Lab 3 – Simulation Study with Diamonds

NAME 1 – varenya3

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please include all requested responses in a document, then save it as a **pdf** when done.
  - You may use this instructions document, or you may create a new document.
  - All responses should be numbered (leaving the original question text is optional!)
- Upload your pdf to **Gradescope** and please **match pages** with the **question number** when prompted to.
- If working with one or two **partners**, be sure to do **both** of these things:
  - Please put all **names and netIDs** at the top of your document (like shown above).
  - Have one person upload the pdf and then ensure **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- For this lab, we're going to explore distributions and the variability of a sample statistic through simulation!
- To do this, we will explore the diamonds dataset saved in the ggplot2 package. This dataset has over 59,000 diamonds catalogued, and we will treat this dataset like it's a population.
- Let's see how much variation we see from sample to sample and how reliable our sample statistics are in different situations!

Step 0

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 3 first: https://stat212-learnr.stat.illinois.edu/
- **Open RStudio** (or RStudio Cloud) to get started and **open** a **new script**.
- We will be using the `diamonds` dataset stored in the tidyverse package. So start by running `library(tidyverse)`
- Open the `diamonds` data by running the code: `View(diamonds)`. Each row represents one diamond from a collection of over 59,000.
- Take a look at the documentation for `diamonds` by running the code: `?diamonds`

Question 1 (5pts): Create a histogram of the `price` variable (*For all histograms in this assignment, use the base R function `hist`).* Also calculate the mean and standard deviation of this variable.

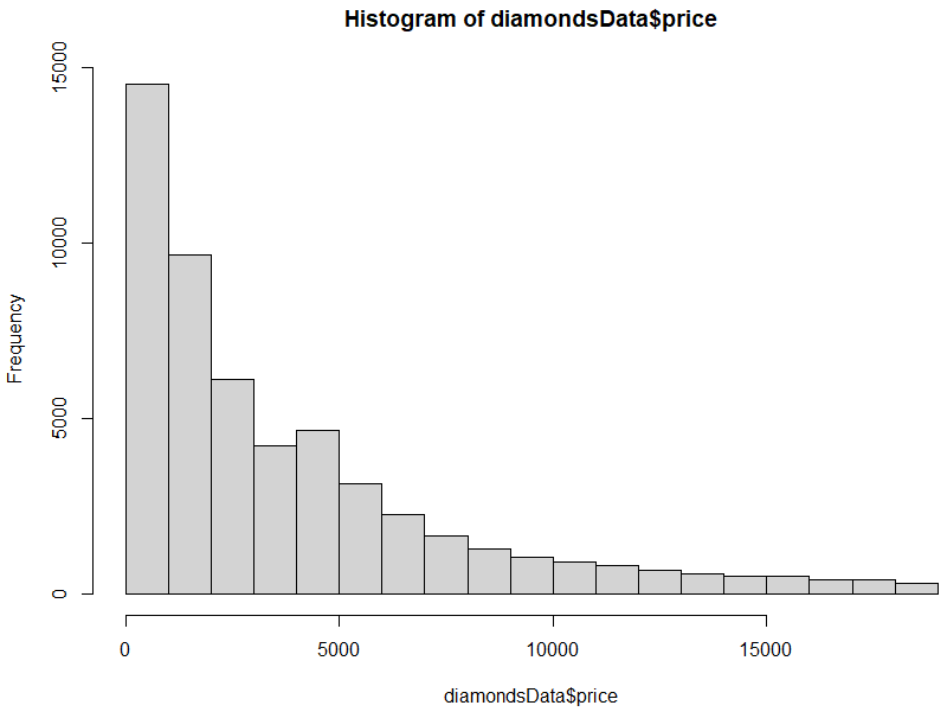
Include the image of your histogram in your report *(you may either save it to your computer and upload it, or include a properly cropped screenshot).*

Include the mean and standard deviation values

Briefly describe the shape of this variable. Is this a symmetric distribution or would you say it's skewed?

```
$ The graph appears to be right-skewed
> dia_price_avg
$ [1] 3932.8
> dia_price_stdev
$ [1] 3989.44

$ "The mean is 3932.799722 and the standard deviation is 3989.439738"
```



Question 2 (5pts): Take a random sample of 50 diamond prices from this dataset and name this vector `fifty_diam` *(If saved properly, you will see this vector of length 50 saved in your global environment!).* Sample without replacement (this will be the default option). Create a histogram of your sample, and then calculate the mean and standard deviation of this sample.

