Homework 6

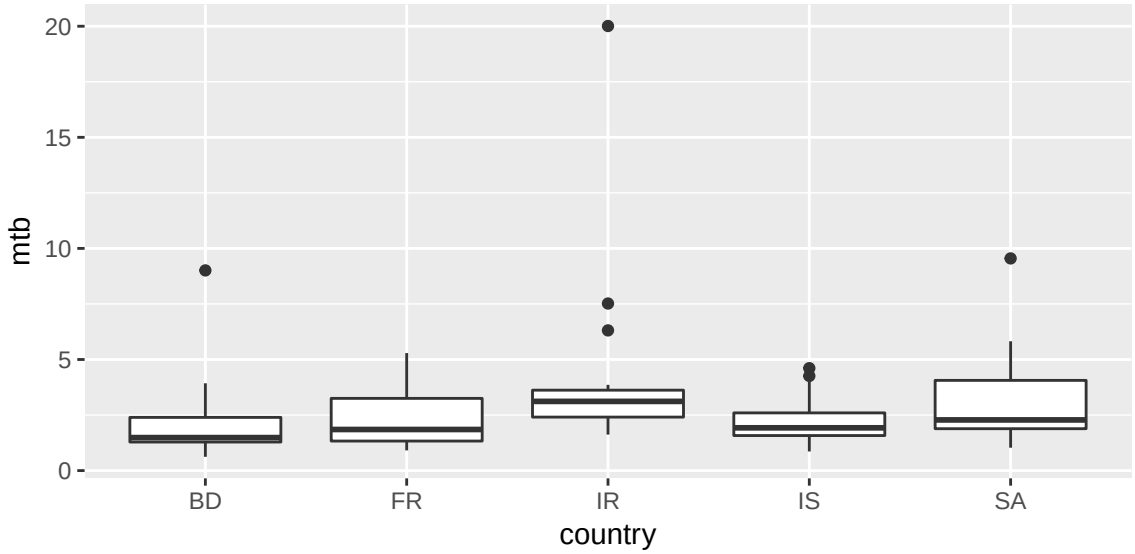
Shun-Lung Chang, Dilip Hiremath

```
library(ggplot2)
library(dplyr)
```

```
# load data
companies <- read.delim("data/Companies.txt")
```

1. (2.5 points) Investigate whether there are any differences between countries in company market to book value (variable mtb). Draw a boxplot by group to visualise the distributions of company market to book value by country. Comment on the plot.

```
ggplot(companies) +
    geom_boxplot(aes(x = country, y = mtb))
```



South Africa and IR comparitively have higher company market to book value. However they are just margianlly larger. IR has the 3 outliers, of which one is quite distinctively away positively. The spread of Germany and South Africa are almost similar.

2. You now investigate the means and medians of the market to book value in more detail.

(a) (1 point) Compute for each country the means and medians of companies' market to book value.

```
m_mtb <- companies %>%
    group_by(country) %>%
    summarise(mean_mtb = mean(mtb, na.rm = TRUE),
```

```
              median_mtb = median(mtb, na.rm = TRUE))
m_mtb
```

```
# A tibble: 5 x 3
  country mean_mtb median_mtb
   <fctr>    <dbl>      <dbl>
1      BD 2.067667      1.485
2      FR 2.379459      1.850
3      IR 4.207778      3.115
4      IS 2.223750      1.925
5      SA 2.986538      2.280
```

The table above shows the means and medians of companies' market to book value in different coutries.

(b) (1.5 points) Compute the ratio between mean market to book value and median market to book value for each country. Which country has the largest ratio, which one the smallest?

```
m_mtb$ratio <- m_mtb$mean_mtb / m_mtb$median_mtb
m_mtb
```

```
# A tibble: 5 x 4
  country mean_mtb median_mtb    ratio
   <fctr>    <dbl>      <dbl>    <dbl>
1      BD 2.067667      1.485 1.392368
2      FR 2.379459      1.850 1.286194
3      IR 4.207778      3.115 1.350811
4      IS 2.223750      1.925 1.155195
5      SA 2.986538      2.280 1.309885
```

Germany has the largest ratio, Israel has the smallest one.

3. Next, you turn to the variance of the market to book value.

(a) (1.5 points) Compute the variance of market to book value for each country. How different are the variances of market to value between the countries?

```
v_mtb <- companies %>%
    group_by(country) %>%
    summarise(var_mtb = mean(mtb, na.rm = TRUE))
v_mtb
```

```
# A tibble: 5 x 2
  country  var_mtb
   <fctr>    <dbl>
1      BD 2.067667
2      FR 2.379459
3      IR 4.207778
4      IS 2.223750
5      SA 2.986538
```

Ireland has the highest variance of 4.207778. All the other countries have relative similarly variances.

