PG06-A1

Lachlan Moody ID27809951

Abhishek Sinha ID31322743

Sen Wang ID30382106

Yiwen Zhang ID31203019

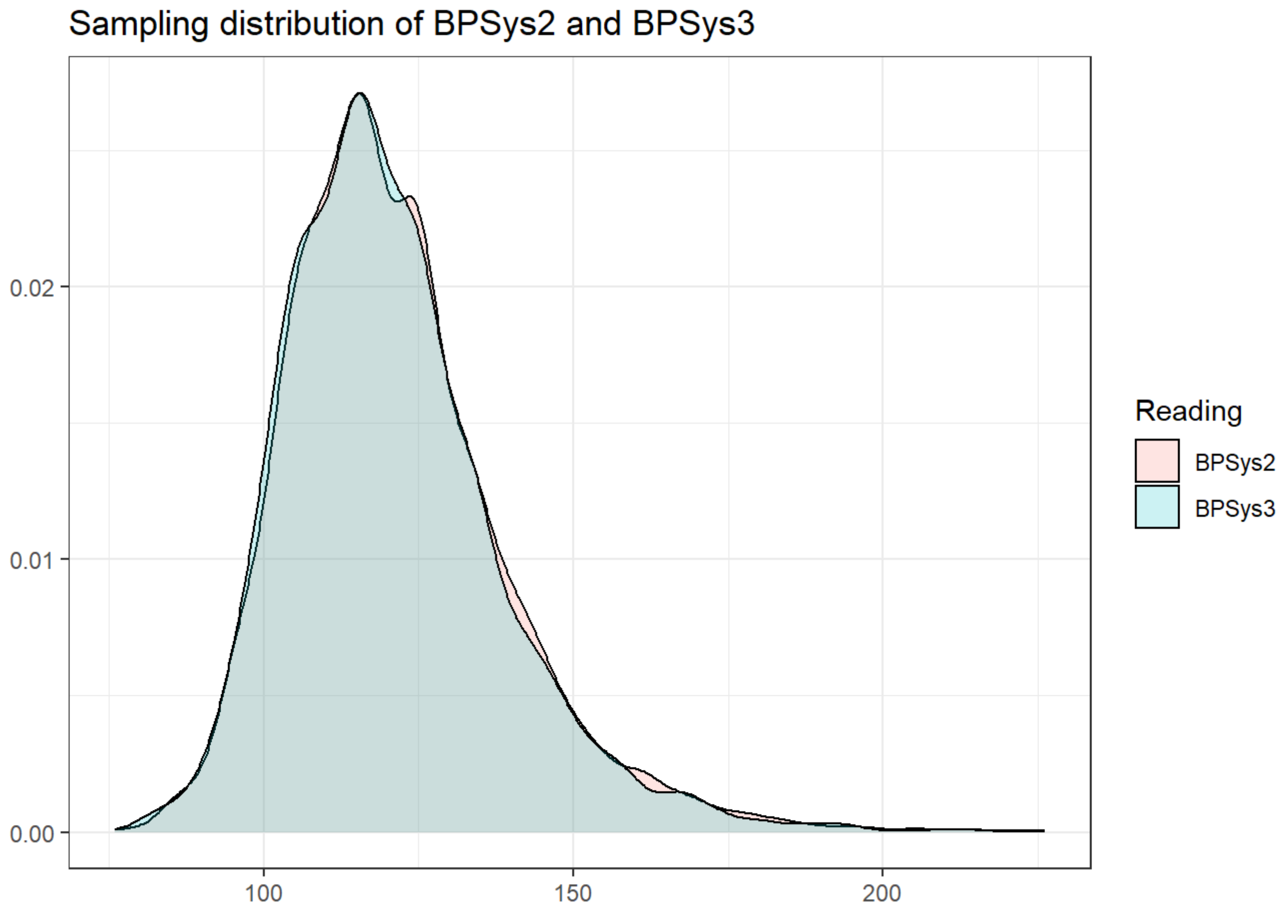
11/09/2020

```
options(digits = 6)

dt <- NHANES %>%
  distinct(ID, .keep_all = TRUE)

dt <- dt %>%
  filter(Age >= 18) %>%
  dplyr::select(Gender,
              Age,
              HomeOwn,
              BPSysAve,
              BPSys2,
              BPSys3)

dt <- dt %>%
  drop_na()

dt1 <- dt %>%
  mutate(Age = as.numeric(Age),
         BPSysAve = as.numeric(BPSysAve),
         BPSys2 = as.numeric(BPSys2),
         BPSys3 = as.numeric(BPSys3))
```

