Homework 1

Spring 2023

KEY

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?)

   We assume each $X_i \sim Bernoulli(\pi_0)$.

   The null hypothesis is $H_0 : \pi_0 = .95$ and the alternative hypothesis is $H_1 : \pi_0 \neq .95$.

   When the null is true, by the Central Limit Theorem (since B is large) we can say:

   $$\bar{X} \sim N\left(mean = 0.95, \quad sd = \sqrt{\frac{0.95(1-0.95)}{B}}\right)$$

and a size $\alpha = 0.05$ significance test will reject H_0 if the P-value is smaller than 0.05.

The P-value is the two-sided tail probability under the null sampling distribution. It will be smaller than $\alpha = 0.05$ if the observed simulated coverage - $\bar{x}$ - is smaller than the 2.5th percentile of the null sampling distribution or larger than the 97.5th percentile.

In other words, we will reject H_0 when

$$\bar{x} < 0.95 - z_{0.025}\sqrt{\frac{0.95(1-0.95)}{B}} = 0.946$$

or

$$\bar{x} > 0.95 + z_{0.025}\sqrt{\frac{0.95(1-0.95)}{B}} = 0.954$$

where $z_{\alpha/2}$ refers to the standard normal quantile with an area of $1-\alpha/2$ below it.

In summary, if Dustin observes a simulated coverage rate smaller than 94.6% or larger than 95.4%, he should be suspicious of whether the confidence interval formula he is using has a 95% confidence level.

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.)

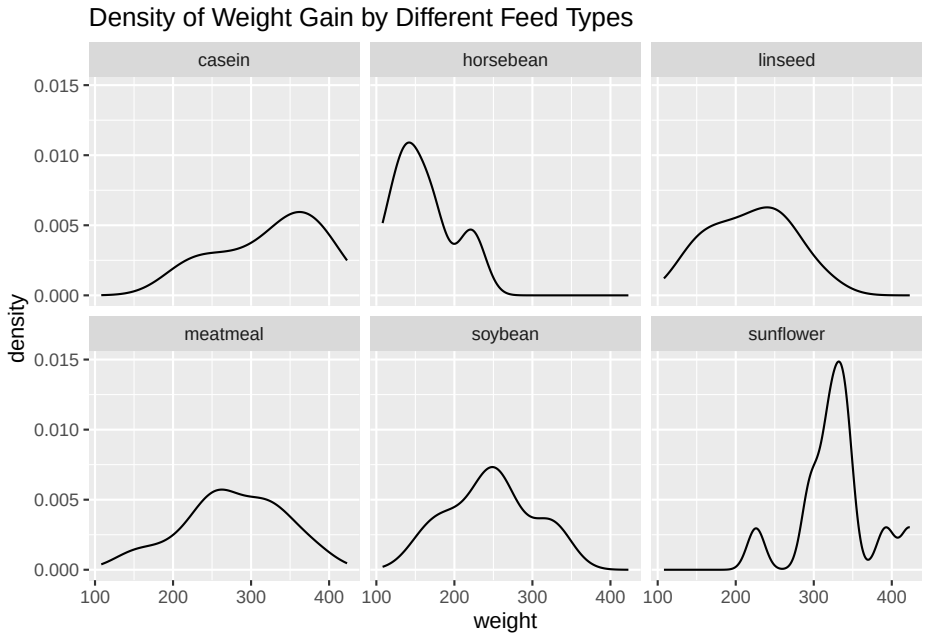
```
library(infer)
options(pillar.sigfig = 6)
#defaults to conf_level = .95 so omitted
#code thanks to Jaiden Attenbury!
chickwts %>%
  group_by(feed) %>%
  do(infer::t_test(x = .,
                   response = weight,
                   mean = mean(weight),
                   alternative = "two.sided",
                   conf_level = 0.95)[6:7])
```

```
## # 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
```

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)

We can be 95% confident, that if one interval is strictly greater (to the right of) than another interval, then that feed will lead to larger weight gain. From above, we can say that for example, casein and sunflower are better than horsebean, linseed, and soybean since their CI's do not intersect. We also can say that any of soybean and meatmeal are better than horsebean.

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)



We see that t-distributions are valid when the data comes from a normal distribution. From these plots we see that many of these feeds (especially casein, horsebean, sunflower) do not have a bell curve shape, and thus could violate the normal assumption.

