# Homework 2

*Shun-Lung Chang, Dilip Hiremath*

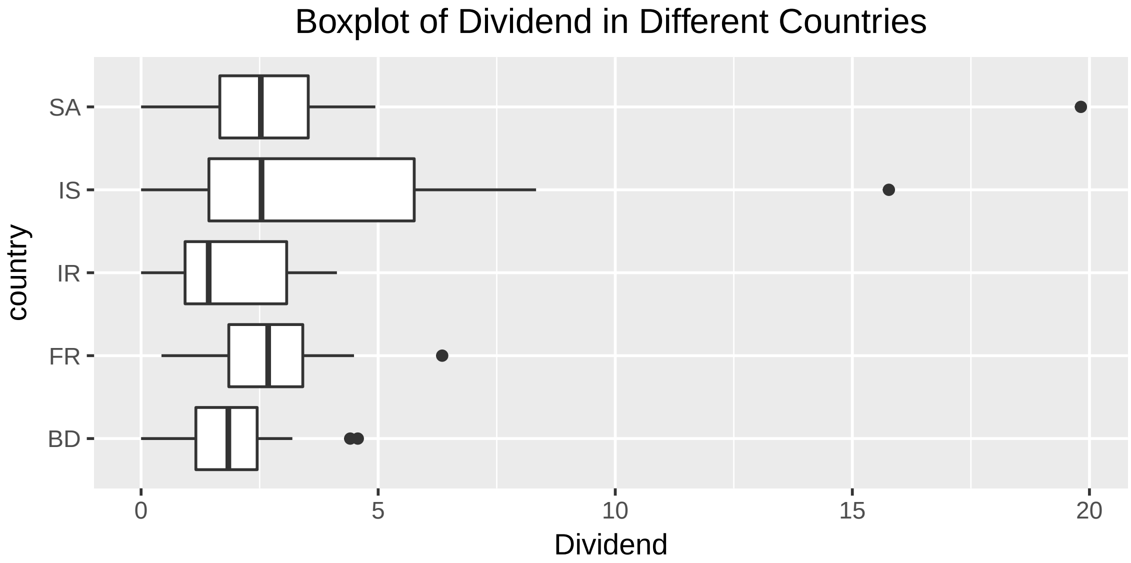
```r
# import required packages
library(magrittr) # pipeline operator
library(dplyr) # data manipulation
library(tidyr) # data manipulation
library(ggplot2) # plot
```

```r
# read data
companies <- read.table("data/Companies.txt",
                        sep = "\t", header = TRUE, stringsAsFactors = FALSE)
```

### 1. (2.5 points) Using a suitable graphical display, investigate whether there are any differences between countries in company dividends. In particular, look at central tendency and spread of the company dividend and provide a short summary of the distributional shape of company dividends for each country as well as the presence of outliers.

The boxplot below shows the dividend among different coutries. All median dividend lie in between 1.4 and 2.6 percent, and the companies in France have the highest dividend. Also, the spread of dividend (measured by interquatile range) for Israel is the highest. The distributions for South Africa, France, and Germany are quite symmetric. Yet, the distributions for Israel and Ireland are right skewed. Lastly, there exist outliers in all countries, except for Ireland. Particularly, the outliers in South Africa and Israel are even higher than 15%.

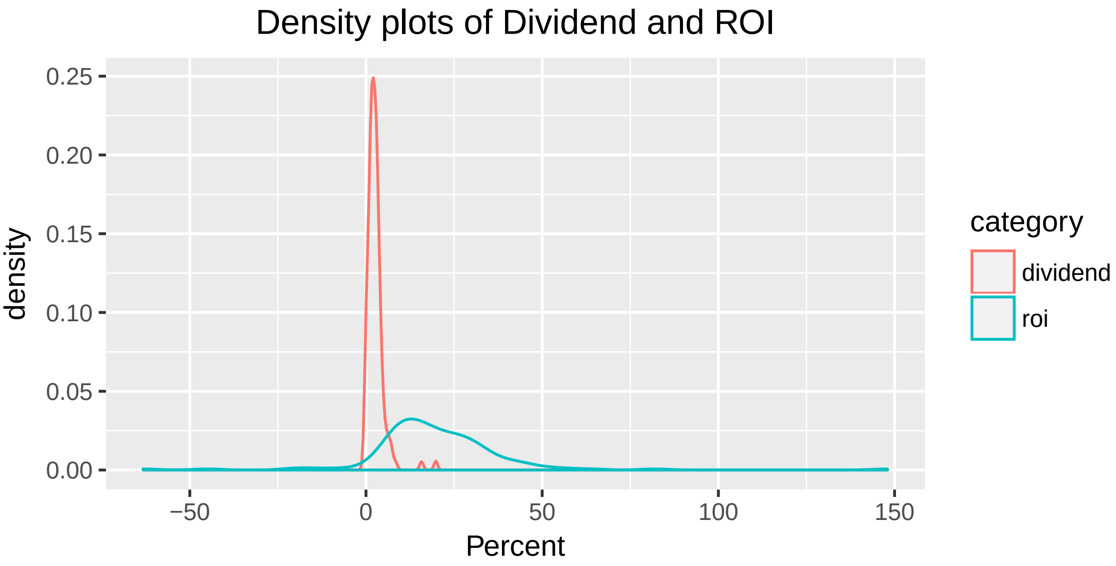
```r
ggplot(data = companies) +
    geom_boxplot(aes(x = country, y = dividend), na.rm = TRUE) +
    coord_flip() +
    labs(title = "Boxplot of Dividend in Different Countries", y = "Dividend") +
    theme(plot.title = element_text(hjust = 0.5))
```



