

STA130 Term Test – Section LEC0201

March 2, 2018

Last Name: Solutions

Last Name: _____

Student number: _____

Tutorial section or teaching assistant's name: _____

Instructions:

- Total marks: 45
- There are 12 pages including this page.
- Backs of pages will not be marked. Write your answers only in the space provided.
- The test is 90 minutes long.
- You are permitted a 1-sided, handwritten, 8.5 x 11 inch aid sheet. You must hand in your aid sheet with your test.
- Calculators are not permitted.
- Questions begin on the next page

Introduction

Throughout this test, we will work with the `diamonds` dataset, which contains data measured on a random sample of diamonds. The dataset contains prices (in 2008 dollars) and attributes of diamonds which might influence their price. These attributes are the “4 Cs”: `carat`, `cut`, `color`, and `clarity`. Carat is a unit of mass equal to 200 mg and is used for measuring gemstones and pearls. Cut grade is an objective measure of a diamond’s light performance, or what we generally think of as sparkle.

Here is a look at the data:

```
glimpse(diamonds)
```

```
## Observations: 53,940
## Variables: 5
## $ carat    <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, ...
## $ cut      <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very G...
## $ color    <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, ...
## $ clarity  <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI...
## $ price    <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339,...
```

```
table(diamonds$cut)
```

```
##
##      Fair      Good Very Good   Premium    Ideal
##      1610      4906      12082      13791      21551
```

```
table(diamonds$color)
```

```
##
##      D      E      F      G      H      I      J
##  6775  9797  9542 11292  8304  5422  2808
```

```
table(diamonds$clarity)
```

```
##
##      I1      SI2      SI1      VS2      VS1      VVS2      VVS1      IF
##      741   9194 13065 12258  8171   5066   3655   1790
```

Question 1

[3 marks] The following three datasets (labelled A, B, and C) were randomly selected from the `diamonds` data set. Each data set has five diamonds (observations) with the variables `carat`, `cut`, and `color`.

A.

Diamond	cut	D	F	G	H	J
1	Fair	NA	NA	NA	NA	1.51
2	Premium	NA	NA	1.27	NA	NA
3	Ideal	0.41	NA	NA	NA	NA
4	Ideal	NA	NA	NA	0.36	NA
5	Premium	NA	1.05	NA	NA	NA

B.

carat	color	1	2	3	4	5
0.33	G	NA	NA	NA	NA	Very Good
0.52	E	NA	Ideal	NA	NA	NA
1.00	E	Fair	NA	NA	NA	NA
1.01	D	NA	NA	NA	Ideal	NA
1.01	E	NA	NA	Very Good	NA	NA

C.

Diamond	carat	cut	color
1	1.00	Very Good	H
2	0.42	Ideal	D
3	0.34	Ideal	E
4	1.00	Fair	E
5	0.71	Fair	I

Only one of these data sets is tidy. Which data set is tidy? Explain why each of the other 2 datasets is not tidy.

