Homework 3

The purpose of this homework is to practice using randomization methods to run hypothesis tests. Please fill in the appropriate code and write answers to all questions in the answer sections, and then submit a compiled pdf with your answers through Gradescope by 11pm on Sunday, September 24th.

As always, if you need help with the homework, please attend the TA office hours which are listed on Canvas and/or ask questions on Ed Discussions. Also, if you have completed the homework, please help others out by answering questions on Ed Discussions, which will count toward your class participation grade.

Part 1: Examining discrimination in hiring

A study by Bertrand and Mullainathan in the American Economic Review examined discrimination in the job market. In the study, the researchers randomly generated information such as years of experience and education details to create a realistic-looking resumes. They then randomly assigned a name to the resume that would communicate the applicant's gender and race. The first names chosen for the study were selected such that the names would predominantly be recognized as belonging to Black or White individuals. For example, Lakisha was a name that their survey indicated would be interpreted as a Black woman, while Greg was a name that would generally be interpreted to be associated with a White male. They then sent the resumes out to jobs posted in Boston and Chicago newspapers and observed whether there were differences in how frequently employers gave a callback based on the perceived race and gender of the applicant.

In the first exercise, you will use the data collected by Bertrand and Mullainathan to run a hypothesis test to assess whether there is evidence that applicants that are perceived to be White have a higher callback rate than applicants that are perceived to be Black.

Credit: An analysis of this data was described in the book 'Quantitative Social Science An Introduction' by Imai, and the data comes from the OpenIntro Data Sets.

Part 1.1 (5 points): To start the analysis, please state the null and alternative hypotheses for testing whether applicants whose names are perceived to be White have a higher callback rate than applicants whose names are perceived to be Black using both words and in the appropriate symbols that we have discussed in class. Also, please state whether this study is an observational study or an experiment, and why.

Note: We are using the term "callback rate" to mean the proportion of resumes that received callbacks from potential employers.

Answer:

The null and alternative hypothesis in words

The null hypothesis is that there is no difference in callback rates between applicants who are perceived to be White and applicants who are perceived to be Black. The alternative hypothesis is that there is a difference

in callback rates between applicants who are perceived to be White and applicants who are perceived to be Black.

The null and alternative hypothesis in symbols

$H_0 : \pi_{white} - \pi_{black} = 0$

$H_A : \pi_{white} - \pi_{black} > 0$

This is an observational study because the researchers are not actively controlling the value of any of the variables. Instead they are observing the values as they naturally exist.

Part 1.2 (8 points) : The data from the Bertrand and Mullainathan study is loaded below. To learn more about the data, including seeing a data dictionary, type `?` `resume` in the console after the data has been loaded.

Please complete the following steps to calculate the main observed statistic of interest for our hypothesis test:

1. Create an object called `race` that contains the data in the `race` variable in the `resume` data frame. Also create an object called `callback` that has the data from `callback_received` variable in the `resume` data frame. For the `callback` data, a 1 indicates the resume received a callback from an employer, while a 0 indicates that the resume did not receive a callback.

2. Calculate the proportion of resumes that received callbacks from resumes that had names that were perceived to be White (i.e, the White name callback rate), and the proportion of resumes that received callbacks from resumes that had names that were perceived to be Black (i.e., the Black name callback rate). Print out these rates and then create an object called `obs_stat` that has the difference of the White and Black name callback rates. Report what the value of the observed statistic is in the answer section below using the appropriate symbol we have used in class. From just looking at this statistic value, state whether you believe there will be statistically significant evidence of a difference in these rates.

Hint: there are several ways to do calculate the callback rates. Reviewing the material from the first three classes could be helpful here. Also, depending on the method you use, it could be important to remember that R is case sensitive.

```
# load the data
load("resume.rda")


# 1. Extract the relevant vectors for the resume data frame

race <- resume$race
callback <- resume$received_callback

# 2. Calculate the White and Black name callback rates and the difference of these rates

white <- race[race == "white"]
black <- race[race == "black"]

# callback variable filtered by race
white_callback <- callback[race == "white"]
black_callback <- callback[race == "black"]
```

```
# White callback rates
white_rate <- mean(white_callback)
white_rate
```

[1] 0.09650924

```
# Black callback rates
black_rate <- mean(black_callback)
black_rate
```

[1] 0.06447639

```
obs_stat <- white_rate - black_rate
obs_stat
```

[1] 0.03203285

Answer:

$\hat{p} = 0.03203285$

I do believe that there will be statistically significant evidence of a difference in these rates.

Part 1.3 (15 points) : Now let's use the `rbinom()` function to generate a null distribution that would be expected if the null hypothesis was true. To do this, please complete the following steps:

1. Calculate the overall callback rate that ignores the perceived race of the name on the resume. Save this overall callback rate to an object called `overall_callback_rate`. This is the callback rate that would be expected if the rate was the same for resumes with stereotypical White and Black names (i.e., it is the callback rate consistent with the null hypothesis).

2. Calculate the number of resumes in the Bertrand and Mullainathan data that have stereotypical White names and save this to an object called `size_white`. Likewise, calculate the number of resumes that have stereotypical Black names and save this to an object called `size_black`.

3. Use the `rbinom()` to generate 10,000 simulated callback rates for the resumes with stereotypical White names that are consistent with the null hypothesis and save this an object called `null_white_rates`. Likewise the `rbinom()` to generate 10,000 simulated callback rates for the resumes with stereotypical Black names that are consistent with the null hypothesis and save this an object called `null_black_rates`. Then calculate the null distribution by subtracting the `null_black_rates` from the `null_white_rates` and save this to an object called `null_distribution`.

4. Plot the null distribution as a histogram and add a vertical line at the observed statistic value.

5. In the answer section below, state whether you believe you will be able to reject the null hypothesis based on looking at the plot you created in step 4.

Hints:

1. The `rbinom()` returns counts of the number of successes out of "size" coin flips. You can turn this into rates (i.e., proportions) by dividing the maximum possible value that `rbinom()` can return.

3

2. If you subtract two vectors of numbers `v1` and `v2` then each element of vector `v1` is subtracted from the element of vector `v1` at the same index location.