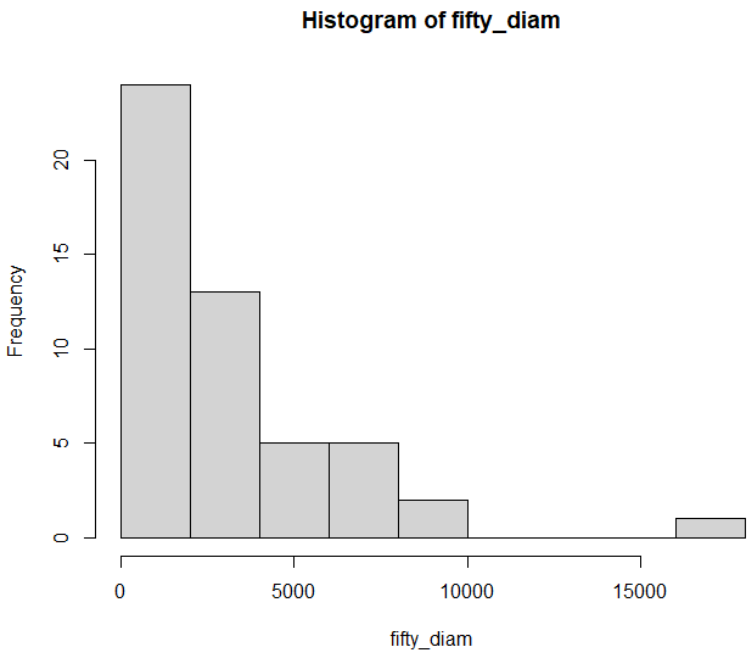
Include the image of your histogram in your report

Include the mean and standard deviation values

What is the *absolute* error of *your* sample mean as an estimate of the true mean? *(for example: if your estimate was 85 and the true value was 100, that would be an absolute error of 15).*

What is the *absolute* error of *your* sample standard deviation (SD) as an estimate of the true SD?

```
- > dia_price_sample_avg
- $ [1] 3150.18
- > dia_price_sample_stdev
- $ [1] 3051.461
- > absolute_error_of_sample_mean
- $ [1] 782.6197
- > absolute_error_of_sample_stdev
- $ [1] 937.9783
-
- $ For this sample the mean is 3150.180000 and the standard deviation is 3051.461441
- $ The absolute error of my sample mean as an estimate of the true mean: 782.619722
- $ The absolute error of my sample SD as an estimate of the true SD is 937.978297
-
-
```



Question 3 (5pts): Next, set up a `for` loop to simulate taking a sample of size 50 at least 10,000 times. Inside your loop, calculate the mean price and save it to a vector called `means`. Here are two tips:

- Remember before the loop to define `means = NULL` so that your loop knows where to save the means.
- Remember inside the loop to include an index indicator with your means vector so that the vector fills iteratively for each iteration of the loop.
- Try running the loop 10 times to ensure it works. This should be instantaneous. Then try running it 10,000 times. The loop should only take a few seconds to complete at 10,000 simulations, so if you wait more than a minute, click the stop button and see if something is defined incorrectly.

After successfully running your simulation, create a histogram of your `means` vector. Again, continue to use the `hist()` function for all histograms in this lab.