Question 1

Q1.a

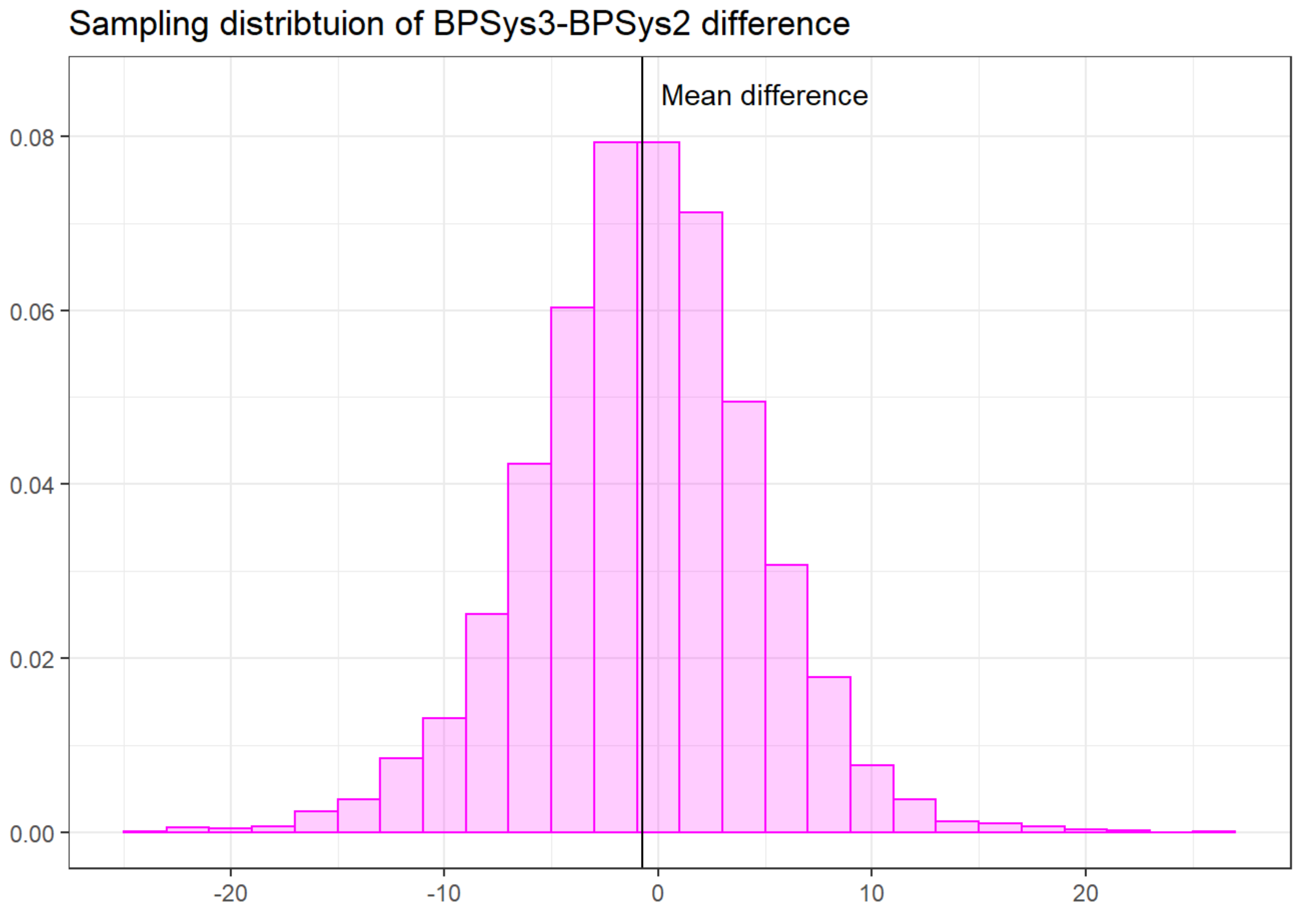
```
dt1 %>%
  pivot_longer(cols = BPSys2:BPSys3, names_to = "Reading") %>%
  ggplot(aes(x = value, y = ..density.., fill = Reading)) +
  geom_density(alpha = 0.2)  +
  theme_bw() +
  labs(x = "", y = "") +
  ggtitle("Sampling distribution of BPSys2 and BPSys3")
```



A kernal density plot was produced to compare the two variables BPSys2 and BPSys 3 with the reading on the x-axis and the relative density on the y-axis. The two x-variables refer to consecutive readings of systolic blood pressure measured in mm Hg. Each reading was also coloured to compare the two. This graphic was chosen as it produces smoother distributions than a histogram, while still capturing the shape of the data. This allows for the peak, or various peaks, to be easily observed. Interestingly, it appears the shape the of BPSys2 and BPSys3 density curves are nearly identical with a peak at around 115 and a smaller second peak at 125. Both plots also have a slight right skew. This suggests that distribution of each variable is very similar within the data set.

Q1.b

```
dt1_diff <- dt1 %>%
  mutate(Diff = BPSys3 - BPSys2)

dt1_diff %>%
  ggplot(aes(x = Diff, y= ..density..)) +
  geom_histogram(colour = "magenta", fill = "magenta", alpha = .2, binwidth = 2) +
  geom_vline(xintercept = mean(dt1_diff$Diff), colour = "black") +
  annotate("text", x = 5, y = .085, label = "Mean difference", colour = "black") +
  theme_bw() +
  labs(x = "", y = "") +
  ggtitle("Sampling distribtuion of BPSys3-BPSys2 difference")
```


Sampling distribtuion of BPSys3-BPSys2 difference

The difference between these two readings was calculated as BPSys3-BPSys2. These differences were then plotted as a histogram with a kernal density plot overlaid on top. This is shown in the graphic above with the difference in mm Hg on the x-axis and the density on the y axis. Additionally, a vertical line of the mean difference was included. From this, it can be observed that the mean difference is very close to 0, only slightly below it, supporting that the two distributions are very similar. Furthermore, the data appears to be normally distributed around the mean, indicating the spread of data is very similar as well.