3. The arguments to `rbinom(num_sims, size, prob)` are:

- `num_sims`: the number of simulations to run.
- `size`: the number of "coin flips" in each simulation.
- `prob`: the probability of getting heads on each coin flip.

```r
# 1. calculate the `overall_callback_rate`

overall_callback_rate <- mean(callback)


# 2. create size_white and size_black objects
size_white <- length(white)
size_black <- length(black)


# 3. create the null distribution

null_white_rates <- rbinom(1e4, size_white, overall_callback_rate) / size_white
null_black_rates <- rbinom(1e4, size_black, overall_callback_rate) / size_black
null_distribution <- null_white_rates - null_black_rates

# 4. plot the null distribution

hist(null_distribution, xlab = "Callback difference",
     main = "Distribution of Callback Rates for Null Hypothesis")
abline(v = obs_stat, col = "red")
```
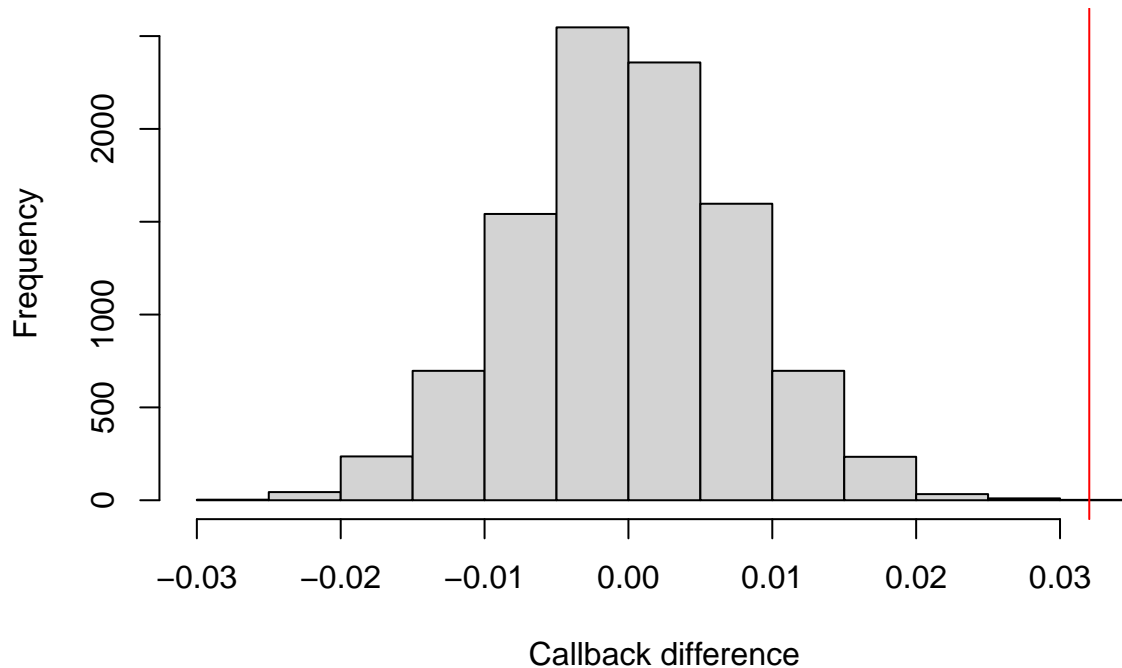
Distribution of Callback Rates for Null Hypothesis

Answer

Based on the above plot, I do believe that I will be able to reject the null hypothesis.

Part 1.4 (5 points): Now, use the `null_distribution` and the `obs_stat` objects to calculate the number of simulations that had a difference of proportion that was as high or higher than the observed statistic. Convert this to a p-value by dividing by the total number of simulations. Describe whether this p-value provides sufficient evidence that resumes with names generally perceived as White receive callbacks at a higher rate than resumes with names generally perceived as Black.

```
p_value <- sum(null_distribution >= obs_stat) / length(null_distribution)

p_value
```

```
## [1] 0
```

Answer:

This p-value does provide sufficient evidence that resumes with names generally perceived as White receive callbacks at a higher rate than resumes with names generally perceived as Black.

Part 1.5 (12 points): Now use the bootstrap to calculate a 95% confidence interval for the callback rate for resumes with names generally perceived as White. In the answer section, report whether your best point

estimate for the proportion of callback for resumes with names generally perceived as Black falls in this confidence interval.

Hint: Creating a vector of the callbacks for White sounding names is a good first step in the analysis.

```r
# create a vector for the callback data for only resumes with stereotypical White names

white_callback <- callback[race == "white"]


# create the bootstrap distribution

boot_dist_white <- NULL
for(i in 1:1e4){
  boot_sample_white <- (sample(white_callback, replace = TRUE))
  boot_dist_white[i] <- mean(boot_sample_white)
}


# plot the bootstrap distribution

hist(boot_dist_white, xlab = "Callbacks for Resumes with White-Sounding Names", main = "Bootstrap Distr
```
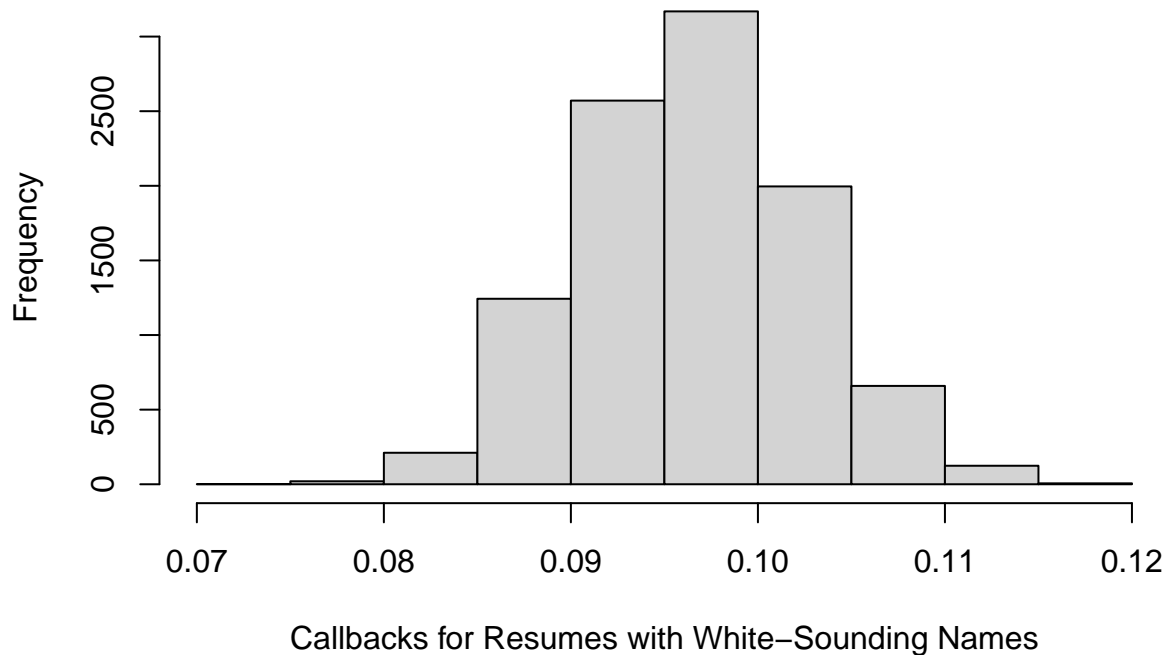
Bootstrap Distribution for White Resume Callbacks

