Data analysis - correlation

Weronika Pyrka

TASK

In the file *sp3.csv*, there are data concerning the quantity of tweets written by the CEO of a certain company each day and the opening price of the company's stock on that day. Does the opening price influence the quantity of tweets? If so, to what extent? Check the statistical significance of the correlation coefficient at a significance level of $\alpha = 0.05$ using a parametric method. Determine the confidence interval using the bootstrap method and perform a permutation test. Compare the obtained results.

Let's start by reading the data from the file.

```
data <- read.csv("sp3.csv")
data
```

```
##    Tweets    Open
## 1      26 184.990
## 2      21 159.635
## 3      27 210.100
## 4      27 192.770
## 5      20 126.370
## 6      20 125.695
## 7      21 127.260
## 8      24 118.960
## 9      21 116.550
## 10     22 103.000
## 11     30 226.040
## 12     19 110.350
## 13     18 181.215
## 14     29 229.770
## 15     24 117.495
## 16     25 208.650
## 17     23 185.050
## 18     22 175.850
## 19     22 173.570
## 20     22 136.000
## 21     28 225.400
## 22     16 146.050
## 23     19 185.060
## 24     26 191.510
## 25     18 122.560
## 26     19 119.950
## 27     27 195.880
```

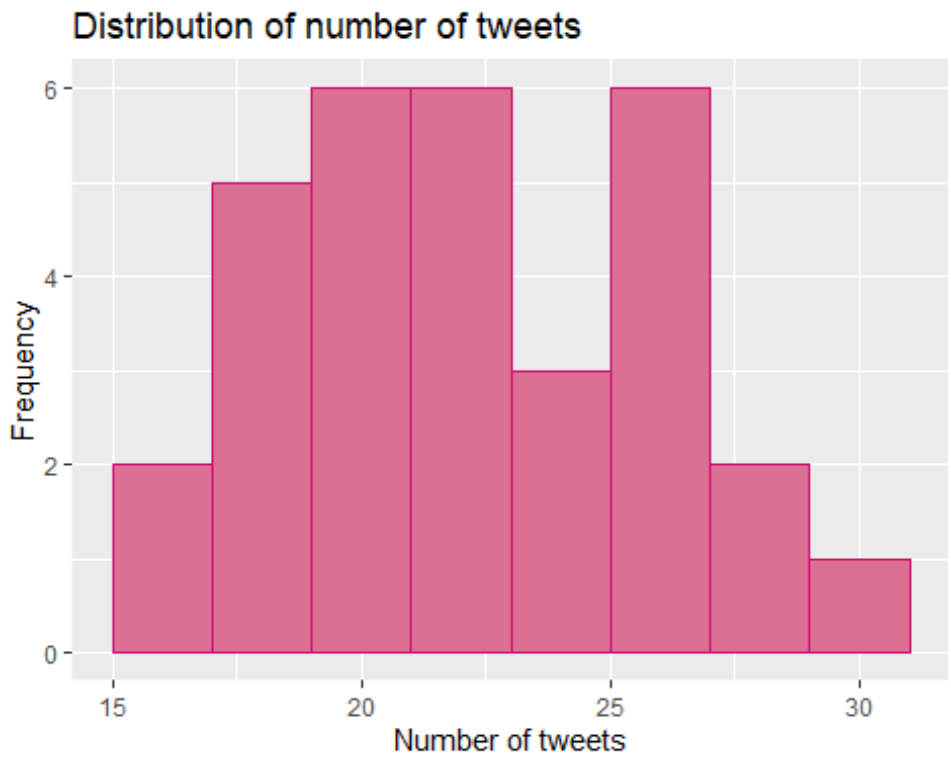
```
## 28     22 128.680
## 29     17 110.510
## 30     26 174.870
## 31     20 122.090
```

The first step before performing the t-test is to check if it can be conducted in this case.

To do this, I check if the sample comes from a normal distribution.

To accomplish this, I draw a histogram of the sample.

```
ggplot(data, aes(x = Tweets)) +
  geom_histogram(binwidth = 2, fill = "palevioletred", color = "deeppink3") +
  labs(x = "Number of tweets", y = "Frequency", title = "Distribution of
number of tweets")
```



Looking at the plot, I can conclude that it might be a sample from a normal distribution or a similar distribution to the normal one.

To confirm this statement, I perform the Shapiro-Wilk test.

```
shapiro_test <- shapiro.test(data$Tweets)
print(paste("The p-value for the Shapiro-Wilk test for the quantity of
tweets:", shapiro_test$p.value))
```

```
## [1] "The p-value for the Shapiro-Wilk test for the quantity of tweets:
0.436872901196105"
```

Check the p-value. Since it is greater than 0.05, it confirms the conclusion drawn after plotting the histogram, that the sample has a distribution similar to the normal distribution.

Therefore, conducting the t-test is valid.

It can proceeded directly to using the t-test.

```
t.test(data$Tweets, var.equal = TRUE)
```

```
##
##  One Sample t-test
##
## data:  data$Tweets
## t = 33.918, df = 30, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.25135 23.97445
## sample estimates:
## mean of x
##   22.6129
```

```
alpha <- 0.05
correlation <- cor(data$Tweets, data$Open)
print(paste("The correlation between the quantity of tweets and the opening
price:", correlation))
```

```
## [1] "The correlation between the quantity of tweets and the opening price:
0.717025159114455"
```

The correlation between the quantity of tweets and the opening price is 0.717025159114455. This indicates a moderate positive relationship between these two variables. It suggests that there is a tendency for the opening price of the stock to increase when the quantity of tweets increase and to decrease when the quantity of tweets decrease.

```
alpha <- 0.05
correlation <- cor.test(data$Tweets, data$Open)

if(correlation$p.value < alpha) {
  print("The correlation coefficient is statistically significant.")
} else {
  print("There is no statistical significance in the correlation
coefficient.")
}
```

```
## [1] "The correlation coefficient is statistically significant."
```

```
bootstrap <- function(data, indices) {
  sample_data <- data[indices, ]
  return(cor(sample_data$Tweets, sample_data$Open))
}

set.seed(123)
```

```r
bootstrap_results <- boot(data, bootstrap, R = 1000)
bootstrap_interval <- boot.ci(bootstrap_results, type = "bca")$bca[, 4:5]

cat("Bootstrap confidence interval:", bootstrap_interval)
```

```
## Bootstrap confidence interval: 0.4290164 0.8583023
```

The bootstrap confidence interval for the correlation coefficient is from 0.429 to 0.858. This means that we can be confident at a 95% confidence level that the true value of the correlation coefficient is within this interval.

This value confirms the statistical significance of the correlation coefficient.

```r
permutation_test <- function(data, indices) {
  permuted_data <- data
  permuted_data$Tweets <- data$Tweets[indices]
  return(cor(permuted_data$Tweets, permuted_data$Open))
}

permutation_results <- replicate(1000, permutation_test(data,
sample(nrow(data))))
permutation_p_value <- mean(permutation_results >= correlation$estimate)

cat("Result of the permutation test:", permutation_p_value)
```

```
## Result of the permutation test: 0
```

The t-test, bootstrap method and permutation test all confirm the statistical significance of the positive association between the quantity of tweets and the opening price of the stock.

It is worth noting that the results of different tests are consistent, which adds confidence to the interpretation of the relationship between the variables.

The correlation coefficient value 0.717 suggests a moderate positive relationship between the quantity of tweets and the opening price of the stock, which may be significant from the perspective of financial market analysis.