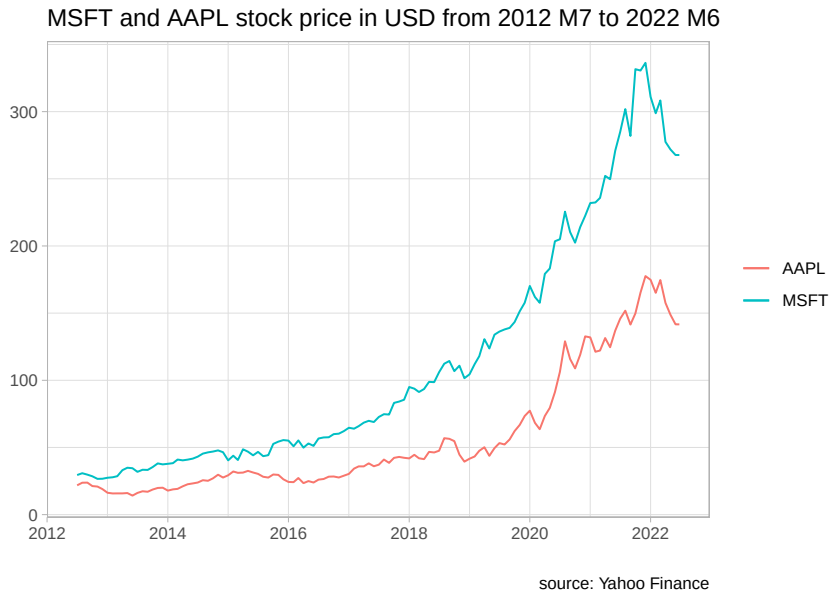
Pearson's vs Spearman's correlation

Nikolaj Nasenko

In this case study we attempt to demonstrate the difference between the applicability of Person's and Spearman's correlation measures on two monthly time series, namely the stock prices of Microsoft (MSFT) and Apple (AAPL). First, let's examine how these two series look like.

```
# load packages
library(quantmod)
library(tidyverse)

# get MSFT and AAPL price from yahoo finance
getSymbols.yahoo(Symbols = c("MSFT", "AAPL"),
                 periodicity = "monthly",
                 from = "2012-06-22",
                 env = .GlobalEnv) %>%
  invisible() # to suppress message

# extract close prices and put them into a single table
data <- tibble(Date = index(MSFT),
               MSFT = as.numeric(MSFT$MSFT.Close),
               AAPL = as.numeric(AAPL$AAPL.Close))

# plot time series
ggplot(data) +
  geom_line(aes(Date, MSFT, color = "MSFT")) +
  geom_line(aes(Date, AAPL, color = "AAPL")) +
  labs(title = "MSFT and AAPL stock price in USD from 2012 M7 to 2022 M6",
       x = "", y = "", caption = "source: Yahoo Finance") +
  theme_light() +
  theme(legend.title = element_blank())
```

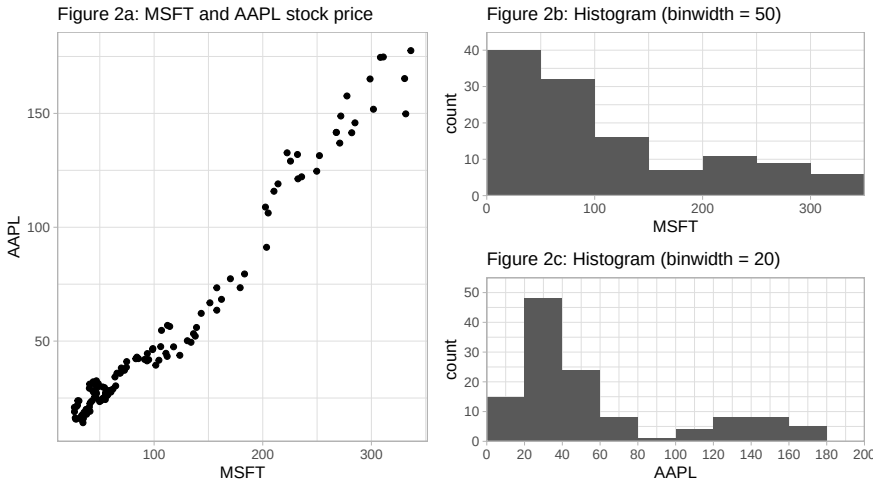


The two series look highly correlated, but let's quantify this observation. Pearson's correlation requires that the underlying relationship between the variables is linear, that the variables are continuous and approximately normally distributed.