Callbacks for Resumes with White–Sounding Names

```r
# create the confidence interval

SE_boot_white <- sd(boot_dist_white)
```

```
CI_lower_white <- mean(white_callback) - 2 * SE_boot_white
CI_upper_white <- mean(white_callback) + 2 * SE_boot_white


# Confidence interval for callback rate for resumes with White-sounding names
c(CI_lower_white, CI_upper_white)
```

[1] 0.08454574 0.10847274

```
# Point estimate for the proportion of callbacks for resumes with Black-sounding names
mean(black_callback)
```

[1] 0.06447639

Answer:

My best point estimate for the proportion of callbacks for resumes with names generally perceived as Black does not fall in this confidence interval.

Part 1.6 (5 points): There is also a formula for calculating the standard error of a proportion which is:

$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Use this formula to create 95% confidence intervals for the callback rate for resumes with names generally perceived as White, and compare it to the 95% confidence intervals you calculated in the previous problem. Describe in the answer section below whether the two methods for calculating confidence intervals lead to similar results.

```
p_hat <- mean(white_callback)

SE_p_hat <- sqrt((p_hat * (1 - p_hat)) / size_white)

new_CI_upper <- p_hat + 2 * SE_p_hat
new_CI_lower <- p_hat - 2 * SE_p_hat

# New confidence interval
c(new_CI_lower, new_CI_upper)
```

[1] 0.0845411 0.1084774

```
# Original confidence interval
c(CI_lower_white, CI_upper_white)
```

[1] 0.08454574 0.10847274

Answer

The two methods for calculating confidence intervals do lead to similar results.

Bonus (0 points):

Explore the data set further and see if you can discover any other interesting results from the study.

Part 2: Do diamond prices differ depending on the quality of the cut?

In 2018, the total global wholesale value of polished diamonds amounted to almost 25.3 billion U.S. dollars. The price of an individual diamond can be affected by several factors including the diamond's color, clarity, size and the quality of how it was cut. In this exercise, we will examine data from a large set of diamonds to assess whether the quality of the cut of a diamond affects the average price.

Part 2.0 (4 points)

The code below loads a data frame called `diamonds` that comes from the ggplot2 package, which is a package that is used to visualize data (we will discuss this package soon). For more information about the data set see this website.

In the exercises below we will compare the mean price of diamonds, for two different levels of the diamond's cut. The levels of cut we are going to compare are the `Fair` cut diamonds (which is the worst cut), and the `Ideal` cut diamonds, which is the best cut. To start this analysis, create a object called `fair_price` that contains the prices for all the diamonds corresponding to the `Fair` cut, and a vector object called `ideal_price` that contains the prices for all the diamonds corresponding to the `Ideal` cut. If done correctly, the `fair_price` vector should have 1,610 elements, and the `ideal_price` vector should have 21,551 elements in it.

Also, report in the answer section below, how many cases and variables the full `diamonds` data frame has (the `dim()` function can help with this) and also whether you think the prices for the `Ideal` cut diamonds will be higher than the prices for `Fair` cut diamonds. Finally, please state whether this study is an observational study or an experiment and why.

```
library(ggplot2)
data(diamonds)

fair_price <- diamonds$price[diamonds$cut == "Fair"]
ideal_price <- diamonds$price[diamonds$cut ==
                                  "Ideal"]

dim(diamonds)
```

```
## [1] 53940    10
```

Answer:

The diamonds data frame has 53,940 cases and 10 variables.

I think that the prices for the Ideal cut diamonds will be higher than the prices for Fair cut diamonds.

This is an observational study because the researchers are not actively controlling the value of any of the variables. Instead they are observing the values as they naturally exist.

In parts 2.1 to 2.5 you will now do the 5 steps to run a hypothesis test!

Part 2.1 (5 points) Let's start our hypothesis testing in the usual way by stating the null and alternative hypotheses using words and symbols. Although you might have a prior expectation that one of the diamonds' cuts will have a higher average price than the other, please state the alternative hypothesis in a non-directional way, such that you are testing whether the average prices of the `Fair` cut and `Ideal` are different, and not that one price is higher than the other. When running your hypothesis test below, also do the analysis in such a way that you are doing a non-directional test (i.e., you will calculate the p-value by looking at both tails of the distribution).

Finally, write down the significance level using the appropriate symbol, and set it to the most commonly used value.

Answer:

we will examine data from a large set of diamonds to assess whether the quality of the cut of a diamond affects the average price

Null hypthesis: There is no difference between the average prices of Fair and Ideal cut diamonds.

Alternative hypothesis: There is a difference between the average prices of Fair and Ideal cut diamonds.

$H_0 : \mu_{fair} - \mu_{ideal} = 0$

$H_A : \mu_{fair} - \mu_{ideal} \neq 0$

Significance level: $\alpha = 0.05$

Part 2.2 (4 points) Now calculate the value of that observed statistic that is appropriate for this analysis. In the answer section, write down the value for this statistic and use the appropriate notation to denote this statistic. Also, visualize the data by creating a side-by-side box plot comparing the two groups. Based on this plot, describe in the answer section whether you believe there is a difference in mean prices.
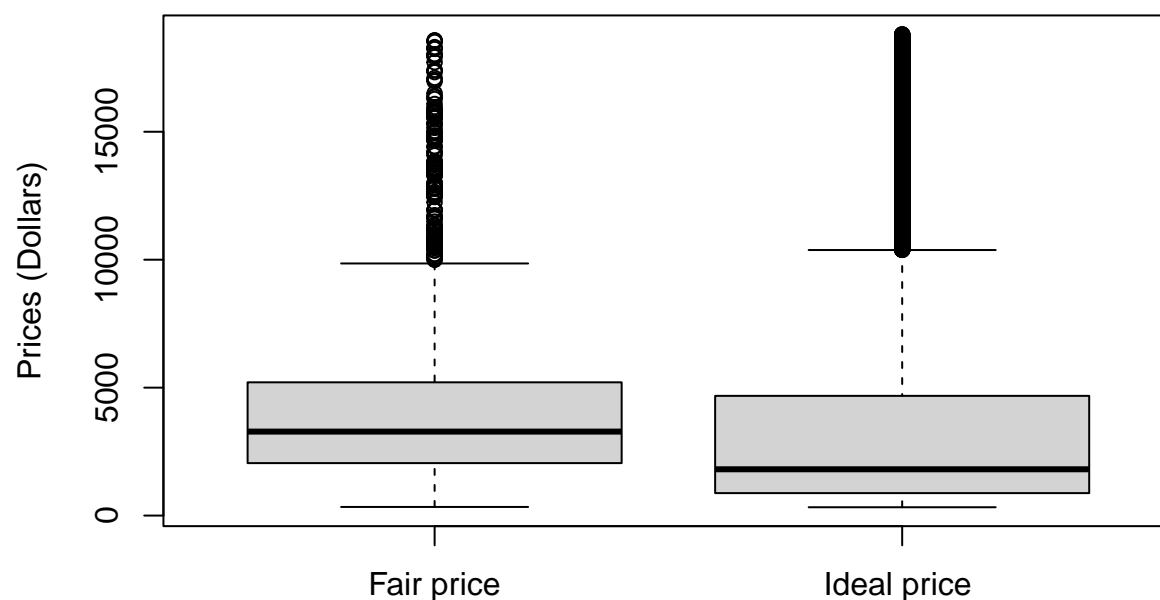
```
obs_stat <- mean(ideal_price) - mean(fair_price)

obs_stat
```

```
## [1] -901.2158
```

```
boxplot(fair_price, ideal_price, ylab = "Prices (Dollars)",
        main = "Boxplots for Fair and Ideal Cut Diamonds",
        names = c("Fair price", "Ideal price "))
```