### 2. (2.5 points) Draw density plots for the variables dividend and roi. Put both plots together in one figure for easy comparison. Point out the interesting features of these distributions.

The density of dividend concentrates between 0 to 5 percent. Yet the density of roi mostly ranges from 0 to 50 percent, showing a wider spread.

```r
companies %>%
    select(dividend, roi) %>%
    gather(category) %>%
    ggplot() +
        geom_density(aes(x = value, color = category)) +
        labs(title = "Density plots of Dividend and ROI", x = "Percent") +
    theme(plot.title = element_text(hjust = 0.5))
```



### 3. You aim at comparing the empirical distribution for return on investment with a theoretical counterpart.

#### (a) (half a point) Compute mean and standard deviation for the variable roi and report them.

The mean and standard deviation for *roi* can be obtained by the following code. The mean of *roi* is 19.79 and the standard deviation of *roi* is 20.03

```r
mean(companies$roi, na.rm = TRUE)
```

```
[1] 19.79081
```

```r
sd(companies$roi, na.rm = TRUE)
```

```
[1] 20.02681
```

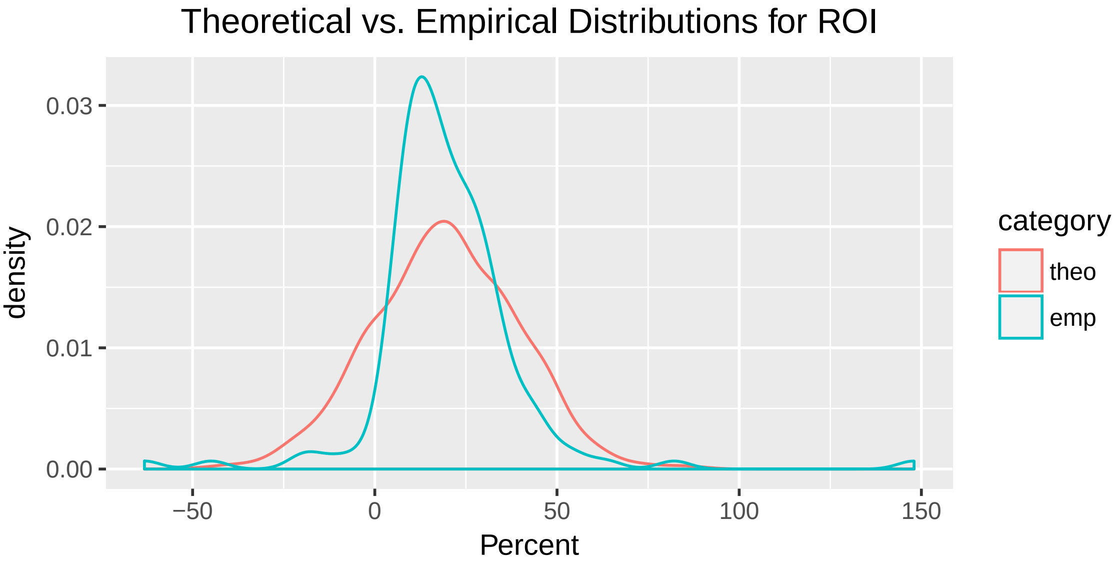
**(b) (1 point) Use these statistics as parameters for a normal distribution. Plot the probability density function for this normal distribution and compare its shape with the shape of the corresponding density plot obtained in Question 2. Use a suitable range for the scores of this variable.**

As can be seen from the plot below, the empirical distribution for roi concentrates more in the center than the theoretical normal distribution (created by random number generator).

```
set.seed(42)

normal <- rnorm(n = 1000, mean(companies$roi, na.rm = TRUE),
                sd(companies$roi, na.rm = TRUE))
df_roi <- rbind(data.frame(category = "theo", value = normal),
      data.frame(category = "emp", value = companies$roi))

ggplot(data = df_roi) +
    geom_density(aes(x = value, color = category))  +
    labs(title = "Theoretical vs. Empirical Distributions for ROI", x = "Percent") +
    theme(plot.title = element_text(hjust = 0.5))
```



**(c) (1 point) Do you find this normal distribution suitable for modelling the corresponding empirical distributions? Why or why not?**

Given the theoretical normal distribution may fail to correctly depict the central tendancy of the empirical distribution, we would not suggest use this theoretical distribution.

