Comments on Assignment 1

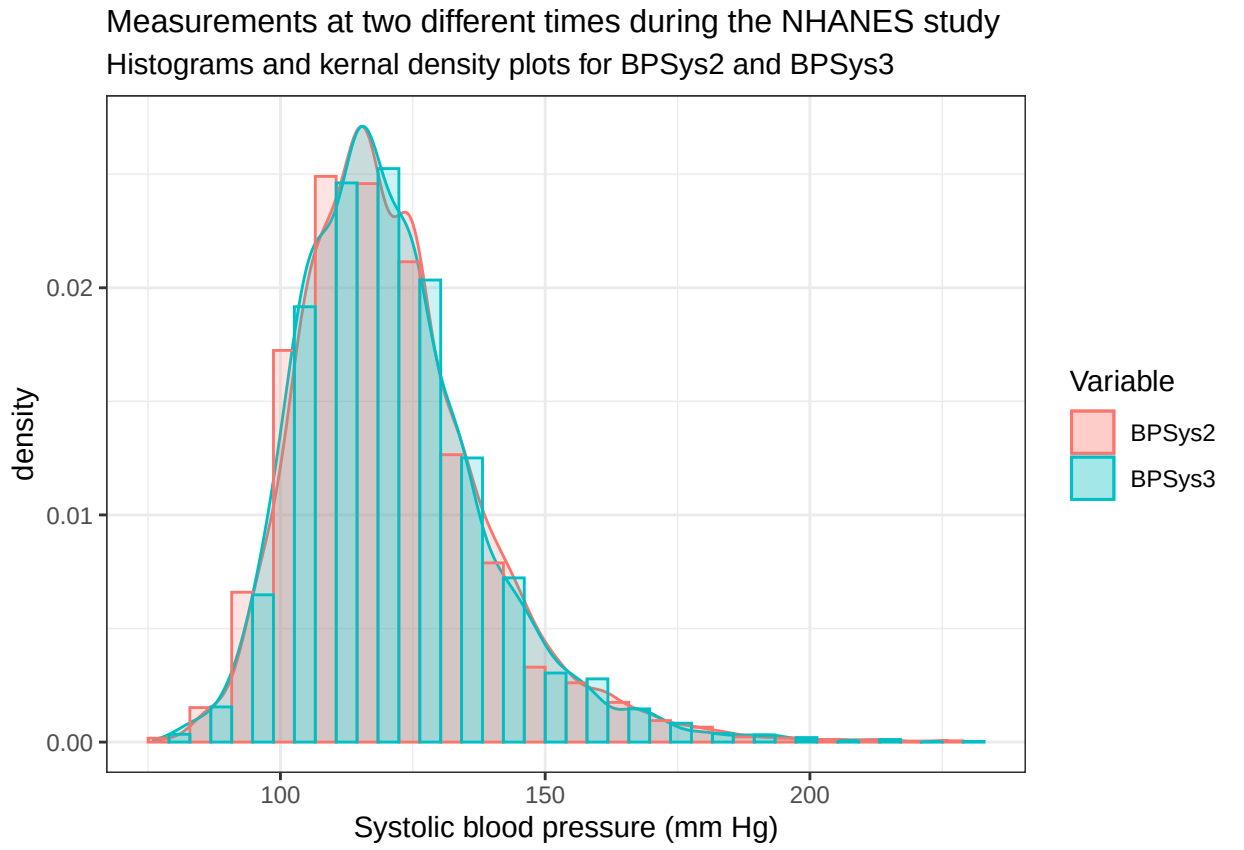
Statistical Thinking 2020

```
options(digits = 6)
library(tidyverse)
library(broom)
library(kableExtra)
theme_set(theme_bw())
# install.packages('NHANES')
library(NHANES)
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, part a

This problem is concerned with comparing the blood pressure measurements of the same individuals at two different time points. An example of a suitable plot for part a is shown below.

```
dt1L <- dt1 %>% pivot_longer(cols = starts_with("BPSys"), names_to = "Variable",
    values_to = "BPSys")

Q1p1 <- dt1L %>% dplyr::filter(Variable != "BPSysAve") %>% ggplot(aes(BPSys,
    ..density.., colour = Variable, fill = Variable)) + geom_density(alpha = 0.2) +
    geom_histogram(alpha = 0.2, position = "dodge", bins = 20) +
    ggtitle("Measurements at two different times during the NHANES study",
        "Histograms and kernal density plots for BPSys2 and BPSys3") +
    theme(plot.title = element_text(size = 12)) + xlab("Systolic blood pressure (mm Hg)")

Q1p1
```



A single plot (compared with two side-by-side plots) is effective at showing similarities and differences between these two samples. A histogram, kernel density plot or both are the most suitable ways to display this type of data, with a bar chart also being acceptable. Although the BPSys2 and BPSys3 variables are only recorded to integer precision, it is common to think of these measurement as being continuous variables. Note that both sample distributions (for BPSys2 and BPSys3) need to be visibly distinct. Colour is a good way to distinguish the two samples, particularly if the two groups are somehow separated (e.g. using "dodge") in the plot construction, though other approaches possible and considered on a case-by-case basis. Axis labels that suit all data shown on the plot help to clarify the measure being compared across samples, while including a legend ensures the reader can quickly identify the variables to compare shown in the plot. A simple title will focus attention on the main purpose of the plot, and correct spelling helps avoid distraction and confirms that the report was carefully written.

The two (very similar) distributions are skewed to the right, indicating that there are a few individuals whose systolic blood pressure was high when measurements were taken. There is greater variability for these large values than is evident for value at the lower end of these distributions. The histogram bin size ("bins") was controlled to smooth out the histogram more than the default number of bins as it seems likely that the integer precision of the original measurements were causes unnecessary variation in the heights of the various histogram bars. Notably, the range of values spans from less than 100 to over 200 millimeters of mercury (mm Hg).

The plot of the histograms shown above is relevant for later parts of this question. We want to know if the two blood pressure measurement populations have the same average values. But we notice when even when plotting the two histograms that the two samples are extremely similar, noting also that they represent two measurements taken on the same person. This alone is sufficient reason to difference the observations for each person and do the paired t-test, as the assignment requested, and submissions should have commented on this.

Question 1, part b

A plot of the sample distribution of the difference $Diff = BPSys3 - BPSys2$ is shown below. Again here a histogram, kernel density plot or both are the most suitable ways to display the difference of the two *BPSys* measurements, though a bar chart would also be acceptable since measurements are recorded as integers. Titles and axis labels should include enough information so that someone who does not know the variable names can still understand what the plot is showing.

Notice that the distribution of the differences is symmetric around a value close to zero, and has a bell shape. Notably the differenced blood pressure measurements have a much smaller range, within -30 to +30 mm Hg. By differencing the individual-specific measurements the estimated variance used to standardise the test-statistic is appropriately reduced, increasing the chance of finding a difference in the population means if one exists.

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

Q1p2 <- dt1 %>% ggplot(aes(BPDiff, ..density..)) + geom_density(colour = "blue",
    fill = "blue", alpha = 0.2, bw = 1) + geom_histogram(bins = 25,
    colour = "blue", fill = "blue", alpha = 0.2) + theme(plot.title = element_text(size = 12)) +
    xlab("Difference in blood pressure measurements (mm Hg)") +
    ggtitle("Histograms of difference in blood pressure measurements on same individual",
        "Diff = BPSys3 - BPSys2")

Q1p2
```