9

Boxplots for Fair and Ideal Cut Diamonds

Answer:

$\bar{x} = 901.2158$

I believe there is a difference between the mean prices of Fair and Ideal cut diamonds.

Part 2.3 (12 points) Now create a null distribution using a permutation test. To do this, combine data from both groups, randomly assign the data to a fake "fair" and "ideal" groups, calculate a null statistic, and repeat 10,000 times to get a null distribution. Also, plot a histogram of the null distribution and add a red vertical line to the plot at the value of the observed statistic. Based on this plot, do you think the average diamond price is the same between the two cuts?

Hint: You can randomize using the sample() function.

```
combined_data <- c(fair_price, ideal_price)
n_fair <- length(fair_price)
total <- length(combined_data)

null_distribution <- NULL
for(i in 1:1e4){
  shuff_data <- sample(combined_data)

  shuff_fair <- shuff_data[1:n_fair]
  shuff_ideal <- shuff_data[(n_fair + 1):total]

  null_distribution[i] <- mean(shuff_fair) -
```
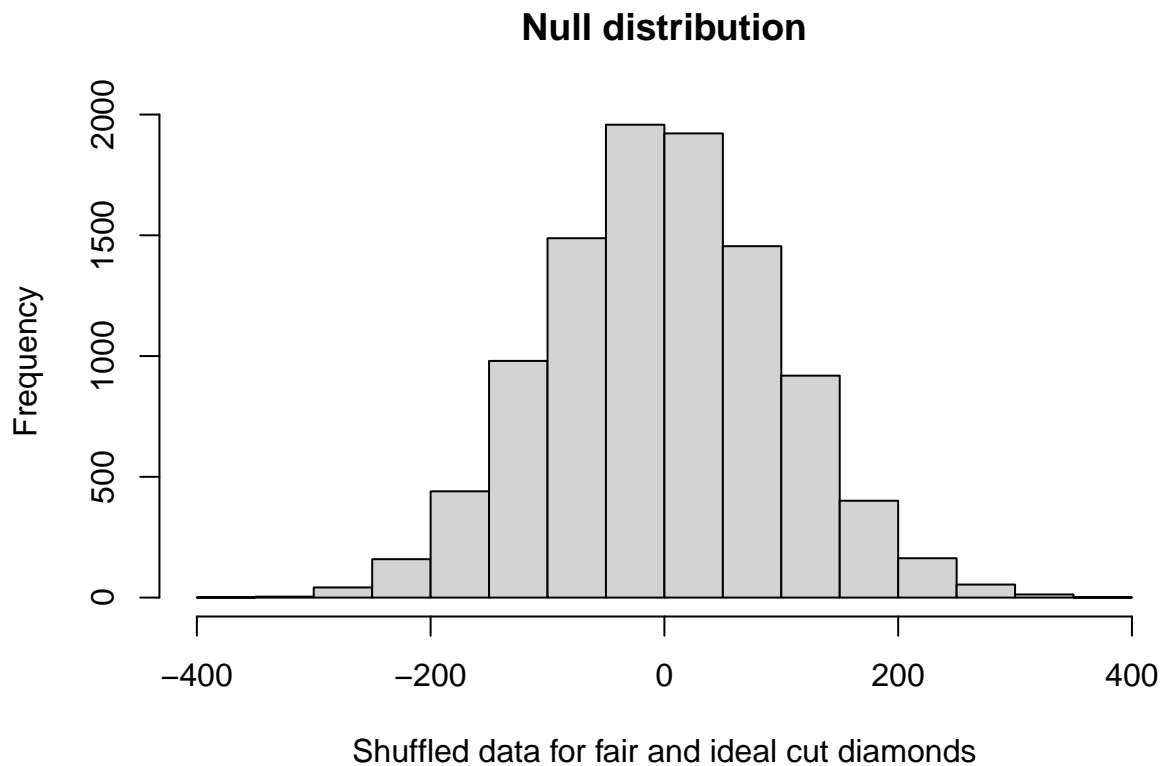
```
    mean(shuff_ideal)
}


hist(null_distribution,
     breaks = 20,
     main = "Null distribution",
     xlab = "Shuffled data for fair and ideal cut diamonds")
abline(v = obs_stat, col = "red")
```

Null distribution

Shuffled data for fair and ideal cut diamonds

Answer

I do not think that the average diamond price is the same between the two cuts.

Part 2.4 (5 points) Now calculate the p-value in the R chunk below. Be sure that you calculate the p-value using both tails (which should be consistent with how you stated the null hypothesis). Is this the p-value you would expect based on the plot of the null distribution and observed statistic in part 2.3?

```
p_val_left <- sum(null_distribution <= obs_stat) /
    length(null_distribution)
p_val_right <- sum(null_distribution >= abs(obs_stat)) / length(null_distribution)

p_val <- p_val_left + p_val_right
```

```
p_val
```

```
## [1] 0
```

Answer

This is the p-value I would expect based on the plot of the null distribution and observed statistic in part 2.3.

Part 2.5 (4 points) Are the results statistically significant? Do you believe there is a difference between these groups? Are these results surprising at all? Please answer these questions below.

Answers:

The results are statistically significant, and I do believe there is a difference between these groups. I was a bit surprised that the observed statistic fell so far outside the range of the null distribution.

