# Data analysis: t-test, confidence intervals, bootstrap

Weronika Pyrka

```r
library(tidyverse)
```

```
## Warning: pakiet 'tidyverse' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'ggplot2' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'tibble' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'tidyr' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'readr' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'purrr' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'dplyr' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'stringr' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'forcats' został zbudowany w wersji R 4.3.2

## Warning: pakiet 'lubridate' został zbudowany w wersji R 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

## TASK 1

Company A produces mobile phones. The packaging of the new model S from company A states that the battery lasts an average of 48 hours. We did not trust company A's claim and left 42 different model S phones playing videos until they discharged. The data collected during this experiment is available in the file *zad2.csv*. Justify that you can use the t-test and use the t-test to verify if company A is not deceiving consumers.

Before starting the task, the data from the file is loaded.

```r
battery <- read.csv("zad2.csv")
battery
```

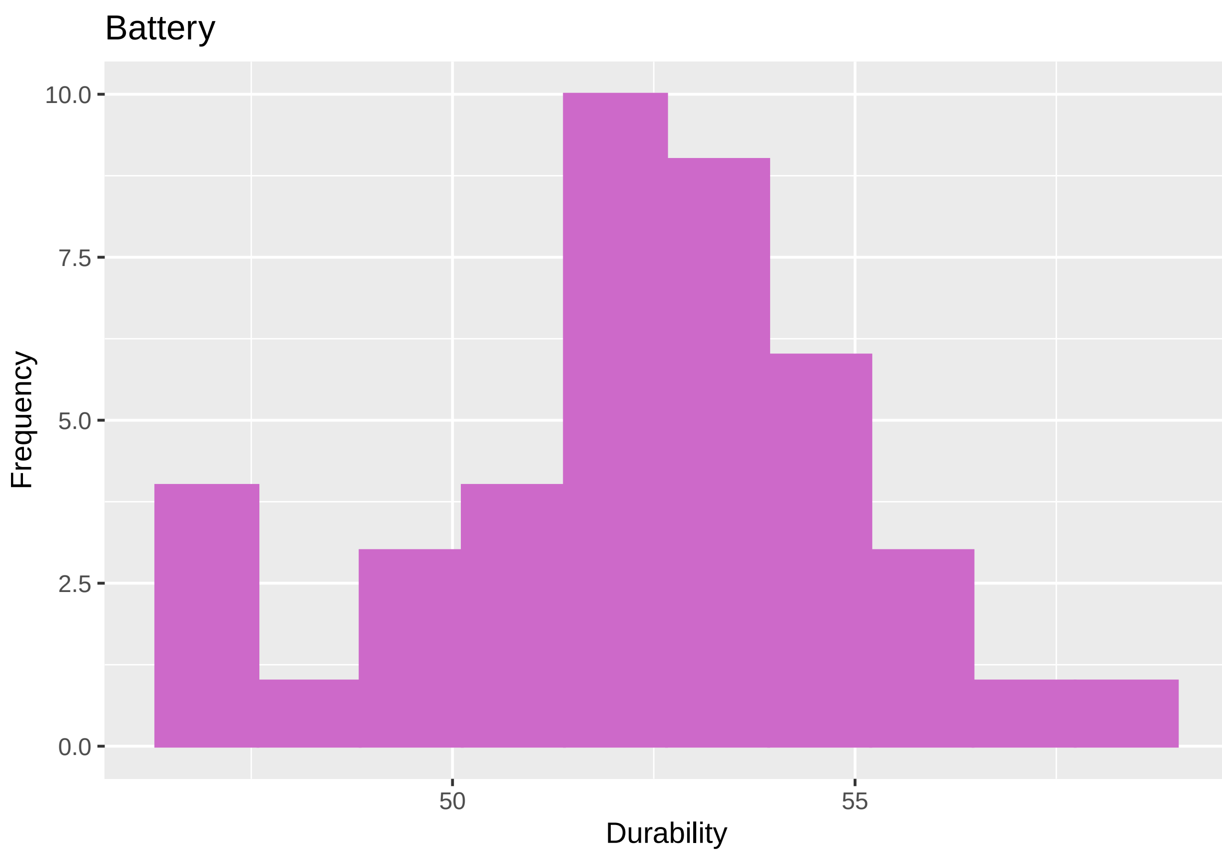
```
##    number durability
## 1       1      52.05
## 2       2      53.69
## 3       3      51.76
## 4       4      55.40
## 5       5      55.68
## 6       6      53.12
## 7       7      51.74
## 8       8      48.08
## 9       9      49.13
## 10     10      53.60
## 11     11      51.89
## 12     12      54.94
## 13     13      50.83
## 14     14      53.78
## 15     15      52.70
## 16     16      46.68
## 17     17      54.43
## 18     18      51.28
## 19     19      56.55
## 20     20      51.67
## 21     21      54.36
## 22     22      47.56
## 23     23      53.94
## 24     24      53.28
## 25     25      46.98
## 26     26      51.42
## 27     27      58.08
## 28     28      51.79
## 29     29      53.54
## 30     30      50.14
## 31     31      46.66
## 32     32      52.13
## 33     33      52.62
## 34     34      52.22
## 35     35      51.38
## 36     36      54.73
## 37     37      55.27
## 38     38      49.27
## 39     39      53.82
## 40     40      49.45
## 41     41      54.81
## 42     42      53.87
```

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

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

To accomplish this, draw a histogram of the sample.

