Stastistical Thinking Assignment 01

Ruimin Lin

Varsha Ujjinni Vijay Kumar

Siddhant V Tirodkar

Ketan Kabu

Junhao Wang

9/19/2020

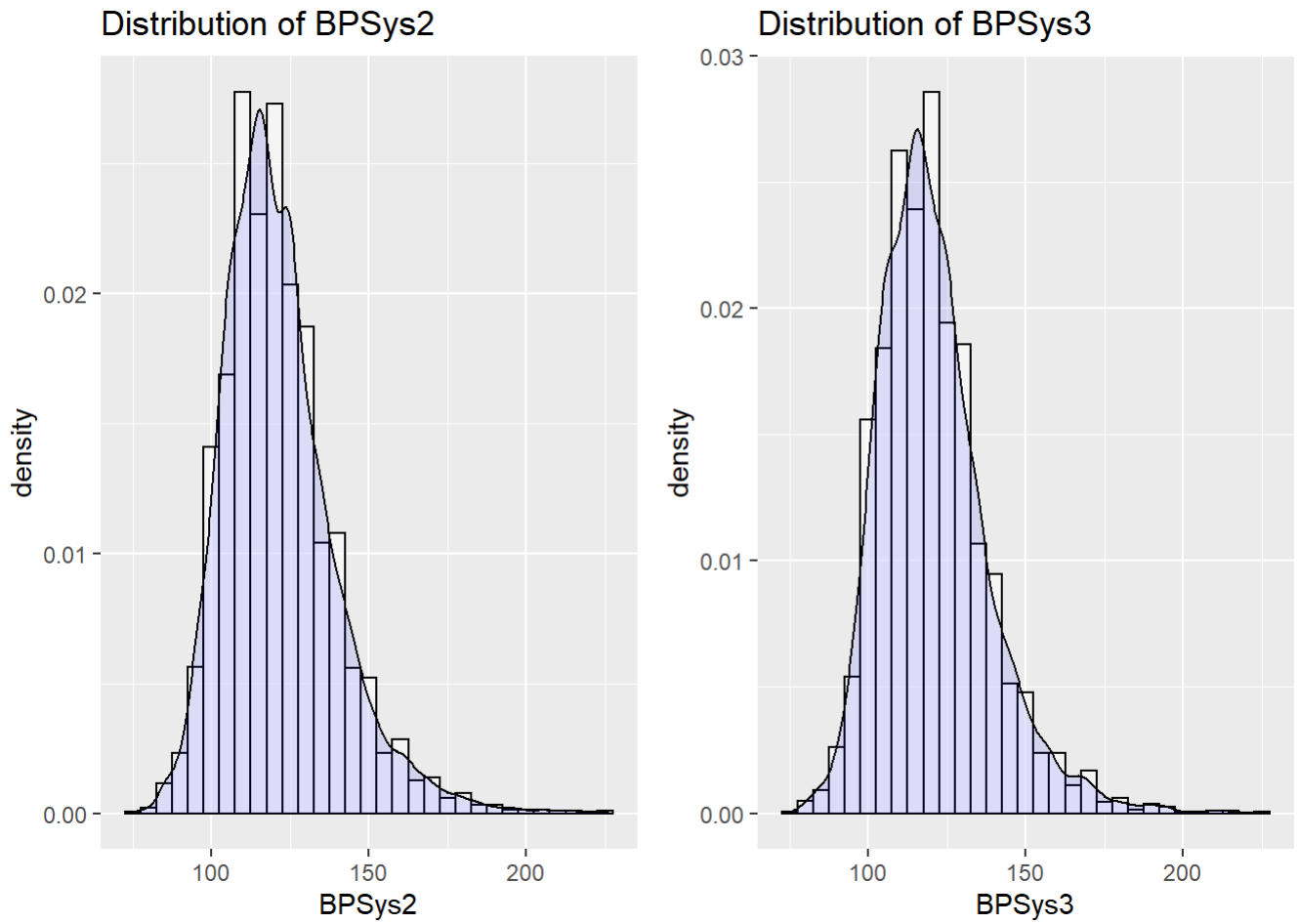
Question 1

Q1.a

Produce a single plot that displays the sample distributions of the two variables, BPSys2 and BPSys3. Explain why the plot is relevant to the comparison of the two variables. Detail the important elements of the plot in your report, and comment on any features apparent from the plot of particular interest.

```
p1_sys2 <- dt1 %>%
  # plotting a graph with the dataset dt1 and x-axis BPSys2
  ggplot(aes(BPSys2)) +
  #adding a histogram to the plot with y-axis as a density plot
  geom_histogram(aes(y = ..density..),
                 alpha = 0.5,
                 binwidth = 5,
                 colour = "black", fill = "white") +
  #adding the density plot to the histogram for the same variable BPSys2 with transparency(alph
a) and color
  geom_density(alpha = 0.1, fill = "blue") + ggtitle("Distribution of BPSys2")

p2_sys3 <- dt1 %>%
  #plotting a graph with the second variable BPSys3 from the dataframe dt1
  ggplot(aes(BPSys3)) +
  #adding a histogram with binwidth of 5 for the variable and an indication for a second geom(de
nsiyt) on the same plot
  geom_histogram(aes(y = ..density..),
                 alpha = 0.5,
                 binwidth = 5,
                 colour = "black", fill = "white") +
  geom_density(alpha = 0.1, fill = "blue") +
  ggtitle("Distribution of BPSys3")
#adding the density plot to the histogram for the same variable BPSys3 with transparency(alpha)
 and color
grid.arrange(p1_sys2, p2_sys3, ncol = 2)
```