Since both series are growing over time, we can expect that linear relationship exists. We can verify this assumption by simply looking at the scatter plot of the two series (viz figure 2a). Both series are continuous, but what about their distribution? Uptrending series like ours usually follow a log-normal distribution. Figures 2b and 2c present a histogram of the stock prices of MSFT and AAPL. We can see that the MSFT price is indeed log-normal. On the other hand, the distribution of the AAPL price is hard to determine, but it is probably not normal either. Spearman's correlation assumes that the variables are ordinal or continuous and have a monotonic relationship (e.g., linear). Thus, in this case, Spearman's correlation is more appropriate measure as it doesn't require normality.

```
library(gridExtra)

# MSFT and AAPL stock price scatter plot
p1 <- qplot(MSFT, AAPL, data = data) +
  theme_light() +
  ggtitle("Figure 2a: MSFT and AAPL stock price") +
  theme(plot.title = element_text(size = 12))

# histogram MSFT price
p2 <- ggplot(data, aes(MSFT)) +
  geom_histogram(binwidth = 50, boundary = 0) +
  theme_light() +
  scale_y_continuous(expand = c(0,0), limits = c(0, 45)) +
  scale_x_continuous(expand = c(0,0)) +
  ggtitle("Figure 2b: Histogram (binwidth = 50)") +
  theme(plot.title = element_text(size = 12))


# histogram AAPL price
p3 <- ggplot(data, aes(AAPL)) +
  geom_histogram(binwidth = 20, boundary = 0) +
```

```
  theme_light() +
  scale_y_continuous(expand = c(0,0), limits = c(0, 55)) +
  scale_x_continuous(expand = c(0,0), breaks = seq(0,220,20), limits = c(0,200)) +
  ggtitle("Figure 2c: Histogram (binwidth = 20)") +
  theme(plot.title = element_text(size = 12))

grid.arrange(p1, p2, p3, ncol = 2, layout_matrix = rbind(c(1,2), c(1,3)))
```



Let's calculate Spearman's correlation and its p-value. A correlation coefficient of 0.97 suggests that the two series are positively and highly correlated. The p-value is much lower than the 5 % significance level, which tells us that we can be confident that the resulting relationship is not a mere coincidence.

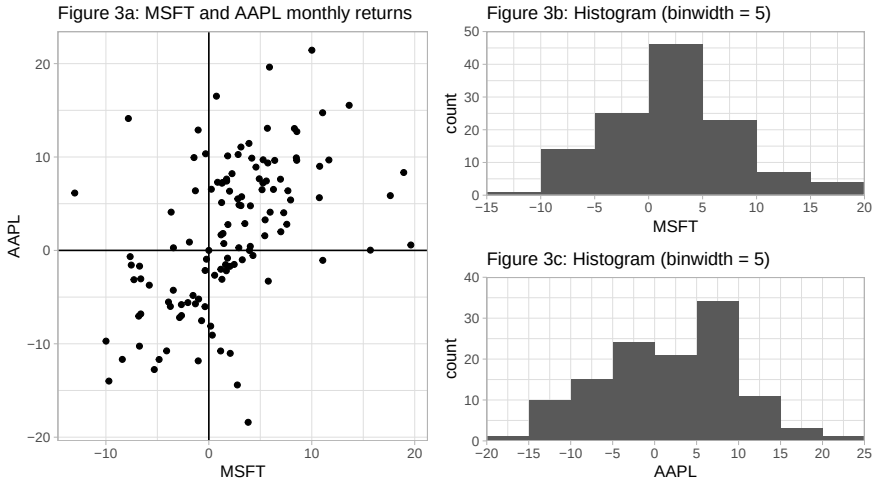
```
cor.test(data$MSFT, data$AAPL, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$MSFT and data$AAPL
## S = 8873.1, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9699463
```