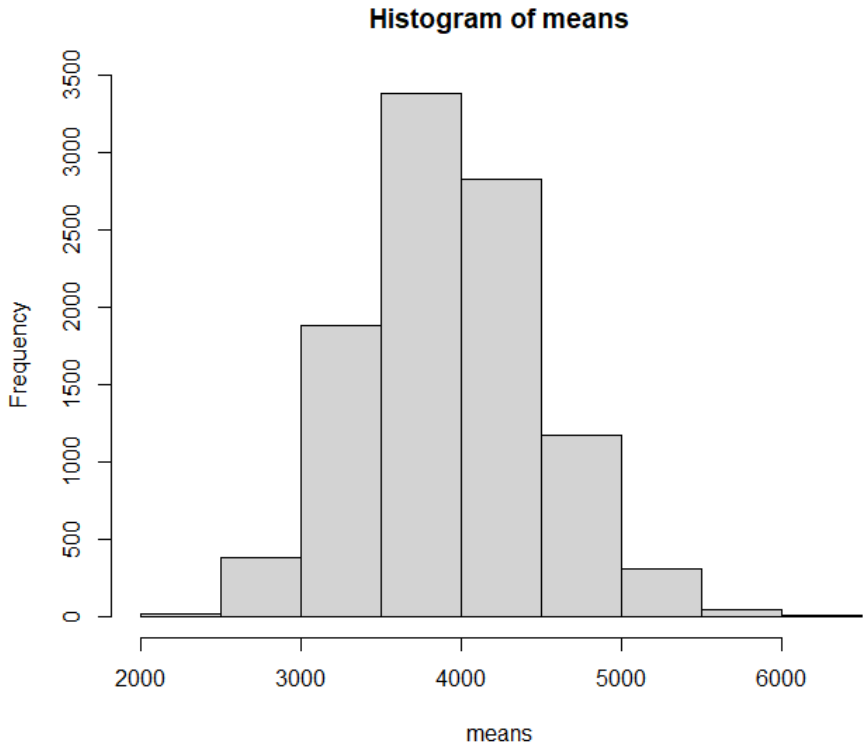
Include the image of your histogram in your report

Include the R code you used to generate this loop

Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? How does this relate to the Central Limit Theorem we learned in class?

```
-  means = NULL
-  sampleSize = 50
-  for (i in 1:10000) {
-    means[i] = mean(sample(x = diamondsData$price, size = sampleSize, replace = TRUE)
   )
-  }
-  means
-  hist(means)
-  print("The graph appears to be a symmetric distribution and relates to the Central
   Limit Theorem in as we take more random samples the graphing of each outcome event
   results in a closer approximation of the theoretical data or Population Distributio
   n in shape. Here it is close to a normally/symmetrically distrubuted graph.")
-  print("Through the Law of Large Numbers, the value of the Sampling Distribution is
   Converging to the Population Mean")
-
```

```
- $ The graph appears to be a symmetric distribution and relates to the Central Limit
  Theorem in that as we take more random samples the graphing of each outcome event,
  the results form a closer approximation Population Distribution in shape.
-
- $ The value of the Sampling Distribution is Converging to the Population Mean based
  on the Law of Large Numbers.
-
-
```



Question 4 (5pts): As you should notice from your histogram, our sample means will vary with each sample we take. Calculate the standard deviation of the `means` vector.

Report the standard deviation of the simulated means

Notice that every time you run your loop again, your vector of sample means will change, and so will your standard deviation of those simulated sample means. What measure is the standard deviation of the simulated means approximating? **Report the <u>name</u> of this measure and calculate the <u>true value</u> for this measure too** *(hint: check the "Distribution of a Sample Statistic" notes!)*

```
- means_sample_avg <- mean(means)
- means_sample_stdev <- sd(means)
- means_expected_error <- (sd(means)) / (sqrt(sampleSize))
- sprintf("For the 'means' vector of samples the average is %f and the standard devia
  tion is %f", means_sample_avg, means_sample_stdev)
- sprintf("The standard deviation %f the simulated means is the 'Standard Error of a
  Sample Mean', and is approximating the 'Expected Error of the Sample Mean' whose tr
  ue value is %f", means_sample_stdev, means_expected_error)
- #
- > means_sample_avg
- $ [1] 3927.925
- > means_sample_stdev
- $ [1] 559.6752
- > means_expected_error
```

```
- $ [1] 79.15003
- $
- $ The average is 3927.924594 and the standard deviation is 559.675235
- $ The standard deviation 559.675235 of the simulated means is 'Standard Error of a
  Sample Mean', and is approximating the 'Expected Error of the Sample Mean' of whose
  true value is 79.150031
-
- $
```

Question 5 (5pts): Repeat question 3, but with a sample size of 10 instead of 50. Call your vector of sample means `means_ten`. After successfully running your simulation, create a histogram of your `means_ten` vector using the `hist` function again.

