HW 3

SDS348 Spring 2021

Rose Hedderman

This homework is due on Feb 15, 2021 at 8am. Submit a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (6 pts)

1.1 (1 pt) Assume diastolic blood pressure is normally distributed in a certain healthy population, with a mean of 82 mmHg and a standard deviation of 11 mmHg. What proportion of this population has a diastolic blood pressure less than 60 mmHg (i.e., what is the probability that a person in this population has a diastolic blood pressure less than 60)?

```
pnorm(60, mean = 82, sd = 11)
```

```
## [1] 0.02275013
```

The probability that a person in this population has a diastolic blood pressure less than 60 is 0.02275013.

1.2 (1 pt) What diastolic blood pressure would put an individual from this population at the 98th percentile?

```
qnorm(0.98, 82, 11)
```

```
## [1] 104.5912
```

A diastolic blood pressure of 104.5912 would put an individual from this population at the 98th percentile.

1.3 (1 pt) What is the probability that a random individual from this population will have a diastolic blood pressure higher than 100? (i.e., what proportion of the population has a diastolic BP greater than 100 mmHg?)

```
pnorm(100, mean = 82, sd = 11, lower.tail = FALSE)
```

```
## [1] 0.05088175
```

The probability that a random individual from this population will have a diastolic blood pressure higher than 100 is 0.0588175.

1.4 (1 pt) What proportion of the population has a diastolic blood pressure bewteen 80 and 90?

```
ans <- pnorm(90, mean = 82, sd = 11) - pnorm(80, mean = 82, sd = 11)
ans
```

```
## [1] 0.3386078
```

The proportion of the population has a diastolic blood pressure bewteen 80 and 90 is 0.3386078.

1.5 (2 pts) Assume this distribution of diastolic blood pressure is for a healthy population. If we observe an individual from an unknown population with a diastolic BP of 110 mmHg, what is the probability of observing an individual with a BP this extreme (i.e., in either direction from the mean) if the individual really came from a normal population with a mean of 82 and standard deviation of 11?

```
1 - pnorm(110, mean = 82, sd = 11)
```

```
## [1] 0.005456779
```

The probability of observing an individual with a BP this extreme is 0.005456779.

Question 2: (10 pts)

2.1 (1 pt) Let’s take a sample of of size 10,000 from a normal distribution with a mean of 82 and a standard deviation of 11 representing diastolic blood pressure. What is the population mean and what is the sample mean? *Note: in order to have reproducible results, set a seed with* `set.seed()`*.*

```
# set a seed to have reproducible results
set.seed(348)
x <- rnorm(10000, mean = 82, sd = 11)
mean(x)
```

```
## [1] 81.67394
```

The population mean is 82 mmHg and the sample mean is 81.67394 mmHg.

2.2 (2 pts) Using `sum()` on a logical vector, how many draws are less than 60? Using `mean()` on a logical vector, what proportion of the total draws is that? How far is your answer from `pnorm()` in 1.1 above?

```
sum(x < 60)
```

```
## [1] 281
```

```
pnorm(60, mean(x), 11)
```

```
## [1] 0.02439869
```

```
diff <- abs(pnorm(60, mean(x), 11) - pnorm(60, mean = 82, sd = 11))
diff
```

```
## [1] 0.00164856
```

281 draws are less than 60 which is a 0.02439869 proportion of the total draws. This answer is only 0.00164856 different from the answer from 1.1.

2.3 (1 pt) What proportion of your sample is greater than 110 or less than 54?

```
pnorm(110, mean(x), sd = 11, lower.tail = FALSE) + pnorm(54, mean(x), sd = 11)
```

```
## [1] 0.01094852
```

The proportion of your sample is greater than 110 or less than 54 is 0.01094852.

2.4 (1 pt) Why are your answers close to what you got above? Why are they not exactly the same?

My answers are close to what I got above because the sample is based off of the population mean. The first question is using the population mean and question 2 is using a sample mean which close but not exactly the same.

