Homework 4

Shun-Lung Chang, Dilip Hiremath

```
library(ggplot2)
```

```
load('data/creditcard.Rdata')
```

1. From your previous analysis you know that HomeOwnership is an important predictor for getting a credit card approved. You hence limit your analysis in this homework to homeowning applicants only. Create a subset of the data just including the homeowners.

```
creditcard <- creditcard[creditcard$HomeOwner == 1, ]
```

(a) (half a point) For this subset, compute the median income.

```
median(creditcard$Income)
```

```
[1] 3.75
```

The median for this subset is 3.75.

(b) (half a point) For this subset, compute the standard deviation of income.

```
sd(creditcard$Income)
```

```
[1] 5.566508
```

The standard deviation is 5.566508.

(c) (half a point) How many applicants are in this subset?

```
nrow(creditcard)
```

```
[1] 295
```

There are 295 applicants in this subset.

(d) (half a point) How many of the applicants in the subset are married?

```
sum(creditcard$MaritalStatus == 1)
```

```
[1] 228
```

228 applicants in the subset are married.

(e) (half a point) How many applicants in the subset own a home?

All the applicants own a home in this subset of the data, hence the number is 295. Since this subset was created based on the condition that the applicant should own a home

2. On a typical working day, your team is able to process 120 credit card applications. To simulate this situation you draw a random sample of size 120 from the data subset generated in Question 1. (In order to make the results reproducible, use set.seed(201803) prior to drawing the sample.) Based on this sample of size 120, you want to test the null hypothesis that the mean income in the population is equal to 4750 USD. [hint: use the command t.test to perform a one-sample t-test to answer this question.]

```
set.seed(201803)
sample_creditcard <- creditcard[sample(nrow(creditcard), 120), ]
```

sample_creditcard has the random sample of 120 from the subset generated in question 1.

(a) (1 point) Based on the result obtained, do you conclude to reject the null hypothesis of the true population mean being equal to 4750 USD?

```
t.test(sample_creditcard$Income, mu = 4.75)
```

```
    One Sample t-test

data:  sample_creditcard$Income
t = 2.1488, df = 119, p-value = 0.03367
alternative hypothesis: true mean is not equal to 4.75
95 percent confidence interval:
 4.832375 6.765792
sample estimates:
mean of x
 5.799083
```

We reject the null hypothesis of the true population mean being equal to 4750 USD, since the p-value is less than 0.05.

(b) (half a point) How large is the test-statistic?

```
t.test(sample_creditcard$Income, mu = 4.75)$statistic
```

```
       t
2.148828
```

The test statistic is 2.148828.

(c) (half a point) How large is the corresponding p-value?

```
t.test(sample_creditcard$Income, mu = 4.75)$p.value
```

```
[1] 0.03367372
```

The corresponding p value is 0.03367372.

(d) (half a point) Does the 95%-confidence interval contain the score 4.75?

```
t.test(sample_creditcard$Income, mu = 4.75)$conf.int
```

```
[1] 4.832375 6.765792
attr(,"conf.level")
[1] 0.95
```

No, the 95%-confidence interval does not contain the score 4.75.

3. To check whether R actually computes the right thing, you decide to double check.

(a) (1 point) You first compute the mean and standard deviation of income in your sample and report these numbers.

```
mean(sample_creditcard$Income)
```

```
[1] 5.799083
```

The observed mean is 5.799083.

```
sd(sample_creditcard$Income)
```

```
[1] 5.348094
```

The observed standard deviation is 5.348.

(b) (half a point) Next you compute the standard error of the mean by dividing the standard deviation of your sample by the square root of the sample size.

```
sde_mean <- sd(sample_creditcard$Income) / sqrt(nrow(sample_creditcard))
sde_mean
```

```
[1] 0.488212
```

The standard error of the mean is observed to be 0.488212.

(c) (1 point) Finally, you compute the test statistic t which is the ratio of the difference between sample mean and hypothetical value and the standard error of the mean.

```
t_value <- (mean(sample_creditcard$Income) - 4.75) / sde_mean
t_value
```

```
[1] 2.148828
```

The test statistic t is observed to be 2.148828.