```r
plot1 <- ggplot(data = battery, aes(x = durability)) +
          geom_histogram(bins=10, fill = "orchid3", color = "orchid3") +
          ggtitle("Battery") + xlab('Durability') + ylab('Frequency')
plot1
```




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.

```r
shapiro.test(battery$durability)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  battery$durability
## W = 0.96515, p-value = 0.2249
```

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.

```r
t.test(battery$durability, mu = 48)
```

```
##
##  One Sample t-test
##
## data:  battery$durability
## t = 10.35, df = 41, p-value = 5.301e-13
## alternative hypothesis: true mean is not equal to 48
## 95 percent confidence interval:
##  51.45556 53.13110
## sample estimates:
## mean of x
##  52.29333
```

Again, I focus on the p-value. The obtained value allows to conclude that the battery durability stated on the phone's box differs from the actual durability.

Additionally, from the conducted test, I can infer that with 95% confidence, the mean battery durability of the phone will be 52.29 hours and more precisely, it will lie within the interval (51.45556;53.13110).

Ultimately, I can assert that Company A is deceiving consumers, but in favor of the customer, as the actual average phone runtime is longer than that stated by the manufacturer on the packaging.

# TASK 2

Company B produces chocolate. After years, the management has decided to change the packaging of their chocolate, which they believe will certainly increase sales. The file *zad3t.csv* contains data on the sales of chocolate with the new packaging in one of the stores in one of the large Polish cities, as well as data on the sales of chocolate with the old packaging in one of the stores in one of the large Polish cities. Using the Student's t-test, check if the management was correct and if the new packaging increased sales.

Before starting the task, the data from the file is loaded.