2.5 (2 pts) Using the code below, take 5 samples of size 10,000 each from a normal distribution with a mean of 82 and a standard deviation of 11. The loop will also calculate the mean for each sample. a) What is the mean of the sample means? How close is this value to the population mean? b) What is the standard deviation of the sample means? How close is this value to the population standard deviation?

```
# Create an empty vector to store the values of the sample means
sample_means <- numeric(0)

# Use a loop to take multiple samples and calculate the sample mean
for (i in 1:5){
  x <- rnorm(10000, 82, 11)
  sample_means <- c(sample_means,mean(x))
}

m_sam <- mean(sample_means)
m_close <- abs(82 - m_sam)
sd_sam <- sd(sample_means)
sd_close <- abs(11 - sd_sam)
m_sam
```

```
## [1] 81.97935
```

```
m_close
```

```
## [1] 0.02064912
```

```
sd_sam
```

```
## [1] 0.1014857
```

```
sd_close
```

```
## [1] 10.89851
```

The mean of the sample means is 82.01294 which is 0.01293928 close to the population mean. The standard deviation of the sample means is 0.1350357 which is 10.86496 close to the population standard deviation.

2.6 (1 pt) Repeat question 2.5 but with 5,000 samples.

```
sample_m <- numeric(0)

# Use a loop to take multiple samples and calculate the sample mean
for (i in 1:5){
  x <- rnorm(5000, 82, 11)
  sample_m <- c(sample_m,mean(x))
}

m_sam <- mean(sample_m)
m_close <- abs(82 - m_sam)
sd_sam <- sd(sample_m)
sd_close <- abs(11 - sd_sam)
m_sam
```

```
## [1] 81.96149
```

```
m_close
```

```
## [1] 0.03850825
```

```
sd_sam
```

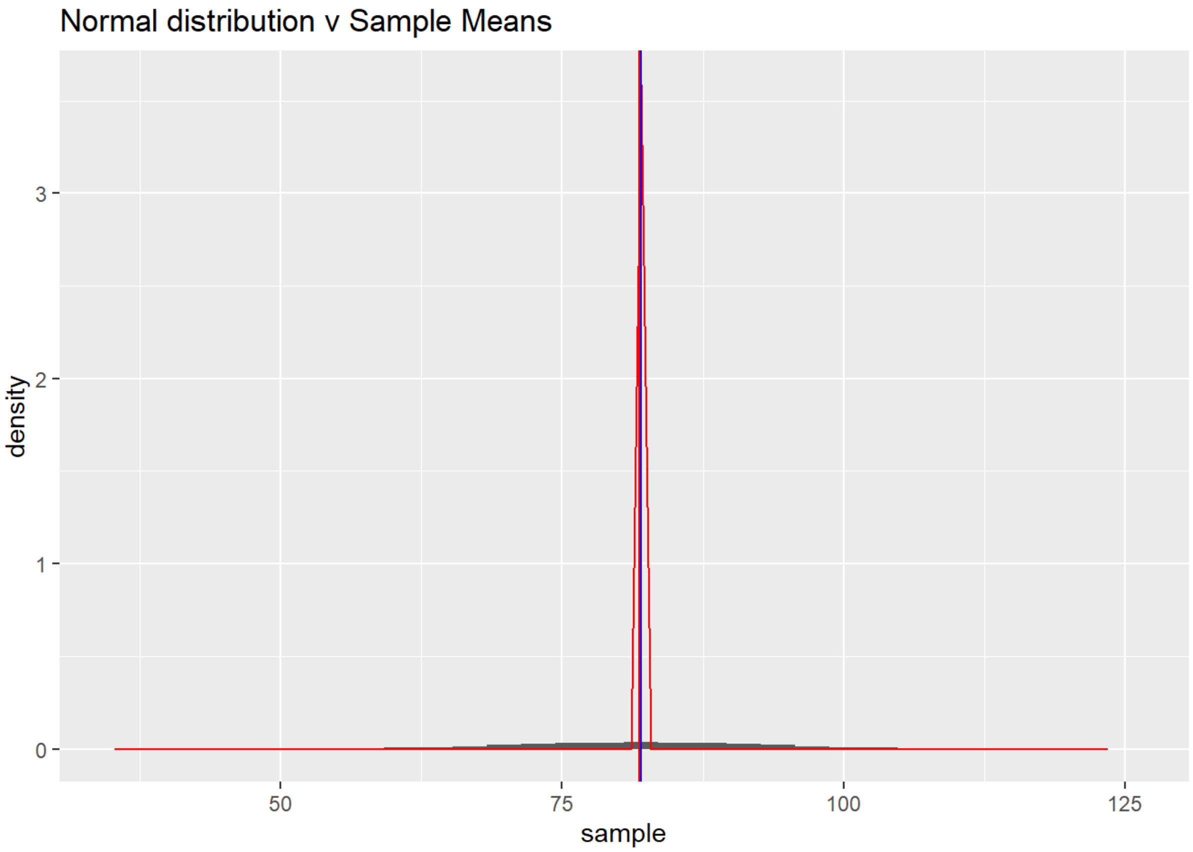
```
## [1] 0.1512267
```

```
sd_close
```

```
## [1] 10.84877
```

