# Homework 1

Spring 2023

Tanner Huck

**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 $\mu_0$. 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 $\mu_0$.

   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 $\pi_0$. What are we hypothesizing about $\pi_0$?)

   Assume that we simulate $B = 10,000$ samples from some distribution and calculate the 95 z-confidence interval of each sample. If we let $X_i$ denote whether the ith z-confidence interval covers the true value, we can let each say $X_i \sim Binom(1, \pi_0)$ where $pi_0$ is the probability that the z-confidence interval covers the true value. Hence each $X_i$ will be 1 if it contains the true value and 0 otherwise. Thus we can say that $\bar{X} = \frac{1}{B}\sum_{i=1}^{B} X_i$ where $X$ denotes the simulated coverage rate.

   Since each $pi_0$ is the probability that the z-confidence interval covers the true value and we are using a 95 z-confidence interval, we know that $\pi_0$ is 0.95. Suppose we now want to test the null hypothesis $H_0 : \pi_0 = 0.95$ against the alternative hypothesis $H_1 : \pi_0 \neq 0.95$ at the $\alpha = 0.05$ confidence interval. We will suspect that that the true coverage rate is not 95% if under the null hypothesis, there is a high probability of landing in the rejection region.

   Using the CLT, since each $X_i$ is a Bernoulli random variable, the distribution of $X$ can be approximated by a normal distribution (assuming large n). Thus we know that,

   $$\mu_0 = \pi_0 = 0.95$$

$$\sigma_0 = \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}} = \frac{\sqrt{0.95 \cdot 0.05}}{\sqrt{10000}}$$

Hence our 95 z-confidence interval is

$$\pi_0 \pm 1.96\sigma_0 = 0.95 \pm 1.96 \times \frac{\sqrt{0.95 \cdot 0.05}}{\sqrt{10000}}$$

$$\approx [0.9457283, 0.9542717]$$

Hence, if the simulated coverage rate is below 0.946 or above 0.954, we should suspect that the true coverage rate 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. )