```r
packaging <- read.csv("zad3t.csv")
packaging
```

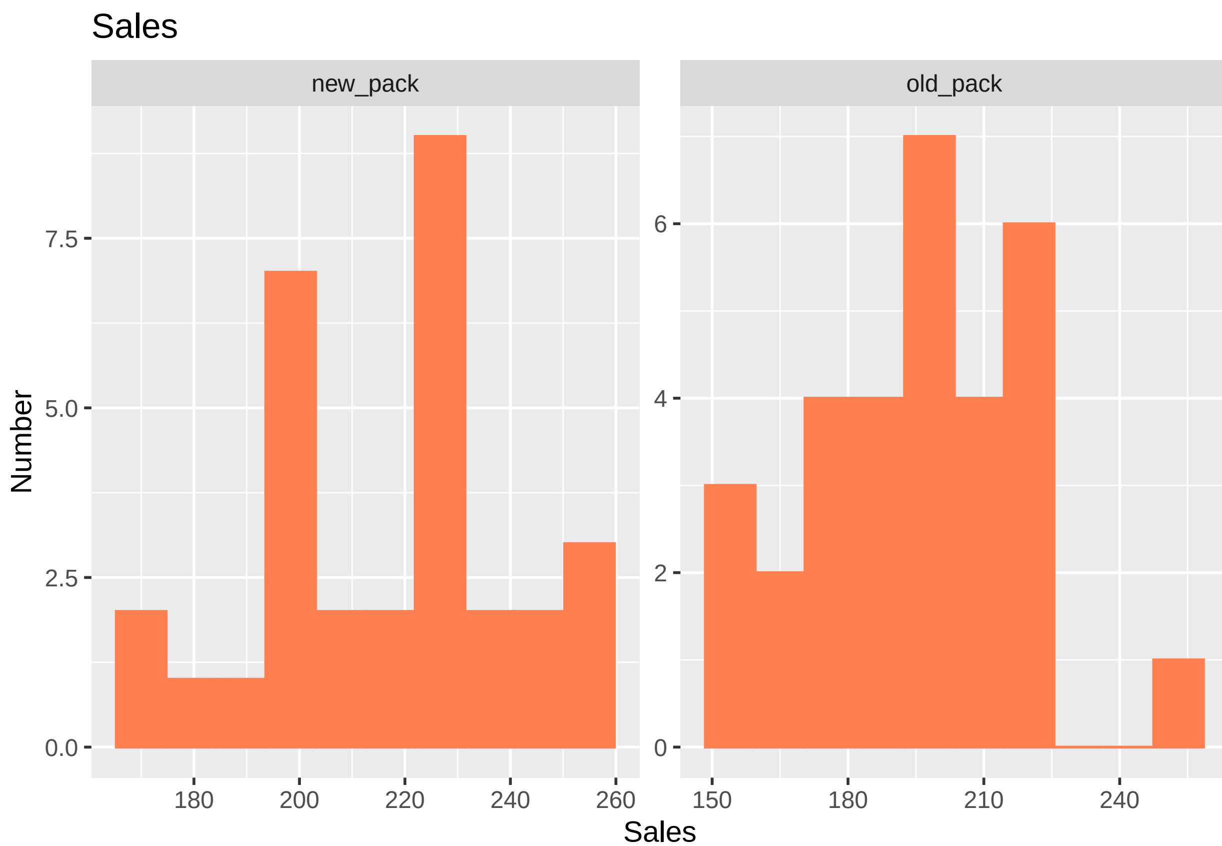
```
##        pack sold
## 1  new_pack  240
## 2  old_pack  220
## 3  new_pack  225
## 4  old_pack  163
## 5  new_pack  172
## 6  old_pack  196
## 7  new_pack  173
## 8  old_pack  166
## 9  new_pack  223
## 10 old_pack  200
## 11 new_pack  200
## 12 old_pack  219
## 13 new_pack  202
## 14 old_pack  158
## 15 new_pack  195
```

```
## 16 old_pack  206
## 17 new_pack  196
## 18 old_pack  172
## 19 new_pack  224
## 20 old_pack  221
## 21 new_pack  234
## 22 old_pack  179
## 23 new_pack  252
## 24 old_pack  151
## 25 new_pack  202
## 26 old_pack  220
## 27 new_pack  257
## 28 old_pack  176
## 29 new_pack  244
## 30 old_pack  153
## 31 new_pack  213
## 32 old_pack  210
## 33 new_pack  221
## 34 old_pack  187
## 35 new_pack  189
## 36 old_pack  182
## 37 new_pack  180
## 38 old_pack  172
## 39 new_pack  230
## 40 old_pack  196
## 41 new_pack  229
## 42 old_pack  196
## 43 new_pack  230
## 44 old_pack  200
## 45 new_pack  205
## 46 old_pack  205
## 47 new_pack  231
## 48 old_pack  250
## 49 new_pack  196
## 50 old_pack  191
## 51 new_pack  253
## 52 old_pack  193
## 53 new_pack  249
## 54 old_pack  216
## 55 new_pack  194
## 56 old_pack  214
## 57 new_pack  228
## 58 old_pack  198
## 59 new_pack  224
## 60 old_pack  225
## 61 new_pack  211
## 62 old_pack  188
```

Next, a histogram of the sales is created.