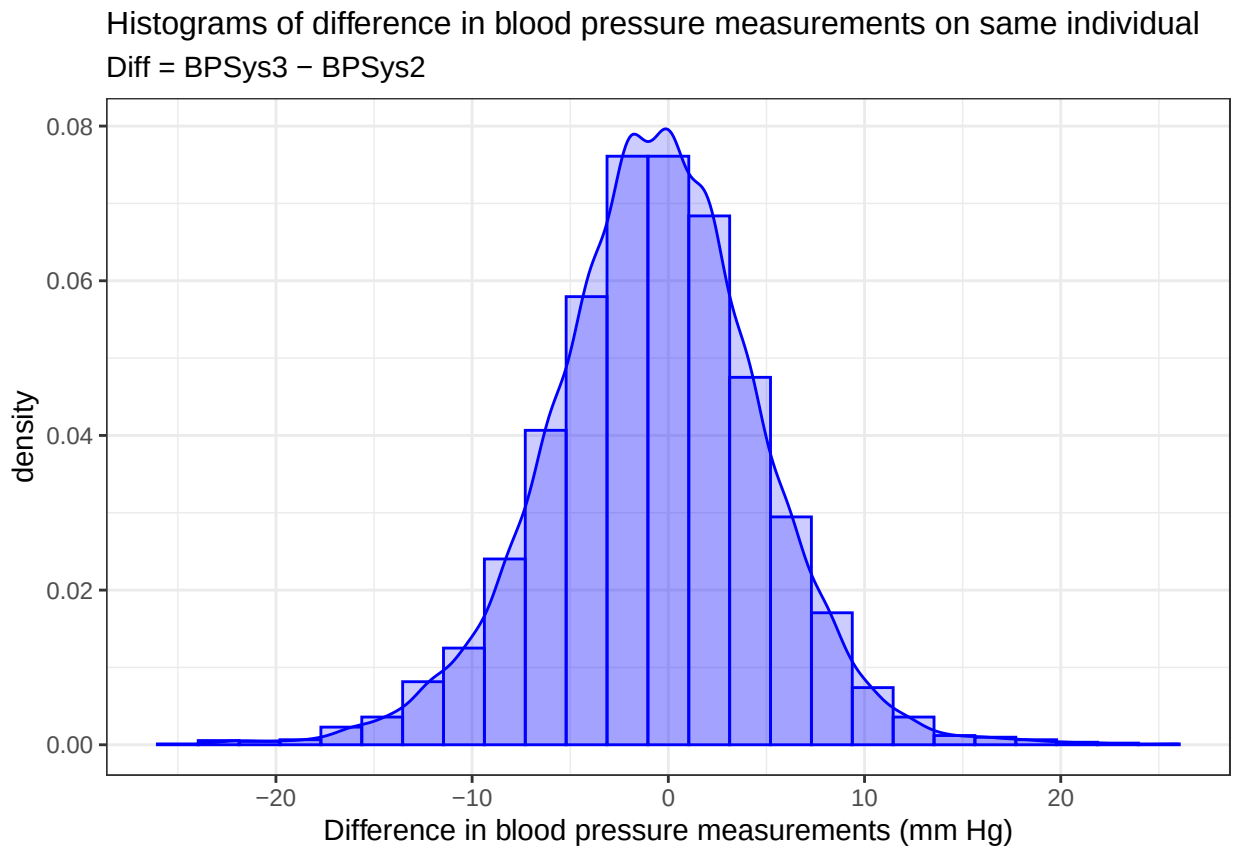


Table 1: Summary statistics for Diff, BPSys2 and BPSys3

Variable	n	mean	sd	se	min	max	IQR
BPDiff	4415	-0.728	5.372	0.0012	-24	26	6
BPSys2	4415	121.323	17.749	0.0040	76	226	20
BPSys3	4415	120.595	17.351	0.0039	76	226	22

Question 1, part c

Several summary statistics for the variables named *BPSys2*, *BPSys3* are shown below. While other summaries could be provided, here focus is given mainly to commonly used measures of location and spread, which are useful for comparing the different distributions. A description of each of the statistics shown in the table below are provided here:

- n: the number of available observations, which is the same for all variables in this case
- *mean*: the sample average value, a measure of location
- *sd*: the sample standard deviation, a measure of spread
- *se*: the estimated standard error of the mean, a measure of variability in the point estimator used for the population mean
- *min*: the minimum value observed, useful to determine the range of observed values
- *max*: the maximum value observed, useful to determine the range of observed values
- *IQR*: the interquartile range, an alternative measure of spread that concentrates on the middle 50% of observations

```
# control the number of digits so not too many showing
options(digits = 4)

# Need to reshape again to include BPDiff
dt2L <- dt1 %>% pivot_longer(cols = starts_with("BP"), names_to = "Variable",
    values_to = "BPSys")

# By having reshaped to a longer format the summary table is
# easier to produce, and fits on the page
dt1S <- dt2L %>% select(Variable, BPSys) %>% filter(Variable !=
    "BPSysAve") %>% group_by(Variable) %>% summarise(n = n(),
    mean = mean(BPSys), sd = sd(BPSys), se = sd(BPSys)/n(), min = min(BPSys),
    max = max(BPSys), IQR = IQR(BPSys))

dt1S %>% kable(caption = "Summary statistics for Diff, BPSys2 and BPSys3") %>%
    kable_styling()
```

Question 1, part d

The Bootstrap is used to produce a 95% confidence interval for the population difference in blood pressure measurements. Although not used as a formal test, it helps to develop an understanding of the variability present in the measurement of this difference to inform the NHANES study developers.