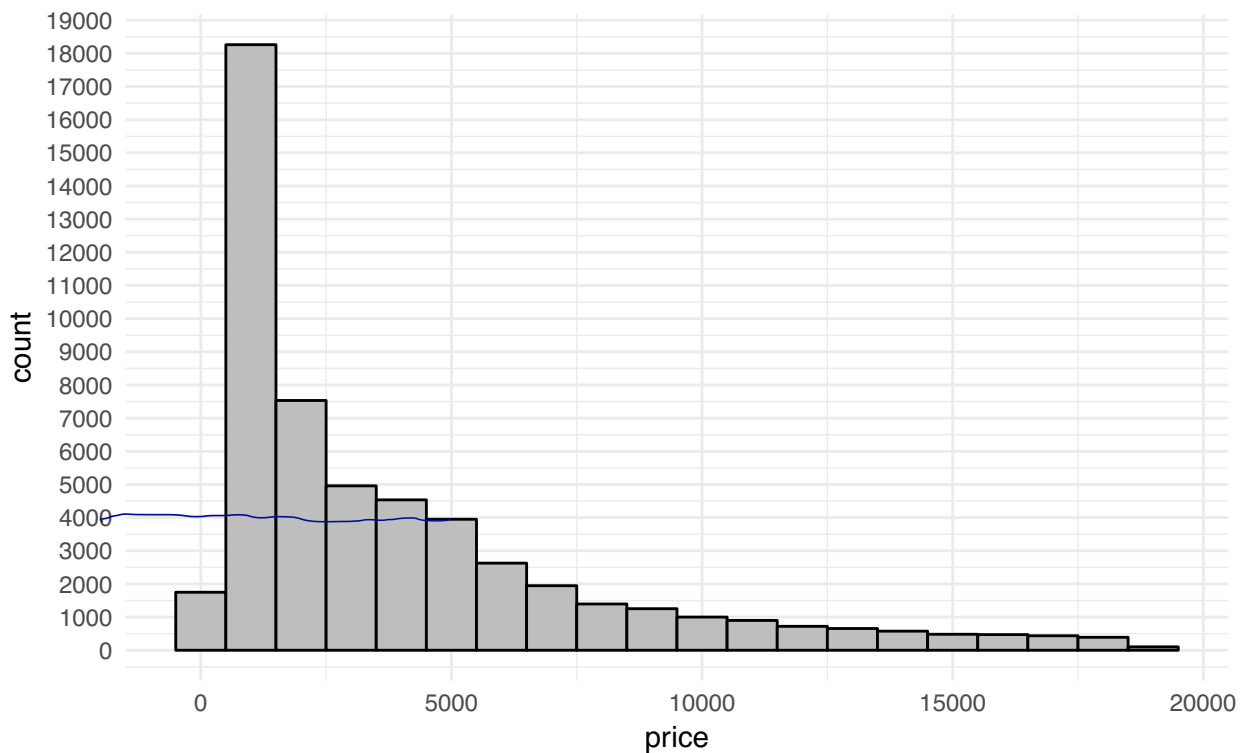
C. is tidy Since; each row corresponds to an observation, each column corresponds to a variable, and each cell is a value

A. and B. are not tidy since each column does not correspond to a variable³.

Question 2

A histogram of diamond prices is shown below.

```
diamonds %>%  
  ggplot(aes(price)) +  
  geom_histogram(colour = "black", fill = "grey", binwidth = 1000) +  
  scale_y_continuous(breaks = seq(0, 20000, by=1000)) +  
  theme_minimal()
```



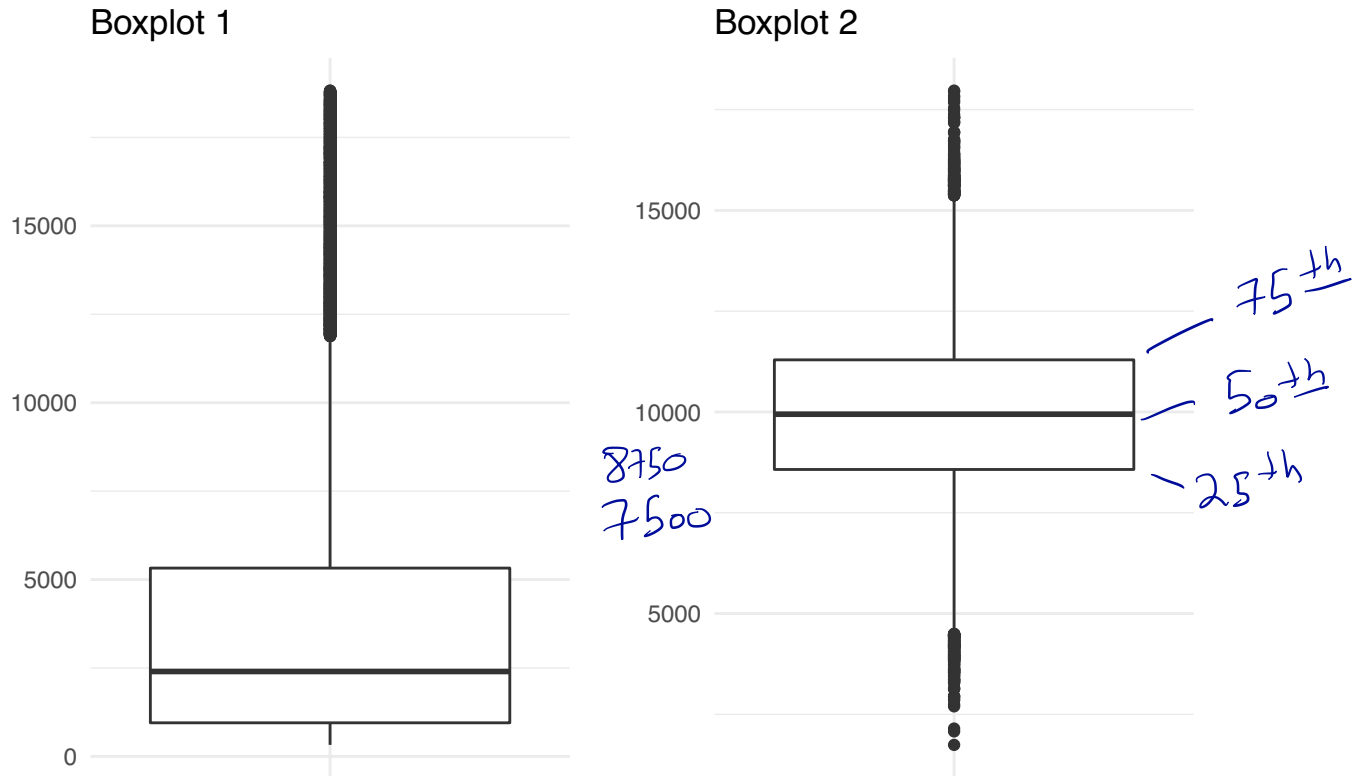
- (a) [4 marks] Let $\hat{f}(x)$ be the histogram estimator. What is $\hat{f}(5000)$? Show your work.

