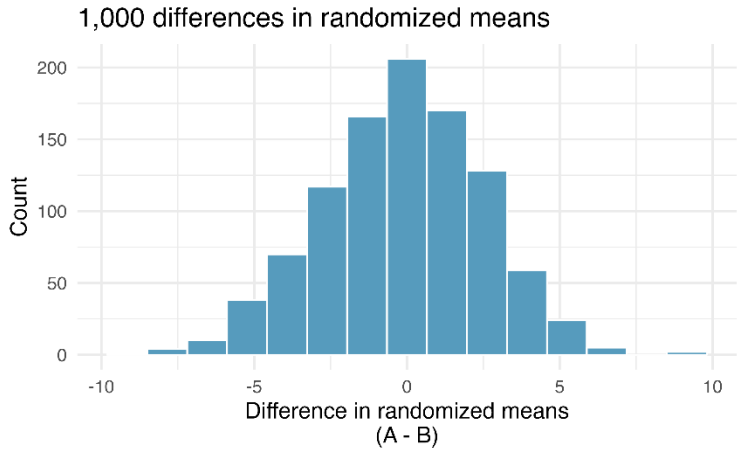


Figure 20.3: Histogram of differences in means, calculated from 1,000 different randomizations of the exam types.

20.1.3 Observed statistic vs. null statistics

The goal of the randomization test is to assess the observed data, here the statistic of interest is $\bar{x}_A - \bar{x}_B = 3.1$. The randomization distribution allows us to identify whether a difference of 3.1 points is more than one would expect by natural variability of the scores if the two tests were equally difficult. By plotting the value of 3.1 on Figure 20.4, we can measure how different or similar 3.1 is to the randomized differences which were generated under the null hypothesis.

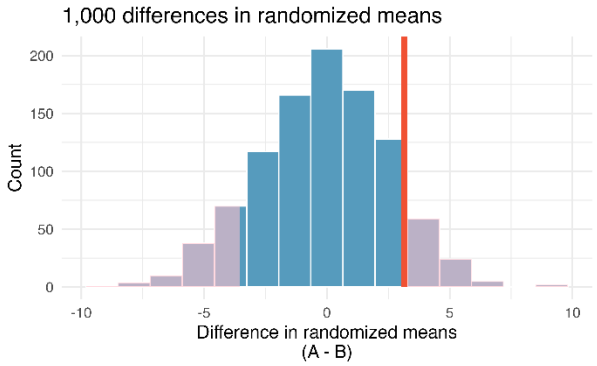


Figure 20.4: Histogram of differences in means, calculated from 1,000 different randomizations of the exam types. The observed difference of 3.1 points is plotted as a vertical line, and the area more extreme than 3.1 is shaded to represent the p-value.

Example1: Use the births14 dataset

This dataset has 1000 observations and 13 variables.

Four cases from this dataset are represented in Table 20.3. We are particularly interested in two variables: weight and habit. The weight variable represents the weights of the newborns and the smoke variable describes which mothers smoked during pregnancy.

Table 20.3: Four cases from the births14 dataset. The empty cells indicate missing data.

fage	mage	weeks	visits	gained	weight	sex	habit
34	34	37	14	28	6.96	male	Nonsmoker
36	31	41	12	41	8.86	female	Nonsmoker
37	36	37	10	28	7.51	female	Nonsmoker
	16	38		29	6.19	male	Nonsmoker

We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who do not smoke? We will use data from this sample to try to answer this question.

1. Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.
2. Check the two conditions for normality:
3. Do NOT use a formulaic approach. Use r-code to perform the hypothesis test computations.

$$H_o: \mu_{smoker} = \mu_{nonsmoker}$$

$$H_a: \mu_{smoker} \neq \mu_{nonsmoker}$$

N=1000 > 30, and we assume the sample was collected randomly and independent subjects. Thus, it passes the conditions for normality. Use the syntax: t.test(y ~ x)

```
t.test(births14_a$weight ~ births14_a$habit)
```

```
        Welch Two Sample t-test

data:  births14_a$weight by births14_a$habit
t = 3.8166, df = 131.31, p-value = 0.0002075
alternative hypothesis: true difference in means between group nonsmoker and group smoker
is not equal to 0
95 percent confidence interval:
 0.2854852 0.8998751
sample estimates:
mean in group nonsmoker    mean in group smoker
               7.269873                6.677193
```

t = 3.8166, df = 131.31, p-value = 0.0002075

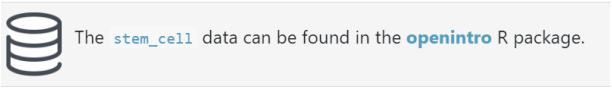
Conclusion

Reject the null. There is very strong evidence that there is a difference in mean birth weights for babies born to smoking and non-smoking mothers.

Example 2: Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack?

Table 20.2 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. (Ménard et al. 2005) Figure 20.8 provides histograms of the two datasets. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery.

Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.



Group	n	Mean	SD
ESC	9	3.50	5.17
Control	9	-4.33	2.76

Table 20.2: Summary statistics of the embryonic stem cell study.

When you only have summary statistics

```
library(BSDA)
tsum.test(3.5, 5.17, 9, -4.33, 2.76, 9)
```

```
        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 4.0081, df = 12.217, p-value = 0.001677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.582009 12.077991
sample estimates:
mean of x mean of y
     3.50     -4.33
```

Conclusion

Zero is not included. We are 95% confident that the true difference in mean change in heart pumping capacity is 3.58 to 12.077 litres/minute higher for the ESC group than the control group.

20.3.3 Observed statistic vs. null statistics

The T score is a ratio of how the groups differ as compared to how the observations within a group vary.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

When the null hypothesis is true and the conditions are met, T has a t-distribution with $df = min(n_1 - 1, n_2 - 1)$.

Conditions:

- Independent observations within and between groups.
- Large samples and no extreme outliers.

But we will not be using this formula – just use R code.

Chapters 20 interactive notes

Chapters 20 interactive notes

R Saidi

Chapter 20 - Difference of Two Groups' Means

load the libraries and dataset

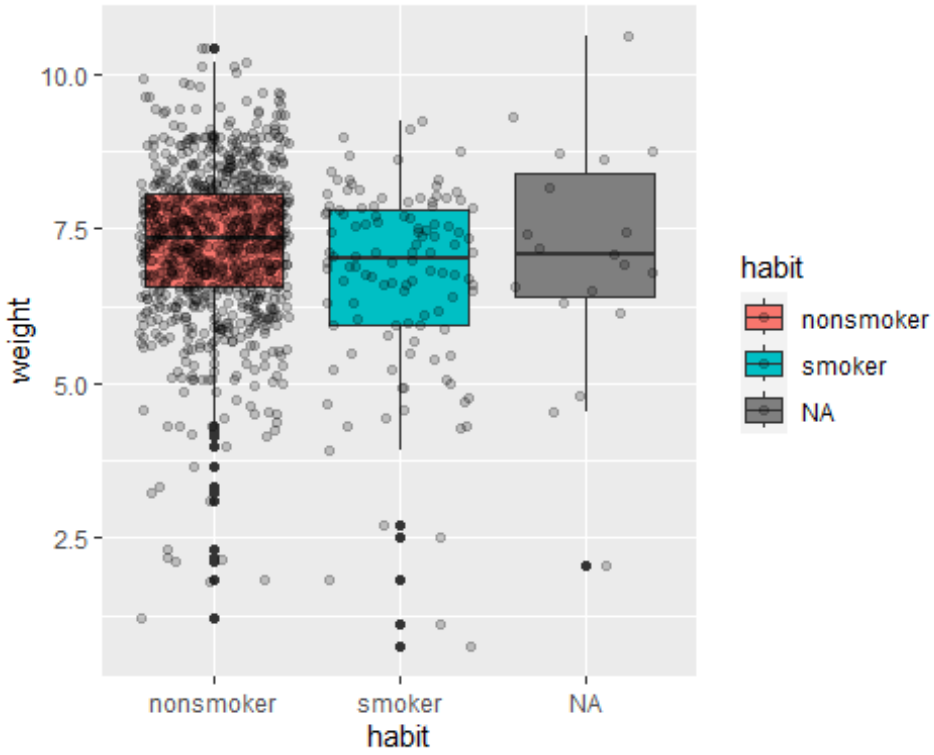
```
library(tidyverse)
library(openintro)
library(tidymodels)

data(births14)
data(stem_cell)
```