To obtain the Boot-strap based interval, we resample many replications of the original dataset by resampling $n = 4415$ *Diff* values *with replacement* and calculate the corresponding sample means. This gives us an approximation to the sampling distribution of the estimator $\bar{Diff}$, which is used to estimate the population mean difference. The 'true' (or 'exact') sampling distribution is a hypothetical quantity that we cannot

observe directly, as it represents the distribution of potential sample means that could occur if an unlimited number of independent samples with the same $n = 4415$ sample size were repeatedly taken from the entire population of possible NHANES respondents. We then use the lower 2.5% and 97.5% quantiles of this distribution as the end points of the 95% confidence interval, which represents a degree of uncertainty in estimating the underlying population mean *Diff* value.

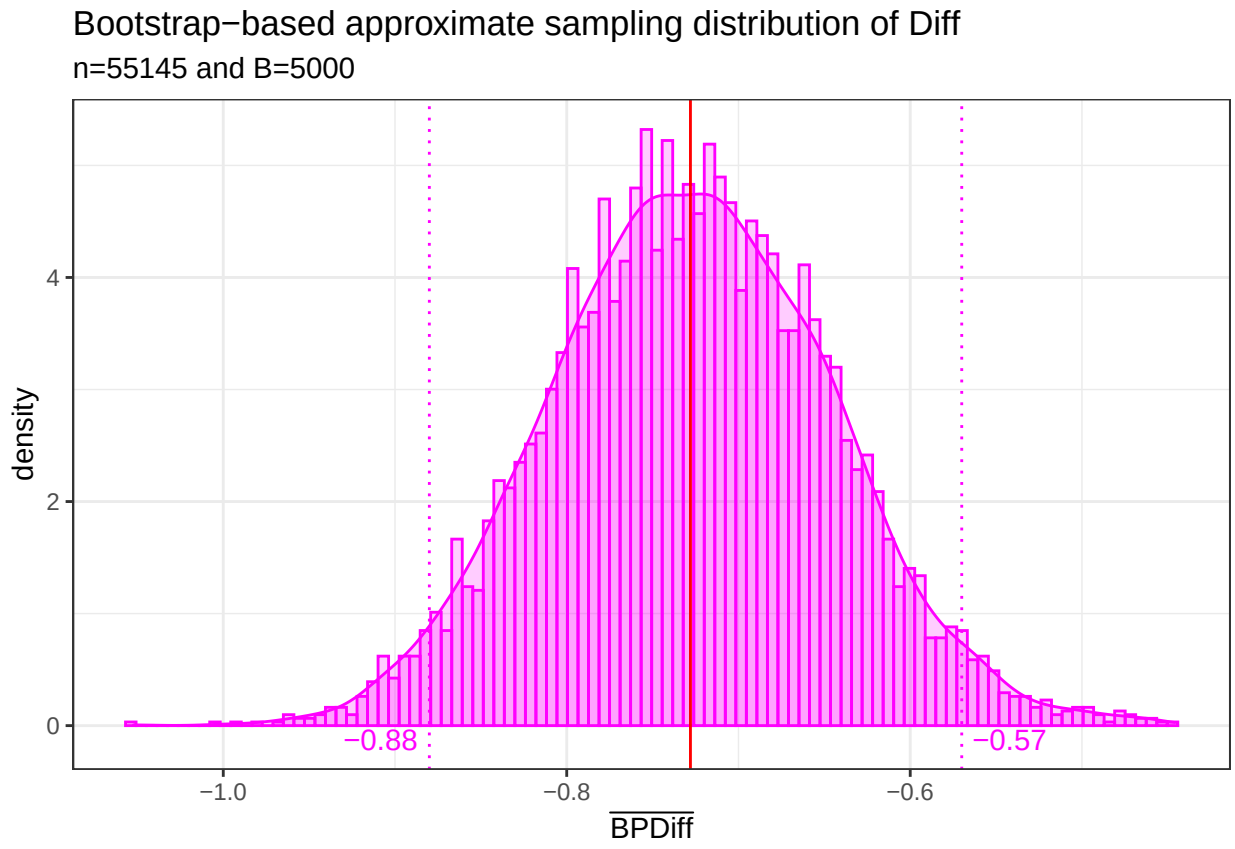
In addition to producing the 95% Bootstrap-based interval, a plot of the bootstrap distribution, which contains the values of the sample means produced from the resampling of $R = 5000$ samples from the original, is produced to get a more complete view of the bootstrap distribution. Note that we set the random seed value for this resampling (using the *set.seed()* **R** function) to ensure that the results produced are able to be reproduced later, if ever needed.

Advantages of the Bootstrap approach are that it does not depend on an asymptotic limit for the approximation to the sampling distribution of the sample average, which in turn only uses the summary statistics $\bar{Diff}$ and its standard error, both reported in part c. Instead, the Bootstrap is able to come up with replicated potential values of $\bar{Diff}$ that are consistent with the realised values of the difference observed in the sample. The bootstrap distribution that is produced reflects the sample size n and the shape of the underlying distribution. It is typically not a symmetric distribution, unlike the normal approximation produced by the Central Limit Theorem (CLT). The empirical distribution of these bootstrap replicated potential mean values is used to construct the confidence interval.

```
# define the bootplot.f function

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()

    p
}
######## end of bootplot.f function #######
```

```
# just need to reuse code from Lab 5 step i.
set.seed(57892)
n <- nrow(dt1)
x <- dt1 %>% pull(BPDiff)
xbar.x <- mean(x)

B <- 5000
xbar_boot <- rep(NA, B)
for (i in 1:B) {
    temp <- sample(x, size = n, replace = TRUE)
    xbar_boot[i] <- mean(temp)
}

## step iv.
boot.CI <- quantile(xbar_boot, c(0.025, 0.975))

## step v.
```

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

## layer the plot
p_xbarboot <- p_xbarboot + geom_vline(xintercept = xbar.x, colour = "red") +
    annotate("text", label = round(boot.CI[1], 2), x = (boot.CI[1] -
        0.025), y = -0.12, colour = "magenta") + annotate("text",
    label = round(boot.CI[2], 2), x = (boot.CI[2] + 0.025), y = -0.12,
    colour = "magenta")