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`.

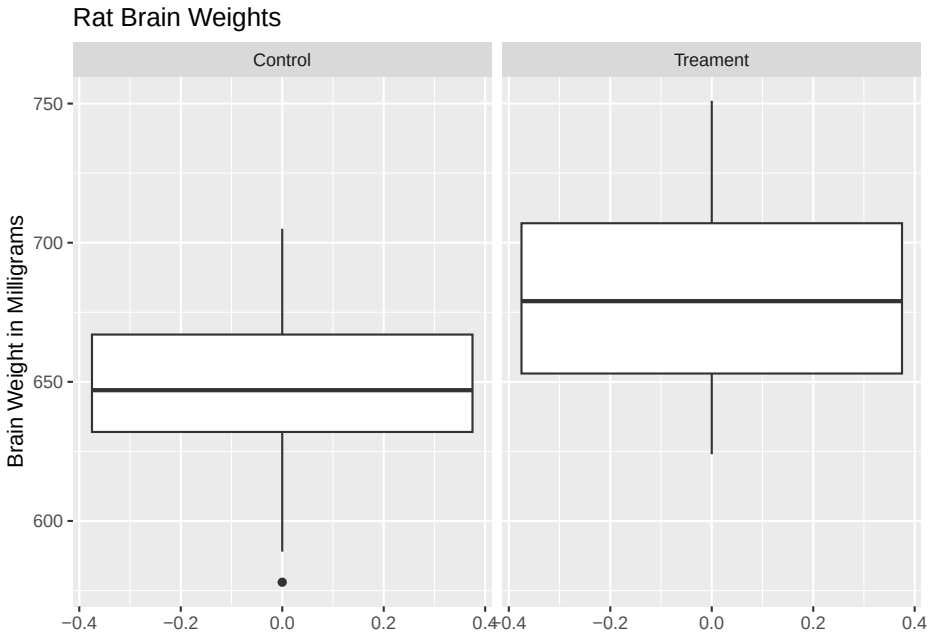
   a. Why did the investigators decide to assign one member of each litter to treatment and another member from the same litter to the control group? What are the advantages?

   The pairing is useful to account for any influence of genetics on the brain weight. Since sibling rats are likely to both be above average (or below average) in terms of their brain weight, by placing one in the treatment and the other in control, we can examine the pure effect of the intervention.

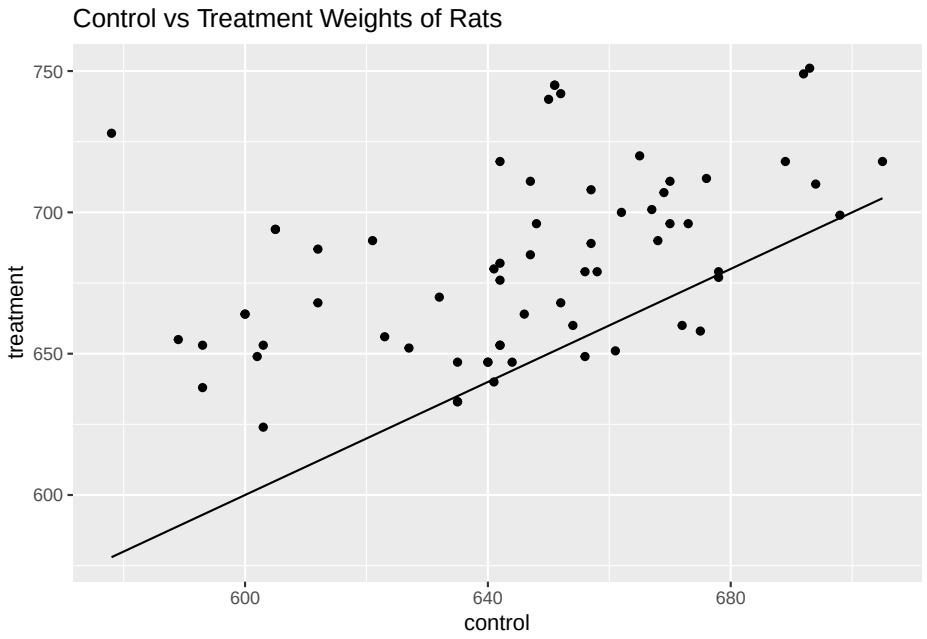
   A second reason for pairing is that the random variable $X - Y$ which is the difference in brain weights will have smaller variability if the X and Y are related.