Difference in distribution of BPSys2 and BPSys3

The density plot shows the systolic blood pressure readings of individuals at two consecutive time points. In case of sample distribution, histograms and density plots are best to view them. The comparison of two variables of interest is essential before doing any statistical procedure, the visualisation of variable offers an idea of how the variable of interest is distributed at a glance. In this case, we can observe that both sample distribution is positively skewed with unimodal distribution, with the same scale on x-axis (approx. 100 - 200). The peak (number of participants who had same systolic blood pressure reading) in BPSys2 originates just after 100, whereas BPSys3 peak is seen just before 125. Overall, the distributions of both variables of interest(BPSys2 and BPSys3) does not have too much difference.

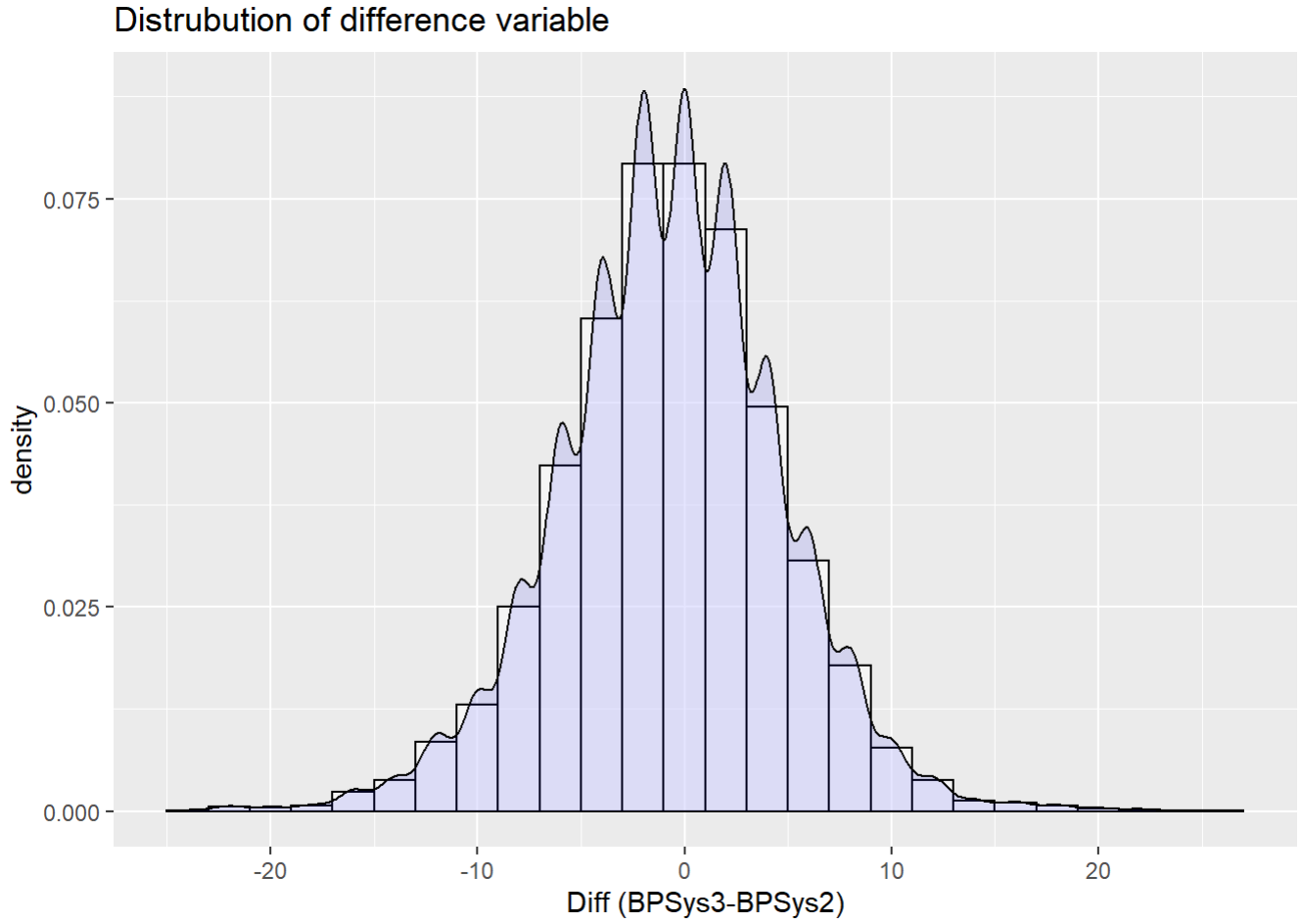
Q1.b

Produce a plot to appropriately display the sample information regarding the distribution of the difference variable Diff = BPSys3 − BPSys2. Detail the important elements of the plot in your report, and comment on any features apparent from the plot of particular interest.

```
dt1diff <- dt1 %>%
  #adding a new column to the dataframe dt1 containing difference between the two variables BPSy
s2 & BPSys3
  mutate(diff = BPSys3 - BPSys2)

p3_diff <- dt1diff %>%
  #plotting a graph with the new added variable diff as a x-axis(independent variable)
  ggplot(aes(diff)) +
  #adding a histogram to the coordinates with binwidth of 2
  geom_histogram(aes(y = ..density..),
                 alpha = 0.5,
                 binwidth = 2,
                 colour = "black", fill = "white") +
  #adding a density plot for the same plot with transparency of 0.1 to view both the geoms
  geom_density(alpha = 0.1, fill = "blue") +
  xlab("Diff (BPSys3-BPSys2)") + ggtitle("Distrubution of difference variable")

#displaying the graph obtained
p3_diff
```