Q1.c

```
summary <- dt1_diff %>%
  pivot_longer(cols = BPSys2:Diff, names_to = "Reading") %>%
  group_by(Reading) %>%
  summarise(Obs = n(),
            Mean = mean(value),
            SD = sd(value),
            Min = min(value),
            Q1 = quantile(value, 0.25),
            Median = median(value),
            Q3 = quantile(value, .75),
            Max = max(value))

summary %>%
  kable(format.args = list(digits = 2, big.mark = ","), caption = "Summary of BPSys2 and BPSy
s3 readings") %>%
  kable_styling(bootstrap_options = c("bordered", "striped"))
```

Summary of BPSys2 and BPSys3 readings

Reading	Obs	Mean	SD	Min	Q1	Median	Q3	Max

Reading	Obs	Mean	SD	Min	Q1	Median	Q3	Max
BPSys2	4,415	121.32	17.7	76	110	118	130	226
BPSys3	4,415	120.60	17.4	76	108	118	130	226
Diff	4,415	-0.73	5.4	-24	-4	0	2	26

Summary statistics for the three variables discussed above, BPSys2, BPSys3 and Diff are included in the table. Firstly, obs refers to the number of observations for each variable which appears to be the same for both BPSys2 and BPSys3 which indicates that every observation has its pair as these were consecutive readings. The mean for the two variables is only separated by 0.73 mmHg in favour of BPSys2 indicating on average that this reading was slightly higher. Next, examining the standard deviation (SD), it can be observed that again the two variables had a very similar spread around the mean, differing by only 0.3. Additionally, the standard deviation of all measured differences was within 5.4 mm Hg. Finally, examining the 5 number summary, 4 out of the 5 observations were identical for the 2 readings, all except for the first quartile (Q1), which was only out by a value of 2 mm Hg. Furthermore, the median difference was in fact 0. This further supports the similarity in the two readings observed earlier.

Q1.d

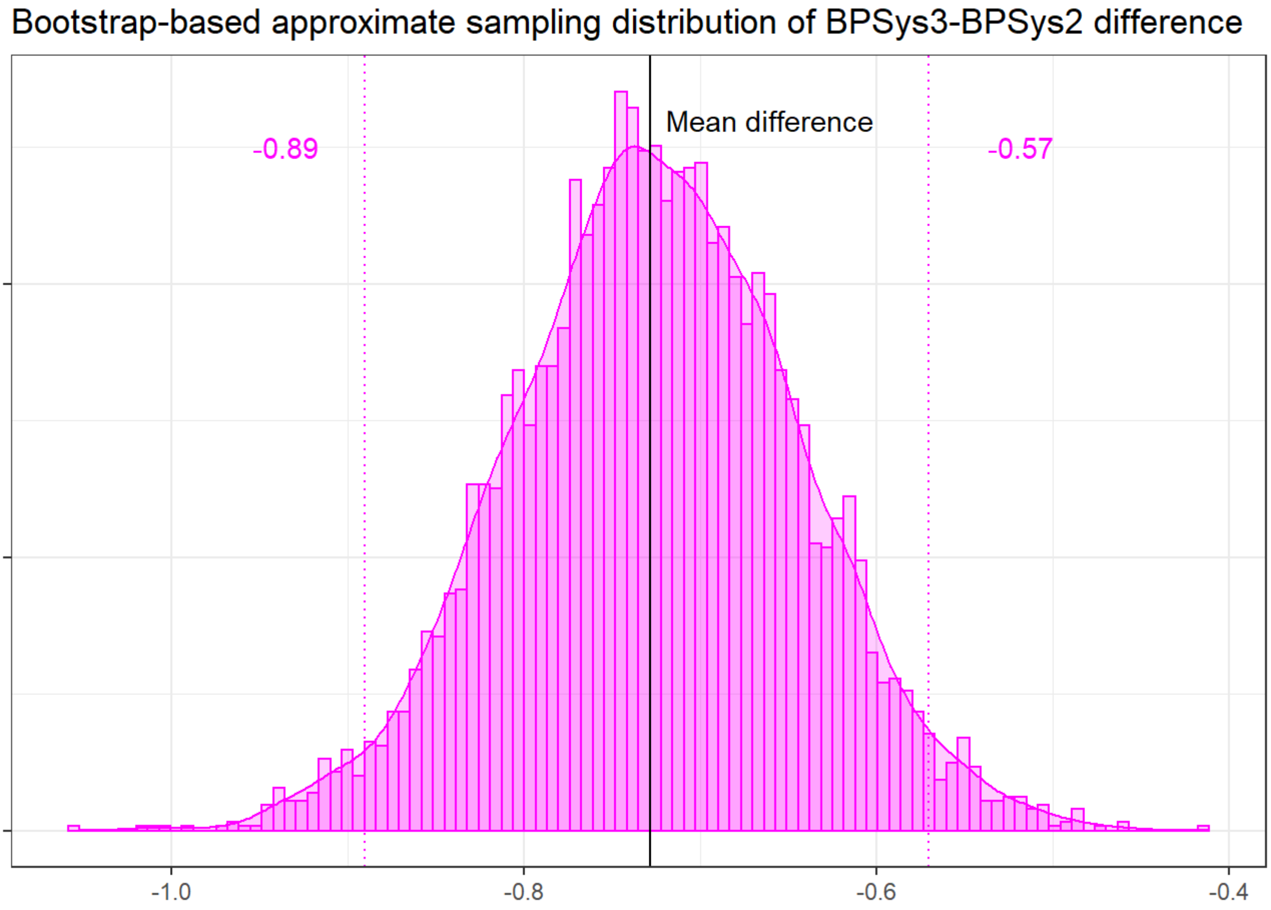
```
set.seed(1000)

n <- nrow(dt1_diff)
B <- 5000

xbar_boot <- rep(NA, B)
for (i in 1:B) {
    temp <- sample(dt1_diff$Diff, size = n, replace = TRUE)
    xbar_boot[i] <- mean(temp)
}

boot.CI <- quantile(xbar_boot, c(0.025, 0.975))
boot.CI
```

```
##      2.5%     97.5% 
## -0.885164 -0.568052
```