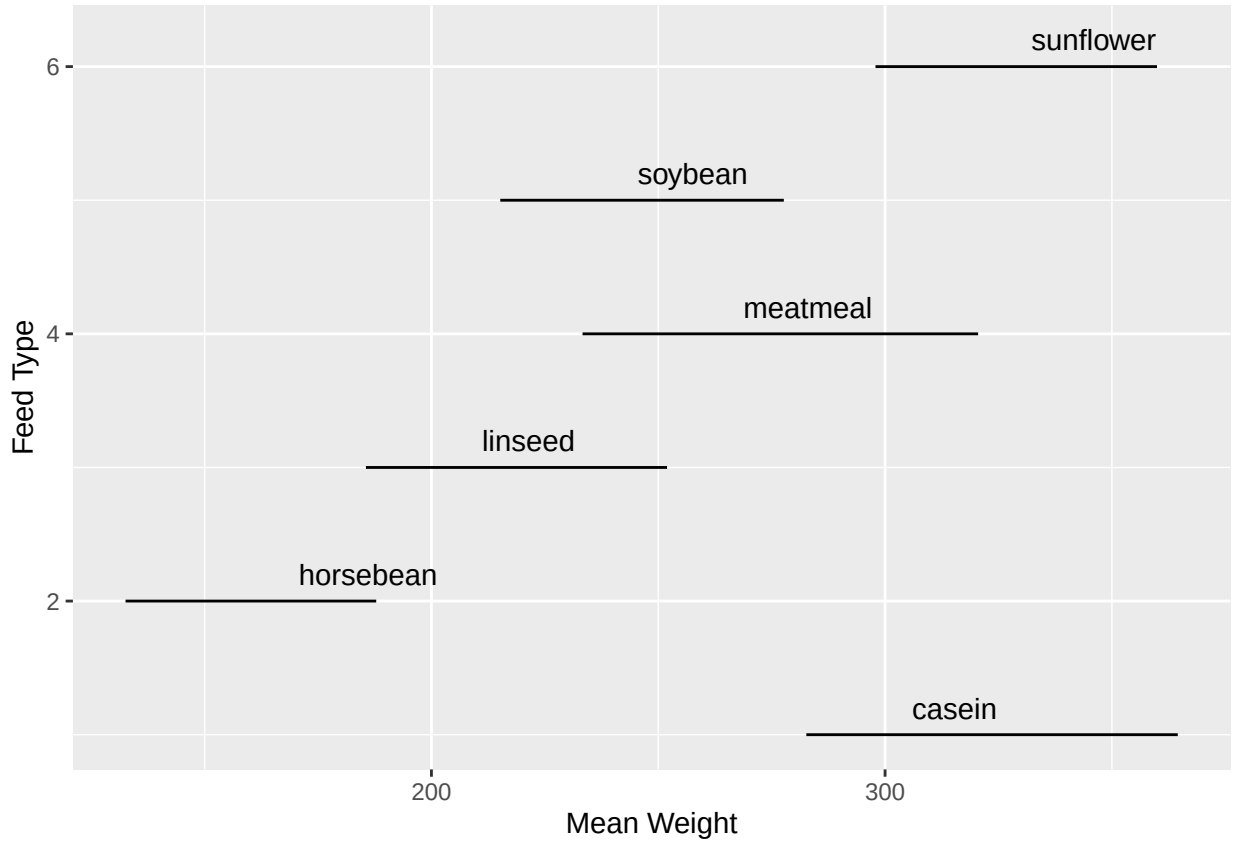
```
# loading data
data(chickwts)
# setting options for better looking table
options(pillar.sigfig = 6)

# creating a tibble for the graph
output_matrix <- tibble(
  Feed = character(),
  Lower_Bound = numeric(),
  Upper_Bound = numeric()
)

# for loop that will run through each type of feed
for (type_feed in levels(chickwts$feed)){
  # subset with feed type
  subset_data <- chickwts %>% filter(feed == type_feed)
  # running t test for subset
  new_row <- t_test(subset_data,
         response = weight,
         alternative = "two-sided",
         conf.level = 0.95)
  # saving t test results into tibble
  output_matrix <- output_matrix %>%
    add_row(Feed = type_feed,
      Lower_Bound = new_row$lower_ci[1],
      Upper_Bound = new_row$upper_ci[1])
}

# show table
output_matrix
```

```
## # A tibble: 6 x 3
##   Feed      Lower_Bound Upper_Bound
##   <chr>           <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)