Distrubution of difference variable

> The density plot for the difference of BPSys2 and BPSys3 demonstrates the presence of multi-modal. From the histogram we can see that the mean of difference of BPSys2 and BPSys3 is centred around 0, however, the shape of the density plot is not behaving as a smooth bell-shaped curve.

Q1.c

Produce a selection of suitable summary statistics for each of the relevant variables from parts a. and b. Describe each statistic produced and explain its relevance.

```
summarySys2 <- dt1 %>%
  #creating a summary with mean,median,standard deviation and quartile ranges for the variable B
PSys2 from the dt1 dataframe
  summarise(n = n(),
            mean = mean(BPSys2),
            median = median(BPSys2),
            SD = sd(BPSys2), IQR = IQR(BPSys2))
summarySys2
```

```
## # A tibble: 1 x 5
##       n  mean median    SD   IQR
##   <int> <dbl>  <dbl> <dbl> <dbl>
## 1  4415  121.    118  17.7    20
```

```
summarySys3 <- dt1 %>%
  #creating a summary with mean,median,standard deviation and quartile ranges for the variable B
PSys3 from the dt1 dataframe
  summarise(n = n(),
            mean = mean(BPSys3),
            median = median(BPSys3),
            SD = sd(BPSys3), IQR = IQR(BPSys3))
summarySys3
```

```
## # A tibble: 1 x 5
##       n  mean median    SD   IQR
##   <int> <dbl>  <dbl> <dbl> <dbl>
## 1  4415  121.    118  17.4    22
```

```
summarydiff <- dt1diff %>%
  #creating a summary with mean,median,standard deviation and quartile ranges for the variable d
iff(BPSys3-BPSys2) from the dt1 dataframe
  summarise(n = n(),
            mean = mean(diff),
            median = median(diff),
            SD = sd(diff), IQR = IQR(diff))

#saving the value of n(samples) as a number
summarydiff$n = as.double(summarydiff$n)
summarydiff
```

```
## # A tibble: 1 x 5
##       n   mean median    SD   IQR
##   <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1  4415 -0.728      0  5.37     6
```

The summary statistic created for BPSys2 and BPSys3 have little difference, its mean and SD only have difference of 0.7280 and 0.3972 respectively, with same median. Such difference also leads to the small summary statistics presented in the summary statistic table for (BPSys3 - BPSys2), the interquartile range tend to be lower compared to the interquartile range for BPSys2 and BPSys3. On the other hand, the low standard deviation implies that the data is clustered around the mean compared to the standard deviation of BPSys2 and BPSys3, where the standard deviation of 17.75 and 17.35 indicates the data is more spread out.

Q1.d

Use a Bootstrap-based approach to produce a 95% confidence interval for the average difference in systolic blood pressure of respondents, as measured by BPSys2 and BPSys3. Report your interval and explain how it was obtained. Include a plot of the empirical Bootstrap sample density in your discussion, explain what it represents as well as how the plot relates to the interval produced.

```
#setting a seed value for computation of bootstrap
set.seed(140920)

#setting the n as the number of values in the sample data dt1diff calculated from previous secti
on
n.q1 <- summarydiff$n

#giving a value for randomly sampling/selecting values from the dataframe here 10,000
B <- 10000
#replicating the values 10,000 times
xbar_boot <- rep(NA,B)
#setting a for loop for the selection process of various values from the dataframe dt1diff with
 the variable diff
for(i in 1:B){
  #this process of for loop for selection of values is with replacement ie. values can be repeat
ed
  dtdiff <- sample(dt1diff$diff, size=n.q1, replace=TRUE)
  #creating an index to store the mean of randomly selected values with replacement
  xbar_boot[i] <- mean(dtdiff)
}

# finding the confidence intervals for the randomly selected values from dataframe xbar_boot
boot.CIq1 <- quantile(xbar_boot, c(0.025, 0.975))
#displaying the calculated lower and upper end points for the graph
boot.CIq1
```