Part 2.6 (5 points) Challenge problem

This is a "challenge problem" that you should try to figure out without getting help from the TAs. Challenges problems might be more difficult than other problems but they won't be worth too many points, so they will not have a large impact on your homework score.

From looking at the observed statistic, we notice that the fair cut diamonds actually have a higher price than the ideal cut diamonds! This is strange since one would expect diamonds that have a better cut would cost more. Try examining the full diamonds data frame and see if you can come up with a reason why it might be the case that fair cut diamonds cost more.

Once you have identified another reason why the prices might be higher for the fair cut diamonds, create a plot that gives evidence that your explanation makes sense, and describe in the answer section below what is causing these strange results. Hint: use boxplots, scatterplots and/or other types of plots you see fit to explain the price difference.

```r
# Approach: I will filter each variable by fair and ideal cuts.
# For each variable I will then determine whether, on average,
  # fair cut diamonds are better than ideal cut diamonds.
# If fair cut diamonds are better for a majority of the variables,
  # I will have an explanation for why fair cut diamonds tend to cost more
    # than ideal cut diamonds.

# vectors for fair and ideal cut diamonds
  # to be used for filtering
fair <- diamonds$cut == "Fair"
ideal <- diamonds$cut == "Ideal"


# A counter. I will add a point to the counter for each variable in which
  # fair cut diamonds are better than ideal cut diamonds.
points <- NULL
```

```r
# Carat: higher carat is better
fair_carat <- diamonds$carat[fair]
ideal_carat <- diamonds$carat[ideal]

points[1] <- mean(fair_carat) > mean(ideal_carat)


# Color: D is best, J is worst
fair_color <- diamonds$color[fair]
ideal_color <- diamonds$color[ideal]

  # Assign 7 points to D, 6 to E, etc. down to 1 point for J
  # Then average the points for fair and ideal cuts and compare
color_pts_fair <- NULL
color_pts_fair[1] <- length(fair_color[fair_color == "D"]) * 7
color_pts_fair[2] <- length(fair_color[fair_color == "E"]) * 6
color_pts_fair[3] <- length(fair_color[fair_color == "F"]) * 5
color_pts_fair[4] <- length(fair_color[fair_color == "G"]) * 4
color_pts_fair[5] <- length(fair_color[fair_color == "H"]) * 3
color_pts_fair[6] <- length(fair_color[fair_color == "I"]) * 2
color_pts_fair[7] <- length(fair_color[fair_color == "J"])

  # Average color points for fair cut diamonds
  # More points means better color
avg_color_fair <- sum(color_pts_fair) / length(fair)


color_pts_ideal <- NULL
color_pts_ideal[1] <- length(ideal_color[ideal_color == "D"]) * 7
color_pts_ideal[2] <- length(ideal_color[ideal_color == "E"]) * 6
color_pts_ideal[3] <- length(ideal_color[ideal_color == "F"]) * 5
color_pts_ideal[4] <- length(ideal_color[ideal_color == "G"]) * 4
color_pts_ideal[5] <- length(ideal_color[ideal_color == "H"]) * 3
color_pts_ideal[6] <- length(ideal_color[ideal_color == "I"]) * 2
color_pts_ideal[7] <- length(ideal_color[ideal_color == "J"])

  # Average color points for ideal cut diamonds
avg_color_ideal <- sum(color_pts_ideal) / length(ideal)

  # Add a point to the counter if on average fair cut diamonds have better colors
points[2] <- avg_color_fair > avg_color_ideal


# Clarity from best to worst: IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
fair_clar <- diamonds$clarity[fair]
ideal_clar <- diamonds$clarity[ideal]

  # Take the same points-based approach from the color variable

clar_pts_fair <- NULL

clar_pts_fair[1] <- length(fair_clar[fair_clar == "IF"]) * 8
clar_pts_fair[2] <- length(fair_clar[fair_clar == "VVS1"]) * 7
```

```
clar_pts_fair[3] <- length(fair_clar[fair_clar == "VVS2"]) * 6
clar_pts_fair[4] <- length(fair_clar[fair_clar == "VS1"]) * 5
clar_pts_fair[5] <- length(fair_clar[fair_clar == "VS2"]) * 4
clar_pts_fair[6] <- length(fair_clar[fair_clar == "SI1"]) * 3
clar_pts_fair[7] <- length(fair_clar[fair_clar == "SI2"]) * 2
clar_pts_fair[8] <- length(fair_clar[fair_clar == "I1"]) * 1

avg_clar_fair <- sum(clar_pts_fair) / length(fair)


clar_pts_ideal <- NULL

clar_pts_ideal[1] <- length(ideal_clar[ideal_clar == "IF"]) * 8
clar_pts_ideal[2] <- length(ideal_clar[ideal_clar == "VVS1"]) * 7
clar_pts_ideal[3] <- length(ideal_clar[ideal_clar == "VVS2"]) * 6
clar_pts_ideal[4] <- length(ideal_clar[ideal_clar == "VS1"]) * 5
clar_pts_ideal[5] <- length(ideal_clar[ideal_clar == "VS2"]) * 4
clar_pts_ideal[6] <- length(ideal_clar[ideal_clar == "SI1"]) * 3
clar_pts_ideal[7] <- length(ideal_clar[ideal_clar == "SI2"]) * 2
clar_pts_ideal[8] <- length(ideal_clar[ideal_clar == "I1"]) * 1

avg_clar_ideal <- sum(clar_pts_ideal) / length(ideal)

points[3] <- avg_clar_fair > avg_clar_ideal


# x, y, and z are dimensions, so bigger is better
  # x
fair_x <- diamonds$x[fair]
ideal_x <- diamonds$x[ideal]

points[4] <- mean(fair_x) > mean(ideal_x)

  # y
fair_y <- diamonds$y[fair]
ideal_y <- diamonds$y[ideal]

points[5] <- mean(fair_y) > mean(ideal_y)

  # z
fair_z <- diamonds$z[fair]
ideal_z <- diamonds$z[ideal]

points[6] <- mean(fair_z) > mean(ideal_z)


# Depth: bigger is better
fair_depth <- diamonds$depth[fair]
ideal_depth <- diamonds$depth[ideal]

points[7] <- mean(fair_depth) > mean(ideal_depth)
```

```r
# Table: bigger is better
fair_table <- diamonds$table[fair]
ideal_table <- diamonds$table[ideal]

points[8] <- mean(fair_table) > mean(ideal_table)


# The average number of points. An average >0.5 indicates that fair cut diamonds are better than ideal
mean(points)
```
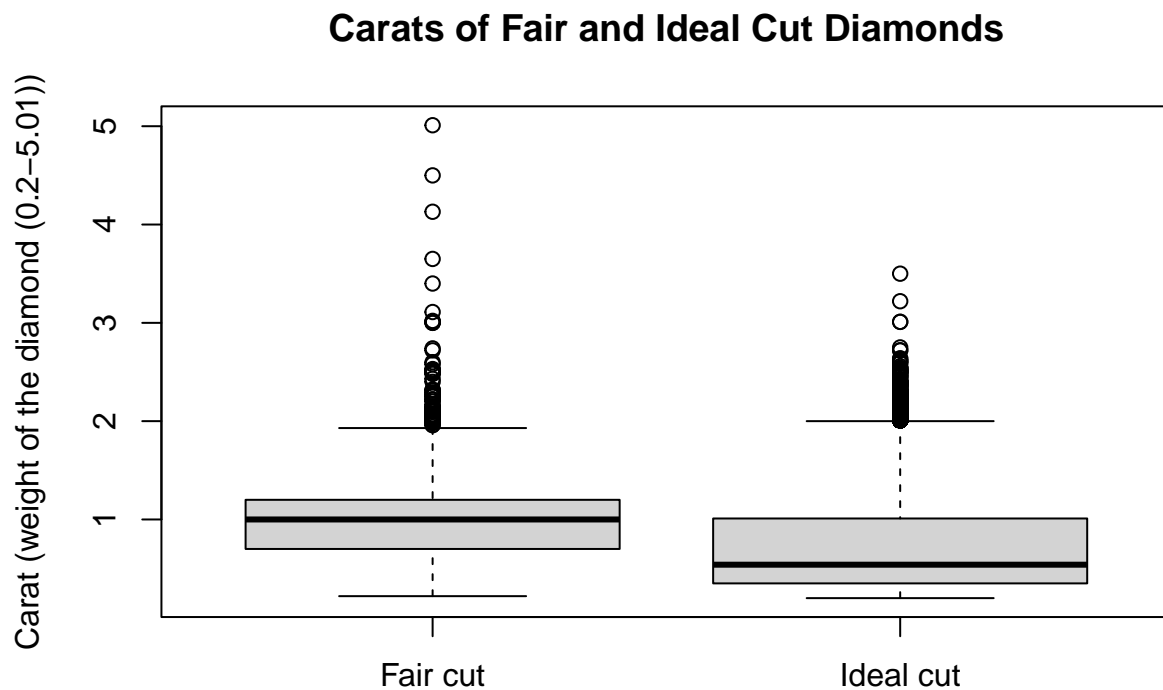
```
## [1] 0.75
```

```r
# Graphs for variables where fair cut diamonds are better than ideal cut diamonds
boxplot(fair_carat, ideal_carat, names = c("Fair cut", "Ideal cut"),
        ylab = "Carat (weight of the diamond (0.2-5.01))",
        main = "Carats of Fair and Ideal Cut Diamonds")
```