Include the image of your histogram in your report

Include the R code you used to generate this loop

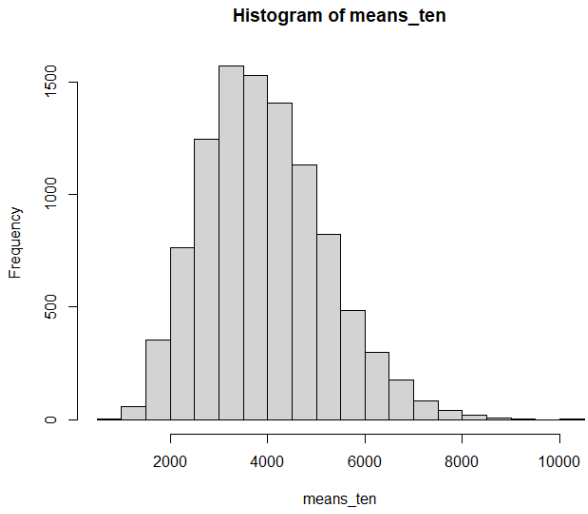
Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? How does this relate to the Central Limit Theorem we learned in class?

Is the standard deviation of the simulated means higher or lower than it was for n = 50?

```
- means_ten = NULL
- means_ten_sampleSize = 10
- for (i in 1:10000) {
-   means_ten[i] = mean(sample(x = diamondsData$price, size = means_ten_sampleSize, r
  eplace = TRUE))
- }
- means_ten
- hist(means_ten)
- print("The graph appears to be a normally distributed graph but slightly right-skew
  ed, relating to the Central Limit Theorem in as we take more random samples the gra
  phing of each outcome event results in a closer approximation of the theoretical da
  ta or Population Distribution in shape. Here it is close to a normally/symmetricall
  y distrubuted graph - but the right-half portion is skewed")
- print("Through the Law of Large Numbers, the value of the Sampling Distribution is
  Converging to the Population Mean")
- means_ten_sample_avg <- mean(means_ten)
- means_ten_sample_stdev <- sd(means_ten)
- means_ten_expected_error <- (sd(means_ten)) / (sqrt(means_ten_sampleSize))
- sprintf("For the 'means' vector of samples the average is %f and the standard devia
  tion is %f", means_ten_sample_avg, means_ten_sample_stdev)
- sprintf("The standard deviation %f the simulated means is the 'Standard Error of a
  Sample Mean', and is approximating the 'Expected Error of the Sample Mean' whose tr
  ue value is %f", means_ten_sample_stdev, means_ten_expected_error)
- means_sample_MINUS_ten_sample_stdev = (sd(means_ten)) - (sd(means))
- sprintf("The standard deviation of the 'new' simulated means is lower than is was f
  or 'old' n = 50; Respectively the St.Dev's are %f and %f, with a difference of %f",
  means_ten_sample_stdev, means_sample_stdev , means_sample_MINUS_ten_sample_stdev)
```

- **The graph appears to be a normally distributed graph but slightly right-skewed, relating to the Central Limit Theorem in as we take more random samples the graphing of each outcome event results in a closer approximation of the theoretical data or Population Distribution in shape. Here it is close to a normally/symmetrically distrubuted graph - but the right-half portion is skewed.**
- **Through the Law of Large Numbers, the value of the Sampling Distribution is Converging to the Population Mean**

- **For the 'means' vector of samples the average is 3940.896040 and the standard deviation is 1255.158361**
- **The standard deviation 1255.158361 the simulated means is the 'Standard Error of a Sample Mean', and is approximating the 'Expected Error of the Sample Mean' whose true value is 396.915924**
- **The standard deviation of the 'new' simulated means is lower than is was for 'old' n = 50; Respectively the St.Dev's are 1255.158361 and 559.675235, with a difference of 695.483125**



-

Question 6 (5pts): We spent some time exploring the behavior of the sample mean, but now let's look at the **sample median**! Redo question 3 with a sample size of 50, but now calculate the sample median inside your loop. Call your vector of sample medians `medians_fifty`. After successfully running your simulation, create a histogram of your `medians_fifty` vector using the `hist` function again.