The 95% confidence interval for the difference in BPSys2 and BPSys3 printed above was obtained using the bootstrap method. To achieve this the data was sampled 5000 times with replacement using the function in the code chunk above to create many simulated samples without needing to repeat the original test. The average of each of these samples was then saved to a new data frame before the 2.5% and 97.5% quantiles were obtained using the quantile function. Thus, the confidence interval produced was from -0.885 to -0.568 (approx).

A plot of the empirical Bootstrap sample density is shown in the figure below with the difference recorded in each sample on the x-axis and the frequency on the y-axis. Additionally, the observed mean difference has again been plotted as a vertical line as well as the values for the 95% confidence interval obtained above. This shows the distribution of the bootstrap samples produced previously, with 95% of the values recorded falling between the two pink lines (the 95% confidence interval). From this it appears that the observed mean difference is in fact significant as it falls within this interval.

```
bootplot.f <- function(stat.boot, bins = 50) {
    df <- tibble(stat = stat.boot)
    CI <- round(quantile(stat.boot, c(0.025, 0.975)), 2)
    p <- df %>% ggplot(aes(x = stat, y = ..density..)) + geom_histogram(bins = bins,
        colour = "magenta", fill = "magenta", alpha = 0.2) +
        geom_density(fill = "magenta", colour = "magenta", alpha = 0.2) +
        geom_vline(xintercept = CI, colour = "magenta", linetype = 3) +
        theme_bw() +
      labs(x = "", y = "")
    p
}

p_xbarboot <- bootplot.f(xbar_boot, bins = 100)

p_xbarboot +
  geom_vline(xintercept = mean(dt1_diff$Diff)) +
  annotate("text", label = round(boot.CI[1], 2), x = (boot.CI[1] - 0.05), y = 5, colour = "ma
genta") +
  annotate("text", label = round(boot.CI[2], 2), x = (boot.CI[2] + 0.05), y = 5, colour = "ma
genta")  +
  annotate("text", x = -.66, y = 5.2, label = "Mean difference") +
  ggtitle("Bootstrap-based approximate sampling distribution of BPSys3-BPSys2 difference")
```



Q1.e

```
ttest <- t.test(x = dt1_diff$Diff) %>%
  tidy()

ttest %>%
  select(conf.low, conf.high) %>%
  kable(caption = "CLT Estimates for BPSys3-BPSys2 difference", format.args = list(digits = 4
)) %>%
  kable_styling(bootstrap_options = c("bordered", "striped"))
```

CLT Estimates for BPSys3-BPSys2 difference

conf.low	conf.high
-0.8865	-0.5695

As an alternative to the Bootstrap approach, a CLT based confidence interval was produced above using a t test. Interestingly, when rounded to two decimal places, this interval -0.89 to -0.57 is identical to the one produced by the Bootstrap method and, more importantly, produces the same outcome of confirming a significant difference between the two readings. The benefit of using a CLT approach is that, supposing the sample size is large enough, the sample mean will approximate the mean of the population. Whereas, using the Bootstrap method allows for a confidence interval to be calculated for any measure very simply and can be used to repeatedly resample data to get new sets. Additionally, this method can estimate parameters for smaller sample sizes which the CLT can not.