Carats of Fair and Ideal Cut Diamonds


```r
boxplot(fair_x, ideal_x, names = c("Fair cut", "Ideal cut"),
        ylab = "Length (mm)",
        main = "Lengths of Fair and Ideal Cut Diamonds")
```
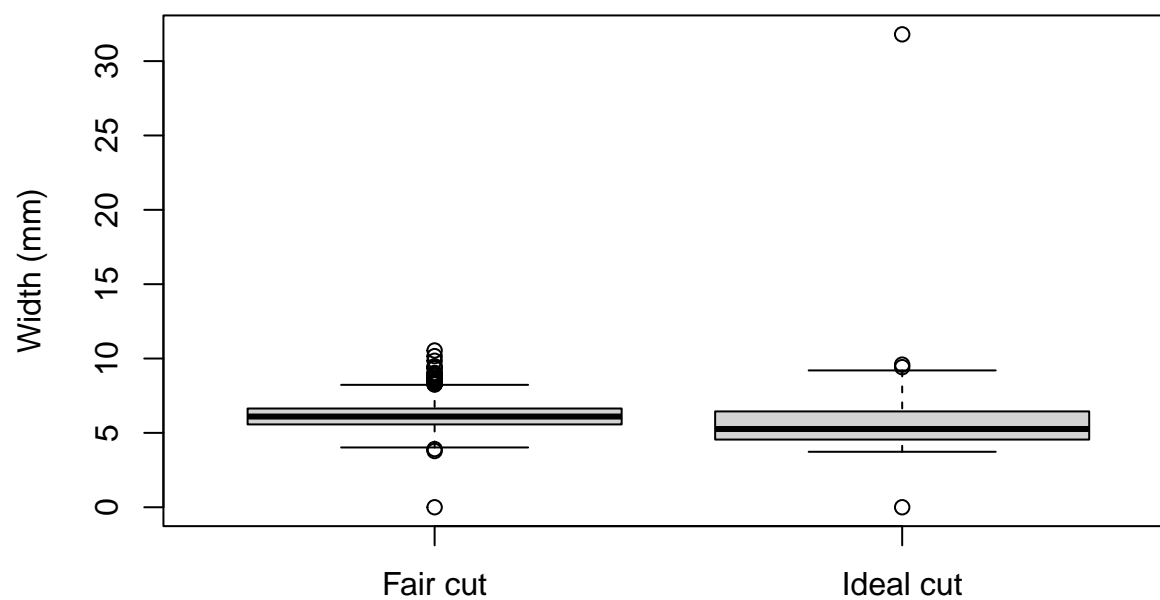
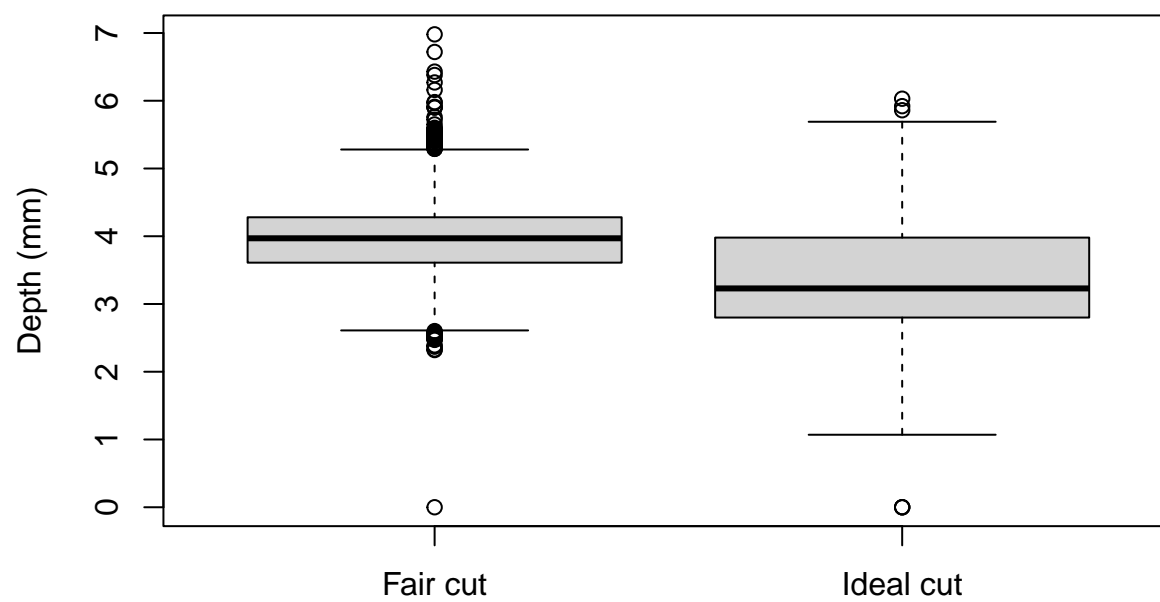## Lengths of Fair and Ideal Cut Diamonds
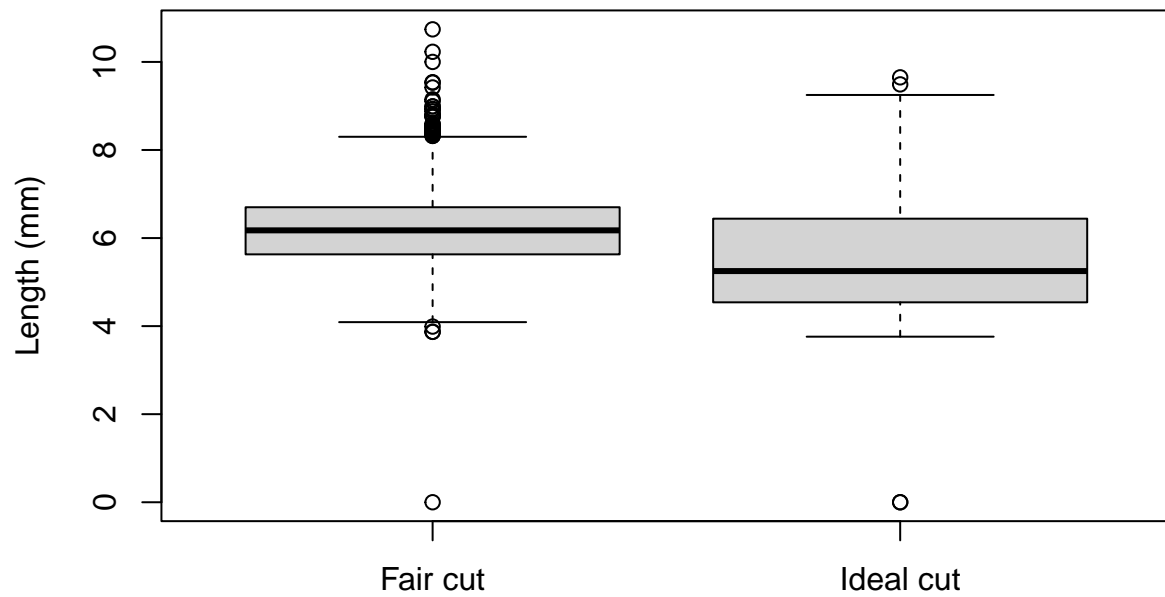

```r
boxplot(fair_y, ideal_y, names = c("Fair cut", "Ideal cut"),
        ylab = "Width (mm)",
        main = "Widths of Fair and Ideal Cut Diamonds")
```