Include the image of your histogram in your report

Include the R code you used to generate this loop

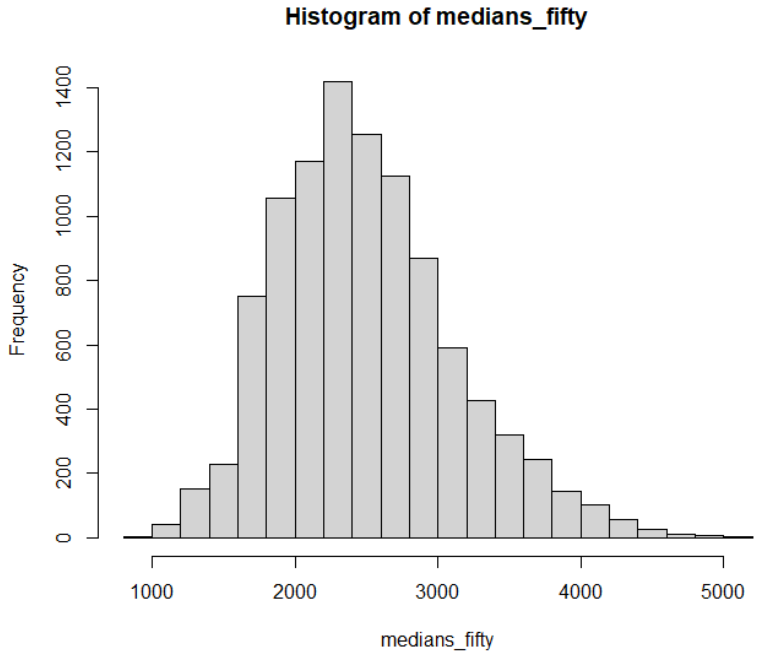
Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? Do you have any predictions for what would happen if we repeated this simulation again, but with a much larger sample size?

Calculate and report the standard deviation of the `medians_fifty` vector. *This is the expected error in a randomly generated sample median as an estimate of the true median.*

```
-   medians_fifty = NULL
-   medians_fifty_sampleSize = 50
-   for (i in 1:10000) {
-     medians_fifty[i] = median(sample(x = diamondsData$price, size = medians_fifty_sam
    pleSize, replace = TRUE))
-   }
-   medians_fifty
-   hist(medians_fifty)
-   print("The graph appears to be a normally distributed graph but slightly right-skew
    ed, relating to the Central Limit Theorem in as we take more random samples the gra
    phing of each outcome event results in a closer approximation of the theoretical da
    ta or Population Distribution in shape. Here it is close to a normally/symmetricall
    y distrubuted graph - but the right-half portion is skewed")
```

```
- print("Through the Law of Large Numbers, the value of the Sampling Distribution is
  Converging to the Population Mean. I predict that if we repeated this simulation ag
  ain with a much larger sample size we would observe a more-normally distributed gra
  ph, or at least one which shows a much more precise skew/curve on the right half")
- medians_fifty_sample_avg <- mean(medians_fifty)
- medians_fifty_sample_stdev <- sd(medians_fifty)
- sprintf("For the 'medians_fifty' vector of samples the average is %f and the standa
  rd deviation is %f", medians_fifty_sample_avg, medians_fifty_sample_stdev)
- print("This standard deviation is the expected error in a randomly generated sample
  median as an estimate of the true median.")
```

- **"The graph appears to be a normally distributed graph but slightly right-skewed, relating to the Central Limit Theorem in as we take more random samples the graphing of each outcome event results in a closer approximation of the theoretical data or Population Distribution in shape. Here it is close to a normally/symmetrically distrubuted graph - but the right-half portion is skewed**
- **Through the Law of Large Numbers, the value of the Sampling Distribution is Converging to the Population Mean. I predict that if we repeated this simulation again with a much larger sample size we would observe a more-normally distributed graph, or at least one which shows a much more precise skew/curve on the right half**
- **For the 'medians_fifty' vector of samples the average is 2495.116700 and the standard deviation is 615.629535**
- **This standard deviation is the expected error in a randomly generated sample median as an estimate of the true median.**



-

Question 7 (5pts): Repeat question 6, but with a sample size of 500.