4. Now, you compare the empirical results with the corresponding theoretical distribution.

(a) (1 point) Compute the 2.5% quantile and the 97.5% quantile of the t-distribution with 119 degrees of freedom. Does the test statistic fall inside this range?

```
qt(c(0.025,0.975), df = 119)
```

```
[1] -1.9801  1.9801
```

The 2.5% quantile and the 97.5% quantile of the t-distribution with 119 degrees of freedom is -1.9801 and 1.9801 respectively. No, the test statistic does not fall inside this range.

(b) (1.5 points) Compute the probability that a random variable that follows a t-distribution with 119 degrees of freedom takes on values that are in absolute values larger than the observed test-statistic i.e. P ($T \geq |2.1488|$).

```
pt(-abs(t_value), df = 119) * 2
```

```
[1] 0.03367372
```

The probability that a random variable that follows a t-distribution with 119 degrees of freedom takes on values that are in absolute values larger than the observed test-statistic is 0.03367372.

5. Now, you simulate a full years work of your team, by drawing a total of 220 samples of size 120 from the income variable in the credit card data set.

```
sample_size <- 120
no_of_samples <- 220
sample <- matrix(0, nrow = sample_size, ncol = no_of_samples)
for (i in 1:no_of_samples) {
    index <- sample(length(creditcard$Income), size = sample_size,replace = FALSE)
    sample[, i] <- creditcard[index, "Income"]
}
```

(a) (1 point) Compute the median income for each sample. Report the median of the sample medians as well as the interquartile range of the sample medians.

```
sample_medians <- apply(sample, 2, median, na.rm = TRUE)
median(sample_medians)
```

```
[1] 4
```

Median of the sample medians is 4.

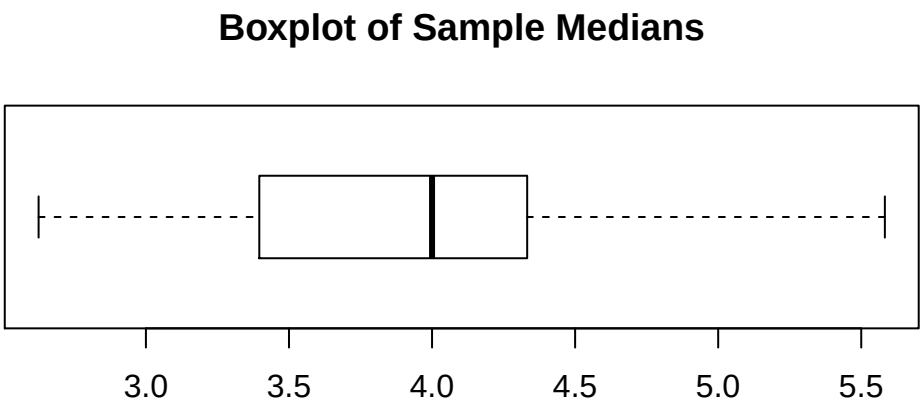
```
IQR(sample_medians)
```

```
[1] 0.925625
```

Interquartile range of the sample medians is 0.925625

(b) (1 point) Draw a boxplot of the sample medians. Based on this plot comment on the sampling distribution of the median income!

```
boxplot(sample_medians, horizontal = TRUE, main = 'Boxplot of Sample Medians')
```



The box plot is not symmetric since the difference between the first quantile and the median is larger than that between the third quantile and the median. Also there are many more outliers about 4.3 than below 3.2.

(c) (half a point) Compute the 0.025-quantile and the 0.975-quantile of your sampling distribution of the median income.

```
quantile(sample_medians, probs = c(0.025, 0.975))
```

```
 2.5% 97.5%
    3     5
```

The 0.025-quantile and the 0.975-quantile of our sampling distribution of the median income is 3 and 5 respectively.

6. Using the data obtained in Question 5 compute the following:

(a) (1 point) Compute the mean income for each sample. Report the mean of the sample means as well as the standard deviation of the sample means.

```
sample_means <- apply(sample, 2, mean, na.rm = TRUE)
mean(sample_means)
```

```
[1] 5.802627
```

The mean of the sample means is 5.802627.