This dataset is a sample of 1000 observations.

Check basic assumptions

Use a side-by-side boxplot to compare the two groups' distributions

```
ggplot(births14, aes(x=habit, y=weight, fill = habit))+
  geom_boxplot()+
  geom_jitter(alpha = .2)
```

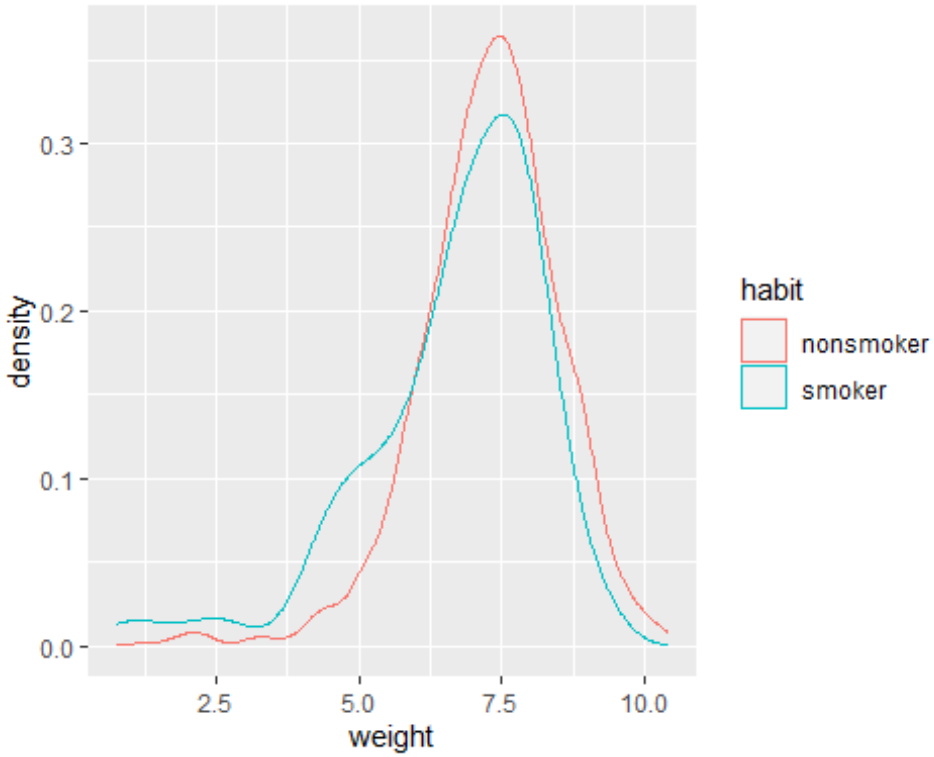


Notice there are some na values. Let's remove those.

```
births14_a <- births14 |>
  filter(!is.na(habit))
unique(births14_a$habit)

[1] "nonsmoker" "smoker"

births_table <- births14_a |>
  group_by(habit) |>
  count()
births_table

# A tibble: 2 × 2
# Groups:   habit [2]
  habit         n
  <chr>     <int>
1 nonsmoker   867
2 smoker      114
```

Now replot, but use density plots

```
ggplot(births14_a, aes(weight, color = habit))+
  geom_density()
```



We can see the density plots are skewed left, but the sample size is 1000, so it passes the CLT)

t.test for difference of means

Use a t.test for difference of means to compare mean birth weights of babies born to smoking mothers versus non-smoking mothers.