$$\begin{aligned}\hat{f}(x) &= \frac{\# \{x_i \text{ in same bin as } x\}}{h \cdot n} \\ &= \frac{4000}{(1000 \cdot 53940)}\end{aligned}$$

- (b) [1 mark] Describe the shape of the histogram.

Skewed to the right.

(c) Below are two boxplots, one of which is the boxplot of diamond prices.



i. [2 marks] Which boxplot is the boxplot of diamond prices? How do you know?

Boxplot 1. The median in boxplot 1 corresponds to the histogram, and the whisker above 75th percentile is longer compared to boxplot 2.

ii. [1 mark] For Boxplot 2, give an estimate of the value such that at least 75% of the data values are less than it and 25% of the data values are larger than it.

1125 is an estimate of the 75th percentile.

iii. [1 mark] What type of plot would be appropriate for examining the distribution of clarity?

Bar chart.

Question 3

[2 marks] For each set of R code below, circle the answer that provides the best explanation of what is accomplished by the R commands.

(a) R commands and output:

```
myfunction <- function(n){  
  results <- sample(diamonds$cut, n, replace = FALSE)  
  (sum(results == "Good") / n)*100  
}  
myfunction(100)
```

indicates without replacement

Sample 100 diamonds.

```
## [1] 11
```

- A. Obtain a random sample, **with replacement**, of the cuts of 100 diamonds, and calculate the percentage in the sample with a “Good” cut. 11% of the diamonds in the sample have a good cut.
- ☒ B. Obtain a random sample, **without replacement**, of the cuts of 100 diamonds, and calculate the percentage in the sample with a “Good” cut. 11% of the diamonds in the sample have a good cut.
- C. Obtain a random sample, **without replacement**, of the cuts of n diamonds, and calculate the percentage in the sample with a “Good” cut. 11% of the diamonds in the sample have a good cut.
- D. Obtain a random sample, **without replacement**, of the cuts of n diamonds, and calculate the percentage in the sample with a “Good” cut. 20% of the diamonds in the sample have a good cut.
- E. Obtain a random sample, **without replacement**, of the cuts of 100 diamonds, and calculate the percentage in the sample with a “Good” cut. 11 of the diamonds in the sample have a good cut.

(b) R commands and output:

```
x <- sum(!(diamonds$cut == "Fair" | diamonds$color == "E"))  
y <- sum((diamonds$cut == "Fair" | diamonds$color == "E"))  
z <- sum((diamonds$cut == "Fair" & diamonds$color == "E"))
```

```
##      x      y      z  
## 42757 11183   224
```