## 4. You aim at comparing the empirical distribution for dividend with a theoretical counterpart.

The mean and standard deviation for *dividend* can be obtained by the following code. The mean of *dividend* is 2.649 and the standard deviation of *roi* is 2.46

**(a) (half a point) Compute mean and standard deviation for the variable dividend and report them.**

```
mean(companies$dividend, na.rm = TRUE)
```

```
[1] 2.648841
```

```
sd(companies$dividend, na.rm = TRUE)
```

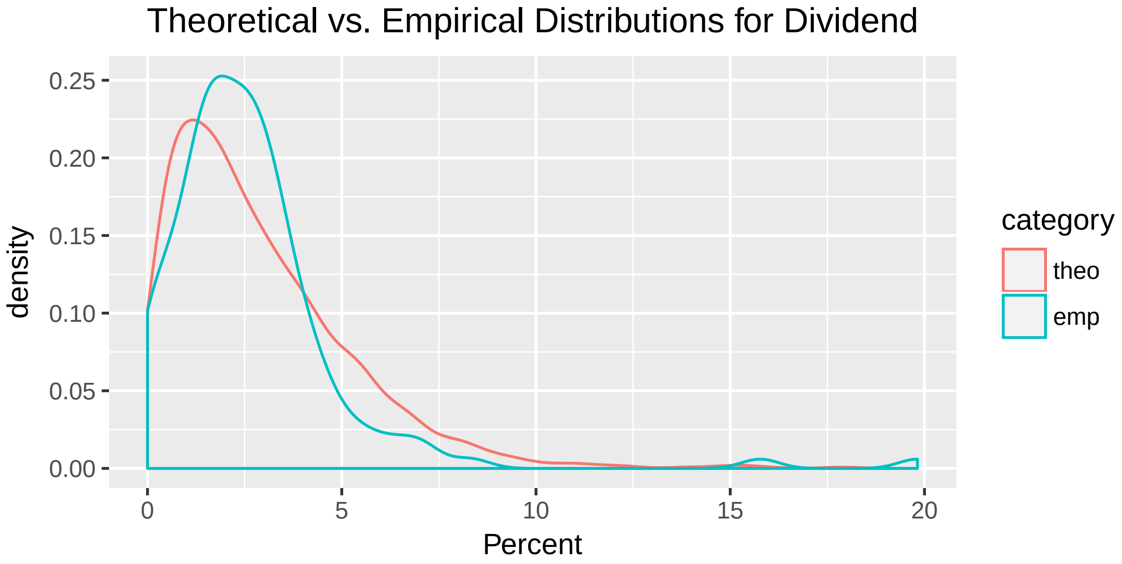
```
[1] 2.459952
```

**(b) (1.5 points) Due to the skewness of the empirical distribution you select a $\chi^2$-distribution to model the data. The $\chi^2$-distribution has one parameter, called degree of freedoms, $\nu$. An integer number which is identical to the theoretical mean of the distribution. Round the computed mean to the nearest integer and use this as parameter for your $\chi^2$- distribution. Plot the probability density function for this $\chi^2$-distribution and compare its shape with the shape of the corresponding density plot obtained in Question 2. Use a suitable range for the scores of this variable.**

As can be seen from the plot, the main difference of the two distributions is that the numbers of values lower or greater than 5. That is, there exist more values lower than 5 in the empirical density, and more values higher than 5 in the theoretical one.

```
set.seed(42)
chisq <- rchisq(1000, df = round(mean(companies$dividend, na.rm = TRUE)))

df_dividend <- rbind(data.frame(category = "theo", value = chisq),
      data.frame(category = "emp", value = companies$dividend))
ggplot(data = df_dividend) +
    geom_density(aes(x = value, color = category)) +
    labs(title = "Theoretical vs. Empirical Distributions for Dividend", x = "Percent") +
    theme(plot.title = element_text(hjust = 0.5))
```