Include the image of your histogram in your report

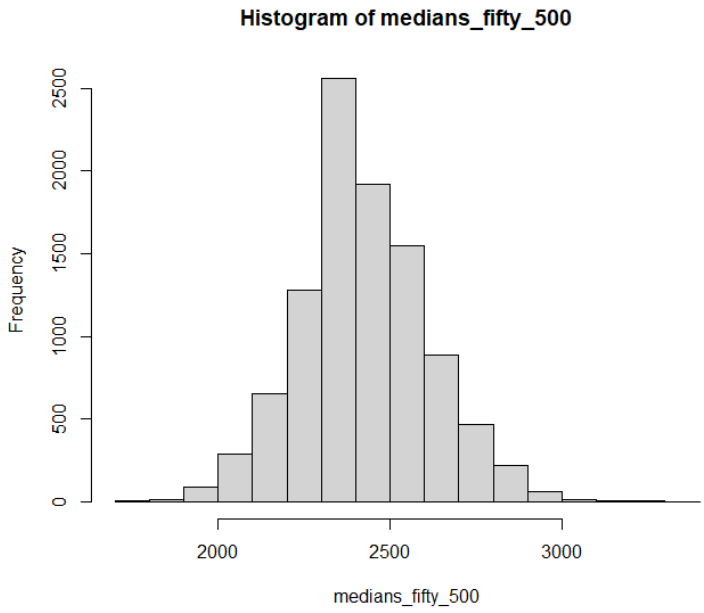
Include the R code you used to generate this loop

Briefly describe the shape of your histogram. How has the shape changed in comparison to the distribution of sample medians when we took samples of size 50?

Calculate and report the standard deviation of your newest vector of medians. **How does this expected error compare to when we had samples of size 50? Is this expected or surprising to you?**

```
- medians_fifty_500 = NULL
- medians_fifty_sampleSize_500 = 500
- for (i in 1:10000) {
-   medians_fifty_500[i] = median(sample(x = diamondsData$price, size = medians_fifty
  _sampleSize_500, replace = TRUE))
- }
- medians_fifty_500
- hist(medians_fifty_500)
- print("The graph appears to be a normally distributed graph but yet again slightly
  right-skewed, relating to the Central Limit Theorem in as we take more random sampl
  es the graphing of each outcome event results in a closer approximation of the theo
  retical data or Population Distribution in shape. Here it is close to a normally/sy
  mmetrically distrubuted graph - but the right-half portion is skewed")
- print("Compared to the previous sample size which was 10 times less, this new histo
  gram is closer to an ideally symmetrical representation")
- medians_fifty_500_sample_avg <- mean(medians_fifty_500)
- medians_fifty_500_sample_stdev <- sd(medians_fifty_500)
- sprintf("For the 'medians_fifty_500' vector of samples the average is %f and the st
  andard deviation is %f", medians_fifty_500_sample_avg, medians_fifty_500_sample_std
  ev)
- medians_stdev_diff <- medians_fifty_sample_stdev - medians_fifty_500_sample_stdev
- sprintf("This standard deviation is the expected error in a randomly generated samp
  le median as an estimate of the true median. It is lower than the previous standard
  deviation by %f, and is expected since more sampling results in closer, more accura
  te, more precise data points", medians_stdev_diff)
```

- **The graph appears to be a normally distributed graph but yet again slightly right-skewed, relating to the Central Limit Theorem in as we take more random samples the graphing of each outcome event results in a closer approximation of the theoretical data or Population Distribution in shape. Here it is close to a normally/symmetrically distrubuted graph - but the right-half portion is skewed**
- **Compared to the previous sample size which was 10 times less, this new histogram is closer to an ideally symmetrical representation**
- **For the 'medians_fifty_500' vector of samples the average is 2418.717300 and the standard deviation is 186.744921**
- **This standard deviation is the expected error in a randomly generated sample median as an estimate of the true median. It is lower than the previous standard deviation by 428.884613, and is expected since more sampling results in closer, more accurate, more precise data points**