## add titles and change axis labels, save plot as object p2
## (changes here)
p2 <- p_xbarboot + xlab(expression(bar(BPDiff)))
p2 <- p2 + ggtitle("Bootstrap-based approximate sampling distribution of Diff",
    "n=55145 and B=5000")

## step vi.
p2
```




```
options(digits = 2)
boot.CI
```

```
##  2.5% 97.5%
## -0.88 -0.57
```

The resulting Bootstrap-based 95% confidence interval is given by (-0.88, -0.57).

Table 2: Paired t-test output

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-0.73	-9	0	4414	-0.89	-0.57	One Sample t-test	two.sided

Question 1, part e

The CLT-based approach to produce a 95% confidence interval for the average difference in systolic blood pressure measurement, is derived from the approximation:

$$\frac{(\bar{Diff} - \mu_{Diff})}{s_{Diff}/\sqrt{n}} \overset{approx}{\sim} N(0,1)$$

The interval itself is given by:

$$\left[\bar{Diff} + z_{0.025}\frac{s_{Diff}}{\sqrt{n}},\ \bar{Diff} + z_{0.975}\frac{s_{Diff}}{\sqrt{n}}\right].$$

```
correct_test <- t.test(dt1$BPDiff) %>% tidy()
correct_test %>% kable(caption = "Paired t-test output") %>%
    kable_styling()
```

To obtain this interval in **R** we use the one-sample *t.test* function (where the one sample is comprised of the values in *Diff*) and extract the *conf.low* and *conf.high* output values. The CLT-based interval is thus (-0.89, -0.57).

In general, the Bootstrap approach provides us with an alternative means to evaluate the sampling variability that may influence the estimator of the population quantity of interest (here the difference in the two population means). Rather than relying on the mathematical result known as the CLT to provide the approximate sampling distribution (for which we need to have a very "large" sample size to feel confident that the approximation is good enough), resampling techniques provide a visual representation of the sampling variation that could result from independent replicates of the study.

On the other hand, the CLT-based approximation is very easy to implement, and is often a good starting point. Many people have learned about t-tests and confidence intervals, and will undstand how such an interval is constructed.

In this particular setting, it is important to consider whether the Bootstrap-based or the CLT-based interval is more appropriate, however we note that the intervals themselves actually differ only slightly:

CLT: (-0.89, -0.57)

Bootstrap: (-0.88, -0.57)

This similarity of the confidence intervals can be explained by the shape of the bootstrap distribution (shown in the figure above), which is roughly symmetric, and somewhat bell-shaped. Though the "tails" of the bootstrap distribution may be slightly longer than for a normal distribution, since we are trimming 2.5% from each end (and not anything smaller) any differences in the extreme values are not relevant to the 95% confidence interval. In this case, we might report the CLT-based interval (as this will be better understood by a greater number of potential readers of the report) and/or avoid reporting the interval bounds with a large number of decimal places.

Question 1, part f

Simply, the measures BPSys2 and BPSys3 are not independent because a pair of each come from the same individual. There are differences between individuals that determine one's predisposition to having higher or

Table 3: Independent samples t-test output (part 1)

estimate	estimate1	estimate2	statistic	p.value	parameter
-0.728	120.6	121.3	-1.949	0.0514	8823

Table 4: Independent samples t-test output (part 2)

conf.low	conf.high	method	alternative
-1.46	0.0043	Welch Two Sample t-test	two.sided

lower blood pressure overall that will influence each of these measurements in the same way. These common influences mean that the two measurements are not independent.
It is critical to use the differenced data, to look only at the sample of differences within individual from one measurement date to another.

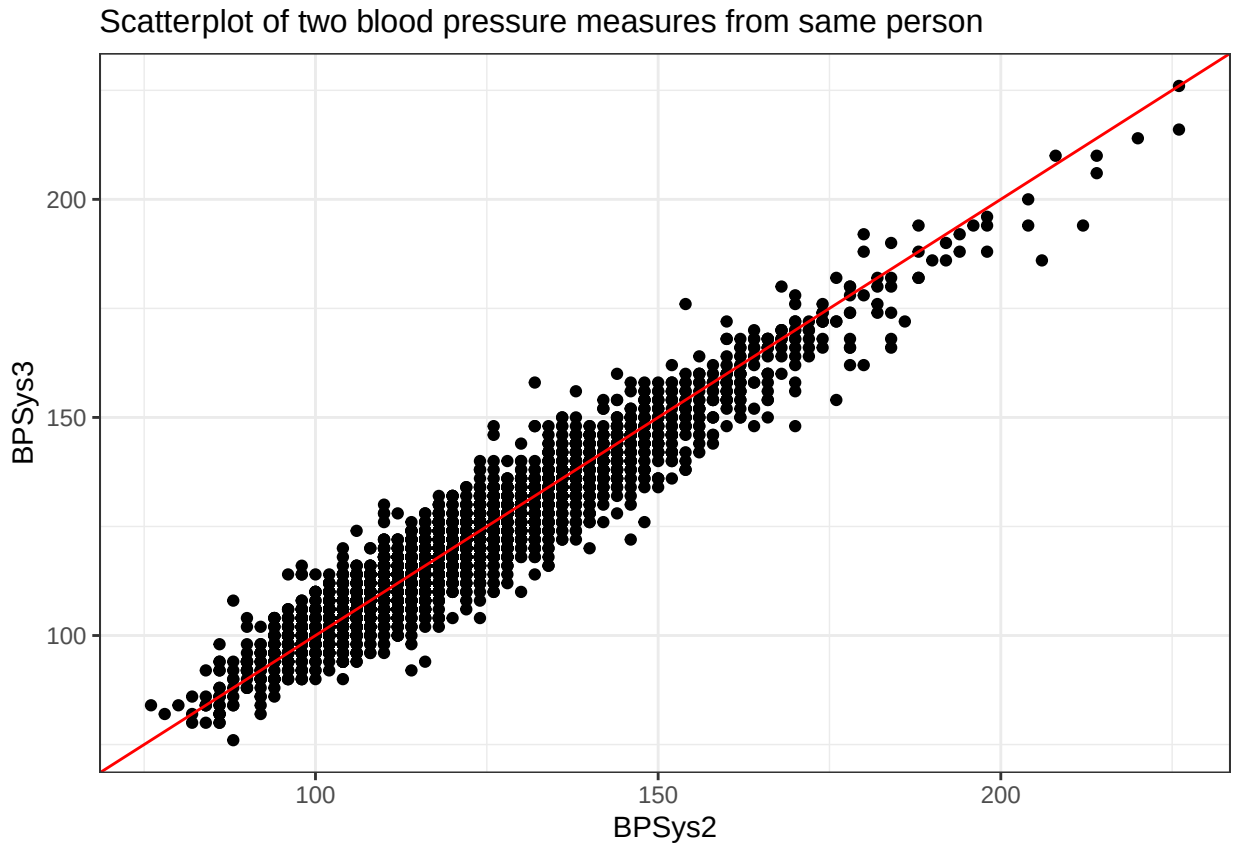
```
options(digits = 4)
incorrect_test <- t.test(dt1$BPSys3, dt1$BPSys2) %>% tidy()

incorrect_test %>% select(estimate, estimate1, estimate2, statistic,
    p.value, parameter) %>% kable(caption = "Independent samples t-test output (part 1)") %>%
    kable_styling()
```