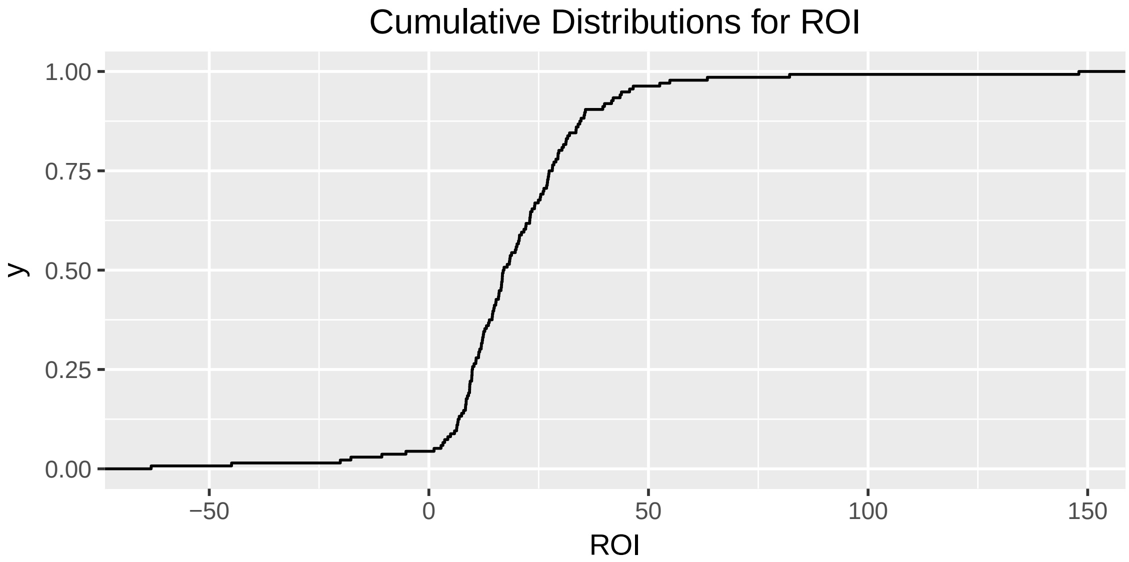
**(c) (half a point) Do you find this $\chi^2$-distribution suitable for modelling the corresponding empirical distributions? Why or why not?**

We may not suggest use this distribution, since as we pointed out in the previous question, values in *dividends* mostly lie between 0 to 5 percent. And the theoretical distribution would not be able to precisely represent the central tendancy.

### 5. (2.5 points) Draw a plot of the empirical cumulative distribution function (ECDF) for the variable roi. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the ECDF plot?

The probability slowly accumulates at the beginning and the end, and sharply increases between 0 and 50 percent, since the most values in *roi* are between 0 and 50 percent. In addition, since it is a plot of cumulative distribution function, the value never drops.

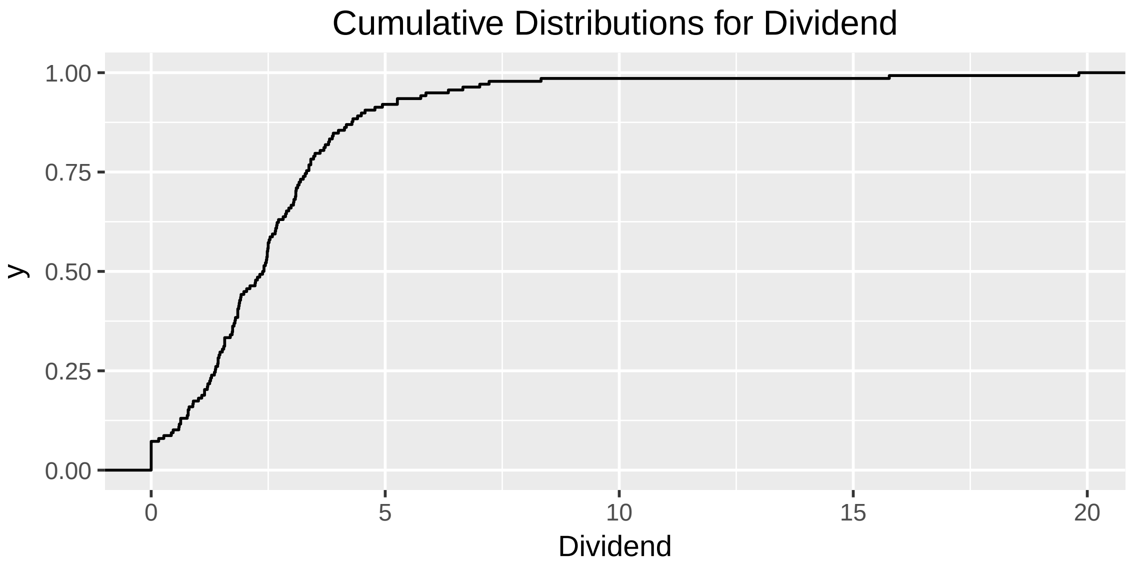
```r
ggplot(data = companies) +
    stat_ecdf(aes(x = roi)) +
    labs(title = "Cumulative Distributions for ROI", x = "ROI") +
    theme(plot.title = element_text(hjust = 0.5))
```



### 6. (2.5 points) Draw a plot of the empirical cumulative distribution function (ECDF) for the variable dividend. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the ECDF plot?