Use the syntax: t.test(y ~ x)

```
t.test(births14_a$weight ~ births14_a$habit)


    Welch Two Sample t-test

data:  births14_a$weight by births14_a$habit
t = 3.8166, df = 131.31, p-value = 0.0002075
alternative hypothesis: true difference in means between group nonsmoker and group smoker
is not equal to 0
95 percent confidence interval:
 0.2854852 0.8998751
sample estimates:
mean in group nonsmoker    mean in group smoker
               7.269873                6.677193
```

t = 3.8166, df = 131.31, p-value = 0.0002075

Alternate code using t_test

```
?t_test

starting httpd help server ... done
```

```
# use package: infer (in tidymodels)
births14_a |>
  t_test(response = weight,
         explanatory = habit,
         null = 0,
         alternative = "two-sided",
         conf_int = FALSE)
```

```
Warning: The statistic is based on a difference or ratio; by default, for
difference-based statistics, the explanatory variable is subtracted in the
order "nonsmoker" - "smoker", or divided in the order "nonsmoker" / "smoker"
for ratio-based statistics. To specify this order yourself, supply `order =
c("nonsmoker", "smoker")`.
```

```
# A tibble: 1 × 5
  statistic  t_df  p_value alternative estimate
      <dbl> <dbl>    <dbl> <chr>          <dbl>
1      3.82  131. 0.000208 two.sided      0.593
```

Alternate code using randomization

use specify, hypothesize, generate, calculate

Calculate the difference in observed means

Use specify and calculate

```
# Calculate observed difference in means
diff_mean_obs <- births14_a |>
  # Specify the response and explanatory variables
  specify(weight ~ habit) |>           # syntax is y ~ x
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))
diff_mean_obs
```

```
Response: weight (numeric)
Explanatory: habit (factor)
# A tibble: 1 × 1
   stat
  <dbl>
1 0.593
```

```
diff_ht_mean <- births14_a |>
  specify(weight ~ habit) |>                    # syntax is y ~ x
  hypothesize(null = "independence") |>         # set the null
  generate(reps = 1000, type = "permute") |>    # shuffle 1000 times
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))
# Specify to calculate a difference in means and what order of subtraction to use
```

View a histogram of randomized distribution under the null

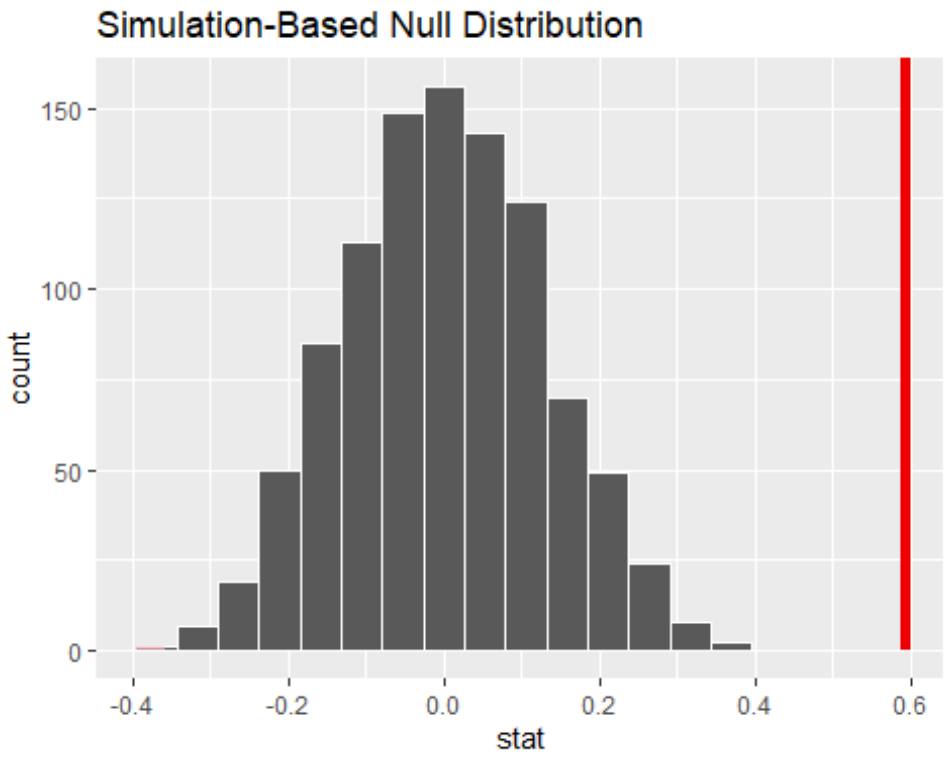
```
ggplot(diff_ht_mean, aes(stat)) +
  geom_histogram() +
  geom_vline(xintercept = pull(diff_mean_obs), color = "red")
```