```
# graphing the different confidence intervals found in part a
ggplot(data = output_matrix) +
  geom_segment(mapping = aes(x = Lower_Bound,
                             xend = Upper_Bound,
                             y = 1:6,
                             yend = 1:6)) +
  labs(x = "Mean Weight",
       y = "Feed Type") +
  annotate("text",
           x = output_matrix$Lower_Bound,
           y = c(1:6)+0.2,
           label = paste0(output_matrix$Feed),
           hjust = -1.25)
```

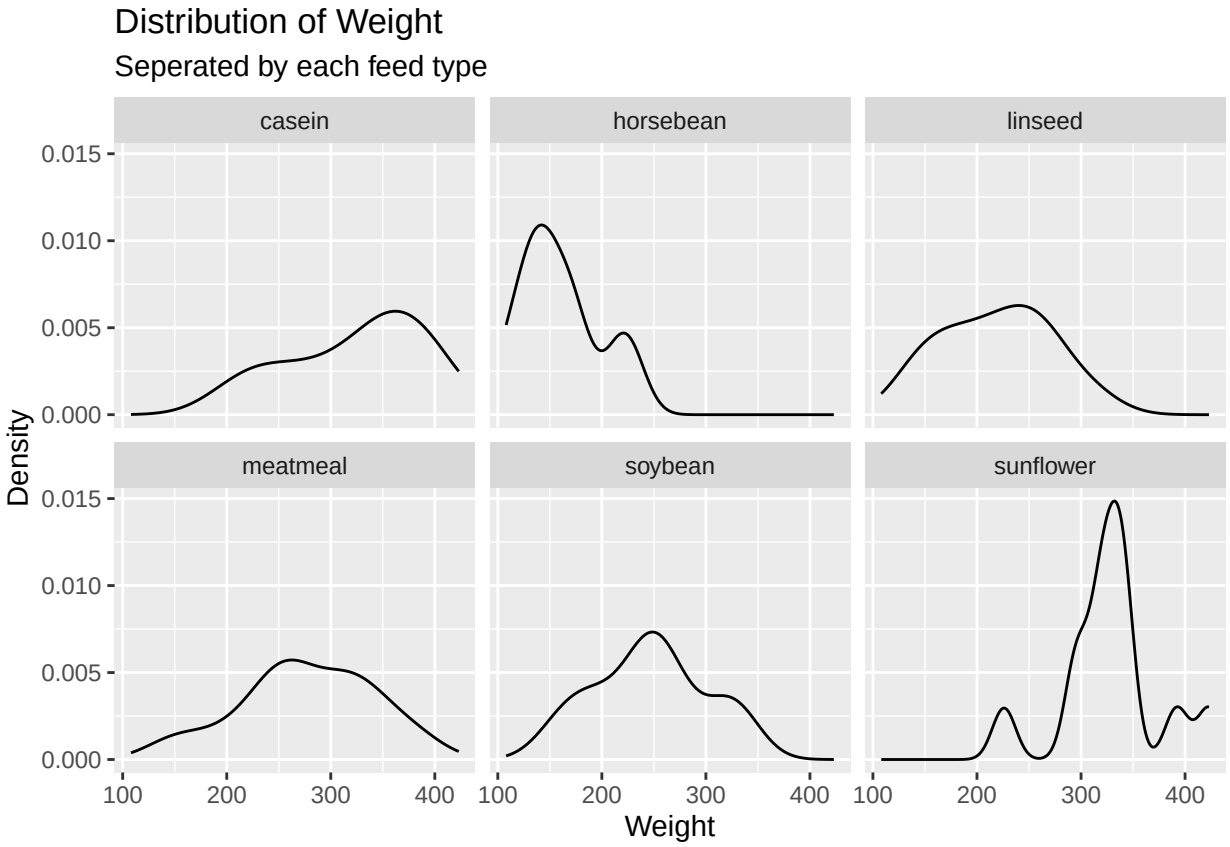



As we can see from the graph, certain diets tend to have chicks with a higher average weight. Looking at each line on the graph, we are 95% confident that each line will contain the true mean weight for its respective feed type. Thus we can clearly see that some feed types like horse bean and linseed will typically have a smaller true mean than the sunflower and casein feed types. In fact, it looks like sunflower and casein feed

types have the highest average weights. However, we do not know if sunflower or casein is better, this is because we are 95% confident that the average weight will be in someone in the graphed line, but the lines have much overlap with each other. For example the sunflower true weight could be in the lower end of the line and the casein true weight could be in the higher end. Hence we are unsure which feed type is the best, but we know that some feed types are better than others.

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 = chickwts,
       mapping = aes(x = weight)) +
  geom_density() +
  facet_wrap(facets = vars(feed) ) +
  labs(title = "Distribution of Weight",
       subtitle = "Seperated by each feed type",
       x = "Weight",
       y = "Density")
```




We can see that a t-distribution may not be appropriate because many of the feed types are not approximately normally distributed, like hoursebean and sunflower. We also know that there are not many trials for each feed type.

