Homework 1

Spring 2023

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday April 12 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.

Exercises

1. (Simulation noise) Dustin is doing simulations to see how well the 95% z-confidence interval covers the true value of the population mean μ_0. Dustin simulates $B = 10,000$ samples, each of size n, from a population distribution, and for each sample he calculates the z-confidence interval, and then notes whether the confidence interval contains the true value for μ_0.

Let X_i denote whether the ith z-confidence interval covers the true value, then

$$\bar{X} = \frac{1}{B}\sum_{i=1}^{B} X_i$$

denotes the simulated coverage rate.

How high or low must the simulated coverage rate be for Dustin to suspect that the true coverage rate is not 95%? Explain. Assume we are using the usual threshold of significance $\alpha = 0.05$. (Hint: Each X_i is a Bernoulli random variable with success probability π_0. What are we hypothesizing about π_0?)

Setup:
If we let X_i denote whether the ith z-confidence interval covers the true value, then the simulated coverage rate is $\bar{X} = \frac{1}{B}\sum_{i=1}^{B} X_i$ where $B = 10000$. Hence it follows that $X_i \sim Binom(n = 1, \pi_0)$, where π_0 is the true coverage rate of the z-confidence interval. Since we want to find the values of π_0 that would make us suspect the true coverage rate is not 95%, it follows that we are looking at a two sided alternative. Thus our competing hypotheses are: $H_0 : \pi_0 = 0.95,\ H_1 : \pi_0 \neq 0.95$. Since $X_i \sim Bernoulli(\pi_0)$ it follows that $E[X_i] = \pi_0$, thus we are concerned about the mean of the distribution. Since $B = 10000$ is large enough for $\bar{X}$ to be approximately normal, we will run a Z test to assess these hypotheses. In particular, 10000 is sufficiently large for the central limit theorem to hold, thus $\bar{X} \approx Norm\left(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{10000}}\right)$.

Furthermore, under the assumption that the null hypothesis is true, $\bar{X} \approx Norm\left(0.95, \sqrt{\frac{0.95 \cdot 0.05}{10000}}\right)$. To find how high or low the simulated coverage rate must be for Dustin to suspect that the true coverage rate is not 95%, we must find the critical values/endpoints of the rejection region. We can do this using the qnorm function in R.

Calculation:

```
lower <- qnorm(p = 0.025, mean = 0.95, sd = sqrt(0.0475 / 10000))
upper <- qnorm(p = 0.975, mean = 0.95, sd = sqrt(0.0475 / 10000))
```

Based on the above calculation, if the simulated coverage rate Dustin finds is above 0.9542716 or lower than 0.9457284, Dustin has significant evidence that the true coverage probability is not 95%.

2. (Chick weights) The `chickwts` dataframe in the **fastR2** package presents results from an experiment in which chickens are fed six different diets. If we assume that the chickens were randomly sampled from some population and also were assigned to the feed groups at random, then for each feed, we can consider the chickens fed that feed to be a random sample from the (conceptual) population that would result from feeding all chickens that particular feed.

a. For each of the 6 feeds, compute 95% confidence intervals for the mean weight of chickens fed that feed. (Use `t_test` from the package **infer** to print the results neatly. Set `options(pillarsig.fig = 6)` to format the printing of the resulting tibble.)

Read in the data:

```
chick_data <- chickwts
```

Find the intervals:

```
options(pillar.sigfig = 6)

grouped <- chick_data %>%
                group_by(feed) %>%
                do(infer::t_test(
                  x = .,
                  response = weight,
                  mean = mean(weight),
                  alternative = "two.sided",
                  conf_level = 0.95)[6:7]
)

grouped
```

```
## # A tibble: 6 x 3
## # Groups:   feed [6]
##   feed      lower_ci upper_ci
##   <fct>        <dbl>    <dbl>
## 1 casein     282.644  364.523
## 2 horsebean  132.569  187.831
## 3 linseed    185.561  251.939
## 4 meatmeal   233.308  320.510
## 5 soybean    215.175  277.682
## 6 sunflower  297.888  359.946
```

For completeness, here is a different way to get the same interval:

```
chick_data %>%
  group_by(feed) %>%
  summarise(lower = mean(weight) - qt(p = 0.975,
                                      df = length(feed) - 1) * sd(weight) / sqrt(length(feed)),
            upper = mean(weight) + qt(p = 0.975,
                                      df = length(feed) - 1) * sd(weight) / sqrt(length(feed)))
```

b. From a visual examination of the six intervals, is there convincing evidence that some diets are better

(lead to more weight gain) than others? Why or why not? (You will learn about the Analysis of Variance method to answer this question in STAT 421)

Since we assume that the chickens were randomly sampled from some super-population and also were assigned to the feed groups at random, then for each feed, we can consider the chickens fed that certain feed to be a random sample from the conceptual super-population that would result from feeding all chickens that particular feed. Thus, we can think of any differences between chickens in a group as occurring in all groups.