Histogram of medians_fifty_500
Frequency
0
500
1000
1500
2000
2500
2000
2500
3000
medians_fifty_500

```
cat("\014")
shell("cls")
#install.packages("tidyverse")
library(tidyverse)
library(dplyr)
library(ggplot2)
View(diamonds)
?diamonds
head(diamonds)
str(diamonds)
######################################################################
#1
cat("\014")
shell("cls")
#create data frame
diamondsData <- diamonds
#simple R histogram of Prices
hist(diamondsData$price)
#mean and standard deviation
dia_price_avg <- mean(diamondsData$price)
dia_price_stdev <- sd(diamondsData$price)
sprintf("The mean is %f and the standard deviation is %f", dia_price_avg, dia_price_stdev)
print("The graph appears to be right-skewed")
######################################################################

#variance of Prices
var(diamondsData$price)
#Simple 5-number summary of Prices
summary(diamondsData$price)
#take the data frame and give a summary of mean and st_dev of Prices
diamondsData %>% summarize(mean = mean(price), std_dev = sd(price))
#split the data into subsets of Cut and thencompute Price mean/summary statistics for each, omitting
any NA values if there are
aggregate(diamonds$price, by=list(diamonds$cut), FUN=mean, na.rm=TRUE)
######################################################################


######################################################################
#2
cat("\014")
shell("cls")
fifty_diam <- sample(x = diamondsData$price, size = 50, replace = FALSE)
length(fifty_diam)
hist(fifty_diam)
dia_price_sample_avg <- mean(fifty_diam)
dia_price_sample_stdev <- sd(fifty_diam)
absolute_error_of_sample_mean = abs(mean(diamondsData$price) - mean(fifty_diam))
absolute_error_of_sample_stdev = abs(sd(diamondsData$price) - sd(fifty_diam))
sprintf("For this sample the mean is %f and the standard deviation is %f", dia_price_sample_avg,
dia_price_sample_stdev)
sprintf("The absolute error of my sample mean as an estimate of the true mean is %f",
absolute_error_of_sample_mean)
sprintf("The absolute error of my sample standard deviation (SD) as an estimate of the true SD is
%f", absolute_error_of_sample_stdev)
######################################################################


######################################################################
#3
cat("\014")
shell("cls")
means = NULL
sampleSize = 50
for (i in 1:10000) {
```

```
  means[i] = mean(sample(x = diamondsData$price, size = sampleSize, replace = TRUE))
}
means
hist(means)
print("The graph appears to be a symmetric distribution and relates to the Central Limit Theorem in
as we take more random samples the graphing of each outcome event results in a closer approximation
of the theoretical data or Population Distribution in shape. Here it is close to a
normally/symmetrically distrubuted graph.")
print("Through the Law of Large Numbers, the value of the Sampling Distribution is Converging to the
Population Mean")
#####################################################################


#####################################################################
#4
cat("\014")
shell("cls")
means_sample_avg <- mean(means)
means_sample_stdev <- sd(means)
means_expected_error <- (sd(means)) / (sqrt(sampleSize))
sprintf("For the 'means' vector of samples the average is %f and the standard deviation is %f",
means_sample_avg, means_sample_stdev)
sprintf("The standard deviation %f the simulated means is the 'Standard Error of a Sample Mean', and
is approximating the 'Expected Error of the Sample Mean' whose true value is %f", means_sample_stdev,
means_expected_error)
#####################################################################


#####################################################################
#5
cat("\014")
shell("cls")
means_ten = NULL
means_ten_sampleSize = 10
for (i in 1:10000) {
  means_ten[i] = mean(sample(x = diamondsData$price, size = means_ten_sampleSize, replace = TRUE))
}
means_ten
hist(means_ten)
print("The graph appears to be a normally distributed graph but slightly right-skewed, relating to
the Central Limit Theorem in as we take more random samples the graphing of each outcome event
results in a closer approximation of the theoretical data or Population Distribution in shape. Here