```
incorrect_test %>% select(conf.low, conf.high, method, alternative) %>%
    kable(caption = "Independent samples t-test output (part 2)") %>%
    kable_styling()
```

Submissions should note that the mean difference is equal to the difference in means, so the two confidence intervals from the correct and incorrect t.test output are both centered at the same place, though the incorrect interval interval is much wider, and the resulting confidence interval for the difference in means that results covers zero. Without controlling for the additional variation from person to person, it is not possible to see the difference between the two average means, from a statistical perspective, if the noise around the measures is too large.

To extend this discussion a bit further, if you were to investigate the data beyond the specific tasks requested in the assignment, you might notice that plotting say BPSys3 against BPSys2 (shown below) results in a scatterplot that shows strong dependency between these two variables.

```
Q1scatter <- dt1 %>% ggplot(aes(x = BPSys2, BPSys3)) + geom_point() +
    geom_abline(intercept = 0, slope = 1, colour = "red") + ggtitle("Scatterplot of two blood pressure m
    theme(plot.title = element_text(size = 12))

Q1scatter
```



The above plot highlights the fact that there are greater differences among BP measurements of different individuals than there are between BP measurements taken on the same person (at least, it seems, from this data). This was also noticeable from the summary statistics reported in part c. By differencing first we actually reduce the variance in the estimator of $\mu_3 - \mu_2$, where μ_2 is the population average measurement of systolic blood pressure on the second occasion and μ_3 is the population average measurement of systolic blood pressure on the third occasion.

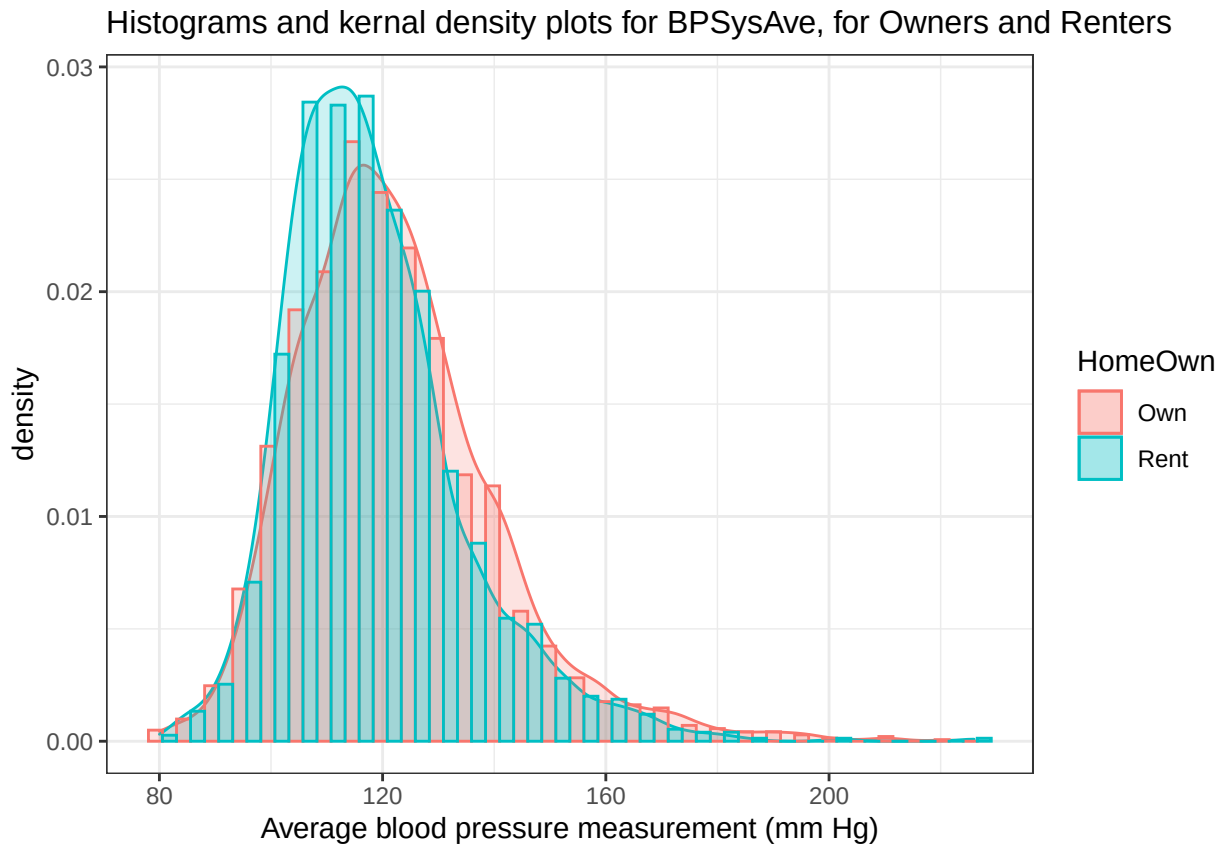
Treating the two populations as *independent* (which they are not, as we can see above they are positively correlated), we overestimate the standard error of the estimator of $\mu_3 - \mu_2$, and in this case produce a confidence interval that is substantially wider than the one produced using the difference measurements. Indeed, the incorrect 95% confidence interval covers zero, which could lead the NHANES study developers to erroneously think that the difference in the two measures are not (statistically) distinguishable from zero.

Question 2, part a

We are now interested in exploring the difference between the average systolic blood pressure, as measured by *BPSysAve*, for people who own their own home (*HomeOwn*="Own") versus people who are renting (*HomeOwn*="Rent"). In particular we will consider whether the population means of the distributions of these two separate groups are equal.

First we produce a single plot, shown below, 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". Immediately now we see must greater differences between the two measurements than we saw in Question 1. Both distributions are still skewed to the right, and appear to take on a similar range of values, however the mode of the "Renters" group is to the left of the mode for the "Owners" group. While the two density estimates overlap a fair bit, the "Renters" distribution seems to be more concentrated and a bit to the left of the "Owners" distribution. It is not obvious from the plot whether the two populations have different means.