(b) (1 point) Compute the ratio between the largest and the lowest variance.

```
max(v_mtb$var_mtb) / min(v_mtb$var_mtb)
```

```
[1] 2.035037
```

The ratio between the largest and the lowest variance is 2.035037.

4. (2.5 points) Compute an ANOVA test and check whether the mean market to book value of companies is the same for all countries! Provide an answer in plain English, reporting all the relevant statistical numbers.

```
mtb_aov <- aov(mtb ~ country, data = companies)
anova(mtb_aov)
```

```
Analysis of Variance Table

Response: mtb
           Df Sum Sq Mean Sq F value   Pr(>F)
country     4  63.86 15.9651  3.8693 0.005286
Residuals 130 536.39  4.1261
```

THe p-value is significantly small and hence the null hypothesis can be rejected in favour of the alternate hypothesis that the the mtb varies are not all the same across country.

5. (2.5 points) Use Tukey's HSD post-hoc test to find out which countries actually differ in mean companies' market to book value.

```
TukeyHSD(mtb_aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mtb ~ country, data = companies)

$country
            diff        lwr        upr     p adj
FR-BD  0.3117928 -1.0688011  1.6923867 0.9708603
IR-BD  2.1401111  0.4647293  3.8154929 0.0050675
IS-BD  0.1560833 -1.3828531  1.6950198 0.9986271
SA-BD  0.9188718 -0.5868224  2.4245659 0.4447728
IR-FR  1.8283183  0.2134605  3.4431761 0.0179608
IS-FR -0.1557095 -1.6285254  1.3171065 0.9983843
SA-FR  0.6070790 -0.8309670  2.0451250 0.7696094
IS-IR -1.9840278 -3.7361840 -0.2318716 0.0179387
SA-IR -1.2212393 -2.9442718  0.5017932 0.2911220
SA-IS  0.7627885 -0.8278913  2.3534682 0.6751901
```

From the TurkeyHSD post hoc test we see that countries pair Israel-BD, Israel-France and Israel-IR differ given the small p-values.

6. (2.5 points) According to the significant differences between the countries mean companies' market to book value, how many clusters do the countries form? Which country belongs to which cluster?

According the table in task 5, Ireland differs relatively significantly from the other coutries, while the differences among the other four countries are not statistically significant. Thus, we have two clusters:

- (France, Germany, Isreal, South Africa)
- (Ireland)

7. (2.5 points) Due to the skewness of the data and given the small sample sizes, your supervisor insists that you double check your results with a non-parametric alternative to the ANOVA test. Hence, you apply the Kruskal-Wallis test to your data. Perform the analysis and report your results.

```
mtb_kw <- kruskal.test(mtb ~ country, data = companies)
mtb_kw
```

```
    Kruskal-Wallis rank sum test

data:  mtb by country
Kruskal-Wallis chi-squared = 17.47, df = 4, p-value = 0.001566
```

The relatively low p-value in the test results indicates we could conclude that the averages of market to book value are different across countries.

8. (2.5 points) As post hoc test you now use the pairwise Wilcoxon rank sum test with Bonferroni correction. Which countries differ in mean companies market to book value?

```
pairwise.wilcox.test(companies$mtb, companies$country, p.adjust.method = "bonf")
```

```
    Pairwise comparisons using Wilcoxon rank sum test

data:  companies$mtb and companies$country

   BD      FR      IR      IS
FR 1.00000 -       -       -
IR 0.00083 0.15074 -       -
IS 1.00000 1.00000 0.04595 -
SA 0.06710 1.00000 1.00000 1.00000

P value adjustment method: bonferroni
```

As the result indicate, Ireland differs significantly in mean companies market to book value.

9. (2.5 points) According to the significant differences between the countries mean companies' market to book value in the pairwise Wilcoxon test, how many clusters do the countries form? Which country belongs to which cluster?

According the p-values in task 8, we can form two clusters:

- (France, Germany, Isreal)
- (Ireland, South Africa)

But, there also exist no significant difference between France and South Africa. Therefore, we would say these two clusters are not far away from each other in mean companies' market to book value.