```
plot2 <- ggplot(packaging, aes(x = sold)) +
           geom_histogram(bins = 10, fill = "coral", color = "coral") +
           ggtitle("Sales") + xlab("Sales") + ylab("Number") +
  facet_wrap(~pack, scales = 'free')
```

```
plot2
```




From the obtained plots, I can conclude that the distribution is similar to a normal distribution.

Next, I rearrange the data to separate the new packaging from the old packaging. This will increase the clarity of the results obtained.

```
packaging <- packaging %>%
  group_by(pack) %>%
  mutate(row = row_number()) %>%
  pivot_wider(names_from = pack, values_from = sold) %>%
  select(new_pack, old_pack)

packaging
```

```
## # A tibble: 31 x 2
##    new_pack old_pack
##       <int>    <int>
##  1      240      220
##  2      225      163
##  3      172      196
##  4      173      166
##  5      223      200
##  6      200      219
##  7      202      158
##  8      195      206
```

```
##  9      196      172
## 10      224      221
## # i 21 more rows
```

Perform the Shapiro-Wilk test to confirm whether the data originates from a normal distribution.

```r
shapiro.test(packaging$old_pack)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  packaging$old_pack
## W = 0.97916, p-value = 0.789
```

```r
shapiro.test(packaging$new_pack)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  packaging$new_pack
## W = 0.96591, p-value = 0.4142
```

Based on the analysis of the p-value, I confirm the assumption that the sample comes from a distribution similar to the normal distribution.

Next, I calculate the variances of both packaging types.

```r
var(packaging$old_pack)
```

```
## [1] 554.4129
```

```r
var(packaging$new_pack)
```

```
## [1] 554.9398
```

Both variances are approximately the same, so proceed to perform the t-test.

```r
t.test(packaging$new_pack, packaging$old_pack, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  packaging$new_pack and packaging$old_pack
## t = 3.7693, df = 60, p-value = 0.0003761
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.58240 34.51438
## sample estimates:
## mean of x mean of y
##  216.8387  194.2903
```

Here again, I pay attention to the p-value. The obtained value is less than 0.05, indicating that the statement that the sales of both packaging types are the same is untrue. With a 95% confidence level, it can be said that the difference between the means lies within the interval (10.58240;34.51438). The "mean of x" element represents the sales of the new packaging, while "mean of y" represents the sales of the old packaging. Comparing these values, I conclude that the management was correct and the change to the new packaging increased chocolate sales.

# TASK 3

Use the bootstrap method to perform the above tests and compare the results.

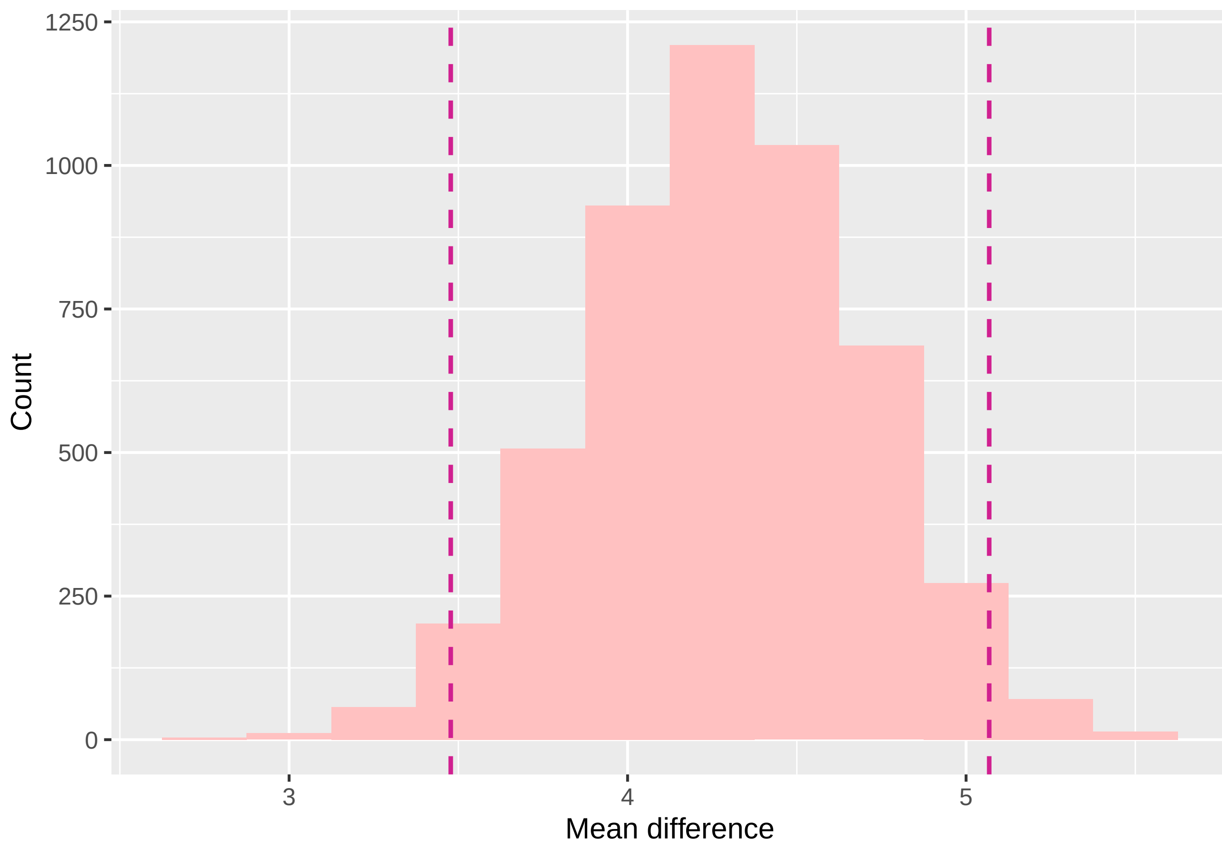
Task 1:

```
n <- length(battery$durability)
mu <- 48
bootstrap_stat <- rep(0, 5000)

for (i in 1:5000) {
  sample_data <- sample(battery$durability, size = n, replace = TRUE)
  bootstrap_stat[i] <- mean(sample_data) - mu
}

bootstrap_stat <- tibble(mean_diff = bootstrap_stat)

plot_1 <- ggplot(bootstrap_stat, aes(x = mean_diff)) +
           geom_histogram(binwidth = 0.25, fill = "rosybrown1") +
           xlab("Mean difference") + ylab("Count") +
           geom_vline(xintercept = c(quantile(bootstrap_stat$mean_diff, 0.025),
                                     quantile(bootstrap_stat$mean_diff, 0.975)),
                      linetype = 'dashed', lwd = 0.8, color = 'violetred')
plot_1
```




```
interval <-  c(quantile(bootstrap_stat$mean_diff, 0.025),
               quantile(bootstrap_stat$mean_diff, 0.975))
interval
```

```
##     2.5%    97.5%
## 3.477595 5.068351
```

Analyzing the information obtained from the bootstrap method, I arrived at conclusions consistent with those drawn from performing the t-test. Specifically: the battery life of the phone differs from what is stated on the packaging; the manufacturers from company A are misleading consumers and the durability of the battery is better than indicated. The confidence intervals from the bootstrap method and the t-test are almost identical. The conclusions I draw from the bootstrap method are the same as those from the t-test.