```
Q2p1 <- dt1 %>% filter(HomeOwn != "Other") %>% ggplot(aes(x = BPSysAve,
    ..density.., colour = HomeOwn, fill = HomeOwn)) + geom_density(alpha = 0.2) +
    geom_histogram(alpha = 0.2, position = "dodge") + theme_bw() +
    ggtitle("Histograms and kernal density plots for BPSysAve, for Owners and Renters") +
    theme(plot.title = element_text(size = 12)) + xlab("Average blood pressure measurement (mm Hg)")

Q2p1
```



Question 2, part b

A selection of summary statistics for *BPSysAve* are shown below, for each of the two *HomeOwn* groups of interest. Definitions for these summary statistics are reported in Question 1 part b, his time relevant to the *BPSysAve* variable and to the "Owner" (Own) and "Renter" (Rent) sub-populations of the NHANES population. Note the "Other" group has been excluded, as it is not relevant to the question.

The features seen in the plot above also appear in the summary statistics. Notably, the average value of the *BPSysAve* variable for "Renters" is 4 mm Hg smaller than that of the "Owners" group. Considering the very small standard errors reported, this seems to be a very large difference. In addition, while the min and max values are not vastly different between the two groups, the IQR values show that at least the middle 50% of the "Renters" distribution is more concentrated as it spans a range of values that is 3 mm Hg smaller than does the middle 50% of the "Owners" group. These summary statistics quantify the features we noted from the plot in part a.

Table 5: Summary statistics for BPSysAve variable, by HomeOwner group

HomeOwn	n	mean	sd	se	min	max	IQR
Own	2815	122.4	17.86	0.0063	80	221	22
Rent	1488	118.4	15.93	0.0107	83	226	19

```
dt2S <- dt1 %>% group_by(HomeOwn) %>% filter(HomeOwn != "Other") %>%
    summarise(n = n(), mean = mean(BPSysAve), sd = sd(BPSysAve),
        se = sd(BPSysAve)/n(), min = min(BPSysAve), max = max(BPSysAve),
        IQR = IQR(BPSysAve))

dt2S %>% kable(caption = "Summary statistics for BPSysAve variable, by HomeOwner group") %>%
    kable_styling()
```

Question 2, part c

```
Own <- dt1 %>% filter(HomeOwn == "Own") %>% pull(BPSysAve)
Rent <- dt1 %>% filter(HomeOwn == "Rent") %>% pull(BPSysAve)
Q2c <- t.test(Own, Rent) %>% tidy()
Q2c
```