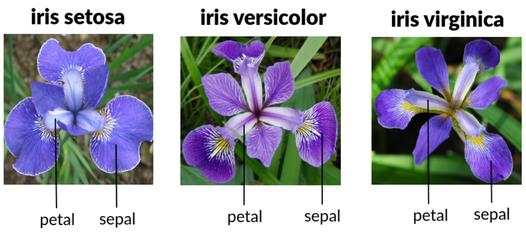
2.7 (2 pts) Using ggplot2, make a histogram of the sample means: set `y=..density..` inside `aes()`. Overlay a normal distribution in red with `stat_function(aes(sample_means), fun=dnorm, args=list(82,11/sqrt(10000)), color=)`. Using `geom_vline(xintercept=, color=)`, add a red line corresponding to the mean of the normal distribution and a blue line corresponding to the mean of the sample means. Label the axes appropriately. How close is the normal distribution to the distribution of the sample means?

```
library(ggplot2)

sample_ok <- numeric(0)

# Use a loop to take multiple samples and calculate the sample mean
for (i in 1:5000){
  so <- rnorm(5000, 82, 11)
  sample_ok <- c(sample_ok,mean(so))
}

df <- as.data.frame(so)
ggplot(df, aes(so)) +
  geom_histogram(aes(y = ..density..)) +
  stat_function(aes(sample_ok), fun=dnorm, args=list(82,11/sqrt(10000)), color= 'red') +
  geom_vline(xintercept= mean(so), color= 'red') +
  geom_vline(xintercept= mean(sample_ok), color= 'blue') +
  labs(x = "sample", y = "density") +
  ggtitle("Normal distribution v Sample Means")
```



The two distributions are very close. The red and blue inercept lines representing sample means have a small difference.

Question 3: (9 pts)



The dataset `iris` contains information about the measurements (in centimeters) of sepals and petals for 50 flowers from 3 different species (see picture above). The first few observations are listed below.

```
library(dplyr)

# Save dataset in environment
myiris <- iris

# Look at the 10 first rows
head(iris,10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```

3.1 (2 pts) Using `dplyr` functions (do not use any `[]` or `$`), show the top 5 flowers that had the highest petal length, sorted from greatest to least top length, with the variables `Species` and `Petal.Length` only. Which species had the top 5 flowers in terms of petal length?

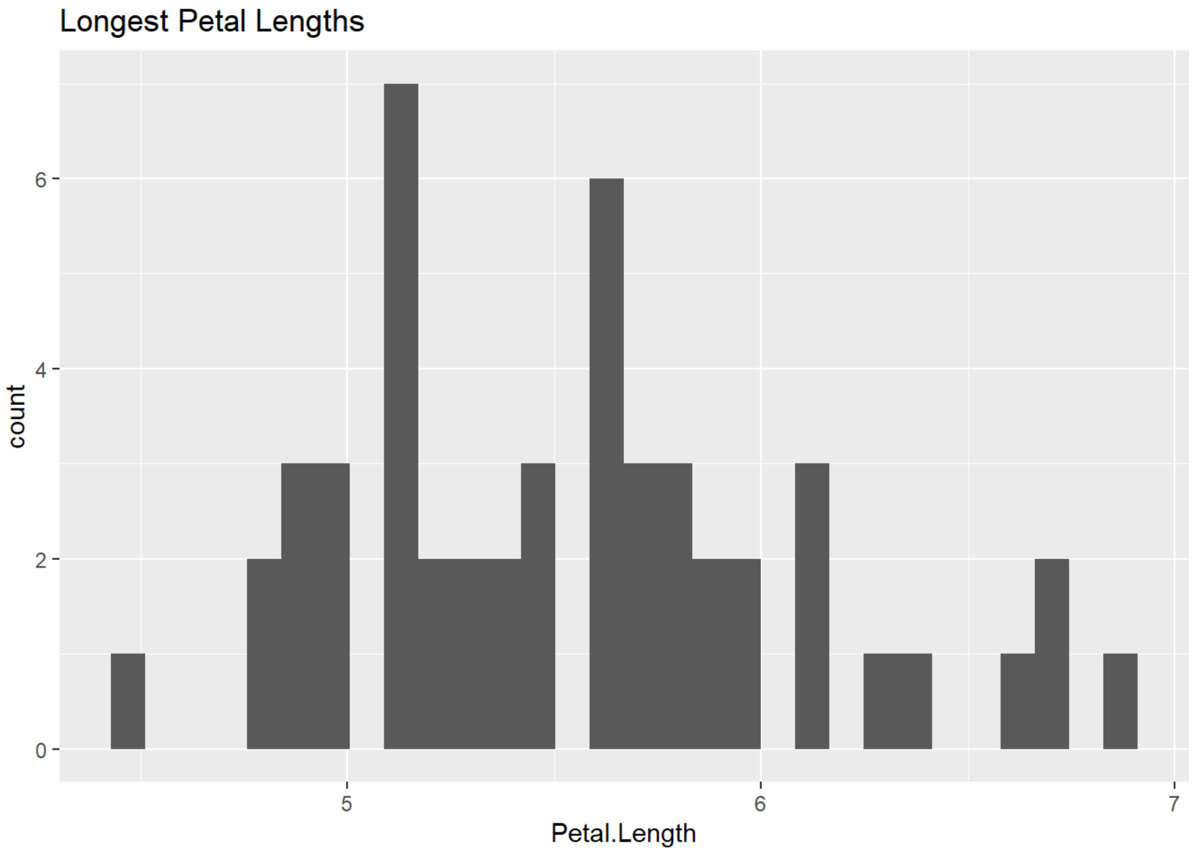
```
one <- myiris %>%
  select(3,5) %>%
  arrange(desc(Petal.Length))
head(one, 5)
```