Since we are 95% confident that each interval contains the true mean weight of a chicken being fed each respective feed, we must realize that in order to assess if some feeds are better for weight gain than others by looking at the 95% confidence intervals, we must look at each interval compared to the others and see if and where they overlap. For example, when comparing casein and sunflower to horsebean, we can see that the interval for casein and sunflower have means way larger than horsebean, as well as having confidence intervals that are much farther away than horsebean with zero overlap. Thus, we have convincing evidence that the means of the latter two groups are larger than horsebean, and thus lead to more weight gain. For example, if we were told that the true mean value of chick weight fell into all of the intervals, we'd know that the means of sunflower and casein have no chance of being remotely close to that of horsebean, and thus we would conclude that the latter two feeds are better at weight gain for chicks than horsebean. Furthermore, even in a worst case scenario where we are told that the true mean of horsebean is above the interval and the true mean of casein and sunflower is below their respective intervals, the spread between the two intervals is so large with no overlap that we would still be safe in assuming that the true mean weight of chickens when fed casein and sunflower is much higher than that of horsebean, and thus better for weight gain.

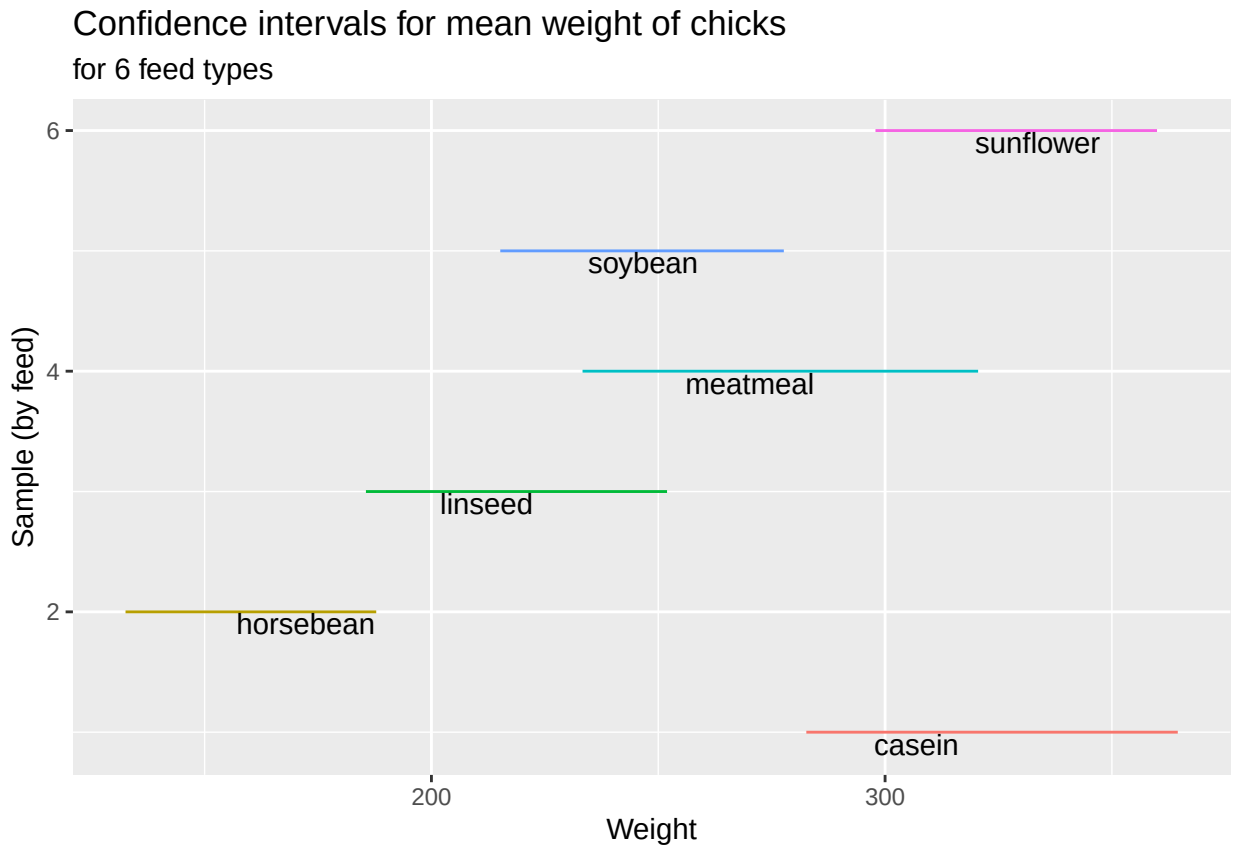
However, this can't be said about all intervals, as the intervals for soybean, linseed, and meatmeal mutually intersect each other in a certain range, so we can't say with any reasonable degree of certainty that any of these feeds are better than the other. This is because even if we were told that the respective true parameter for each feed lies in these intervals, we would still not know **where** in the interval they lie, thus it is very possible that they all land in the section where the intervals mutually intersect, and thus we would know each respective feed wouldn't be better at weight gain than the rest.

In summary, we have evidence that some feeds are better than others, like casein and sunflower both being better than horsebean and linseed, while at the same time we have non-convincing evidence that feeds like soybean, meatmeal, and linseed are better than each other when it comes to the weight gain of chicks.