```
## # 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>
```

```
ObsDiff <- Q2c %>% pull(estimate)
```

We are first asked to estimate the average difference in *BPSysAve* for "Owners" relative to "Renters", and using a CLT-based approach, and to report a 95% confidence interval for this difference. The sample difference itself (sample average for Owners minus average for Renters) is 3.9986 (mm Hg) while the corresponding CLT-based 95% confidence interval for the population difference is given by (2.9538, 5.0434).

In addition, from the output of the *t.test()* function used to produce the confidence interval, we find that the test of:

$$H_0 : \mu_{Owners} = \mu_{Renters} \quad \text{vs.} H_1 : \mu_{Owners} \neq \mu_{Renters}$$

where μ_{Owners} and $\mu_{Renters}$ correspond to the population mean *BPSYsAve* measurements. Owing to the extremely small reported p-value of 7.9114×10^{-14}, we reject H_0 and conclude that the population average *BPSysAve* measurement for each of the "Owners" and "Renters" are not equal.
Given the plot from part a. and the summary statistics from part b., it is not surprising to see such a small p-value for difference in population means, though perhaps there is some surprise given that there are some similarities in the two sample distributions. We note, however, that the sample sizes are relatively large, and as noted above this results in the standard errors of the corresponding sample means being very small, likely the feature giving rise to the very small p-value.

Question 2, part d

The code presented in this question is recognised as (partial) code for a *Randomisation test*. However, unlike similar code used in Week 3 Lab, here we are testing for the difference of two independent sample means. This test uses random permutations of the data, breaking any existing dependency that could possibly exist between the *HomeOwn* categories (either "Owners" or "Renters") and the measured *BPSysAve* variable. For each random permutation of the data produced, a new sample is obtained and a value of the outcome *RDiff* is able to be calculated. The resulting empirical distribution produced from this set of R permutations that represents the distribution of the the hypothetical difference between *BPSysAve* values under H_0. By comparing the position of the observed difference in *BPSysAve* values (= 3.9986 (mm Hg)) to this null distribution we can work out the p-value, and hence the conclusion, of the permutation test of the same hypotheses detailed in part c.

Question 2, part e

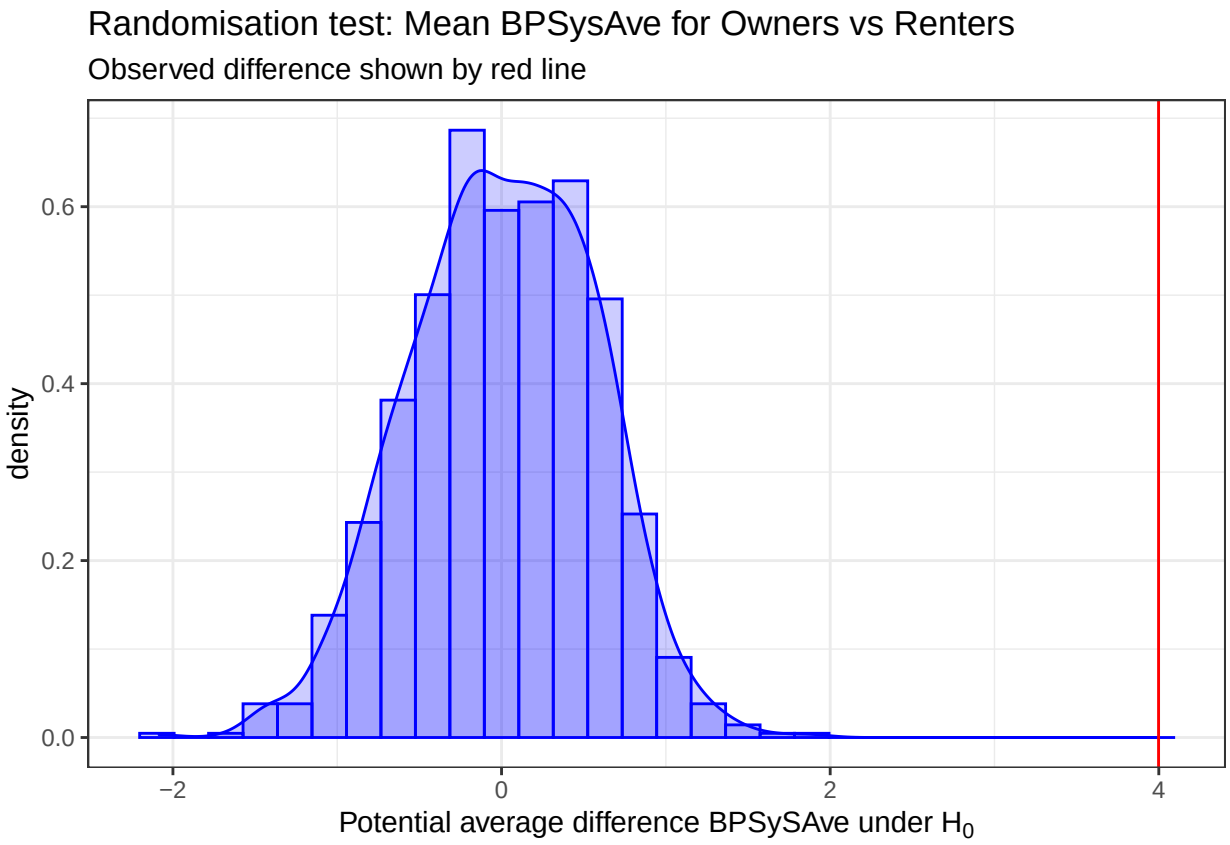
Some additional elements are required to complete the provided code, and these are shown in the code chunk below.

- First we note the reduced dataset (in object *dt1*) is used, and the "Other" group from the *HomeOwn* variable is excluded from the analysis.
- The resulting sample size n is defined along with the number of permutation replications R. (While $R = 1000$ is not very large, the differences in the means here are so large we really don't need any finer precision.)
- An empty array named *RDiff* is created to store the *BPSysAve* differences produced from the permutations.
- Next, the *set.seed()* function is again used here, as before to ensure that the detailed results are able to be reproduced later, if ever needed. (Submissions should explicitly include this line. However, it will be sufficient to mention in the text that since the *set.seed()* function is used earlier in their .Rmd file it is possible to reproduce the output, if needed, even without adding an additional *set.seed()* function here.)
- A new tibble named *Rdt2* is created as a copy of the original *dt2* tibble. This copy is used so that the original is not overwritten with a permutation of the data.
- Next is a for-loop that iterates over the index variable r, taking on values from 1 to R. This produces the permutation sample for the test. For each r value:
  - the *BPSysAve* variable is permuted and replaces the previous *BPSysAve* values in the *Rdt2* tibble. Since only the values of this variable are permuted, any connection between the individual *BPSysAve* values and the *HomeOwn* variable values is broken;
  - the "Owners" and "Renters" group sample average *BPSysAve* values are calculated; and
  - the difference in the average of the *BPSysAve* measurements, for "Owners" minus "Renters", is stored in the r^{th} component of the *RDiff* array.

- Next, a new tibble is created from the *RDiff* array. (This is done to enable the use of the *ggplot2* package functions.)

A plot of the permutation sample distribution is produced, showing both a histogram overlaid with the kernel density estimate, and a vertical red line showing the actual observed difference in *ObsDiff*, 3.9986 (mm Hg). The outcome of the test is determined by the proportion of randomised samples that are *as or more extreme* than the observed difference. As this red line is completely removed from the entire permutation sample distribution, the corresponding p-value of the test must be *less than 1/1000*. (1/1000 is the value that would be used if the observed difference was, in absolute value, the largest value seen in the randomised sample. But since even this has not occurred, the true p-value must be even smaller!) This is certainly small enough to reject H_0, at the $\alpha = 5\%$ level, and conclude that the true average blood pressure measurement for "Owners" is not the same as that for "Renters".

```r
dt2 <- dt1 %>% filter(HomeOwn != "Other")
n <- nrow(dt2)  # completed
R <- 1000  # completed

RDiff <- array(dim = R)  # added
set.seed(2020.9)  # added

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]) %>%
        as.numeric()
}

# lines below all added
RDiff_tbl <- tibble(Diff = RDiff, r = 1:R)
Q2p3 <- RDiff_tbl %>% ggplot(aes(x = Diff, y = ..density..)) +
    geom_histogram(colour = "blue", fill = "blue", alpha = 0.2) +
    geom_density(colour = "blue", fill = "blue", alpha = 0.2) +
    geom_vline(xintercept = ObsDiff, colour = "red") + xlab(expression(paste("Potential average differe
    H[0]))) + ggtitle("Randomisation test: Mean BPSysAve for Owners vs Renters",
    "Observed difference shown by red line")
Q2p3
```



Question 2, part f

If cases are restricted to only include men aged between 35 and 44 (inclusive) the strength of evidence (i.e. the p-value for the test) changes. To implement the same test on the different data, we need to first filter the data to appropriately restrict the *Age* and *Gender* of the respondents. This is done in the code chunk below.