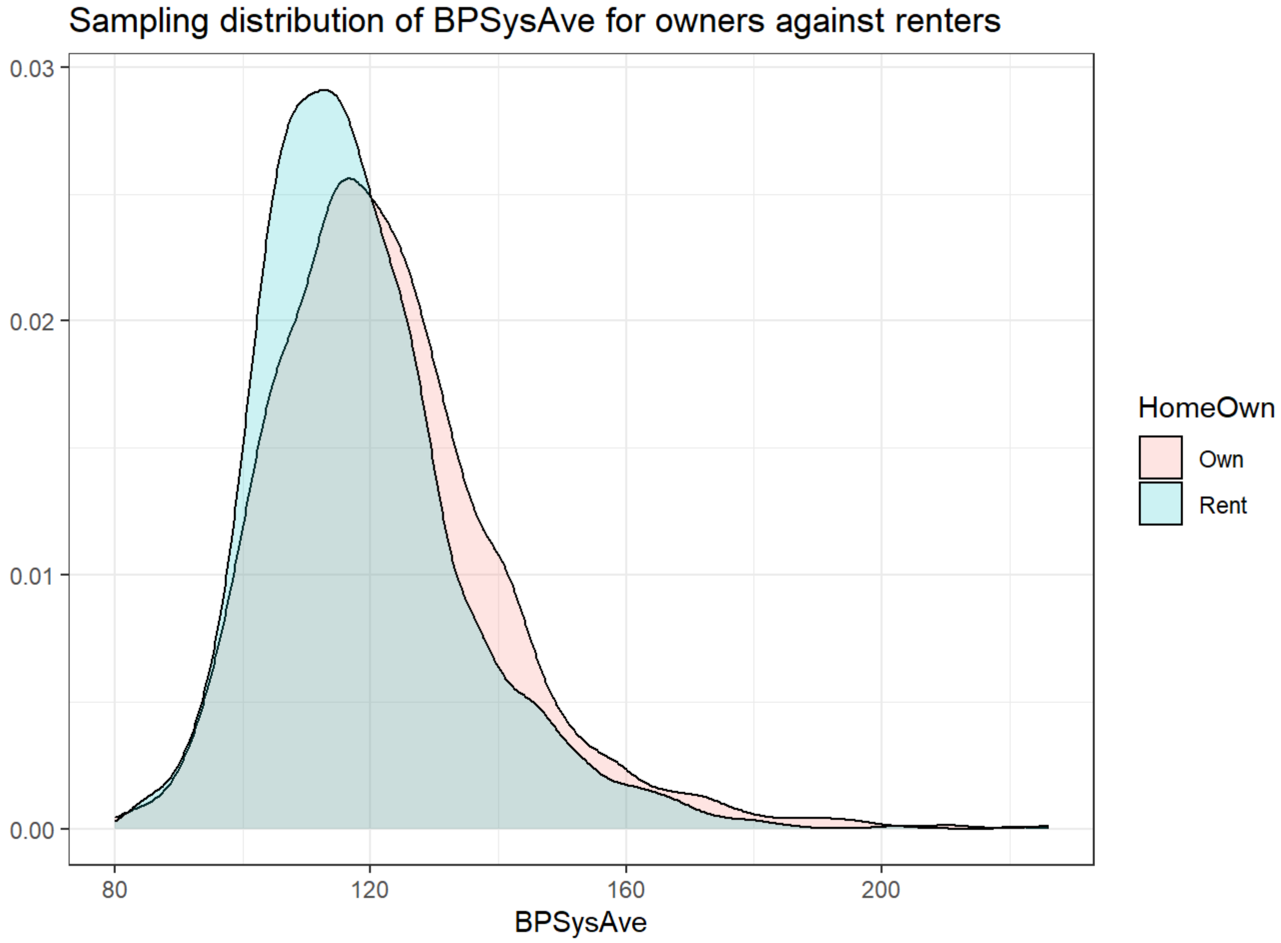
Q1.f

The two measures are not independent as BPSys2 and BPSys3 are consecutive readings of the same measure, systolic blood pressure, taken from the same person. Thus it is reasonable to assume that the any prior reading would impact on a future reading. It is important to take this dependence into account as, if it were assumed the two populations were independent, a different confidence interval would have been calculated. This is because, in that situation, only the average overall difference between groups would've been used to compare differences. However, as they are dependent, the object of interest is the difference within the pairs rather than within the two populations.

Question 2

Q2.a

```
dt1_home <- dt1 %>%
  filter(HomeOwn %in% c("Own", "Rent")) %>%
  select(Gender, Age, HomeOwn, BPSysAve)

dt1_home %>%
  ggplot(aes(x = BPSysAve, y = ..density.., fill = HomeOwn)) +
  geom_density(alpha = 0.2)  +
  theme_bw() +
  labs(y = "") +
  ggtitle("Sampling distribution of BPSysAve for owners against renters")
```



A kernal density plot was produced to compare the BPSysAve for home owners and renters with the reading on the x-axis and the relative density on the y-axis. The x-variable refers to the average reading of systolic blood pressure measured in mm Hg. Each curve was coloured according to tenure type. This plot was produced by filtering the data to only include the two ownership type of interest. Compared to the comparison in question 1, here two independent samples are being compared. As a result, there appears to be some difference in the distributions of each. Renters (the blue curve), have a slightly lower average systolic blood pressure than owners and also have a taller distribution with less spread. Both samples are relatively right skewed.