```
##   Petal.Length   Species
## 1          6.9 virginica
## 2          6.7 virginica
## 3          6.7 virginica
## 4          6.6 virginica
## 5          6.4 virginica
```

The virginica species had the top 5 flowers in terms of petal length.

3.2 (2 pts) Using `dplyr` and `ggplot2` functions (do not use any `[]` or `$`), construct a histogram of `Petal.Length` for the species that had the top 5 flowers with the longest petals (found in 3.1). Describe the distribution (e.g., use the words symmetric, skewed, the center is around _, …).

```
two <- one %>%
  filter(Species == 'virginica')
ggplot(two, aes(Petal.Length)) +
  geom_histogram() +
  ggtitle("Longest Petal Lengths")
```



This distribution is bimodal at a Petal Length around 5.2 and 5.7 cm. This distribution is also skewed right.

3.3 (1 pt) Using `dplyr` functions (do not use any `[]` or `$`), are there any cases in the dataset for which the ratio of sepal length to sepal width exceeds the ratio of petal length to petal width?

```
filter(myiris, Sepal.Length/Sepal.Width > Petal.Length/Petal.Width)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 1          6.9         3.1          5.1         2.3 virginica
```

```
three_sepal <- 6.9/3.1
three_petal <- 5.1/2.3
three_sepal
```

```
## [1] 2.225806
```

```
three_petal
```

```
## [1] 2.217391
```

There is one case in the dataset where the sepal ratio is 2.225806 which exceeds the petal ratio of 2.217391.

3.4 (2 pts) Using `dplyr` functions (do not use any `[]` or `$`), create a new variable `petal_ratio` as the ratio of petal length to petal width then calculate the mean and standard deviation of this ratio for each species. Based on these summary statistics, does `petal_ratio` seem to differ across species? *No need to conduct hypothesis testing but informally state if they seem to differ and if so, how.*