```
dt2f <- dt1 %>% filter(between(Age, 35, 44), Gender == "male")
dt2fS <- dt2f %>% group_by(HomeOwn) %>% summarise(mean = mean(BPSysAve))
dt2fS <- dt2fS %>% summarise(obsDiff = mean[1] - mean[2])
ObsDiff2 <- as.numeric(dt2fS)
ObsDiff2
```

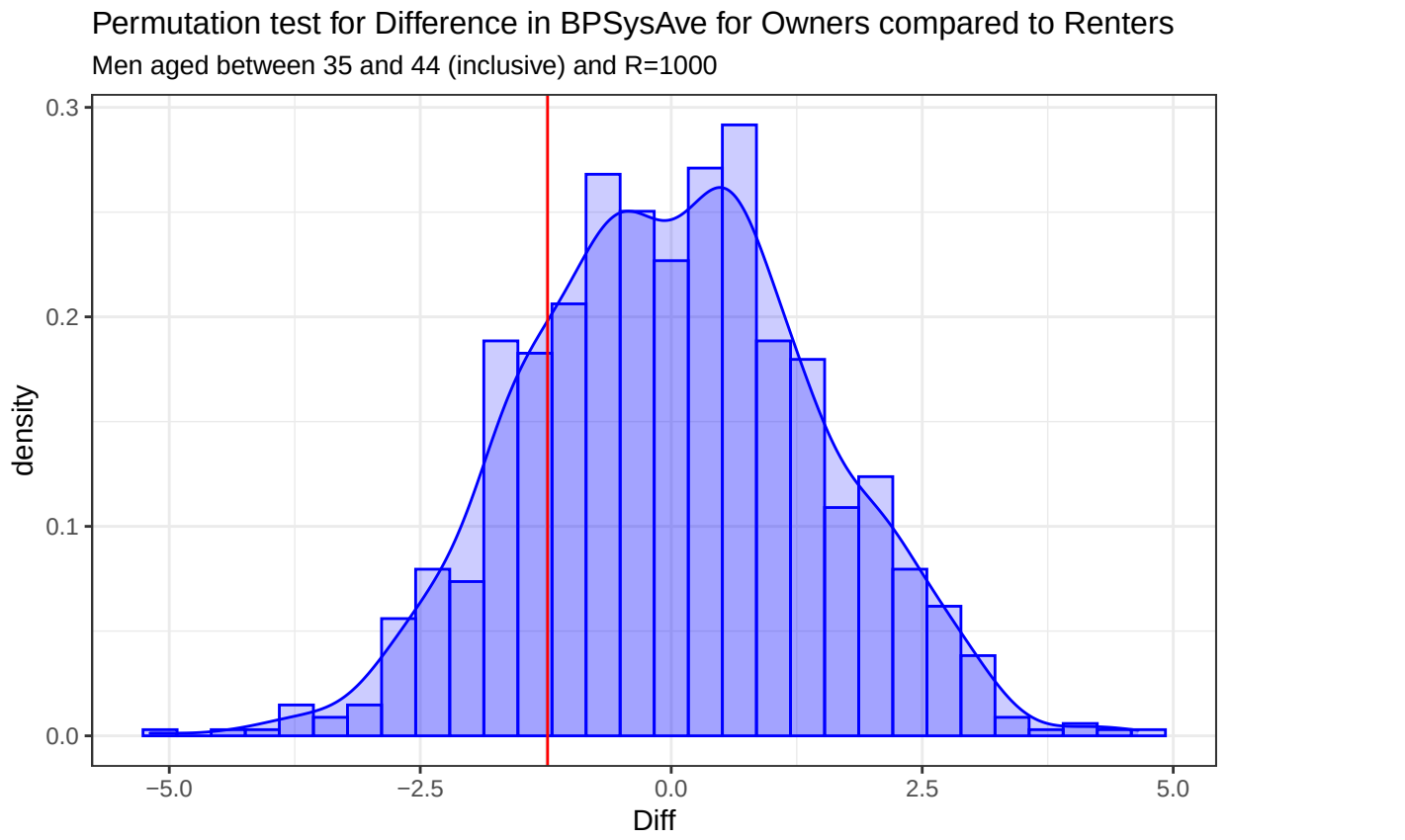
```
## [1] -1.232
```

Next we implement the same test, making note of the

```
n <- nrow(dt2f)
R <- 1000
RDiff2 <- array(dim = R)
set.seed(2020.9)
Rdt2f <- dt2f
for (r in 1:R) {
    Rdt2f <- Rdt2f %>% mutate(BPSysAve = sample(dt2f$BPSysAve,
        n, replace = FALSE))
```

```
    Rdt2fS <- Rdt2f %>% group_by(HomeOwn) %>% summarise(mean = mean(BPSysAve))
    RDiff2[r] <- Rdt2fS %>% summarise(Diff = mean[1] - mean[2]) %>%
        as.numeric()
}
```

```
RDiff_tbl2 <- tibble(Diff = RDiff2, r = 1:R)
Q2f <- RDiff_tbl2 %>% ggplot(aes(x = Diff, y = ..density..)) +
    geom_histogram(alpha = 0.2, colour = "blue", fill = "blue") +
    geom_density(alpha = 0.2, colour = "blue", fill = "blue") +
    geom_vline(xintercept = dt2fS$obsDiff, colour = "red") +
    theme_bw() + ggtitle("Permutation test for Difference in BPSysAve for Owners compared to Renters",
    "Men aged between 35 and 44 (inclusive) and R=1000") + theme(plot.title = element_text(size = 12),
    plot.subtitle = element_text(size = 10))
Q2f
```



The most important part here is the p-value calculation, which is not immediately known from the plot. We need to find the proportion of mean differences for the two groups (in the *RDiffs* object) from the permutations whose absolute value is larger than the absolute value of the observed *ObsDiff2* value. This is what we mean by 'as or more extreme than' in a two-sided test.

```
indc <- rep(0, R)
indc[abs(RDiff_tbl2$Diff) >= abs(dt2fS$obsDiff)] <- 1
pval <- mean(indc)
```

To compute the p-value: - A vector named *indc* (for "indicator") of length equal to R and containing all zero values is constructed. - Every element of *indc* for which the absoluted value of the permuted *RDiff2*

variable is greater than the absolute value of the observed *ObsDiff2* value is replaced with a "1". - The mean value of this *indc* vector represents the proportion of the R permuted samples where the difference in sample average *BPSysAve* measurements is "more extreme" (either on the positive side or on the negative side) than the observed difference in sample average *BPSysAve* measurements. Since this mean *indc* value is equal to 0.403, we cannot reject H_0, where

$$H_0 : \mu_{Owners} = \mu_{Renters} \quad \text{vs.} H_1 : \mu_{Owners} \neq \mu_{Renters},$$

where now μ_{Owners} and $\mu_{Renters}$ corresponds only to *Male* "Owners" and "Renters", respectively, aged between 35 and 44 years of age, inclusively. In this case, the "strength of evidence", or p-value, is much larger than the significance level $\alpha = 0.05\%$.

Why the difference in conclusions? Presumably the major differences between the two distributions of "Owners" and "Renters" that was apparent in parts a-e relate mainly to differences in blood pressure measurements between people of different genders and people of different ages. By restricting the data, we effectively control these aspects. What remains suggests that home ownership itself is not related to differences in blood pressure - something that might have been asserted earlier without careful consideration of the sampling population used in this study.