The below plot shows in particular which intervals "overlap" with each other, and provides some clarity to the lengthy explanation above.

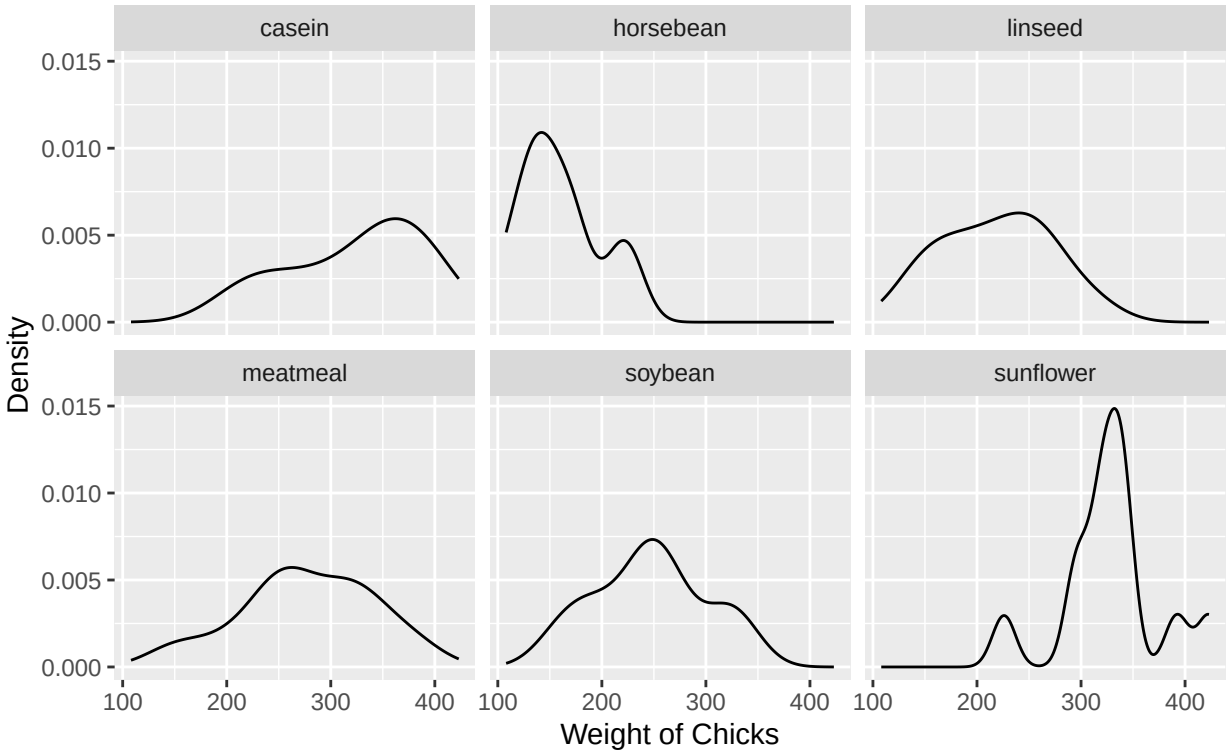
Plot of intervals:

```
ggplot(data = grouped) +
  geom_segment(mapping = aes(x = lower_ci, xend = upper_ci, y = 1:6, yend = 1:6,
                             color = feed)) +
  labs(title = "Confidence intervals for mean weight of chicks",
       subtitle = "for 6 feed types",
       x = "Weight",
       y = "Sample (by feed)") +
  annotate("text",
         x = grouped$lower_ci,
         y = (1:6) - 0.1,
         label = paste0(grouped$feed),
         hjust = -0.8) +
  theme(legend.position = "none")
```



c. Are there any features of the data that might suggest that a t-distribution may not be entirely appropriate? The following incomplete code should help you make a density plot of the weight distribution by the feed. (I want to see references to what you learned from the simulation in Problem Set 1)

```
ggplot(data = chick_data,
       mapping = aes(x = weight)) +
  geom_density() +
  facet_wrap(facets = vars(feed) ) +
  labs(title = "Density Plots of the Weight of Chicks",
       subtitle = "For 6 different feeds",
       x = "Weight of Chicks",
       y = "Density")
```



One of the key assumptions when running a t-test is that the underlying distribution that our data are being drawn from is normal. However, in Problem Session 1 we found that even when our data isn't drawn from a normal distribution, as the number of data points increases, the distribution of the statistic $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S}$ becomes more and more like the t-distribution. This happens because by the central limit theorem the distribution of $\bar{X}$ becomes approximately normal, and thus $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S}$ converges to a standard normal just like the t distribution does.

However, looking at the above density plots, the distributions of chick weights when fed casein, or horsebean are not symmetric, while the distribution of chick weights when fed sunflower is somewhat symmetric but definitely not normal. The distributions of chick weights when fed linseed, meatmeal, and soybean are the only distributions that could be seen as approximately normal. Thus it would be inappropriate to assume that all of the distributions of weight by feed follow a normal distribution.