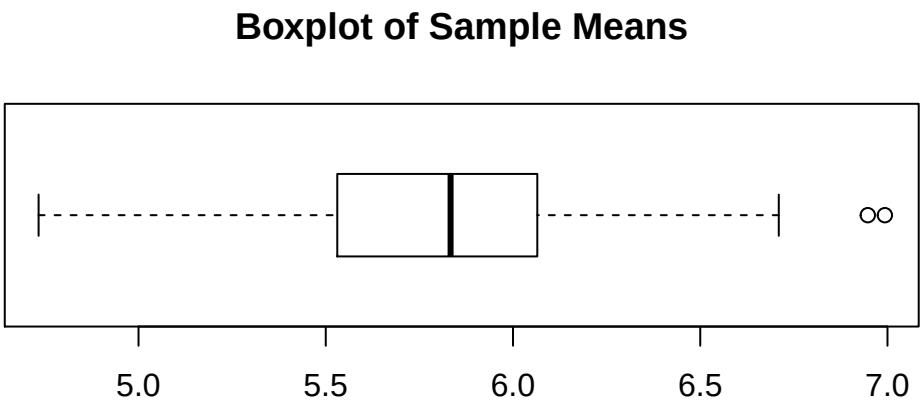
```
sd(sample_means)
```

```
[1] 0.3820323
```

The standard deviation of the sample means is 0.3820323.

(b) (1 point) Draw a boxplot of the sample means. Based on this plot comment on the sampling distribution of the mean income!

```
boxplot(sample_means, horizontal = TRUE, main = 'Boxplot of Sample Means')
```



The distribution is positively skewed. There are more observations below 5.5 than above 6.1. There are two outliers that are close to 7.0 value.

(c) (half a point) Compute the 0.025-quantile and the 0.975-quantile of your sampling distribution of the mean income.

```
quantile(sample_means, probs = c(0.025, 0.975))
```

```
    2.5%    97.5%
5.008343 6.456191
```

The 0.025-quantile and the 0.975-quantile of our sampling distribution of the median income is 5.008343 and 6.456191 respectively.

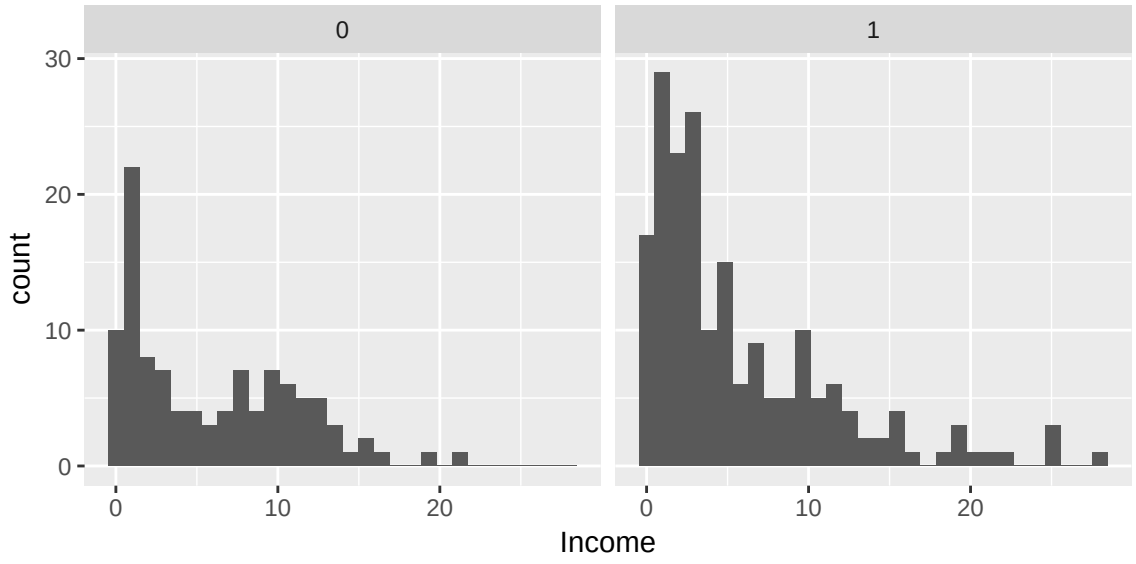
7. (2.5 points) Coming back to your subset data of homeowners (see Question 1), you want to investigate if there is a gender bias in income. For that, briefly describe in plain English the distributions of income, separately for males and females (variable Gender). Use relevant numerical summaries as well as one graphical representation for each of the two distributions.

```
by(creditcard$Income, creditcard$Gender, summary,
   quantiles = c(0,.25,.5,.75,1))
```