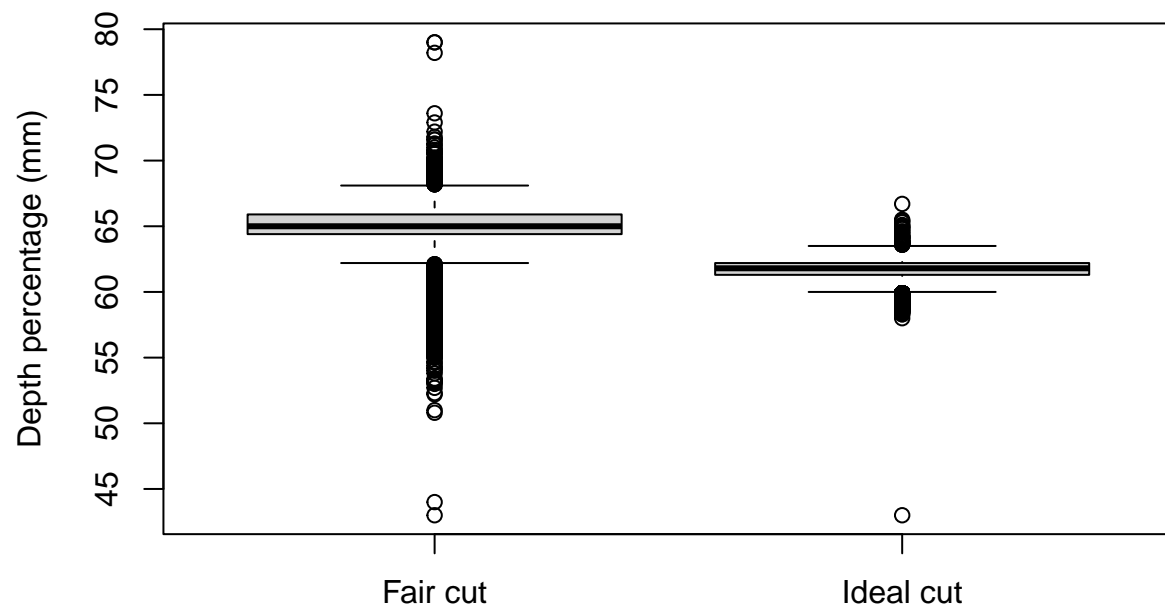**Widths of Fair and Ideal Cut Diamonds**


```
boxplot(fair_z, ideal_z, names = c("Fair cut", "Ideal cut"),
        ylab = "Depth (mm)",
        main = "Depths of Fair and Ideal Cut Diamonds")
```