Before we can completely make a decision on if the data might suggest that a t-distribution may not be entirely appropriate, we need to look at the sample sizes for each feed type.

```
chick_data %>%
  group_by(feed) %>%
  summarise(sampple_size = length(weight))
```

```
## # A tibble: 6 x 2
##   feed      sampple_size
##   <fct>            <int>
## 1 casein              12
## 2 horsebean           10
## 3 linseed             12
## 4 meatmeal            11
## 5 soybean             14
```

```
## 6 sunflower           12
```

As can be seen from the above data frame, the sample sizes for each feed range between 10 and 14.

Due to the fact that sample sizes are small, and the fact that the distribution of the weights are clearly not normal for half of the feeds, these features might suggest that a t-distribution may not be entirely appropriate to use for inference for all of the feeds.

3. (Psychology of Rats) Does the psychological environment affect the anatomy of the brain? The subjects for one study came from a genetically pure strain of rats. From each litter, one rat was selected at random for the treatment group and one for the control group. Both groups got the same food and drink – as much as they wanted. But each animal in the treatment group lived with 11 others in a cage, furnished with playthings which were changed daily. Animals in the control group lived in isolation, with no toys. After a month, all animals were sacrificed and their cortex weights measured in milligrams. The data set is in the file `brain-weights.csv`.

a. Why did the investigators decide to assign one member of each litter to treatment and another member from the same litter to the control group? What are the advantages?

The investigators decided to assign one member of each litter to treatment and another member from the same litter to the control group in order to analyze the treatment effect of psychological environment on the anatomy of the brain. In particular, they paired these rats up to run a paired t-test. Since each rat is a sibling meaning they share similar DNA, it is reasonable to pair these two rats together for experimenting. Furthermore, due to these similar environmental characteristics that each rat experienced with its sibling, we will see a positive association between each rat and its sibling. The advantages of pairing these rats together is that it lets the investigators compare the two treatments, those being living together versus isolation, while keeping the confounding variables the same. Thus, if they see a difference between cortex weights, then they can conclude that it must be due to differences in the effect of the treatment.

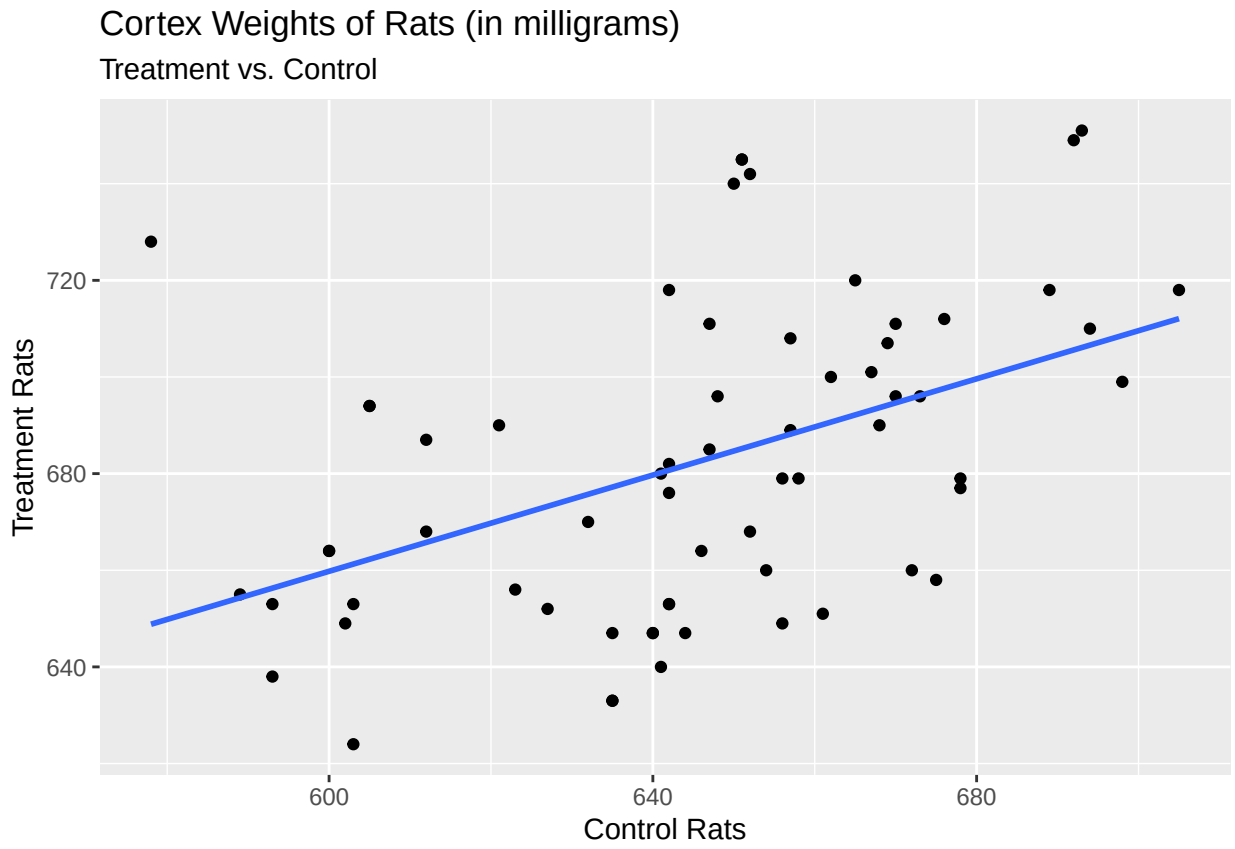
b. Explore these data by making a scatterplot, a boxplot, and calculating some summary statistics. Write briefly about what you are looking for in these plots. (be sure to show your code and output - sans error/warning messages; label your plots; keep your explanation pointed - this means just talk about what's important.)