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,

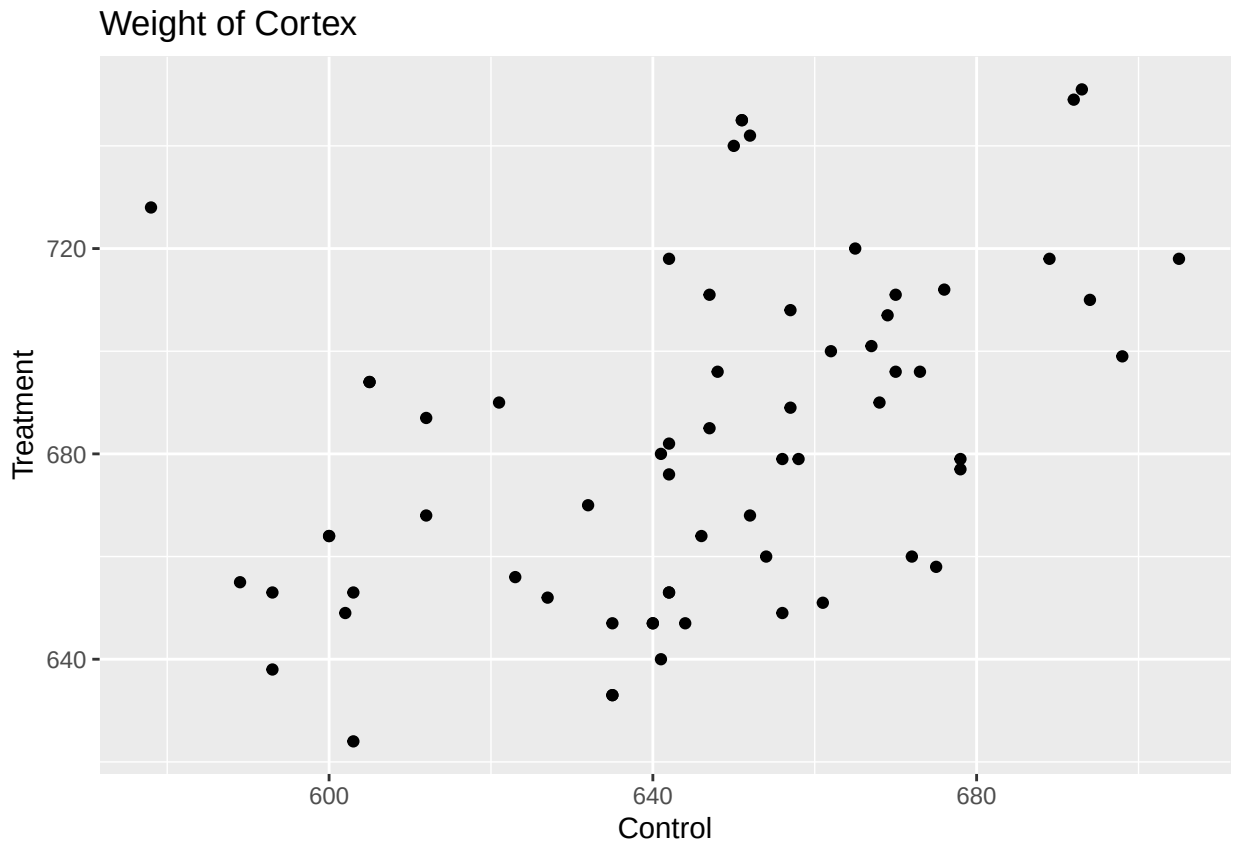
furnished with playthings which were changed daily. Animals in the control group lived in isolation, with no toys. After a month, all animals were sacrificed and their cortex weights measured in milligrams. The data set is in the file `brain-weights.csv`.

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?

In choosing one member of each litter for the treatment and one member for the control groups, one advantage is to minimize differences between rats. If the rats are randomly chosen from the same litter, it gives us more confidence in saying that the rats are independent and identically distributed.