The values between 0 to 5 percent contribute mostly to the rise and again, the value never drops in a cumulative frequency plot.

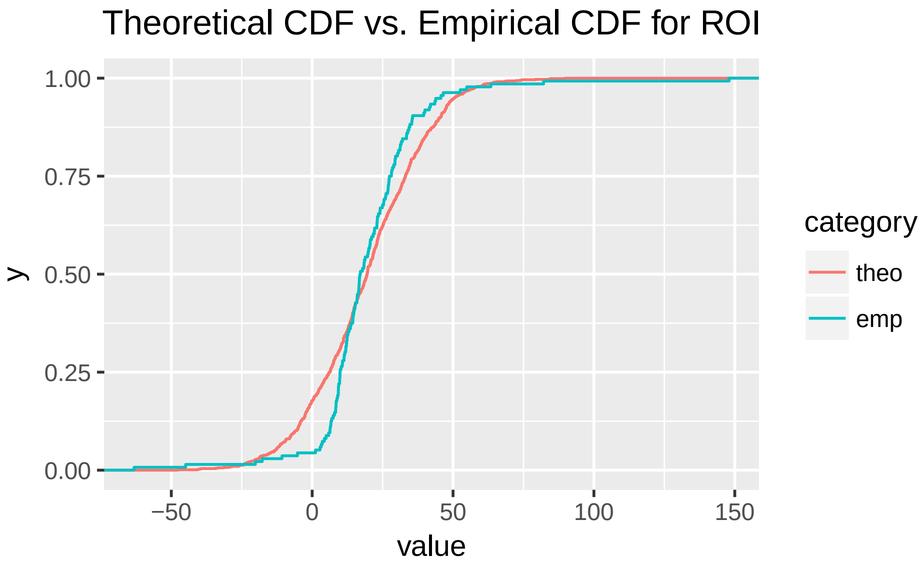
```r
ggplot(data = companies) +
    stat_ecdf(aes(x = dividend)) +
    labs(title = "Cumulative Distributions for Dividend", x = "Dividend") +
    theme(plot.title = element_text(hjust = 0.5))
```



**7. (2.5 points) Now, draw the corresponding cumulative distribution functions (CDF) for the densities derived in Question 3 and Question 4. Which of the interesting features of the distribution that you spotted in the density plot above, do you also find here? Are there any additional features that you spot in the CDF plot? Compare the CDFs with the corresponding ECDFs!**

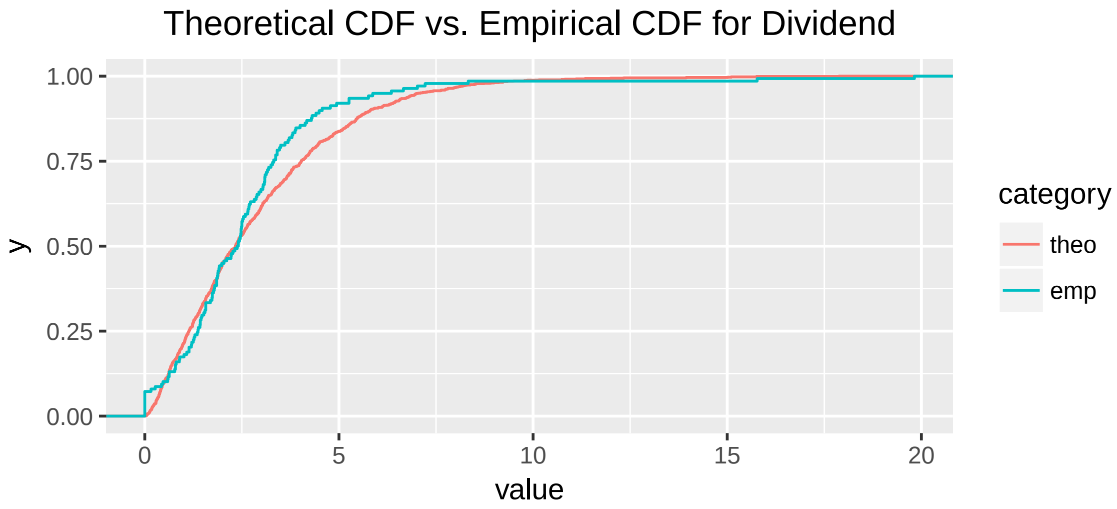
As we pointed out in question 3, the empirical density for *roi* increases mainly in the interval between 0 and 50, while the sharp rise in the theoretical one lies in a larger interval, which is around -25 to 75 percent.

```r
ggplot(df_roi) +
    stat_ecdf(aes(x = value, color = category)) +
    ggtitle("Theoretical CDF vs. Empirical CDF for ROI") +
    theme(plot.title = element_text(hjust = 0.5))
```