Read in the data:

```
rat_data <- read_csv("brain_weights.csv")
```

Scatterplot:

```
ggplot(data = rat_data,
       mapping = aes(x = control, y = treatment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Control Rats",
       y = "Treatment Rats",
       title = "Cortex Weights of Rats (in milligrams)",
       subtitle ="Treatment vs. Control")
```



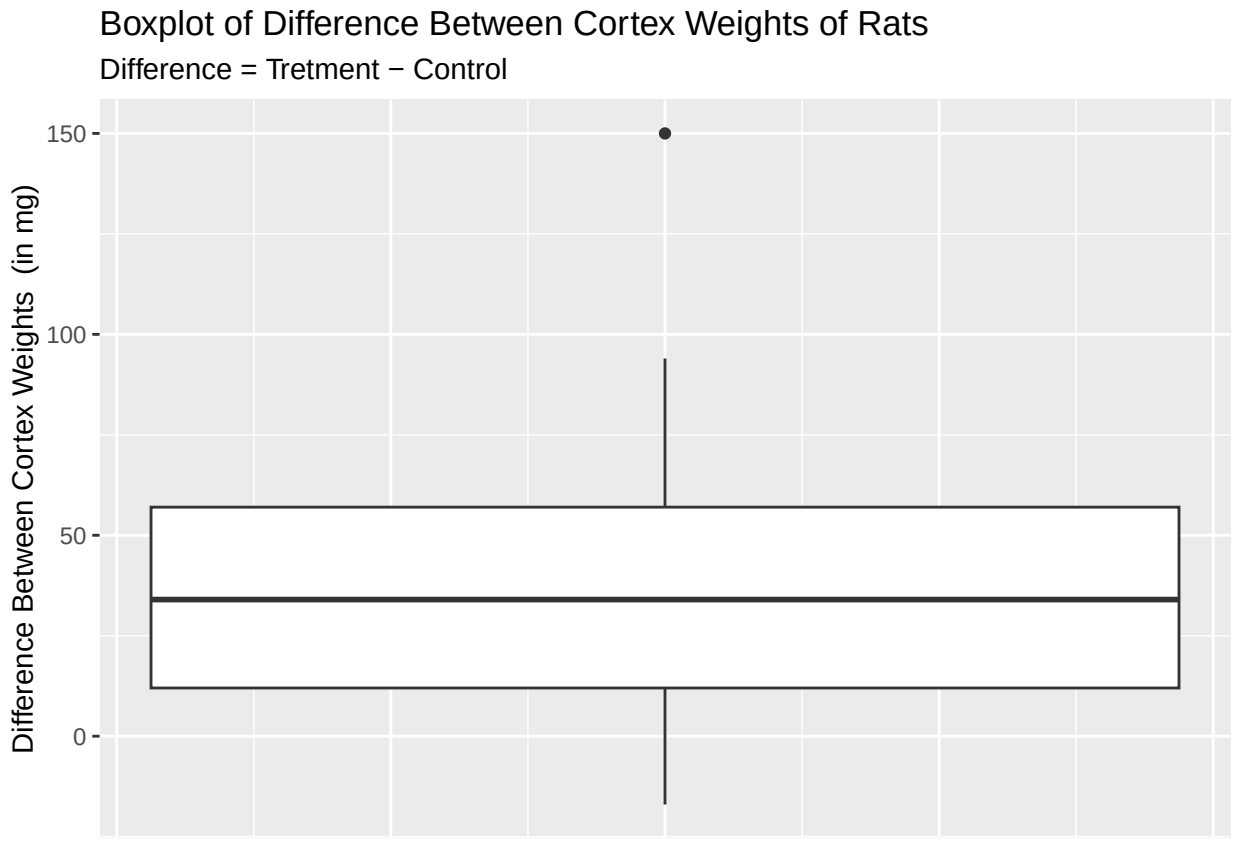
The above scatterplot shows that there is some positive correlation between the cortex weight in the pairs of rats in treatment and control. However, since we are concerned about the difference between these two weights, in order to assess if the treatment is increasing these weights, we will need a better way to visualize this difference.