Q2.b

```
summary <- dt1_home %>%
  group_by(HomeOwn) %>%
  summarise(Obs = n(),
            Mean = mean(BPSysAve),
            SD = sd(BPSysAve),
            Min = min(BPSysAve),
            Q1 = quantile(BPSysAve, 0.25),
            Median = median(BPSysAve),
            Q3 = quantile(BPSysAve, .75),
            Max = max(BPSysAve))

summary %>%
  kable(format.args = list(digits = 4, big.mark = ","), caption = "Summary of BPSysAve for ow
ners against renters") %>%
  kable_styling(bootstrap_options = c("bordered", "striped"))
```

Summary of BPSysAve for owners against renters

HomeOwn	Obs	Mean	SD	Min	Q1	Median	Q3	Max
Own	2,815	122.4	17.86	80	110	120	132	221
Rent	1,488	118.4	15.93	83	107	116	126	226

The table above provides summary statistics for BPSysAve across the two ownership types. As can be seen by the ‘Obs’ column, more owner than renters were included in the study and on average they had a higher average blood pressure with more spread around the mean. The five number summaries are quite close, with owners scoring higher than renters in all categories except for the maximum. This indicates that it was a renter who recorded the highest average reading, even though on average this group was lower.

Q2.c

```
own <- dt1_home %>%
  filter(HomeOwn == "Own") %>%
  pull(BPSysAve) %>%
  as.numeric()

rent <- dt1_home %>%
  filter(HomeOwn == "Rent") %>%
  pull(BPSysAve) %>%
  as.numeric()

ttest <- t.test(x = own, y = rent) %>%
  tidy()

ttest %>%
  select(estimate,
         p.value,
         lower = conf.low,
         upper = conf.high) %>%
  kable(caption = "CLT estimates of BPSysAve for owners against renters", format.args = list
(digits = 4)) %>%
  kable_styling(bootstrap_options = c("bordered"))
```

CLT estimates of BPSysAve for owners against renters

estimate	p.value	lower	upper
3.999	0	2.954	5.043

For a two sided hypothesis test for this scenario the two options would be:

$$H_0 = BPSysAve_{own} = BPSysAve_{rent}$$

$$H_1 = BPSysAve_{own} \neq BPSysAve_{rent}$$

To test these hypotheses, a two sided t test was performed, the results of which are printed above. This test produced an estimate of 3.999 and a 95% confidence interval of 2.954 to 5.043. From this outcome the null hypothesis would be rejected in favour of the alternative hypothesis. This conclusion was reached as the 95% confidence interval did not contain 0, as this would indicate that there is a chance the observed difference was not significant and rejecting the null hypothesis would be a type 1 error.

Q2.d

```
# 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
```

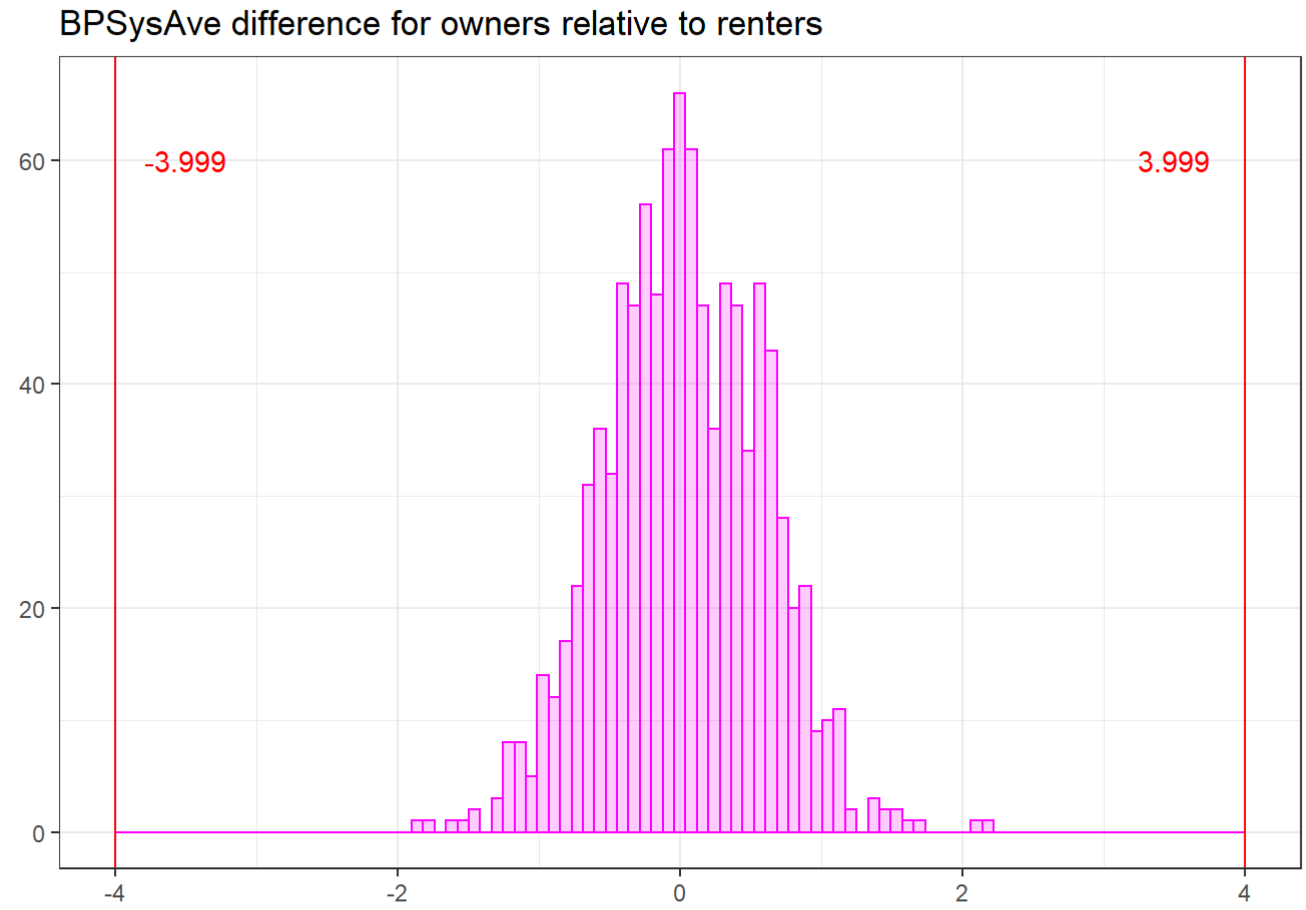
The code in the chunk above, with some alterations made later on can be used to test the hypotheses detailed in part c. The first three lines of code subsets the data to only include home owners and renters, sets an n value for the number of cases and a R valur for the number of replications. Following that, the data is saved to a new variable Rdt2 before being included in the for loop function. The first line within the function samples the BPSysAve variable n number of times without replacement. This effectively breaks the association observed originally between ownership type and average systolic blood pressure. This newly sampled data is then summarised by ownership type calculating a mean for each. Finally, the difference between these two means is calculated and stored in the RDiff object. This process is repeated R number of times. Overall this is produces a randomisation test. Once this has been performed, a 95% confidence interval can be constructed from the random permutations to see if the observed difference falls within its range as this would indicate the oberserved difference was not signifcant.