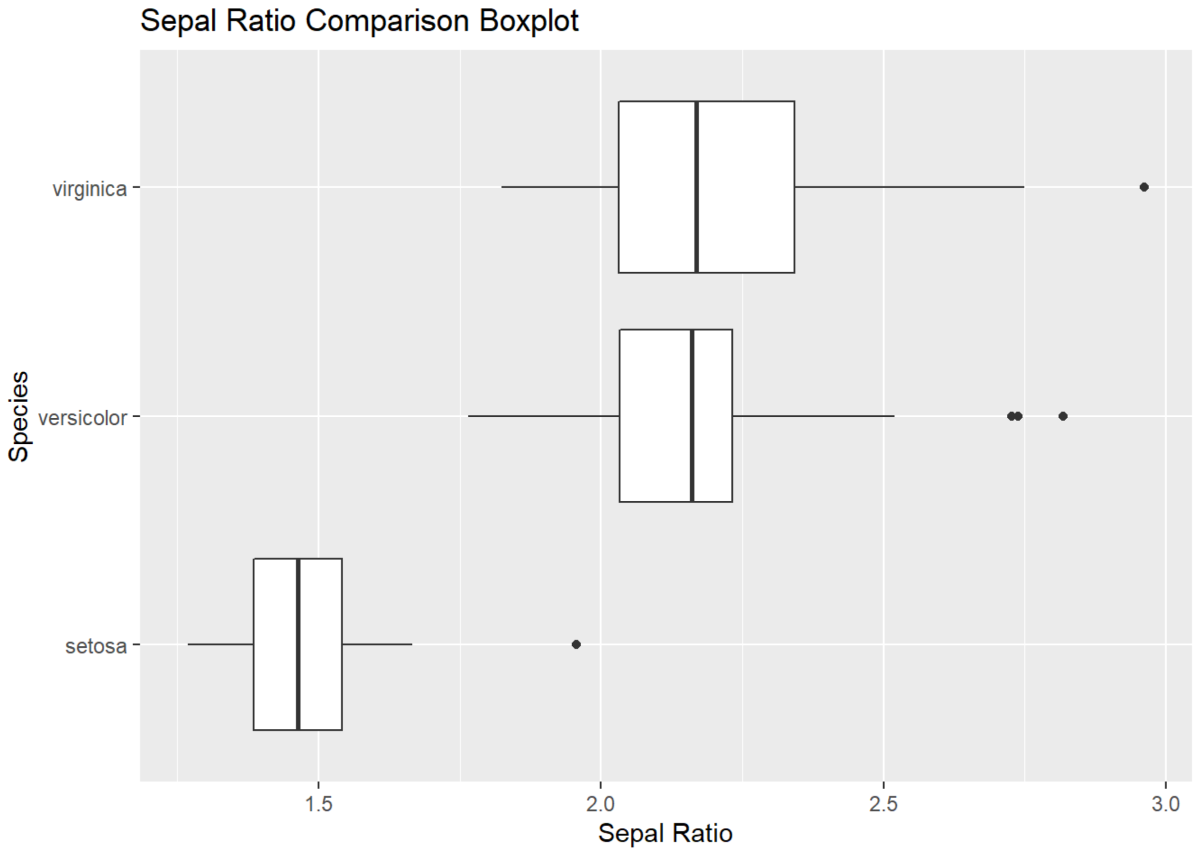
```
four <- mutate(myiris, petal_ratio = Petal.Length/Petal.Width)

four %>%
  group_by(Species) %>%
  summarise(mean_pr=mean(petal_ratio, na.rm=T),
            sd_pr=sd(petal_ratio, na.rm=T) )
```

```
## # A tibble: 3 x 3
##   Species    mean_pr sd_pr
## * <fct>        <dbl> <dbl>
## 1 setosa        6.91 2.85
## 2 versicolor    3.24 0.312
## 3 virginica     2.78 0.407
```

The petal_ratio summary statistics do seem to differ across species.For example, the mean petal ratio for setosa is 6.908 while the mean petal ratio for virginica is 2.780662.

3.5 (2 pts) Using `dplyr` and `ggplot2` functions (do not use any `[]` or `$`), create a new variable `sepal_ratio` as the ratio of sepal length to sepal width then construct boxplots representing `sepal_ratio` for each species. Based on these visualizations, does `sepal_ratio` seem to differ across species? *No need to conduct hypothesis testing but informally state if they seem to differ and if so, how.*

```
five <- mutate(myiris, sepal_ratio = Sepal.Length/Sepal.Width)

ggplot(five, aes(sepal_ratio, Species)) +
  geom_boxplot()+
  labs(x = "Sepal Ratio") +
  ggtitle("Sepal Ratio Comparison Boxplot")
```



According to the boxplot visualizations, sepal_ratio summary statistics do seem to differ across species. The sepal_ratio looks similar for the virginica and versicolor species, but very different for the setosa species.

```
##        sysname        release        version       nodename        machine
##      "Windows"       "10 x64"  "build 18363"     "ROSE-XPS"       "x86-64"
##          login           user effective_user
##        "roseh"        "roseh"        "roseh"
```