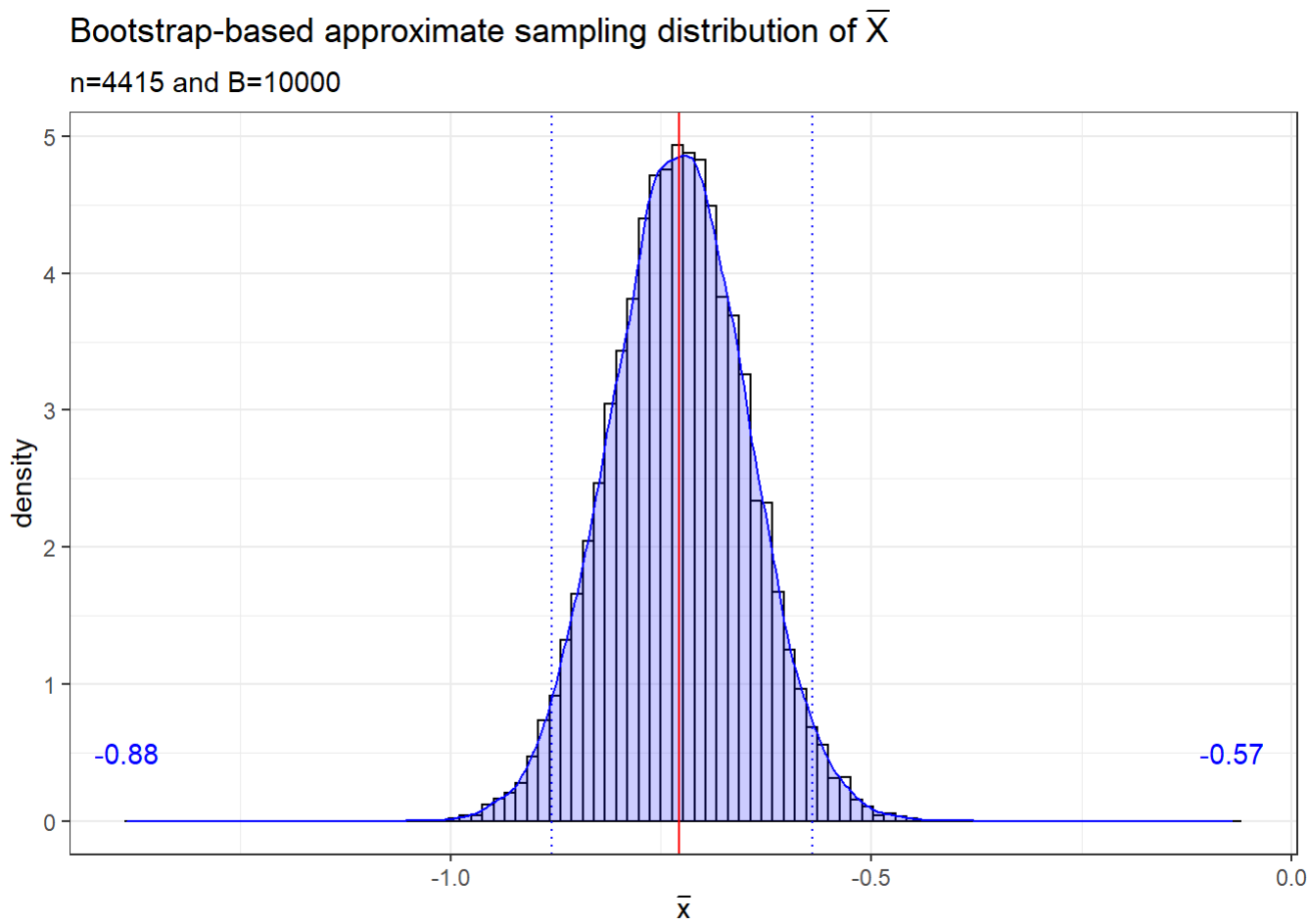
```
##      2.5%     97.5%
## -0.884723 -0.570781
```

```
set.seed(140920)
#creating a bootstrap function for plot with arguments as the variable and binwidth
bootplot.f<- function(stat.boot, bins=50){
  #storing the values as a tibble df
  df <- tibble(stat = stat.boot)
  #finding the confidence interval and rounding the values to 2 decimal points
  CI <- round(quantile(stat.boot, c(0.025, 0.975)),2)
  #generating the plot with x-axis as the values and including a density geom into the same plot
  p <- df %>% ggplot(aes(x=stat, y=..density..)) +
    #plottinga histogram plot with the binwidth same as the arguments
    geom_histogram(bins=bins, colour="black", fill="white", alpha=0.2) +
    geom_density(fill="blue", colour="blue", alpha=0.2) +
    #adding a verticle line to indicate the upper and lower end confidence interval points on th
e same map
    geom_vline(xintercept = CI, colour = "blue", linetype=3) +
    theme_bw()
  #displaying the plot
  p
}

#setting the mean value for the plot from the summary drawn previously
mu.true <- summarydiff$mean
#setting the standard deviation value for the plot from the summary drawn previously
sig.true <- summarydiff$SD
# producing a random normal distribution with mean and SD from previously set values
x <- rnorm(n=n.q1, mean=mu.true, sd=sig.true)
#using the bootstrap fucntion to plot the graph with arguments stat = xbar_boot and binwidth 100
p_xbarboot <- bootplot.f(xbar_boot, bins=100)
#calculating the mid value between the either end points of the distribution
#adding layers to set up a range for viewing the population density
len <- (max(xbar_boot) - min(xbar_boot))/3
xxmax <- (max(xbar_boot) + sqrt(n.q1)*len)
xxmin <- (min(xbar_boot) - sqrt(n.q1)*len)

p_xbarboot <- p_xbarboot +
  #adding the values of either end points ie confidence intervals
  annotate("text",label=round(boot.CIq1[1],2), x=(boot.CIq1[1]-0.5),
           y=0.5,colour="blue") +
  annotate("text",label=round(boot.CIq1[2],2), x=(boot.CIq1[2]+0.5),
           y=0.5,colour="blue") +
  #drawing the verticle line for mean value in red
  geom_vline(xintercept=mu.true, colour="red")
# adding a x title
p4boot <- p_xbarboot + xlab(expression(bar(x)))
#adding a title to the plot
p4boot <- p4boot + ggtitle(expression(paste("Bootstrap-based approximate sampling distribution o
f ", bar(X))),"n=4415 and B=10000")

p4boot
```



Bootstrap-based approximate sampling distribution of X bar

```
#displaying the plot
```

Eplain what it represents as well as how the plot relates to the interval produced.