Depths of Fair and Ideal Cut Diamonds


```r
boxplot(fair_depth, ideal_depth, names = c("Fair cut", "Ideal cut"),
        ylab = "Depth percentage (mm)",
        main = "Total Depth Percentage of Fair and Ideal Cut Diamonds")
```
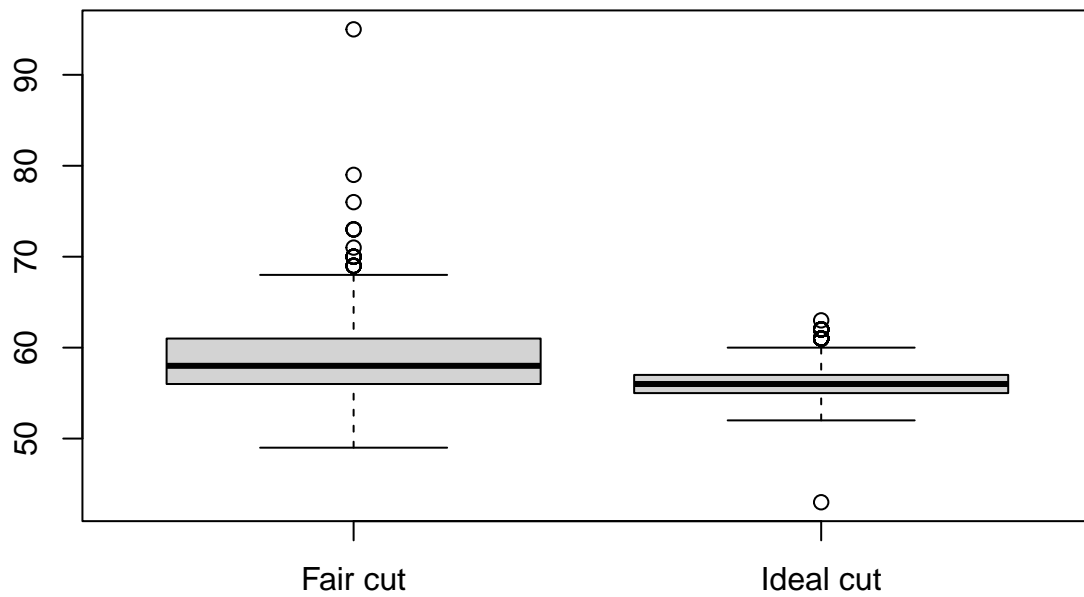
Total Depth Percentage of Fair and Ideal Cut Diamonds


```r
boxplot(fair_table, ideal_table, names = c("Fair cut", "Ideal cut"),
        ylab = "Width of Top of Diamond Relative to Widest Point (43-95)",
        main = "Tables of Fair and Ideal Cut Diamonds")
```
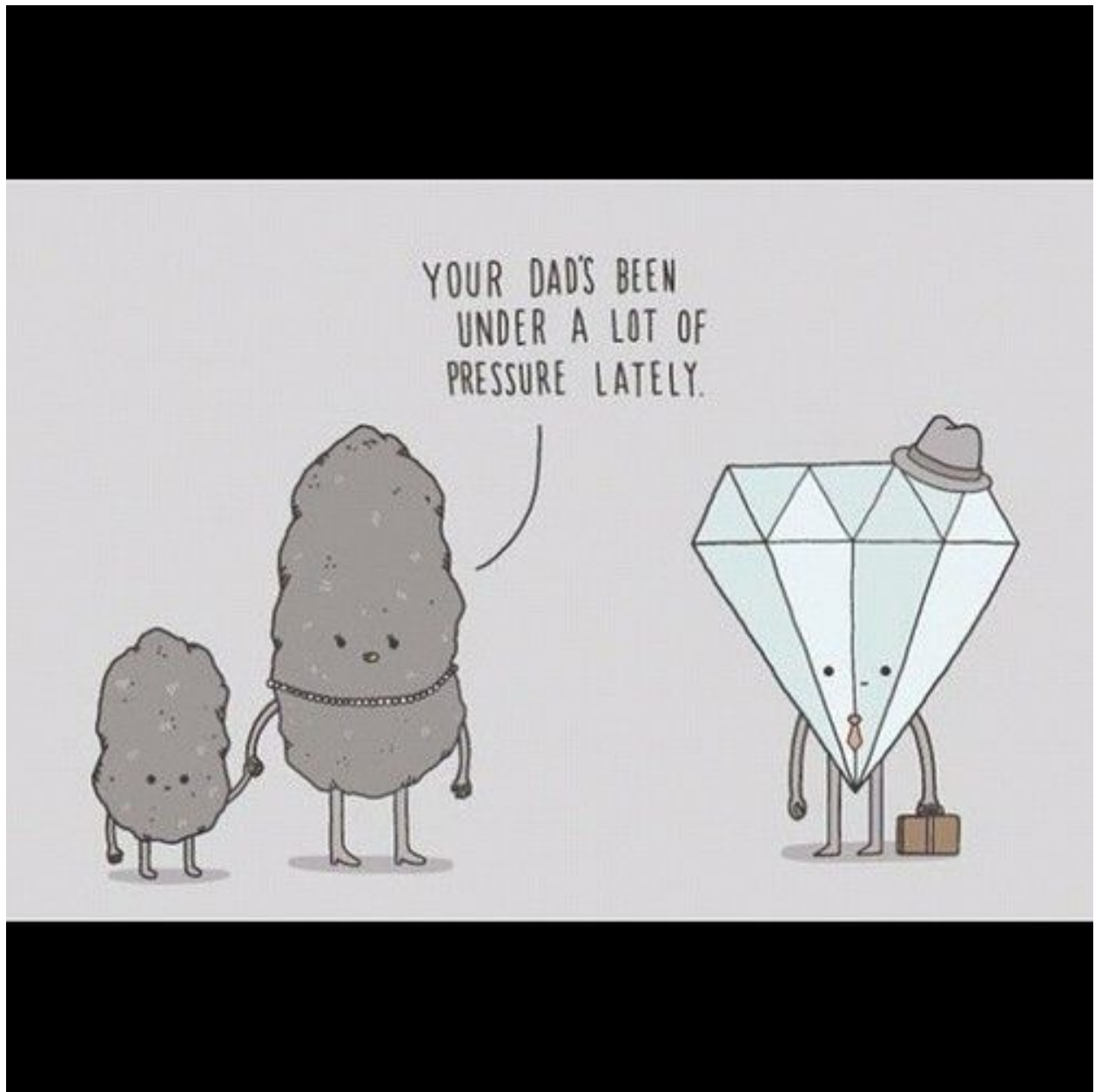
Tables of Fair and Ideal Cut Diamonds

Answers:

For six of the eight remaining variables (carat, x, y, z, depth, and table), fair cut diamonds outperform ideal cut diamonds on average. The only two variables for which ideal cut diamonds are on average better are color and clarity. It therefore makes sense that fair cut diamonds would cost more on average.

Reflection (3 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 3.