10. Now, you look at companies' dividends for two sectors, namely automobiles and chemicals.

```
sub_comp <- companies %>%
    filter(sector %in% c('AUTOMOBILES', 'CHEMICALS'))
```

(a) (1.5 points) Compare the mean dividend for companies in these two sectors using the appropriate t-test. Summarize your analysis.

```
var.test(dividend ~ sector, data = sub_comp)
```

```
    F test to compare two variances

data:  dividend by sector
F = 0.76532, num df = 6, denom df = 7, p-value = 0.7603
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1495184 4.3588830
sample estimates:
ratio of variances
         0.7653245
```

```
t.test(dividend ~ sector, data = sub_comp, var.equal = TRUE)
```

```
    Two Sample t-test

data:  dividend by sector
t = 0.13036, df = 13, p-value = 0.8983
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.459875  1.647375
sample estimates:
mean in group AUTOMOBILES   mean in group CHEMICALS
                  2.64000                   2.54625
```

We first investigate the variances of the two sectors, and the high p-value shows that the variances of the two sectors are not significantly different. Given the same variances, we conduct a t-test to assess whether the mean dividend differs in two sectors. The high p-value suggests that the mean dividend does not differ significantly in two sectors.

(b) (1 point) To double-check, you also run a non-parametric alternative to the t-test.

```
wilcox.test(dividend ~ sector, exact = FALSE, data = sub_comp)
```

```
    Wilcoxon rank sum test with continuity correction

data:  dividend by sector
W = 32, p-value = 0.6854
alternative hypothesis: true location shift is not equal to 0
```

The high p-value in the Wilcoxon signed-rank test indicates that the mean dividend does not differ significantly in two sectors.

11. You now want to generate a bootstrap confidence interval for the difference in means in dividend between companies in the automobiles sector and companies in the chemicals sector.

(a) (1 point) In a first step you draw 10000 samples of size 10 from the observed dividends for companies in the two sectors. Draw with replacement and use set.seed(20180323)! Report the summary statistics of your differences in the sample means in companies' dividend between the two sectors.

```
set.seed(20180323)
sample_mean_auto <- sapply(1:10000,
                           FUN = function(x) {
                               mean(sample(sub_comp[sub_comp$sector == 'AUTOMOBILES',
                                                    'dividend'], 10, replace = TRUE))})

sample_mean_chem <- sapply(1:10000,
                           FUN = function(x) {
                               mean(sample(sub_comp[sub_comp$sector == 'CHEMICALS',
                                                    'dividend'], 10, replace = TRUE))})
```

```
summary(sample_mean_auto)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.328   2.386   2.648   2.644   2.893   3.896
```

```
summary(sample_mean_chem)
```