Q2.e

```
dt2 <- dt %>% filter(HomeOwn!="Other")
n = nrow(dt2)
R = 1000

# student to add
# student to add

RDiff <- array(dim = R)
set.seed(1000)

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
RDiff <- as.data.frame(RDiff) %>%
  pivot_longer(cols = 1:1000) %>%
  select(value)

xobs <- summary$Mean[1] - summary$Mean[2]



pval <- RDiff %>%
  filter(abs(value) >= abs(xobs)) %>%
  nrow()/R

pval
```

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

```
RDiff %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 100, colour = "magenta", fill = "magenta", alpha = 0.2) +
  geom_vline(xintercept = xobs, colour = "red") +
  annotate("text", x = xobs-.5, y = 60, label = round(xobs, digits = 3), colour = "red") +
  geom_vline(xintercept = -xobs, colour = "red") +
  annotate("text", x = -xobs+.5, y = 60, label = round(-xobs, digits = 3), colour = "red")  +
  labs(x = "",y = "") +
  ggtitle("BPSysAve difference for owners relative to renters") +
  theme_bw()
```



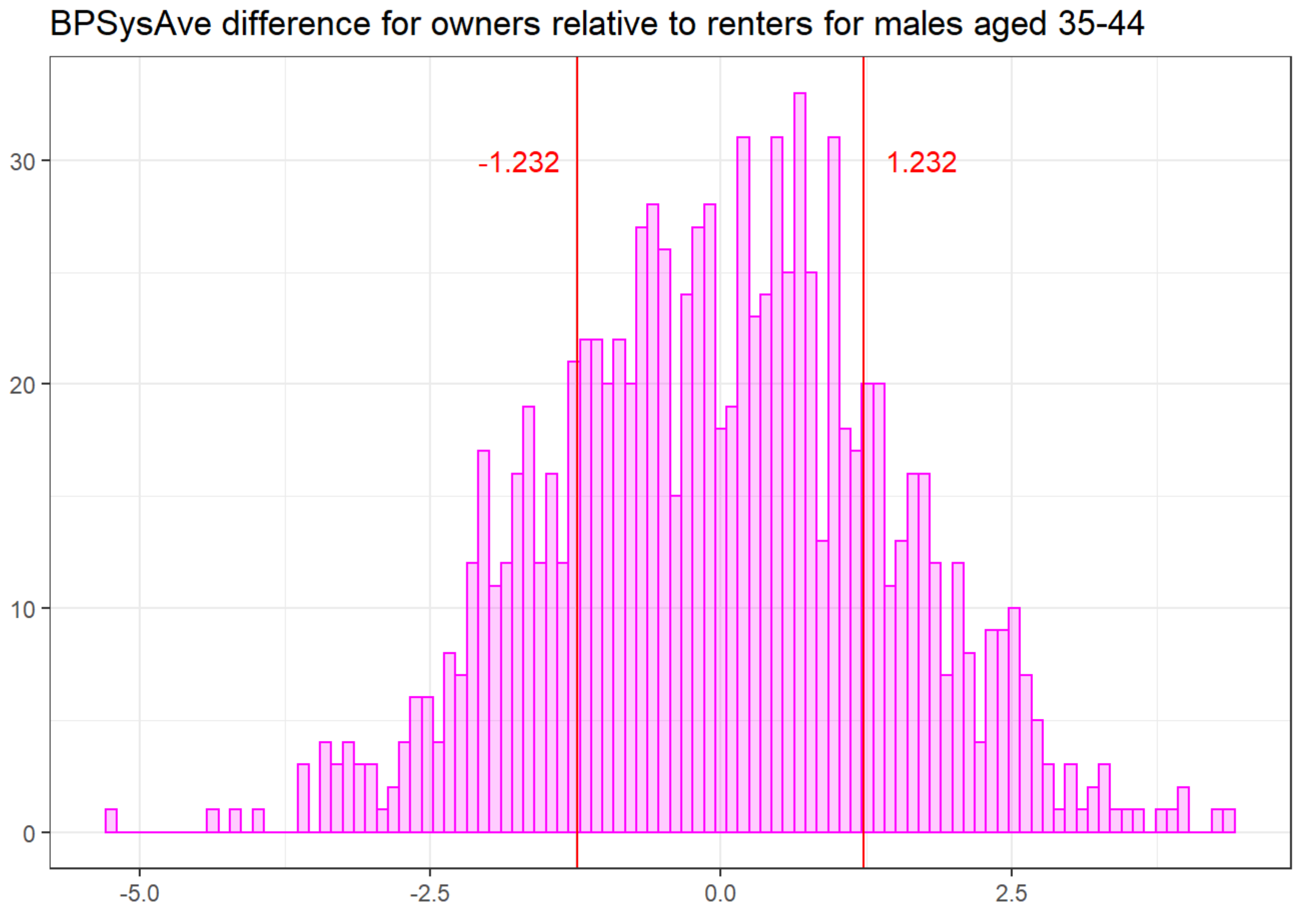
The code from 2.d has been supplemented to carry out the full test. From this a p value was calculated by taking the number of observations more extreme than the observed difference and dividing it by the number of randomisations performed (in this case 1000). The code printed above displays that in fact 0 of the randomisations produced a value this extreme. Note that this was a two-sided test so the absolute values were used. This would lead to the null hypothesis being rejected in favour of the alternative with 95% strength of evidence. This can be observed also in the graphic displayed above. This plot shows the distribution of the randomisation differences with the recorded difference on the x-axis and the frequency on the y-axis. The values corresponding to the observed difference are printed in red. As can be seen, none of the tests produced a difference anywhere near as large as the one observed, with most centering around a difference of 0. This again would lead to the rejecting of the null in favour of the alternative in that the difference in BPSysAve for home owners and renters is not 0.

Q2.f

```
dt2 <- dt %>% filter(HomeOwn!="Other",
                     Gender == "male",