   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.)

```
library(patchwork)
brain_dat <- read_csv("brain_weights.csv")
new_df <- data.frame(weights = c(brain_dat$treatment,brain_dat$control))
new_df$group <- c(rep("Treament",nrow(brain_dat)), rep("Control",nrow(brain_dat)))
ggplot(data=new_df)+
  geom_boxplot(aes(y=weights))+
  facet_wrap(~group)+
  labs(title = "Rat Brain Weights",
       y = "Brain Weight in Milligrams")
```




```
ggplot(data=brain_dat)+
  geom_point(aes(x=control,y=treatment))+
  geom_function(fun = function(x) x)+
  labs(title = "Control vs Treatment Weights of Rats")
```



- From the boxplot, it seems that the treatment weights are larger than the control. We see this because both the median, and much of the treatment box are higher than the controls.
- We see this similar pattern in the scatterplot, where the points mostly lie above the line $y = x$ indicating, that the treatment has more weight. The scatterplot also shows the strong positive relationship in brain weights of siblings from the same family.

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.

- 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.)
- 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.

Method 1:

We have $H_0 : \pi_0 = .5$ and $H_1 : \pi_0 > .5$

We have the sum of our signs as:

```
brain_dat_sign <- brain_dat %>% filter(treatment != control)
brain_dat_sign$sign <- brain_dat$treatment > brain_dat$control
sum(brain_dat_sign$sign)
```

```
## [1] 57
```

We can conduct a large sample test of π_0 as described in section 1 of chapter 21. Or we can base our test on X - the number of pairs where the treatment weight is higher.

This is the approach we will take here. Specifically, assuming each pair of rats is independent of the other, $X \sim Binom(n = 65, prob = \pi_0)$.

The P-value will be the probability of seeing an observed number of 57 or higher (since the alternative is one sided) under the null sampling distribution.

$$X \sim Binom(n = 65, \pi_0 = .5)$$

We have this P-value as:

```
pbinom(56,size=65,prob = .5,lower.tail = F)
```

```
## [1] 1.581622e-10
```

Method 2:

```
brain_dat %>% mutate(diff = treatment - control) %>%
  infer::t_test(response = diff,mu = 0,alternative = "greater")
```

```
## # A tibble: 1 x 7
##   statistic  t_df     p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>       <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1   9.19007    64 1.31904e-13 greater      36.9538  30.2426      Inf
```

We see in both cases the p-value is incredibly small, leading us to reject the null in each case.

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

One advantage of the sign test is that, computationally, it is much easier to do. This lets us find p-values and conduct hypothesis tests without too much calculation. It does, however, lose the magnitude of the differences which may be important. By discarding the size of the difference, we lose information about the data, and thus may have a less powerful test. By losing information, we may not be able to reject the null hypothesis even when we should.

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 $Var(Z) = E\left[(Z - E\,[Z])^2\right]$)

We have:

$$\begin{aligned} Var(Z) &= Var(\frac{Y}{\sigma_y} - c\frac{X}{\sigma_X}) \\ &= Var(\frac{Y}{\sigma_y}) + Var(c\frac{X}{\sigma_X}) - 2 * Cov(\frac{Y}{\sigma_y}, c\frac{X}{\sigma_X}) \\ &= \frac{Var(Y)}{\sigma_y^2} + c^2\frac{Var(X)}{\sigma_x^2} - \frac{2c}{\sigma_y\sigma_x}Cov(X,Y) \\ &= 1 + c^2 - 2c\rho \end{aligned}$$

From here we are thus minimizing $f(c) = 1 + c^2 - 2c\rho$.

Taking the first derivative and setting to zero, we get $f'(c) = 2c - 2\rho = 0$. This yields a critical point, $c^* = \rho$. We see that $f''(c) = 2 > 0$ so this is a minimum.

b. What is the minimal value of this variance?

Plugging in our values we have that the minimum of $Var(Z)$ will be $1 - \rho^2$. Since $-1 \leq \rho \leq 1$, the variance thus takes a minimum when $\rho = 1$ or $\rho = -1$. If this is the case we will have that $Var(Z) = 0$