We created a loop that generates 10,000 bootstrap samples from difference data we computed using BPSys3-BPSys2, and hence we compute the 95% bootstrap confidence interval from the sample.The 95% bootstrap confidence interval is -0.8847 and -0.5708,indicating that we are 95% confident that the difference between third reading and second reading of systolic blood pressure is between -0.8847 and -0.5708 mm Hg.

The empirical Bootstrap sample density plot represents the result of 10,000 replicated Bootstrap samples by re-sampling observed values with replacement. The bootstrap sample eliminates the multimodal issue occurred previously, showing a smoother density curve. The plot also demonstrated the 95% bootstrap sample interval by a pair of blue dotted lines with corresponding annotation. The red vertical line indicates $\bar{x}$ of original sample, and blue vertical line indicates the bootstraped sampled $\bar{x}$.

Q1.e

Use a CLT-based approach to produce a 95% confidence interval for the average difference in systolic blood pressure measurements, corresponding to part d. Report this alternative interval, and compare it to the one obtained using the Bootstrap method in part d. Explain the relative benefits of each approach used to produce the competing confidence intervals.

```
#testing the same variable diff from datfarame dt1diff with CLT based testing using the t.test()
ttest.out <- t.test(x = dt1diff$diff) %>%
  #making the values into a tidy format
  tidy()
#finding the confidence intervals
CLT.CIq1 <- c(ttest.out$conf.low, ttest.out$conf.high)
CLT.CIq1
```

```
## [1] -0.886483 -0.569463
```

The CLT-based 95% confidence interval produced the result that we are 95% confident the difference between third and second reading of systolic blood pressure is between -0.8865 and -0.5695 mm Hg. Comparing the result with bootstrap sample confidence interval we find out that there's little difference between 2 approach (difference of 0.0018 and 0.0013 respectively). CLT states that the sampling distribution will be approximately normally distributed if the sample size is large enough (n > 30), in this case our sample size is 4415 which fulfills the requirement. CLT is a theoretical approach, where, if we can determine that the sample size is >30, we will assume the sample distribution is normal. On the other hand, bootstrap sampling is the process of replicate "hypothetical" data sets by re-sampling observed values with replacement provides, that the bootstrap approach takes the repeated sample into practice and provides the sample distribution for repeated samples, and therefore, confidence interval can be calculated even with unknown sampling distributions.

Q1.f

Explain why the measures BPSys2 and BPSys3 are not independent and why it is important to take the dependence in to account. What would be the result for each of parts d. and e. if it were to be assumed that the two populations (for the two measurements) were independent?