- ☒ A. 11183 diamonds have a “Fair” cut or “E” colour; 42757 diamonds are neither a “Fair” cut nor “E” color; 224 diamonds have a “Fair” cut and “E” colour.
- B. 42757 diamonds have a “Fair” cut or “E” colour; 11183 diamonds are neither a “Fair” cut nor “E” color; 224 diamonds have a “Fair” cut and “E” colour.
- C. 11183 diamonds have a “Fair” cut or “E” colour; 53716 diamonds are neither a “Fair” cut nor “E” color; 224 diamonds have a “Fair” cut and “E” colour.
- D. 11183 diamonds have a “Fair” cut or “E” colour; 42757 diamonds are neither a “Fair” cut nor “E” color; 53716 diamonds have a “Fair” cut and “E” colour.
- E. 42757 diamonds have a “Fair” cut or “E” colour; 42757 diamonds are neither a “Fair” cut nor “E” color; 224 diamonds have a “Fair” cut and “E” colour.

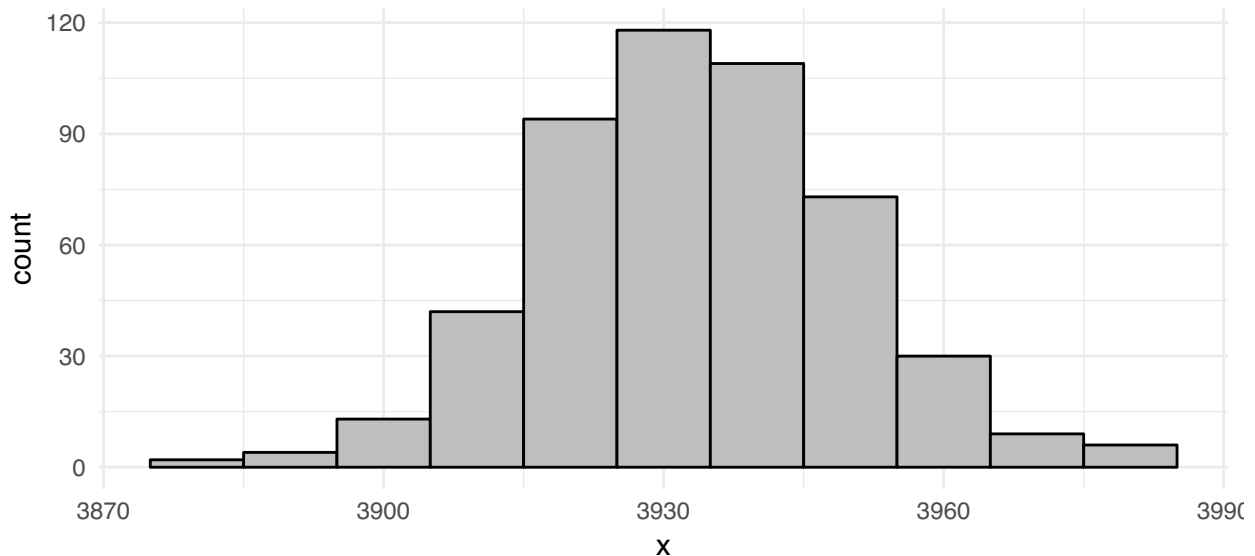
Question 4

Consider the following R code and the plot it produces:

```
N <- 500
M <- as.numeric(diamonds %>% summarize( n() ))
x <- rep(NA, N)
for (i in 1:N) {
  y <- diamonds %>% sample_n(size = M, replace = TRUE)
  x[i] <- as.numeric(y %>% summarize(mean(price)))
}

dat <- data_frame(x)

dat %>% ggplot(aes(x)) +
  geom_histogram(binwidth = 10, colour = "black", fill = "grey") +
  theme_minimal()
```



```
# information for confidence interval 1
quantile(dat$x, c(0.075, 0.925))
```

```
##      7.5%    92.5%
## 3910.790 3957.642
```

```
# information for confidence interval 2
quantile(dat$x, c(0.05, 0.95))
```

```
##      5%     95%
## 3907.630 3961.328
```

Answer the questions on the next page based on this R code.

(a) [1 mark] The histogram shows the results of a simulation. What is the purpose of this simulation? Circle the correct answer.

- A. Examine the mean of price for the population.
- B. Examine how price varies from sample to sample for a samples of a certain size.
- C. Examine the sampling distribution of the mean for a samples of a certain size.
- D. Examine the distribution of a test statistic for an hypothesis test.
- ☒ E. Examine the bootstrap sampling distribution of the mean.

(b) [3 marks] The code includes the R objects N, M, and x. Explain the role that each of these objects plays in the code.