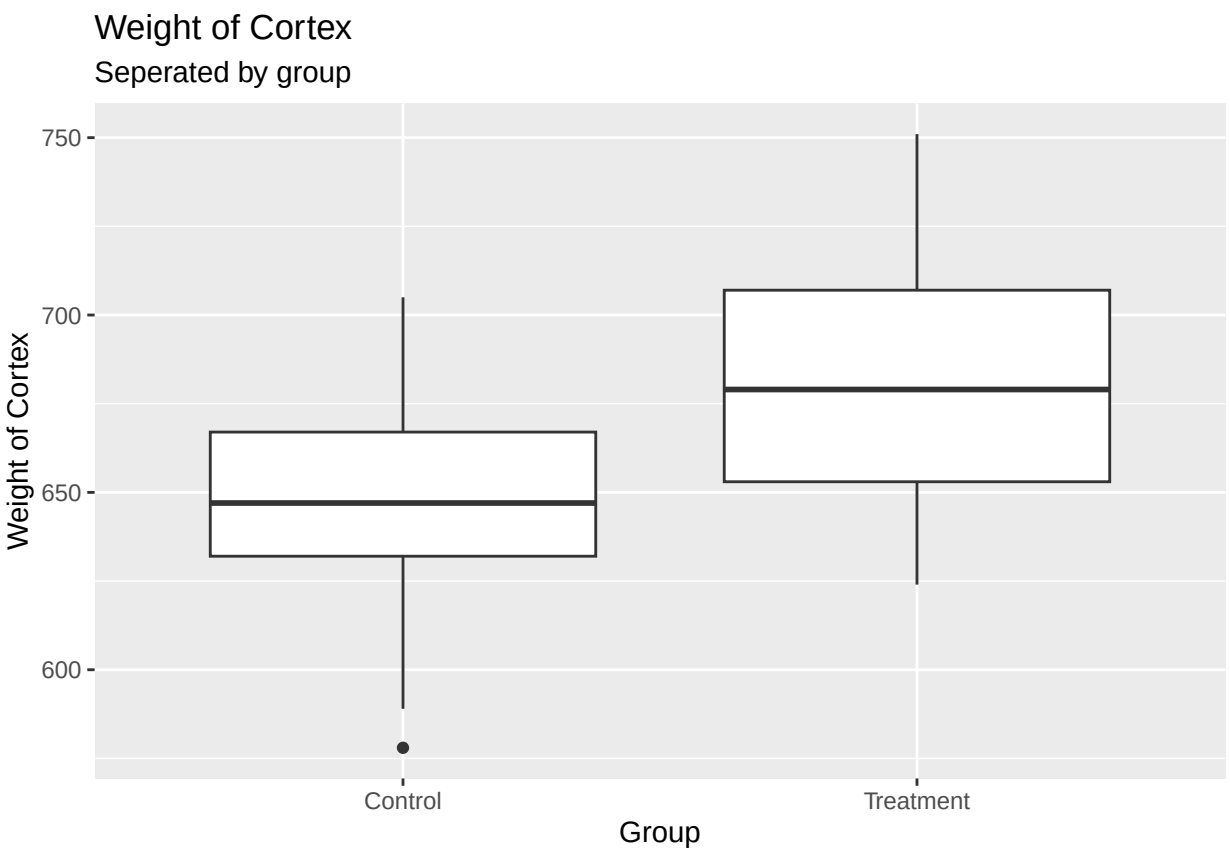
b. Explore these data by making a scatterplot, a boxplot, and calculating some summary statistics. Write briefly about what you are looking for in these plots. (be sure to show your code and output - sans error/warning messages; label your plots; keep your explanation pointed - this means just talk about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

```
rats <- read.csv("brain_weights.csv")

#scatter ploat
ggplot(data=rats) +
  geom_point(aes(x=control, y=treatment)) +
  labs(x = "Control",
       y = "Treatment",
       title = "Weight of Cortex")
```