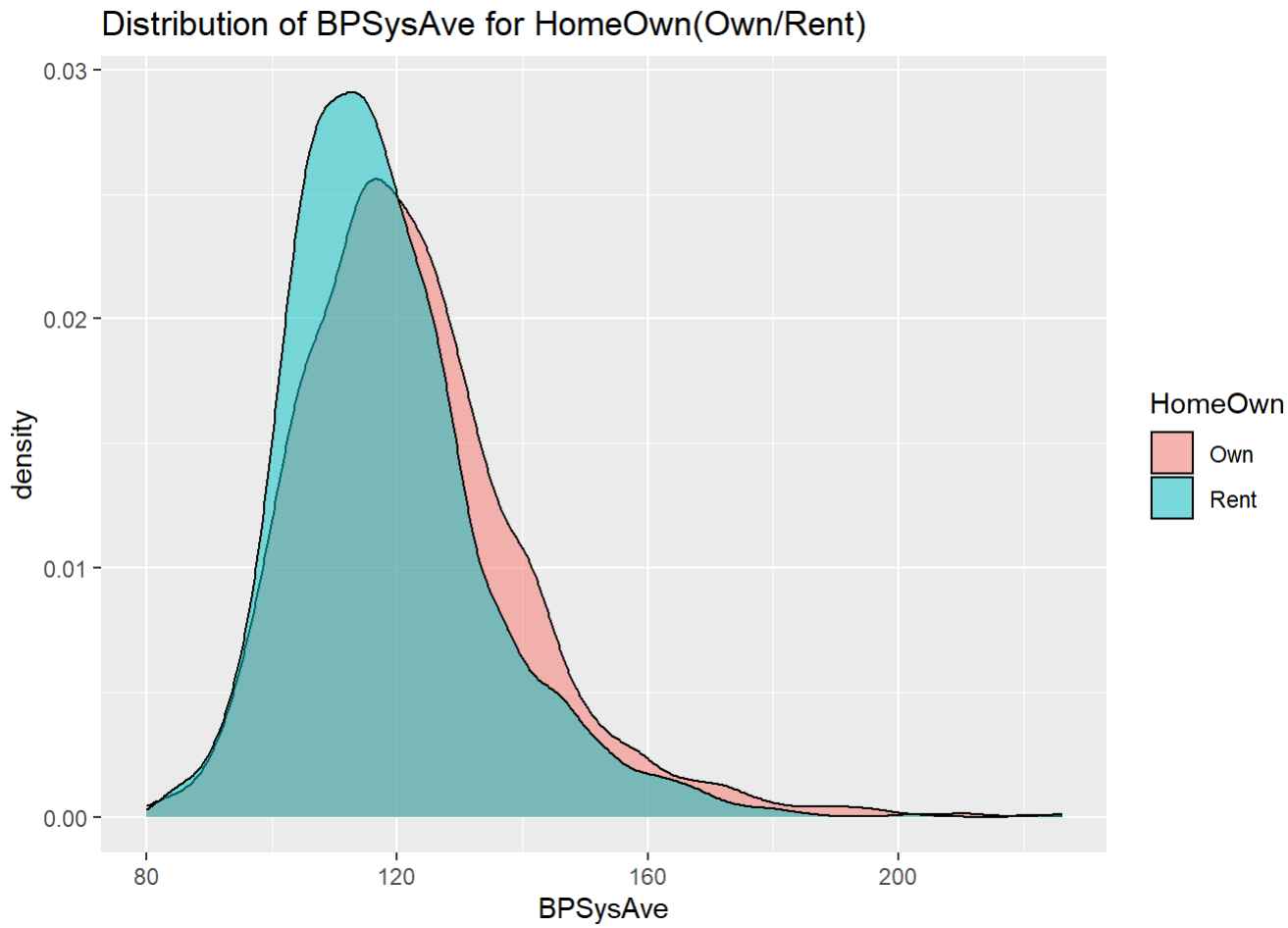
For question 1, we are looking at the difference between BPSys3 and BPSys2, the third and second reading of systolic blood pressure, that compares the different variables from same data set. In other words, the readings of systolic blood pressure is on a same individual that takes the examination, therefore the second reading of systolic blood pressure (BPSys2) has a natural correspondence with the third reading of systolic blood pressure (BPSys3). And therefore, measures of BPSys2 and BPSys3 are considered dependent, and paired. It is important to take into account the dependence as we are looking for the difference of two readings on each individuals. The confidence interval for two approach may result in larger difference if we assumed that the two populations were independent, which means that the measure of BPSys2 is not correspond to the measure of BPSys3 (e.g. Records the measure of Mark's second reading and Emily's third reading), and the result will be biased and not reliable.

Question 2

Q2.a

Produce a single plot that displays the sample distribution of the BPSysAve variable, for each of the two HomeOwn groups of interest, i.e. the "Owners" and the "Renters". Detail the important elements of the plot in your report, and comment if anything seems of particular interest. Explain the steps undertaken to produce the plot.

```
p4dist <- dt1 %>%
  #filtering out the other value from the variable HomeOwn in the dataframe dt1
  dplyr::filter(HomeOwn != "Other") %>%
  #plotting a graph to depict the distribution of Blood pressure amongst the renters and owners
  ggplot(aes(BPSysAve, fill = HomeOwn)) +
  #adding geom_density layer to the plot coordinates
  geom_density(alpha = 0.5) + ggtitle("Distribution of BPSysAve for HomeOwn(Own/Rent)")
#displaying the plot
p4dist
```



Distribution of BPSysAve for HomeOwn(Own/Rent)

Detail the important elements of the plot in your report, and comment if anything seems of particular interest. Explain the steps undertaken to produce the plot.

From the density plot we observed that the density for the renters are higher than the density of owners. Both sample distributions are positively skewed, and the rent sample distribution is skewer than the owned. The interest of the plot is to examine the difference of combined systolic blood pressure reading based on the housing status of each participant.

The variable HomeOwn, consists of 3 groups, Own, Rent and Other, since we are only interested in owners and renters, the `other` group should be eliminated first using `dplyr::filter`. Therefore, the sample distribution of BPSysAve is plotted and differentiated using `fill = HomeOwn`, and the distribution of BPSysAve for two HomeOwn groups is produced.

Q2.b

Produce a selection of summary statistics for BPSysAve, for each of the two HomeOwn groups of interest. Comment on anything interesting you find in these summaries.

```
summaryown <- dt1 %>%
  #filtering the dataframe dt1 for only the value 'Own' under the variable HomeOwn
  dplyr::filter(HomeOwn == "Own") %>%
  #using the summarise() from dplyr package to find the summary statistics for the variable filt
ered above
  summarise(nO = n(),
            #finding the mean,median,standard deviation and inter-quartile range
            meanO = mean(BPSysAve),
            medianO = median(BPSysAve),
            SDO = sd(BPSysAve), IQRO = IQR(BPSysAve))

summaryown %>%
  #using the kable() to produce a good table for the summary stats
  kable() %>%
  kable_styling()
```