N : The number of Simulations 500

M: The number of observations 53,940

x: Vector of length 500, each element
contains the mean of the bootstrapped
sample.

(c) [6 marks] Quantitative information for two confidence intervals is indicated by R comments in the code next to the information. Use this information to fill in the blanks in the following statements:

Confidence interval 1:

A 85 % confidence interval for the mean price of diamonds is: 3910.79 - 3957.642

Confidence interval 2:

A 90 % confidence interval for the mean price of diamonds is: 3907.63 - 3961.328

$(1 - 0.05 \times 2) \times 100$

Question 5

A diamond's clarity refers to the presence of (or absence of) blemishes in the diamond that are visible under 10 times magnification. Clarity grades range from Internally Flawless (diamonds which are completely free of blemishes) to Imperfect 3 (diamonds which possess large, heavy blemishes that are visible to the naked eye). A diamond is given clarity grade "IF" if it is Internally Flawless, and a grade of "VVS1" or "VVS2" if only an expert can detect flaws under 10 times magnification.

We will investigate whether there is a difference in price between diamonds with clarity grade "IF" and diamonds with clarity grade either "VVS1" or "VVS2".

- (a) [3 marks] The first step in comparing clarity categories "IF" to "VVS1" and "VVS2" is to create a new variable that has only two categories of clarity: "IF" and "VVS1-VVS2" (this second category includes all diamonds that are either "VVS1" or "VVS2").

Here is R code and output that produces two data frames, `dat1` and `dat2`. Only one of these data frames is the data frame we need.

```
dat1 <- diamonds %>%  
  filter(clarity == "IF" | clarity == "VVS1" | clarity == "VVS2" ) %>%  
  mutate(clarity2 = ifelse(clarity == "IF", "IF", "VVS1-VVS2"))  
dat1 %>% count(clarity2)
```

```
## # A tibble: 2 x 2  
##   clarity2      n  
##   <chr> <int>  
## 1      IF  1790  
## 2 VVS1-VVS2 8721
```

```
dat2 <- diamonds %>%  
  select(clarity) %>%  
  mutate(clarity2 = ifelse(clarity == "IF", "IF", "VVS1-VVS2"))  
dat2 %>% count(clarity2)
```

```
## # A tibble: 2 x 2  
##   clarity2      n  
##   <chr> <int>  
## 1      IF  1790  
## 2 VVS1-VVS2 52150
```

Which data frame, `dat1` or `dat2`, correctly creates the object `clarity2` with two categories where one category contains diamonds with clarity "IF" and the other category contains diamonds with category "VVS1-VVS2"? Briefly explain what is wrong with the other data frame.

dat1 is the correct data frame. dat2 includes diamonds with all clarities. So, the category VVS1-VVS2 contains diamonds with all clarities except VVS1.