Is there any way we can utilize the Person's correlation? We can take the logarithm of the two series, which should solve the non-normality issue, or convert prices to a monthly returns (see Figure 3). We can see now that MSFT and AAPL returns appear to be normally distributed, although only weakly linearly dependent in comparison to prices.

```
# converting prices to monthly returns
data.returns <- data %>%
  mutate(MSFT = ((MSFT/lag(MSFT)) - 1) * 100,
         AAPL = ((AAPL/lag(AAPL)) - 1) * 100)

# MSFT and AAPL returns scatter plot
```

```
p4 <- qplot(MSFT, AAPL, data = data.returns) +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = 0) +
  theme_light() +
  ggtitle("Figure 3a: MSFT and AAPL monthly returns") +
  theme(plot.title = element_text(size = 12))

# histogram of MSFT returns
p5 <- ggplot(data.returns, aes(MSFT)) +
  geom_histogram(binwidth = 5, boundary = 0) +
  theme_light() +
  scale_y_continuous(expand = c(0,0), limits = c(0,50)) +
  scale_x_continuous(expand = c(0,0), breaks = seq(-20,20,5)) +
  ggtitle("Figure 3b: Histogram (binwidth = 5)") +
  theme(plot.title = element_text(size = 12))

# histogram AAPL returns
p6 <- ggplot(data.returns, aes(AAPL)) +
  geom_histogram(binwidth = 5, boundary = 0) +
  theme_light() +
  scale_y_continuous(expand = c(0,0), limits = c(0,40)) +
  scale_x_continuous(expand = c(0,0), breaks = seq(-20,25,5)) +
  ggtitle("Figure 3c: Histogram (binwidth = 5)") +
  theme(plot.title = element_text(size = 12))

grid.arrange(p4,p5,p6, ncol = 2, layout_matrix = rbind(c(1,2), c(1,3)))
```



Person's correlation coefficient takes the value 0.5 and can vary between 0.35 and 0.62, which supports our previous observation. The returns of both stocks are only weakly dependent. The p-value is well below 5 % level, so the probability that we end up with a similar or stronger relationship on randomly drawn data is close to zero.

```
cor.test(data.returns$MSFT, data.returns$AAPL)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  data.returns$MSFT and data.returns$AAPL
## t = 6.1918, df = 118, p-value = 8.932e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3467400 0.6194616
## sample estimates:
##       cor
## 0.4952043
```

Note that the aforementioned list of Pearson's and Spearman's correlation assumptions is not exhaustive. There are other things to consider, especially the presence of outliers, that may significantly affect our results.

In conclusion, we began by taking historical stock prices over the past ten years. We figured out that we can measure the strength of the relationship between MSFT and AAPL prices using Spearman's correlation coefficient. Then, after converting prices into monthly returns, we found, by utilizing Person's correlation coefficient, that MSFT and AAPL returns are positively but weakly related to one another. In other words, based on our findings, we conclude that in the long run and in the short run, MSFT and AAPL prices tend to move in the same direction. In the long run, the AAPL price move effect on MSFT price is stable. In the short run, however, the AAPL price move reaction on the MSFT price move is unstable (Figure 4).

```
# MSFT and AAPL stock price regression line
p7 <- p1 + geom_smooth(method = "lm", se = FALSE, size = 1, color = "darkblue") +
  ggtitle("Figure 4a: MSFT and AAPL stock price") +
  theme(plot.title = element_text(size = 12))

# MSFT and AAPL monthly returns regression line
mod <- lm(AAPL~MSFT, data = data.returns)

p8 <- p4 + geom_abline(intercept = coef(mod)[1] - seq(0,4,0.2),
                       slope = coef(mod)[2] + seq(0,4,0.2),
                       color = "darkblue", size = 0.5) +
  ggtitle("Figure 4b: MSFT and AAPL monthly returns") +
  theme(plot.title = element_text(size = 12))

grid.arrange(p7, p8, ncol = 2)
```