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

Calculate the p-value

```
diff_ht_mean |>
  get_p_value(obs_stat = diff_mean_obs, direction = "two-sided")
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
based on the number of `reps` chosen in the `generate()` step.
```



Example using stem_cell data

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 20.2 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study.

Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

Group	n	Mean	SD
ESC	9	3.50	5.17
Control	9	-4.33	2.76

When you only have summary statistics to work with

Use the function tsum.test(x-bar1, s1, n1, x-bar2, s2, n2) in the package BSDA.

```
library(BSDA)
```

```
Warning: package 'BSDA' was built under R version 4.3.3
```

```
Loading required package: lattice
```

```

Attaching package: 'lattice'
```

```
The following objects are masked from 'package:openintro':

    ethanol, lsegments
```

```
Attaching package: 'BSDA'

The following object is masked from 'package:datasets':

    Orange
```

```
tsum.test(3.5, 5.17, 9, -4.33, 2.76, 9)
```

```

    Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 4.0081, df = 12.217, p-value = 0.001677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.582009 12.077991
sample estimates:
mean of x mean of y
     3.50     -4.33
```

Bootstrap code

```
diff_mean_ci <- births14_a |>
  # Specify weight vs. habit
  specify(weight ~ habit) |>
  # Generate 1500 bootstrap replicates
  generate(reps = 1500, type = "bootstrap") |>
  # Calculate the difference in means, nonsmoker then smoker
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))
```

Now calculate the 95% CI

```
# Calculate the 95% CI via percentile method
diff_mean_ci |>
  get_confidence_interval(level = 0.95, type = "percentile")
```

```
# A tibble: 1 × 2
  lower_ci upper_ci
     <dbl>    <dbl>
1    0.292    0.901
```

Conclusion:

Zero is not included. We are 95% confident that the true difference in mean weight for babies is between 0.31 and 0.90 lbs higher for nonsmoking mothers versus smoking mothers.