                     Age >=35 & Age <= 44)
n = nrow(dt2)
R = 1000

# student to add
# student to add

RDiff <- array(dim = R)
set.seed(1000)

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
RDiff <- as.data.frame(RDiff) %>%
  pivot_longer(cols = 1:1000) %>%
  select(value)

diff <-  dt2 %>%
  group_by(HomeOwn) %>%
  summarise(avg = mean(BPSysAve))

xobs <- diff$avg[1] - diff$avg[2]

pval <- RDiff %>%
  filter(abs(value) >= abs(xobs)) %>%
  nrow()/R

pval
```

```
## [1] 0.401
```

```
RDiff %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 100, colour = "magenta", fill = "magenta", alpha = 0.2) +
  geom_vline(xintercept = xobs, colour = "red") +
  annotate("text", x = xobs - .5, y = 30, label = round(xobs, digits = 3), colour = "red") +
  geom_vline(xintercept = -xobs, colour = "red") +
  annotate("text", x = abs(xobs) + .5, y = 30, label = round(abs(xobs), digits = 3), colour =
"red") +
  labs(x = "",y = "") +
  ggtitle("BPSysAve difference for owners relative to renters for males aged 35-44") +
  theme_bw()
```



The test can be repeated as in part 2.e to compare cases only for men aged 35-44 by first filtering the data to match these two conditions. Repeating the steps as before, a p value of 0.401 is obtained and printed above. This would suggest that there is not enough evidence to reject the null hypothesis that there is no difference in BPSysAve between owners and renters for males aged 35-44. This again can be seen graphically by plotting the randomisation differences on the x-axis and the frequencies on the y-axis, and plotting vertical lines corresponding to the observed difference. From this plot we can see that a significant amount of the randomisations produced outcomes more extreme than the observed difference. This suggests that there is eviedence that the observed difference was simply due to random noise and was not based on whether someone was an owner or a renter. This difference in results compared to part 2.e may be partly due to the fact that gender and age is playing a bigger role in the BPSysAve reading than the home ownership status. Thus the difference observed in 2.e may be due to different distributions of gender and age within each tenure type rather than the type itself.