Express each pair as a difference:
In order to assess anything about the differences between the cortex weight in treatment and control, we will need to calculate these differences in our sample data.

```
rat_data_diff <- rat_data %>%
                     mutate(diff = treatment - control)
```

Boxplots:

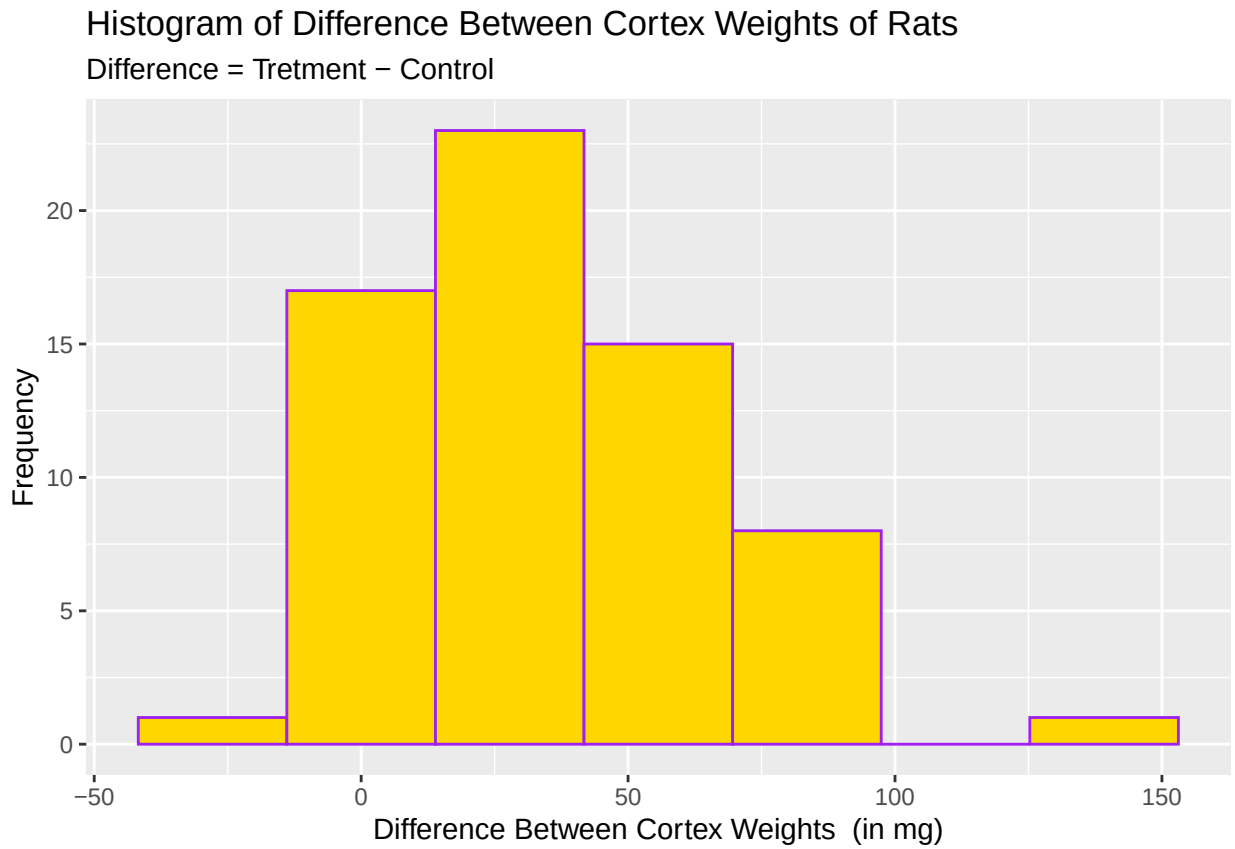
```
ggplot(data = rat_data_diff,
       mapping = aes(y = diff)) +
  geom_boxplot() +
  labs(y = "Difference Between Cortex Weights  (in mg)",
       title = "Boxplot of Difference Between Cortex Weights of Rats",
       subtitle = "Difference = Tretment - Control") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```



The boxplot shows that on average, the weight of the cortex in the treatment group is higher than that of the control group. However, as we can see there is an extreme outlier and the data is fairly spread out. Furthermore, we can see that the difference between the treatment and the control weight of the cortex seems pretty symmetrical.

Histogram:

```
ggplot(data = rat_data_diff,
       mapping = aes(x = diff)) +
  geom_histogram(bins = 7,
                 color = "purple",
                 fill = "gold") +
  labs(x = "Difference Between Cortex Weights  (in mg)",
       y = "Frequency",
       title = "Histogram of Difference Between Cortex Weights of Rats",
       subtitle = "Difference = Tretment - Control")
```



As talked about in the boxplot analysis, the data is fairly symmetric, however looking at the histogram it is more apparent that there is a small right skew which is mainly being controlled by the one major outlier.

Summary statistics:

```
mode <- function(x) {
  uniqx <- unique(x)
  uniqx[which.max(tabulate(match(x, uniqx)))]
}

summary_stats <- rat_data_diff %>%
                          summarise(mean = mean(diff),
                                    sd = sd(diff),
                                    median = median(diff),
                                    iqr = iqr(diff),
                                    mode = mode(diff),
                                    min = min(diff),
                                    max = max(diff))
summary_stats
```