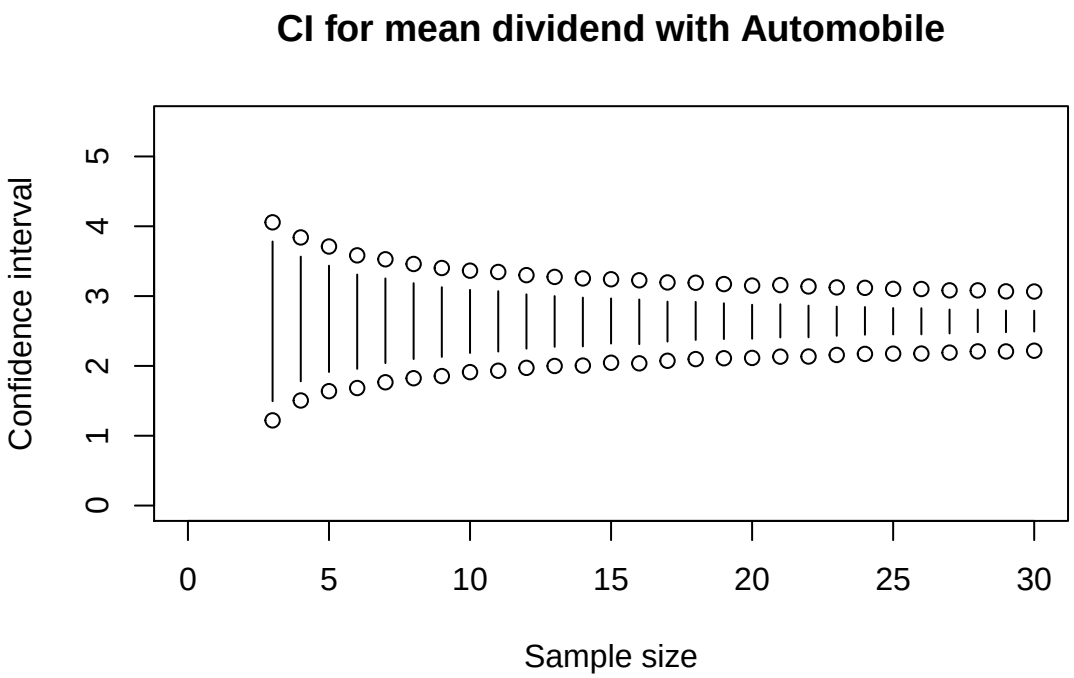
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.121   2.239   2.523   2.550   2.826   4.443
```

Statistics in automibile sector are all greater that that in chemical sector expect the maximal value.

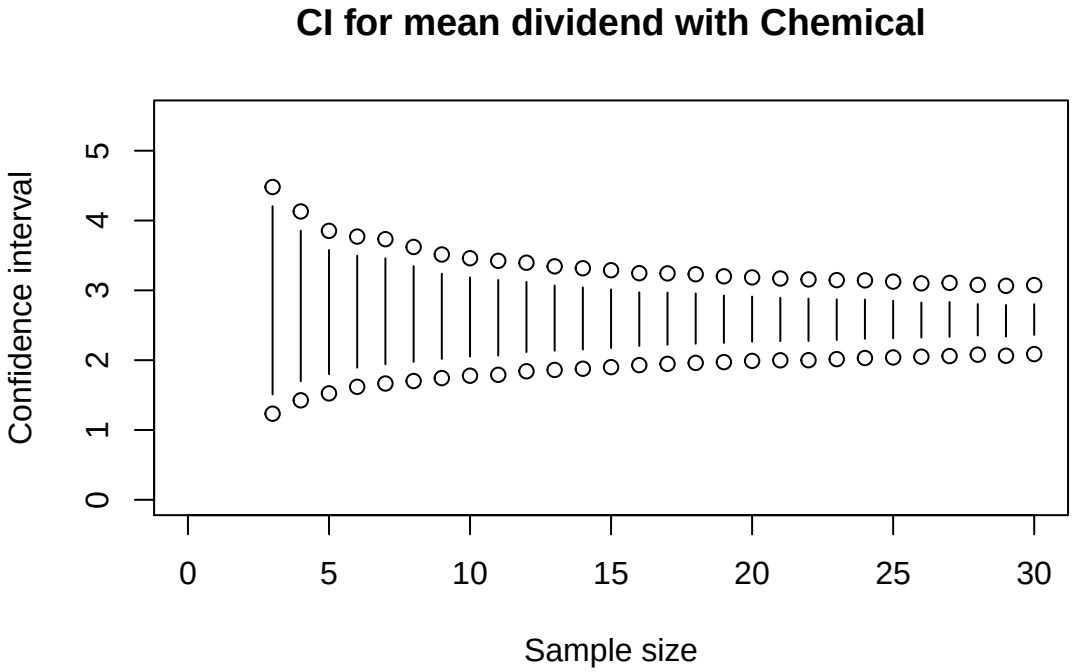
(b) (1.5 points) Next, generate 10000 samples for each sample size k, where k varies from 3 to 30 in steps of 1. Draw with replacement and use set.seed(20180323)! Plot the resulting 95% bootstrap confidence intervals for the differences in the sample means in companies' dividend between the two sectors against the sample size k.

The following two plots show the bootstrap confidence intervals with different sample sizes in two sectors.

```r
set.seed(20180323)
plot(c(0, 30), c(0, 5.5), type = "n", xlab = "Sample size", ylab = "Confidence interval",
     main = "CI for mean dividend with Automobile")
for (k in 3:30) {
    sample_mean <- sapply(1:10000,
                          FUN = function(x) {
                              mean(sample(sub_comp[sub_comp$sector == 'AUTOMOBILES',
                    'dividend'], k, replace = TRUE))})
    points(c(k, k), quantile(sample_mean, c(0.025, 0.975)), type = "b")
}
```




```r
set.seed(20180323)
plot(c(0, 30), c(0, 5.5), type = "n", xlab = "Sample size", ylab = "Confidence interval",
     main = "CI for mean dividend with Chemical")
for (k in 3:30) {
    sample_mean <- sapply(1:10000,
                          FUN = function(x) {
                              mean(sample(sub_comp[sub_comp$sector == 'CHEMICALS',
                    'dividend'], k, replace = TRUE))})
    points(c(k, k), quantile(sample_mean, c(0.025, 0.975)), type = "b")
}
```


CI for mean dividend with Chemical
Confidence interval
0
1
2
3
4
5
0
5
10
15
20
25
30
Sample size