nO	meanO	medianO	SDO	IQRO
2815	122.353	120	17.861	22

```
summaryrent <- dt1 %>%
  #filtering the dataframe dt1 for only the value 'Rent' under the variable HomeOwn
  dplyr::filter(HomeOwn == "Rent") %>%
   #using the summarise() from dplyr package to find the summary statistics for the variable fil
tered above
  summarise(nR = n(),
             #finding the mean,median,standard deviation and inter-quartile range
            meanR = mean(BPSysAve),
            medianR = median(BPSysAve),
            SDR = sd(BPSysAve), IQRR = IQR(BPSysAve))

summaryrent %>%
   #using the kable() from to produce a good table for the summary stats
  kable() %>%
  kable_styling()
```

nR	meanR	medianR	SDR	IQRR
1488	118.355	116	15.9348	19

The summary statistic produced indicates that the sample size of Owner are far more than the sample size of Renters, but on the density plot produced previously, Renter has higher density than Owners. Another interest statistic to notify is that owners tends to have higher mean/median systolic blood pressure than renters.

Q2.c

Estimate the average difference in BPSysAve for “Owners” relative to “Renters”, and using a CLT-based approach, report a 95% confidence interval for this difference and report on the corresponding outcome of the formal two-sided hypothesis test. In your report, detail the form of the null and alternative hypotheses, and explain how you reached the conclusion of the test.

```
dtOwn <- dt1 %>%
  #filtering the variable HomeOwn for only Own
  filter(HomeOwn == "Own") %>%
  #just retrieving the value of BPSysAve for Owners
  pull(BPSysAve)
#storing the value retrieved as a number(integer/double)
dtOwn <- as.numeric(dtOwn)

dtRent <- dt1 %>%
   #filtering the variable HomeOwn for only Rent
  filter(HomeOwn == "Rent") %>%
  #just retrieving the value of BPSysAve for Renters
  pull(BPSysAve)
#storing the value retrieved as a number(integer/double)
dtRent <- as.numeric(dtRent)

#calculating the CLT based testing for the two different types of values under the HomeOwn varia
ble
ttestq2 <- t.test(x = dtOwn, y = dtRent) %>%
  #storing the values in a tidy format
  tidy()
#displaying the table

ttestq2
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic  p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl>
## 1     4.00      122.      118.      7.50 7.91e-14     3340.     2.95      5.04
## # ... with 2 more variables: method <chr>, alternative <chr>
```

We are 95% confident that the difference of averaged systolic blood pressure for Owners and Renters are between 2.9538 and 5.0434 mm Hg.

H0: There is no difference in average systolic blood pressure reading for Owners and Renters. $\mu diff = 0$.

H1: There is a difference in average systolic blood pressure reading for Owners and Renters. $\mu diff \neq 0$.

> Rejection rule: Reject H0 if p-value < 0.05.

> Interpretation: We reject the null hupothesis(H0) since the p-value of 0.00000000000007911 is far less than 0.05. That is, we have found convincing evidence that the averaged systolic blood pressure for Owners and Renters are different.

> Alternatively, we can also interpret as we found convincing evidence that Owners have higher averaged systolic blood pressure than Renters.

Q2.d

Consider the R code provided in the code chunk below (and available in the R script file named A1.R, noted in the Introduction). Explain what the code does, and how it is used to test the relevant hypotheses detailed in part c.

```
# Code chunk for Q2 part d.

#dt2 <- dt %>%
#  filter(HomeOwn!="Other")
#n # student to add
#R # student to add

# student to add
# student to add

#Rdt2 <- dt2

#for (r in 1:R){
#  Rdt2 <- Rdt2 %>%
#    mutate(BPSysAve=sample(dt2$BPSysAve, n, replace=FALSE))
#  Rdt2S <- Rdt2 %>%
#    group_by(HomeOwn) %>%
#    summarise(mean=mean(BPSysAve))
#  RDiff[r] <- Rdt2S %>%
#    summarise(Diff=mean[1]-mean[2])
#}

# student to add additional lines
```

The code first filter the data to exclude observations that belongs to group “Other” in the variable HomeOwn, then state the number of observations and number of random sampling that will be processed. The for each r-th element, it creates a loop of random sampling for BPSysAve with no replacement, and then it is summarised by group HomeOwn, resulted in mean of BPSysAve for Owners and Renters. The last step is calculating the difference in mean of Owners and Renters from the sampled data.

Since the difference between mean of BPSysAve for Owners and Renters is calculated, the confidence interval and hypothesis testing can be constructed.