Task 2:

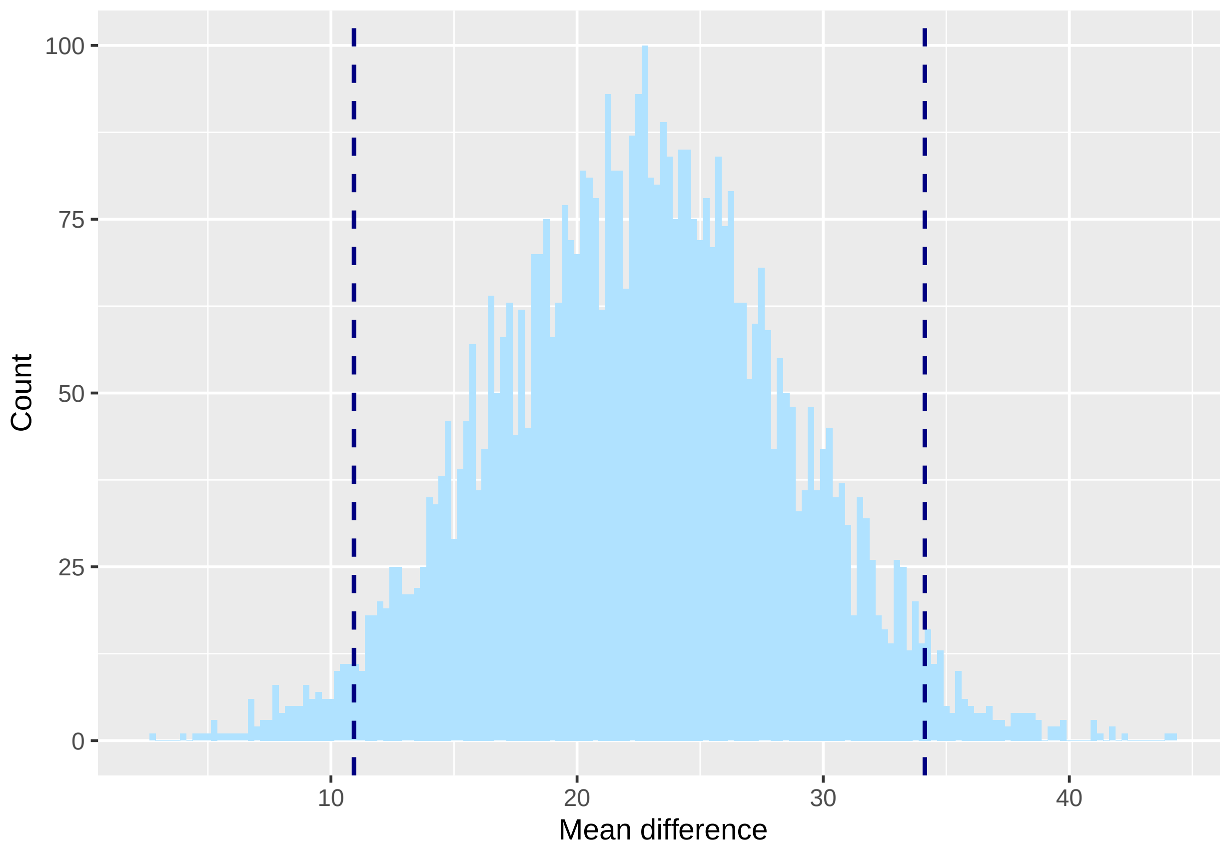
```
n = length(packaging$old_pack)
bootstrap_stat = rep(0,5000)
for(i in 1:5000){
  sample_1 = sample(packaging$new_pack, size=n, replace=TRUE)
  sample_2 = sample(packaging$old_pack, size=n, replace=TRUE)
  bootstrap_stat[i] = mean(sample_1) - mean(sample_2)
}
bootstrap_stat = tibble(mean_diff = bootstrap_stat)

confidence_interval <- c(quantile(bootstrap_stat$mean_diff, 0.025,),
```

```
                                quantile(bootstrap_stat$mean_diff, 0.975))
confidence_interval
```

```
##     2.5%    97.5%
## 10.93548 34.12984
```

```
plot_2 <- ggplot(bootstrap_stat, aes(x = mean_diff)) +
           geom_histogram(binwidth = 0.25, fill = "lightskyblue1") +
           xlab("Mean difference") + ylab("Count") +
           geom_vline(xintercept = c(quantile(bootstrap_stat$mean_diff, 0.025,),
           quantile(bootstrap_stat$mean_diff, 0.975)), linetype = 'dashed',
           lwd = 0.8, color = 'navy')
plot_2
```




The conclusions drawn from the implementation of the bootstrap method for Task 2 are the same as those from the bootstrap method for Task 1. The parameters of the bootstrap method align with those obtained from the t-test and the confidence intervals are nearly identical. The new packaging achieved better sales, confirming that the management was correct in expecting an increase in sales due to the packaging change.