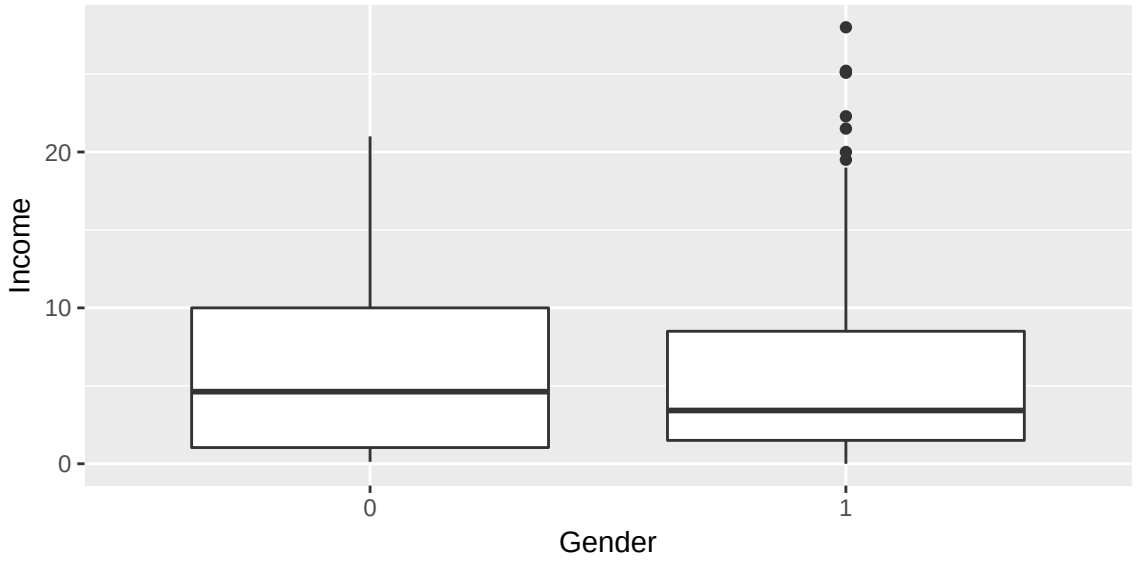
```
creditcard$Gender: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.125   1.040   4.625   5.914  10.000  21.000
------------------------------------------------------------
creditcard$Gender: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.500   3.417   5.683   8.500  28.000
```

We observe that the maximal income for male is 28.000 while that for a female is 21. ALso the minimum value for male is 0.00 while that of female is 0.125. The mean income for female home owners is higher than that for male home owners.

```
ggplot(creditcard) +
    geom_histogram(aes(x = Income)) +
    facet_grid(~ Gender)
```




```
ggplot(creditcard) +
    geom_boxplot(aes(x = Gender, y = Income))
```



As observed from the above representation we see that there are about 6 outliers for male home owners while there are not any outliers for female homeowners. Also the spread of income is more even in females when compared to males. Observations for males are larger in number than for females. The distribution for males is positively skewed. Interquartile range is smaller in males as shown in the boxplot.

8. Again using the subset data for homeowners, you want to see whether the difference in means is large in comparison to the spread of the data.