The scatter plots help us see if there is any relationship between the weights between treatment and control group. We can see that there is a slight linear relationship, but it is not clear. We need to look more closely at the data to see if there is any real relationship.

```
ggplot(data = rats) +
  geom_boxplot(aes(x = "Treatment",
                   y = treatment)) +
  geom_boxplot(aes(x = "Control",
                   y = control)) +
  labs(x = "Group",
       y = "Weight of Cortex",
       title = "Weight of Cortex",
       subtitle = "Seperated by group")
```




The box plot can tell us more specific information than the scatter plot. We can see that there is not a clear difference between the two groups, however, the min, median, and max of the treatment group are slightly higher. We can start to see that the treatment group might have higher weights.

```
# Treatment
summary(rats$treatment)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   624.0   653.0   679.0   682.4   707.0   751.0
```

```
# Control
summary(rats$control)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   578.0   632.0   647.0   645.5   667.0   705.0
```

These results again confirm what we saw in the scatter plot and box plots. We can now see that on average, the treatments group is 40 units greater in weight in comparison to the control group. We can see that every summary value in the treatment group is larger the in the control group.

In conclusion we can see the the rats placed with other rats and toys, on average, had higher cortex weights than the rats in isolation.

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 $\pi_0$ 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 $\mu_d$ is the expected value of $D$.

**Method 1**

```
rats$diff <- rats$treatment - rats$control
rats$binary <- ifelse(rats$diff > 0, 1, 0)
total <- sum(rats$binary)
pval_sign_test <- pbinom(57-1, length(rats$diff), 0.5, lower.tail=F)
pval_sign_test
```

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

From this test we can observe a p-value of about 0. This is very small, hence we can reject the null hypothesis and conclude that there is difference in weight between the treatment and control groups.

**Method 2**

```
diff <- rats$treatment - rats$control
results_paired_t_test <- t.test(diff, alternative = "greater")
results_paired_t_test$p.value
```

```
## [1] 1.319041e-13
```

From this test we can observe a p-value of about 0. This is very small, hence we can reject the null hypothesis and conclude that there is difference in cortex weight between the treatment and control groups, hence rats living with other rats and toys leads to higher cortex weight.

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

One advantage of the paired t-test was that it was easier to code (personally). It can also be more precise than the sign test with smaller sample sizes. However the paired t-test is sensitive to outlines and assumes the difference between the two groups are normally distributed. The sign test makes no assumptions about the difference between groups and the distribution, but is less powerful when the difference is approximately normal.

4. Suppose $X$ and $Y$ are jointly distributed with variances $\sigma_X^2$ and $\sigma_Y^2$, respectively. The correlation coefficient $\rho$ 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 $Var(Z) = E\left[(Z - E\,[Z])^2\right]$ )

First calculating the expected value of $Z$ using linearity of expectation,

$$\begin{aligned} E[Z] &= E[\frac{Y}{\sigma_Y} - \frac{cX}{\sigma_X}] \\ &= \frac{1}{\sigma_Y}E[Y] - \frac{c}{\sigma_X}E[X] \end{aligned}$$

Now finding $Var[Z]$,

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

Then, to find the minimum of $var[Z]$ we have to find its derivative.

$$\begin{aligned} \frac{d}{dc}var[Z] &= \frac{d}{dc}1 - 2c\rho + c^2 \\ &= -2\rho + 2c \\ &\quad \text{now setting this to } 0 \\ 0 &= -2\rho + 2c \\ \rho &= c \end{aligned}$$

Hence $\rho = c$ is a critical point. Then finding the second derivative of the varience,

$$\frac{d}{dc} - 2\rho + 2c = 2$$

Since the second derivative is positive, then the function has a minimum at the critical point $\rho = c$.

b. What is the minimal value of this variance?

Substituting our value of $\rho = c$ into the varence of Z,

$$Var[Z] = 1 - 2\rho^2 + \rho^2 = 1 - \rho^2$$

Thus the minimum value of the variance is $1 - \rho^2$.