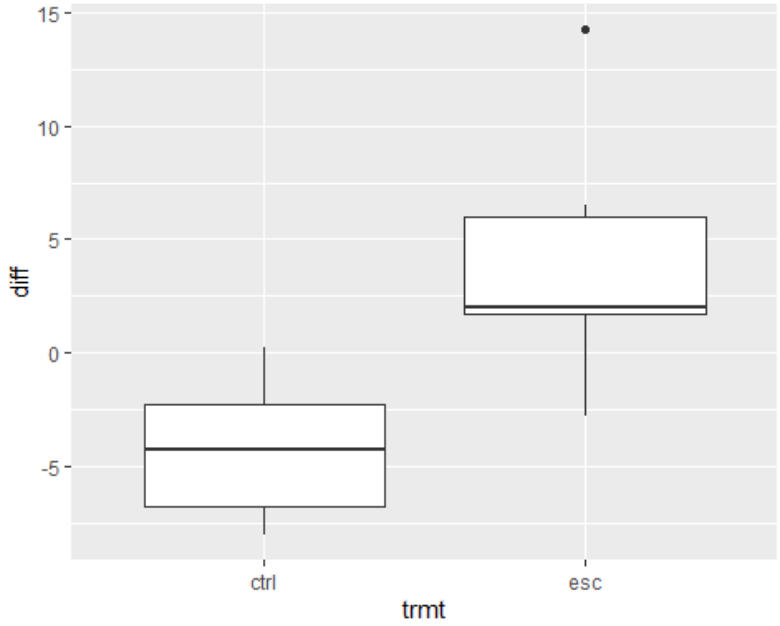
```
stem <- stem_cell |>
  mutate(diff = after - before)
head(stem)
```

```
# A tibble: 6 × 4
  trmt  before after  diff
  <fct>  <dbl> <dbl> <dbl>
1 ctrl    35.2  29.5 -5.75
2 ctrl    36.5  29.5 -7
3 ctrl    39.8  36.2 -3.5
```

```
4 ctrl    39.8  38   -1.75
5 ctrl    41.8  37.5 -4.25
6 ctrl    45    42.8 -2.25
```

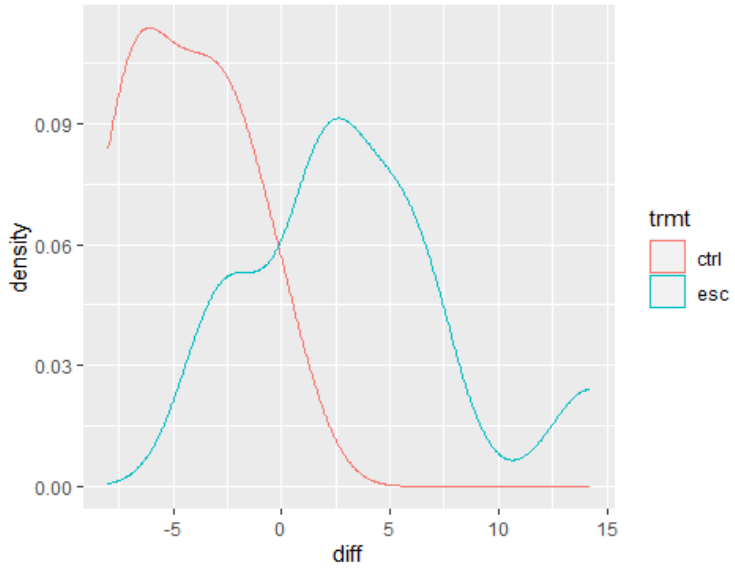
Side-by-side boxplots

```
ggplot(stem, aes(trmt, diff))+
  geom_boxplot()
```



histograms/density plots

```
ggplot(stem, aes(diff, color = trmt))+
  geom_density()
```



When the basic assumptions fail for a t.test

When the sample sizes for each group are small AND the distributions are not bell-shaped.

Non-parametric approach

Wilcox Mann Whitney Test

Null: There is no difference in heart pumping capacity between control and treatment groups.

Alternative: There is a shift in distributions between the control and treatment groups for heart pumping capacity.

```
wilcox.test(stem$diff ~ stem$trmt)
```

```
Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties


    Wilcoxon rank sum test with continuity correction

data:  stem$diff by stem$trmt
W = 6, p-value = 0.002666
alternative hypothesis: true location shift is not equal to 0
```

Conclusion: Because the p-value is very small, there is a right shift in the difference in heat pumping capcity after treatment for the ESC group over the control group.

Homework Chapter 20

1. Review section 20.5 (the chapter review)
2. Suggested problems from textbook section 20.6 exercises: 1, 3, 4, 5, 9, 11, 15