Continuing with the analysis from the previous page, the data frame with the correct version of `clarity2` was renamed to `dat`. Recall that the purpose of the analysis is to investigate whether there is a difference in price between diamonds with clarity grade “IF” and diamonds with clarity grade “VVS1-VVS2”. Some code and output for this investigation are below.

```
mean_data <- dat %>% group_by(clarity2) %>% summarise(means = mean(price))
mean_data
```

```
## # A tibble: 2 x 2
##   clarity2     means
##   <chr>     <dbl>
## 1      IF 2864.839
## 2 VVS1-VVS2 2964.958
```

```
test_stat <- as.numeric(mean_data %>% summarise(test_stat = diff(means)))
test_stat
```

```
## [1] 100.1186
```

```
repetitions <- 500
simulated_stats <- rep(NA, repetitions)
```

```
for (i in 1:repetitions)
{
  sim <- dat %>% mutate(clarity2 = sample(clarity2))

  sim_test_stat <- sim %>%
    group_by(clarity2) %>%
    summarise(means = mean(price)) %>%
    summarise(sim_test_stat = diff(means))

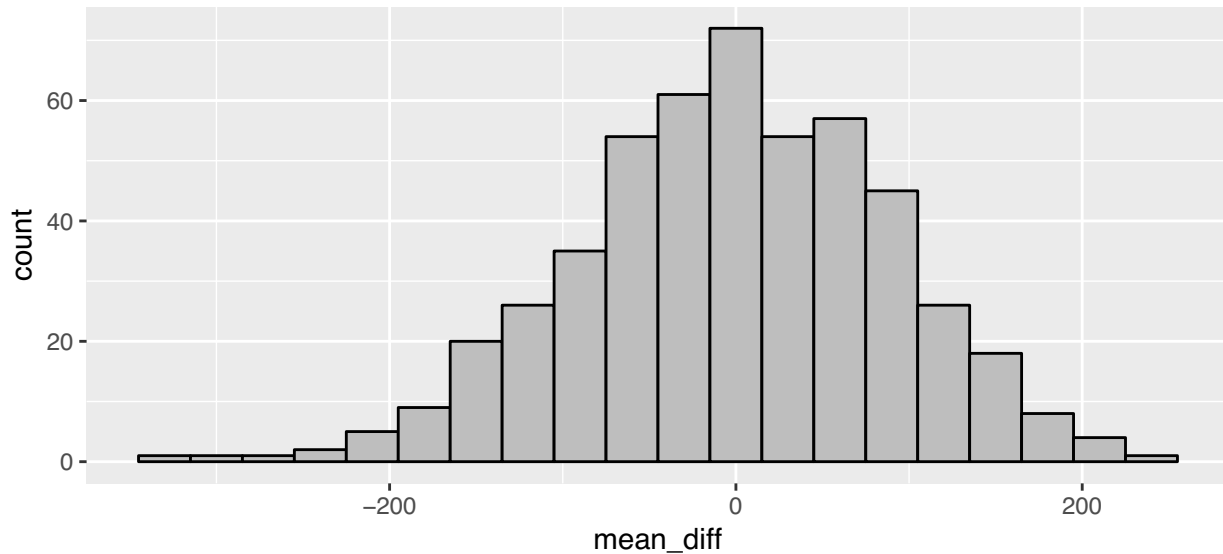
  simulated_stats[i] <- as.numeric(sim_test_stat)
}
```

```
sim <- data_frame(mean_diff = simulated_stats)
```

```
sim %>%
  filter(mean_diff >= abs(test_stat) | mean_diff <= -1*abs(test_stat)) %>%
  summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   133
```

```
sim %>% ggplot(aes(x = mean_diff)) +
  geom_histogram(colour = "black", fill = "grey", binwidth = 30)
```



- (b) [2 marks] What is the null hypothesis for this test? Define the parameters.

$H_0: \mu_{IF} = \mu_{VVS1-VVS2}$. μ_{IF} , $\mu_{VVS1-VVS2}$ are the mean prices of diamonds with clarity IF, VVS1-VVS2 respectively.

- (c) [1 mark] The value of the test statistic in this analysis is 100.1186.

- (d) [2 marks] Explain why the histogram is centred at 0.

The sampling distribution is constructed assuming the null hypothesis is true, and it is the distribution of the difference in the sample means which the null hypothesis assumes is zero.

- (e) [1 mark] The P-value for testing the difference between the mean price of diamonds with clarity "IF" and clarity "VVS1-VVS2" is 133/500 = 0.266

- (f) [1 mark] TRUE or FALSE (Circle one): The P-value for the test is calculated under the assumption that the mean price is the same for diamonds with clarity "IF" and diamonds with clarity "VVS1-VVS2".

- (g) [1 mark] TRUE or FALSE (Circle one): The P-value for the test implies that we have strong evidence that the observed mean difference in price between clarity "IF" and clarity "VVSS1-VVS2" is not due to chance.
- (h) [2 marks] Explain what is wrong with the following explanation of the P-value: "The P-value is the probability that the null hypothesis is true."

The p-value is the probability of observing a value at least as extreme as the test stat. assuming H_0 is true. This is different than the probability that H_0 is true.

- (i) [8 marks] Summarize the results of the hypothesis test in a four to six sentence paragraph, using complete English sentences. The paragraph should address what is being tested, what was observed in the data, the P-value and what it means, and an appropriate conclusion.

Comparing the mean price of diamonds with Clarity IF to VVSS1-VVS2 is the primary question.

The observed difference in price is

\$100.12. Assuming there is no difference in the mean price between the groups, 133 of the 500 (0.266) simulated mean differences were as extreme as the test statistic.

Therefore, we have no evidence

against the hypothesis that the

mean prices are equal for Clarity

IF and VVSS1-VVS2.