(a) (1 points) Calculate the means ($\bar{x}_{inc.f}$, $\bar{x}_{inc.m}$) and the standard deviations ($s_{inc.f}$, $s_{inc.m}$) of income separately. Now calculate the test-statistic of the independent-samples t-test.

```
inc_x_bar <- tapply(creditcard$Income, creditcard$Gender,
                    mean, na.rm = TRUE)
inc_sd <- tapply(creditcard$Income, creditcard$Gender,
                    sd, na.rm = TRUE)
inc_x_bar
```

```
       0        1
5.913714 5.683105
```

```
inc_sd
```

```
       0        1
5.101051 5.819205
```

The means of income for female $\bar{x}_{inc.f}$ is 5.913714 and means of income for male $\bar{x}_{inc.m}$ is 5.683105.

the standard deviation for female income is $s_{inc.f}$ is 5.101051 and standard deviation for male income is $s_{inc.m}$ is 5.819205.

```
n1 <- nrow(creditcard[creditcard$Gender == 0 & !is.na(creditcard$Income), ])

n2 <- nrow(creditcard[creditcard$Gender == 1 & !is.na(creditcard$Income), ])

t_sd <- ((n1 - 1) * inc_sd[1] ^ 2 + (n2 - 1) * inc_sd[2] ^ 2) / (n1 + n2 - 2)

t_value <- (inc_x_bar[1] - inc_x_bar[2]) /
    sqrt(t_sd * (1 / n1 + 1 / n2))
t_value <- unname(t_value)
t_value
```

```
[1] 0.340173
```

Test statistic of the independent-samples t-test is 0.340173.

(b) (1 point) Using a t-distribution with n1 + n2 - 2 degrees of freedom, calculate the probability of a t-distributed random variable being larger than or equal the above calculated t-statistic score.

```
p_larger <- pt(abs(t_value), df = n1 + n2 - 2, lower.tail = FALSE)
p_larger
```

```
[1] 0.366985
```

The probability of a t-distributed random variable being larger than or equal the above calculated t-statistic score of 0.340173 using a t-distribution with 293 degrees of freedom is 0.366985.

(c) (half a point) Based on the results so far, compute the probability under the null hypothesis to obtain a result for the test statistic that is as extreme as the one we have obtained.

```
2 * pt(abs(t_value), df = n1 + n2 - 2, lower.tail = FALSE)
```

```
[1] 0.73397
```

Probability under the null hypothesis to obtain a result for the test statistic that is as extreme as the one we have obtained is 0.73397.

9. (2.5 points) Use the function t-test to check with an independent samples t-test whether income significantly differs between males and females in your subset of homeowners. Assume equal variances for the two groups. State the statistical null hypothesis to be tested as well as the alternative hypothesis.

Take a look at the output and compare it with your results above.

```
f_income <- creditcard[creditcard$Gender == 0 & !is.na(creditcard$Income), 'Income']
m_income <- creditcard[creditcard$Gender == 1 & !is.na(creditcard$Income), 'Income']

t.test(f_income, m_income, mu = 0, var.equal = TRUE)
```

```
    Two Sample t-test

data:  f_income and m_income
t = 0.34017, df = 293, p-value = 0.734
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.103595  1.564813
sample estimates:
mean of x mean of y
 5.913714  5.683105
```

The null hypothesis to be tested is if the difference between mean of income between male and female is equal to 0. Alternatiev hypothesis is the difference in mean is not equal to 0. Since the P value is greater than the 0.05 we cannot reject the null hypothesis in favour of the alternative hypothesis. We also see that the t-statistics value is similar to what we had calculated individually.

10. (2.5 points) Visualise the previous results. Draw a plot for the pdf of the t-distribution with the adequate number of degrees of freedom for the test statistic t. Color the areas under the pdf for all values smaller than t and larger than t.

The following plot shows the area where values are smaller than t and larger than t.

```
d <- data.frame(xt = seq(-5, 5, by = 0.01))
d$dt <- dt(d$xt, df = n1 + n2 - 2)

ggplot(d) +
    geom_path(aes(xt, dt)) +
    geom_linerange(data = d[d$xt > qt(p_larger, n1 + n2 - 2, lower.tail = FALSE), ],
                   aes(xt, ymin = 0, ymax = dt),
                   color = "red") +
```

```
    geom_linerange(data = d[d$xt < qt(p_larger, n1 + n2 - 2), ],
                   aes(xt, ymin = 0, ymax = dt),
                   color = "red") +
    theme_bw()
```