```
## # A tibble: 1 x 7
##      mean      sd median   iqr  mode   min   max
##     <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 36.9538 32.4189     34    45    38   -17   150
```

Again, since we see that the mean, median and mode are all similar we have evidence that the distribution of the differences of the cortex weight between the treatment group and control group is nearly symmetric. The large standard deviation and inter quartile range further show the large spread of the differences.

In conclusion, we have evidence to assume that D_i, the difference between the cortex weight of the treatment group and the control group is approximately normal. In particular, $D_i \sim Norm(\mu_d, \sigma_d)$.

c. The goal is to examine if the treatment increases cortex weight. Two different analytic strategies are described below. Conduct both analyses, and summarize the conclusions.

Explanation of significance level:
In the below methods no α level is explicitly given, thus we will have to make our own call for an appropriate value for α. Since the $P(\text{Type I}) = \alpha$ we will need to assess the importance of a type one error in the context of this experiment. In this specific case a type one error occurs when we say that the treatment increases cortex weight when in reality there is no difference in the average treatment effect. Hence, there could be serious ramifications of falsely sending out a drug for use that actually has no effect, thus we will want a relatively small α level for the below tests. It is standard in clinical trials to have an α level of 0.01. This will be used to conduct the below significance tests.

- Method 1: Dichotomize the data for each pair as "1" if treatment cortex is heavier and "0" otherwise. (Ignore ties in the data if any.) Then use a binomial model to test $H_0 : \pi_0 = 0.5$ versus $H_1 : \pi_0 > 0.5$ where π_0 is the probability that the treatment cortex is heavier. (This method is called a **sign test** since we are recording whether the sign of the difference in weights - treatment minus control - is positive or not.)

Dichotomize the data:

```
rat_data_dich <- rat_data %>%
                      subset(treatment != control) %>%
                      mutate(dichotomize = if_else(treatment > control, 1, 0))
```

Calculate and assess the results:
Now that we've dichotomized the data, we will use a binomial model to test $H_0 : \pi_0 = 0.5$ versus $H_1 : \pi_0 > 0.5$ where π_0 is the probability that the treatment cortex is heavier. Let S denote the number of pairs where the treatment cortex is heavier than the control cortex. Under the assumption that the null hypothesis is true, and under the assumption that the different pairs of rats are independent from each other, $S \sim Binom(n = 65, \pi = 0.5)$. To asses whether or not we believe the the treatment increases cortex weight, we will find the p-value associated to our observed value of S from the sample under the null hypothesis.

It is important to note that since π_0 is the probability that the treatment cortex is heavier, and $H_1 : \pi_0 > 0.5$, values in the upper tail will be evidence in favor of H_1.

```
obs_s <- sum(rat_data_dich$dichotomize)
method1_pval <- pbinom(q = obs_s - 1,  size = 65, prob = 0.5, lower.tail = F)
```

As computed above, our p-value was $1.5816216 \times 10^{-10}$. Since this p-value is very low, at most significance levels we have significant evidence that $\pi_0 > 0.5$ where π_0 is the probability that the treatment cortex is heavier. In particular, we reject at our predefined significance level of 0.01, which shows that the probability that the treatment cortex is heavier is greater than 50%, which is evidence that the treatment increases cortex weight.

- Method 2: Express the data for each pair as the difference, D in cortex weights between the treatment and control animal. Then conduct a paired t-test of $H_0 : \mu_d = 0$ versus $H_1 : \mu_d > 0$ where μ_d is the expected value of D.

Setup:
As we analyzed in the previous problem, it is reasonable to assume that $D_i \sim Norm(\mu_d, \sigma_d)$. Since we are assuming that D_i is normal, it follows that we can use a one sample t-test with D_i as the data and the test statistic $T = \frac{\sqrt{65}(\bar{D}-0)}{S_d}$ where S_d is the sample standard deviation of the difference. Our two hypotheses that we will be testing are $H_0 : \mu_d = 0$ versus $H_1 : \mu_d > 0$ where μ_d is the expected value of D. Under the assumption that the null hypothesis is true, it follows that $T \sim t_{64}$.

Calculate and assess the results:

Now that we have setup the problem, we can run a paired t-test in R and interpret the results from there:

```
method2_pval <- rat_data_diff %>% infer::t_test(response = diff,
                                                mu = 0,
                                                alternative = "greater")
```

As computed above, our p-value was $1.3190414 \times 10^{-13}$. Since this p-value is very low, at most significance levels we have significant evidence that $\mu_0 > 0$. In particular, we reject at our predefined significance level of 0.01. Since we rejected the null hypothesis, we have significant evidence that on average the difference of the cortex weight of a rat in the treatment group and a rat in the control group is above zero, hence we have evidence that the treatment increases cortex weight in rats.

d. What are some advantages/disadvantages of the sign test compared with the paired t-test?

Advantages:
One of the main and obvious advantages of the sign test compared with the paired t-test is that it makes no assumptions about the distribution of the data, it merely performs a condition on the data and classifies the data based on that condition. On the other hand, the paired t-test makes an assumption that the data is normal or at the very least approximately normal. Thus, you can always perform a sign test, while on the contrary, if you don't meet the assumptions for a paired t-test you can't run it or you run the risk of obtaining inaccurate/misleading results by carrying out the test with flawed assumptions.

Disadvantages:
One of the main disadvantages of the sign test compared with the paired t-test is that the sign test merely assigns a value of 1 or 0 to each observation, according to whether it meets some condition while completely ignoring the magnitude of the meeting or failing of said condition. For example, in the above method 1, we completely ignored the magnitude of this difference and instead focused solely on the fact if the difference was greater than zero. Although we haven't talked about it in this class, the problem with ignoring the magnitude of the difference is that this may increase the probability of a type 2 error, and thus decrease the statistical power of the test.

4. Suppose X and Y are jointly distributed with variances σ_X^2 and σ_Y^2, respectively. The correlation coefficient ρ of X and Y is defined by

$$\rho = \frac{Cov\,(X, Y)}{\sigma_X\,\sigma_Y}.$$

a. Consider the random variable

$$Z = \frac{Y}{\sigma_y} - c\frac{X}{\sigma_X}.$$

Show that $c = \rho$ is the value of c which minimizes $Var(Z)$. (Hint: From defining principles $Var(Z) = E\left[(Z - E\,[Z])^2\right]$)

To show $c = \rho$ is the value of c which minimizes $Var(Z)$, we will use the second derivative test. First, we will

use the defining principle of variance and manipulate the equation $Var(Z) = E\left[(Z - E[Z])^2\right]$.

$$\begin{aligned}
Var(Z) &= E\left[(Z - E[Z])^2\right] \\
&= E[((\frac{1}{\sigma_Y}Y - \frac{c}{\sigma_X}X) - E[\frac{1}{\sigma_Y}Y - \frac{c}{\sigma_X}X])^2] \\
&= E[((\frac{1}{\sigma_Y}Y - \frac{c}{\sigma_X}X) - (\frac{1}{\sigma_Y}E[Y] - \frac{c}{\sigma_X}E[X]))^2] \\
&= E[((\frac{1}{\sigma_Y}Y - \frac{c}{\sigma_X}X) + (\frac{-1}{\sigma_Y}E[Y] + \frac{c}{\sigma_X}E[X]))^2] \\
&= E[((\frac{1}{\sigma_Y}Y - \frac{1}{\sigma_Y}E[Y]) - (\frac{c}{\sigma_X}X - \frac{c}{\sigma_X}E[X]))^2] \\
&= E[(\frac{1}{\sigma_Y}(Y - E[Y]) - \frac{c}{\sigma_X}(X - E[X]))^2] \\
&= E[(\frac{1}{\sigma_Y^2}(Y - E[Y])^2 + \frac{c^2}{\sigma_X^2}(X - E[X])^2 - \frac{2c}{\sigma_X\sigma_Y}(X - E[X])(Y - E[Y])] \\
&= \frac{1}{\sigma_Y^2}Var[Y] + \frac{c^2}{\sigma_X^2}Var[X] - \frac{2c}{\sigma_X\sigma_Y}Cov[X, Y] \\
&= 1 + c^2 - 2c\rho \quad (\text{Since } Var[X] = \sigma_X^2,\ Var[Y] = \sigma_Y^2, \text{ and } \rho = \frac{Cov\,(X, Y)}{\sigma_X\ \sigma_Y})
\end{aligned}$$

Now we can think of the equation for $Var[Z]$ as a function of c, then with this we can take derivatives to find the critical value and show that it's a minimum. Let $f(c) = Var[Z]$:

$$\begin{aligned}
f(c) &= 1 + c^2 - 2c\rho \\
f'(c) &= 0 + 2c - 2\rho \\
2c &= 2\rho \\
c &= \rho
\end{aligned}$$

Now, in order to show that $c = \rho$ minimizes $Var[Z]$, we must show that $f''(\rho) > 0$:

$$\begin{aligned}
f'(c) &= 2c - 2\rho \\
f''(c) &= 2 - 0 \\
f''(c) &= 2
\end{aligned}$$

Since $f''(c) = 2$, it follows that $f''(\rho) = 2 > 0$, thus $c = \rho$ minimizes $Var[Z]$.

b. What is the minimal value of this variance?

As shown above, $Var[Z] = 1 + c^2 - 2c\rho$ is minimized when $c = \rho$. Thus the minimal value of this variance is $Var[Z] = 1 + \rho^2 - 2\rho^2 = 1 - \rho^2$.

For completeness, we can compute the hypothetical smallest value that $Var[Z]$ can take. The smallest value that $Var[Z]$ could possibly take occurs when ρ^2 is maximized. Since $\rho \in [-1, 1]$, this implies that $\rho^2 \in [0, 1]$, thus the maximum value of ρ^2 is 1 and the smallest possible value of $Var[Z]$ is $1 - 1^2 = 0$.