Q2.e

Supplement the code provided and produce a two-sided test based on the sampling distribution described in part d. Discuss the rationale for the test, carefully explain how to determine the outcome of the test, and report the corresponding “strength of evidence” that results.

```
#filtering the dataframe for all values other than other in variable HomeOwn
dt2 <- dt %>% filter(HomeOwn!="Other")
#setting the seed for reproducible random sample computation
set.seed(140920)

#setting the sample value to the number of rows in the dataframe dt2f
n <- nrow(dt2)
#giving a value for randomly sampling/selecting values from the dataframe here 5000
R <- 5000
#replicating the values 5000 times
RDiff <- rep(NA,R)

Rdt2 <- dt2
#creating a for loop for calculating the random sampled mean values
for (r in 1:R){
  Rdt2 <- Rdt2 %>%
    #finding the random sample for the BPSysAve, for every row in n without replacement ie. valu
es cannot be repeated
    mutate(BPSysAve=sample(dt2$BPSysAve, n, replace=FALSE))
  Rdt2S <- Rdt2 %>%
    #grouping the HomeOwn values and finding the mean of the variable BPSysAve for each value
    group_by(HomeOwn) %>%
    summarise(mean=mean(BPSysAve))
  RDiff[r] <- Rdt2S %>%
    #finding the difference in means obtained for each set and stored as a array
    summarise(Diff=mean[1]-mean[2])
}
#setting the datatype for the Diff2f to numbers(integer/double)
RDiff = as.numeric(RDiff)
#finding the 95% confidence intervals for the dataframe Diff2f
R.CIq2 <- quantile(RDiff, c(0.025, 0.975))
R.CIq2
```

```
##     2.5%    97.5%
## -1.10402  1.07295
```

```
#calculating the CLT based testing for estimates and p-value
robs <- Rdt2S %>%
  summarise(Diff = mean[1] - mean[2])

pval2e <- sum(RDiff >= robs)/R
pval2e
```

```
## [1] 2e-04
```

H0: There is no difference between the averaged systolic blood pressure for owners and renters. p^O - p^W = 0

H1: There is difference between the averaged systolic blood pressure for owners and renters, with owners have higher averaged systolic blood pressure. p^O - p^W > 0

Decision rule: Reject H0 if p-value < alpha, otherwise, do not reject H0.

The process of the random sampling loop is to compute simulated hypothetical samples, this randomization is performed to simulate what would have happened if the reading of averaged systolic blood pressure had been independent of participants' ownership of house (Owner or Renter).

To find the p-value, we implement the formula $\tilde{p} - value$ = $\frac{number\ of\ xobs^{[r]} - xobs}{R}$, number of xobs[r] in this case will be the final `RDiff` result we obtained, which is the random sampled difference, on the other hand, xobs is the sample difference from the original data. Hence, we arrive that a result where p-value is 0.0002 which is less than the alpha (0.05) and hence we arrive at a result that we reject the null hypothesis and there's difference in the averaged systolic blood pressure for renters and owners in the participants, and owner tend to have higher averaged systolic blood pressure than renters.

Q2.f

If you restrict the cases to only include men aged between 35 and 44 (inclusive) the strength of evidence changes. Applying the same method as in part e., show the test results, report the "strength of evidence" for the conclusion, and discuss any apparent reasons for the difference in the results compared to part e.

```
dt2f <- dt %>%
  #filtering the dataframe dt for males between 34-45 years of age who are renters/owners
  filter(Gender == "male",
         HomeOwn!="Other",
         Age > 34 & Age < 45)
#setting the seed for reproducible random sample computation
set.seed(140920)
#setting the sample value to the number of rows in the dataframe dt2f
n2f <- nrow(dt2f)
#giving a value for randomly sampling/selecting values from the dataframe here 5000
R2f <- 5000
#replicating the values 5000 times
Diff2f <- rep(NA,R2f)

Rdt2f <- dt2f
#creating a for loop for calculating the random sampled mean values
for (r in 1:R2f){
  Rdt2f <- Rdt2f %>%
    #finding the random sample for the BPSysAve, for every row in n without replacement ie. valu
es cannot be repeated
    mutate(BPSysAve=sample(dt2f$BPSysAve, n2f, replace=FALSE))
  Rdt2Sf <- Rdt2f %>%
    #griuping the HomeOwn values and finding the mean of the variable BPSysAve for each value
    group_by(HomeOwn) %>%
    summarise(mean=mean(BPSysAve),.groups = 'drop')
  Diff2f[r] <- Rdt2Sf %>%
    #finding the difference in means obtained for each set and stored as a array
    summarise(Diff=mean[1]-mean[2])
}
#setting the datatype for the Diff2f to numbers(integer/double)
Diff2f = as.numeric(Diff2f)
#finding the 95% confidence intervals for the dataframe Diff2f
R.CI2f <- quantile(Diff2f, c(0.025, 0.975))
R.CI2f
```

```
##     2.5%    97.5%
## -2.74132  2.75772
```

```
#calculating the CLT based testing for estimates and p-value
robs2f <- Rdt2Sf %>%
  summarise(Diff = mean[1] - mean[2])

pval2f <- sum(Diff2f >= robs2f)/R2f
pval2f
```

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

When we filter the sample into male participant aged between 35 and 44 (inclusive), we discovered that the confidence interval is widened compared to the previous result. That we are 95% confident the difference of averaged systolic blood pressure between two groups (Owner and Renter) for male participants aged between 35 and 44 is -2.7413 and 2.7577 mm Hg. On the other hand, the p-value is decreased from 0.0002 to 0. Providing a stronger evidence to reject the null hypothesis, that there is difference on the averaged systolic blood pressure in the renters and owners for male participants aged between 35 and 44. - The reason of the difference in the result may be that the sample size has decreased about a considerable amount due to the restriction.

References

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra (https://CRAN.R-project.org/package=kableExtra)

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom (https://CRAN.R-project.org/package=broom)

Randall Pruim (2015). NHANES: Data from the US National Health and Nutrition Examination Study. R package version 2.1.0. https://CRAN.R-project.org/package=NHANES (https://CRAN.R-project.org/package=NHANES)

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra (https://CRAN.R-project.org/package=gridExtra)