The empirical density for `dividend` increases more between 0 and 5 than the theoretical one, and the reason has already been pointed out in question 4.

```r
ggplot(df_dividend) +
    stat_ecdf(aes(x = value, color = category)) +
    ggtitle("Theoretical CDF vs. Empirical CDF for Dividend") +
    theme(plot.title = element_text(hjust = 0.5))
```



## 8. Now, you investigate whether there is any relationship between volatility and price change of a company's stock.

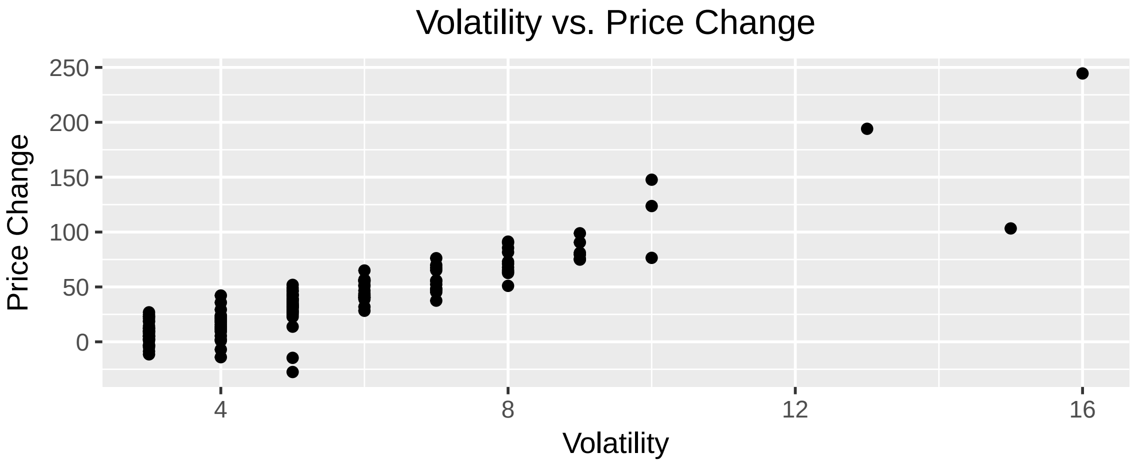
### (a) (1 point) Would you expect any relationship? If yes, which one? If not, why not?

We would assume a positive relationship between these two variables, since high volatility usually implies a higher price change.

### (b) (1 point) Draw a scatter plot of volatility against pricechange and comment on it.

As the plot indicates, *volatility* is positively correlated with *pricechange*.

```r
ggplot(data = companies) +
    geom_point(aes(x = volatility, y = pricechange)) +
    labs(title = "Volatility vs. Price Change", x = "Volatility", y = "Price Change") +
    theme(plot.title = element_text(hjust = 0.5))
```



**(c) (half a point) Calculate the Pearson correlation coefficient and comment on it.**

The coefficient is 0.9003, and suggests a highly positive correlation.

```r
cor(companies$pricechange, companies$volatility)
```

```
[1] 0.9002681
```