it is close to a normally/symmetrically distrubuted graph - but the right-half portion is skewed")
print("Through the Law of Large Numbers, the value of the Sampling Distribution is Converging to the
Population Mean")
means_ten_sample_avg <- mean(means_ten)
means_ten_sample_stdev <- sd(means_ten)
means_ten_expected_error <- (sd(means_ten)) / (sqrt(means_ten_sampleSize))
sprintf("For the 'means' vector of samples the average is %f and the standard deviation is %f",
means_ten_sample_avg, means_ten_sample_stdev)
sprintf("The standard deviation %f the simulated means is the 'Standard Error of a Sample Mean', and
is approximating the 'Expected Error of the Sample Mean' whose true value is %f",
means_ten_sample_stdev, means_ten_expected_error)
means_sample_MINUS_ten_sample_stdev = (sd(means_ten)) - (sd(means))
sprintf("The standard deviation of the 'new' simulated means is lower than is was for 'old' n = 50;
Respectively the St.Dev's are %f and %f, with a difference of %f", means_ten_sample_stdev,
means_sample_stdev , means_sample_MINUS_ten_sample_stdev)
#####################################################################


#####################################################################
#6
cat("\014")
shell("cls")
medians_fifty = NULL
```

```
medians_fifty_sampleSize = 50
for (i in 1:10000) {
  medians_fifty[i] = median(sample(x = diamondsData$price, size = medians_fifty_sampleSize, replace =
TRUE))
}
medians_fifty
hist(medians_fifty)
print("The graph appears to be a normally distributed graph but slightly right-skewed, relating to
the Central Limit Theorem in as we take more random samples the graphing of each outcome event
results in a closer approximation of the theoretical data or Population Distribution in shape. Here
it is close to a normally/symmetrically distrubuted graph - but the right-half portion is skewed")
print("Through the Law of Large Numbers, the value of the Sampling Distribution is Converging to the
Population Mean. I predict that if we repeated this simulation again with a much larger sample size
we would observe a more-normally distributed graph, or at least one which shows a much more precise
skew/curve on the right half")
medians_fifty_sample_avg <- mean(medians_fifty)
medians_fifty_sample_stdev <- sd(medians_fifty)
sprintf("For the 'medians_fifty' vector of samples the average is %f and the standard deviation is
%f", medians_fifty_sample_avg, medians_fifty_sample_stdev)
print("This standard deviation is the expected error in a randomly generated sample median as an
estimate of the true median.")
######################################################################


######################################################################
#7
cat("\014")
shell("cls")
medians_fifty_500 = NULL
medians_fifty_sampleSize_500 = 500
for (i in 1:10000) {
  medians_fifty_500[i] = median(sample(x = diamondsData$price, size = medians_fifty_sampleSize_500,
replace = TRUE))
}
medians_fifty_500
hist(medians_fifty_500)
print("The graph appears to be a normally distributed graph but yet again slightly right-skewed,
relating to the Central Limit Theorem in as we take more random samples the graphing of each outcome
event results in a closer approximation of the theoretical data or Population Distribution in shape.
Here it is close to a normally/symmetrically distrubuted graph - but the right-half portion is
skewed")
print("Compared to the previous sample size which was 10 times less, this new histogram is closer to
an ideally symmetrical representation")
medians_fifty_500_sample_avg <- mean(medians_fifty_500)
medians_fifty_500_sample_stdev <- sd(medians_fifty_500)
sprintf("For the 'medians_fifty_500' vector of samples the average is %f and the standard deviation
is %f", medians_fifty_500_sample_avg, medians_fifty_500_sample_stdev)
medians_stdev_diff <- medians_fifty_sample_stdev - medians_fifty_500_sample_stdev
sprintf("This standard deviation is the expected error in a randomly generated sample median as an
estimate of the true median. It is lower than the previous standard deviation by %f, and is expected
since more sampling results in closer, more accurate, more precise data points", medians_stdev_diff)
######################################################################
```