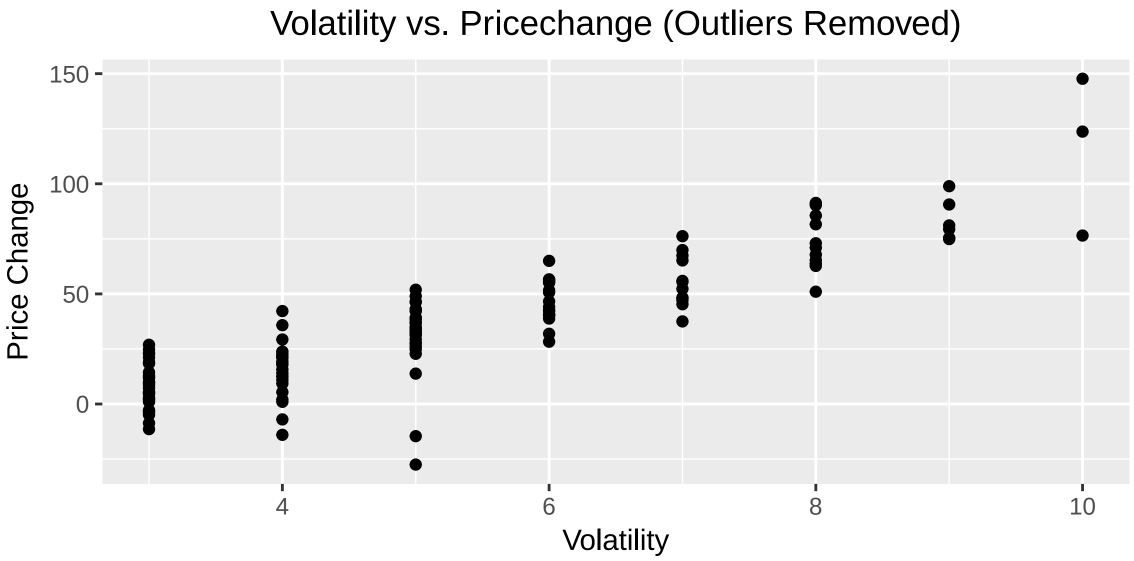
### 9. There seem to be (at least) three observations that fall a bit far from the other points; two having a high price change score and high volatility, the third one only having high volatility.

**(a) (1 point) Draw the scatter plot leaving out these three points.**

The plot is the scatter plot after removing the outliers.

```r
# remove points whose volatility is greater than 12
d <- companies[!companies$volatility > 12, ]

ggplot(data = d) +
    geom_point(aes(x = volatility, y = pricechange)) +
    labs(title = "Volatility vs. Pricechange (Outliers Removed)",
         x = "Volatility", y = "Price Change") +
    theme(plot.title = element_text(hjust = 0.5))
```



**(b) (half a point) Is there now a stronger linear relationship?**

It is difficult to judge whether there is a stronger linear relationship from the plot. But the new Pearson correlation coefficient, which is 0.8861, shows a weaker linear relationship.

**(c) (1 point) Compute the Pearson correlation coefficient anew, ignoring these three companies. How has the correlation coefficient changed?**

The new Pearson correlation coefficient is 0.8861, decreases after the outliers are removed.

```
cor(d$pricechange, d$volatility)
```

```
[1] 0.8861143
```

## 10. Now, you investigate whether there is any relationship between price change over the last 12 months and dividend.

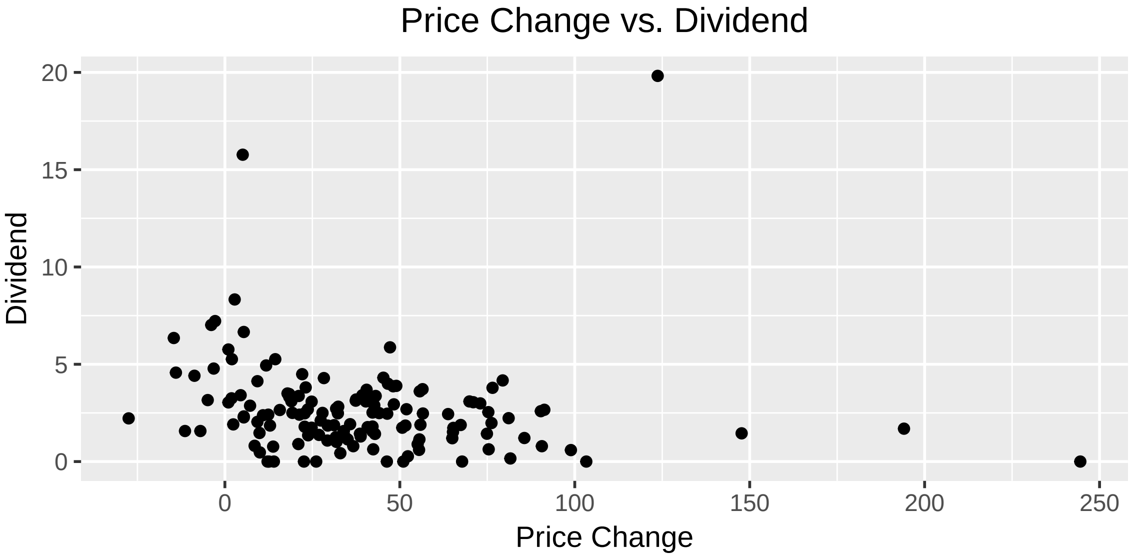
### (a) (half a point) Would you expect a relationship?

We would expect these two variables are negatively correlated, given the higher price change could result in lower dividend.

### (b) (half a point) Draw a scatter plot of price change against dividend and comment on it.

The plot below suggest a slightly negatively correlation.

```
ggplot(data = companies) +
    geom_point(aes(x = pricechange, y = dividend)) +
    labs(title = "Price Change vs. Dividend", x = "Price Change", y = "Dividend") +
    theme(plot.title = element_text(hjust = 0.5))
```



### (c) (half a point) Calculate the Pearson correlation coefficient and comment on it.

The Pearson correlation coefficient is -0.1302, showing a weakly negative correlated relationship.

```
cor(companies$pricechange, companies$dividend, use = "complete.obs")
```

```
[1] -0.130151
```

### (d) (half a point) Calculate Kendall's $\tau$ and Spearman's Rank correlation and comment on them.

The Kendall's $\tau$ coefficient is -0.1822 and the Spearman's Rank correlation coefficient is -0.2711. Both indicate that the two variables are weakly negative correlated.

```
cor(companies$pricechange, companies$dividend, use = "complete.obs", method = "kendall")
```

```
[1] -0.1821568
```

```
cor(companies$pricechange, companies$dividend, use = "complete.obs", method = "spearman")
```

```
[1] -0.2710561
```

**(e) (half a point) Which of the three correlation coefficient deems most appropriate to you in this situation?**

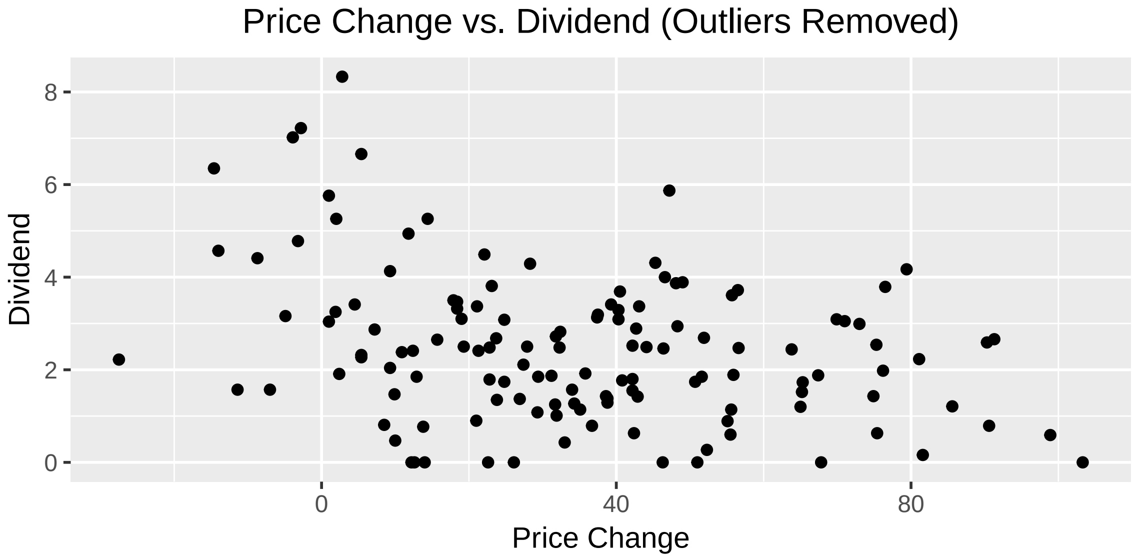
Given these two variables are both continuous, Pearson correlation coefficient would be considered most appropriate.

### 11. You continue investigating the relationship between price change over the last 12 months and dividend. From the revious question you conclude that there are five outliers: two companies with high dividend, three with high price change.

**(a) (half a point) Eliminate these five outliers and draw the plot anew. Has the relationship changed?**

```
d <- companies[!(companies$dividend > 15 |
                        companies$pricechange > 140), ]

ggplot(data = d) +
    geom_point(aes(x = pricechange, y = dividend)) +
    labs(title = "Price Change vs. Dividend (Outliers Removed)",
         x = "Price Change", y = "Dividend") +
    theme(plot.title = element_text(hjust = 0.5))
```



**(b) (1 point) Give the company names for those that you eliminated from your analysis?**

```
companies[(companies$dividend > 15 &
               !is.na(companies$dividend) | companies$pricechange > 140 ), "name"]
```

```
[1] "IDB Holdings"    "Brait SA. (JSE)" "Combined Motor"  "Peregrine"
[5] "PSG Group"
```

**(c) (1 point) Calculate the Pearson correlation coefficient for the restricted set. By how much has the correlation coefficient changed.**

```
cor(d$pricechange, d$dividend, use = "complete.obs")
```

```
[1] -0.3311798
```