Contents

Reham Alanazi

Executive Summary:

This report will examine which supermarket Coles or Woolworths has lower prices. To make our investigation, we need to set basic procedures. The first one, because we have paired data, we will use testing hypotheses for the paired-samples -test. Before doing this test, we should choose the null hypothesis (H_0), and the alternative hypothesis (H_1). H_0 is given by the difference between the means of two supermarkets equal zero, while the alternative H_1 is not equal or greater than or smaller than zero. Then, we should establish the assumptions of this investigation, such as the data normality. Also, set the decision rules that when we reject or fail to reject the null hypothesis. Finally, make the Conclusion, which includes when the test will be statistically significant. The second procedure is making the T-test by R-Studio. We should get the p-value from this test and compare it with the significant level. If the p-value smaller than the significant level and 95% of the CI of the mean difference does not capture H_0, then we reject the null hypothesis. Otherwise, we fail to reject it.

Load Data and Packages:

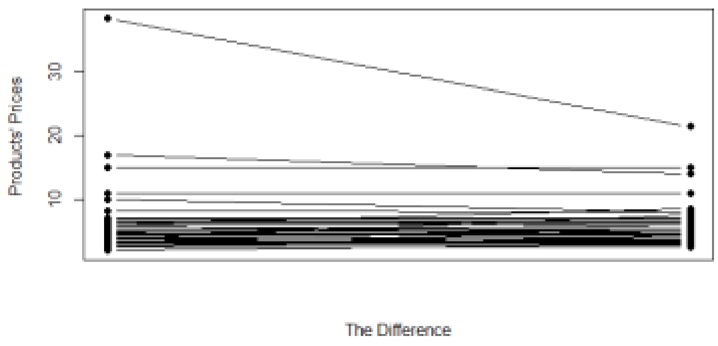
```
# Import the data:

library(readxl)
> datasetcolse_woolworths <- read_excel("C:/Users/great/Desktop/Study/semeste
r3/Statistics/secondAssignment/datasetcolse&woolworths.xlsx")
> View(datasetcolse_woolworths)

# Initialize dataframes for colse and woolworths prices:
> colesPrices = datasetcolse_woolworths$Coles_price;
> woolworthsPrices = datasetcolse_woolworths$Woolworths_price;

Plots:
# matplot
matplot(t(data.frame(datasetcolse_woolworths$Coles_price, datasetcolse_woolwo
rths$Woolworths_price)),
         type = "b", pch = 19, col = 1,
         lty = 1, xlab = "The Difference",
         ylab = "Products' Prices ", xaxt = "n"
 )
# qqPlot
> library(car)
pricesDifference$d %>% qqPlot(dist="norm")
[1] 35 24
```

```
# Conduct Paired sample t-test
#  x̄1  − x̄2
>t.test(pricesDifference$Coles_price, pricesDifference$Woolworths_price,paire
d = TRUE, m=0, conf.level = 0.95 )
# x̄2  − x̄1
>t.test( pricesDifference$Woolworths_price,pricesDifference$Coles_price,paire
d = TRUE, m=0, conf.level = 0.95 )
```

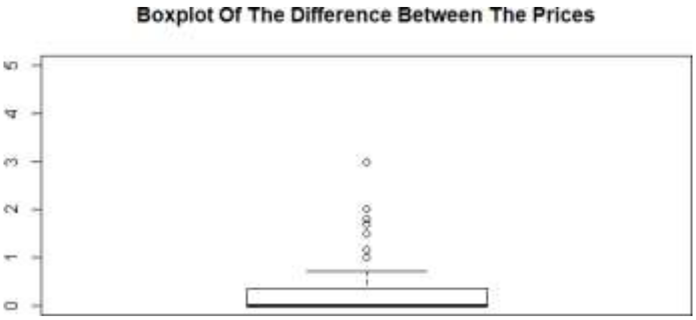
Summary Statistics:

```
# The Coles summary

  Min    Q1  Median    Q3   Max   Mean    SD    n  Missing
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
     2  3.42    4.9     6  38.5  5.96  5.50    50       0
  > IQR(pricesDifference$Coles_price)
    [1] 2.575
 # The Woolworths summary
  Min    Q1   Median    Q3   Max   Mean    SD    n   Missing
  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
   2.5  3.42    4.5  6.22  21.5  5.52  3.49    50       0
  > IQR(pricesDifference$Woolworths_price)
    [1] 2.8

# The difference summary
    Min    Q1   Median    Q3   Max   Mean    SD   n   Missing
  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
  -3.75     0      0 0.312    17 0.440   2.61    50       0

  > IQR(pricesDifference$d)
  [1] 0.3125
```

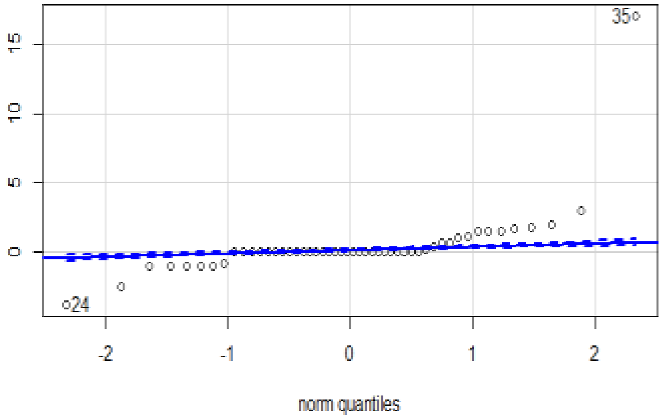
- ***# Visualization:***
    - Matplot describes the difference between two samples. The first column in the right side represents Coles data, while the second column in the left side



represents Woolworths data. We can see most the products have the same prices, just a few of them are different.



- boxplot represents the minimum, maximum, median, first quartile and third quartile of the distributed data. There are outliers above the third quartile (out of the expected range) it seems to be not a normal distribution, but the size is higher than 30, so we can assume the data is normal.



- The qqplot looks weird the mean reason is that the range of the data is from -3.75 to 17 and most of them small values.

Hypothesis Test:

In this situation, we will use a paired sample t-test. This test is sufficient to examine the difference between two samples. The data is randomly chosen matched products from each store, so the hypothesis is for dependent / paired data which we need to test one sample.

Hypotheses for the paired (dependent) samples -test:

H_0 : μΔ = 0 (which means the different between μ1 of Coles data and μ2 of Woolworths data =0).

H_A: μΔ ≠ 0, HA : μΔ < 0, HA : μΔ > 0 (if the μΔ smaller than 0 then Coles prices lower than Woolworths, but if μΔ greater than 0 them Coles has prices higher than Woolworths.

Assumptions:

- Comparing the population average of the difference between two distinct prices.
- They are normally distributed; the data size is greater than 30 (n>30) for both Coles and Woolworths, so the sample of the difference between them will have the same size (50 > 30).

Decision Rules:

Reject H_0 :If p-value < 0.05.
If 95% CI of the mean difference does not capture H_0
Otherwise, fail to reject.

Conclusion:

Test will be statistically significant if we reject H_0, because we will accept the alternative value.
Otherwise, the test is not statistically significant.

Interpretation:

- T-tests by R.

-
  ```
  # Conduct Paired sample t-test:
  #the mean of x1-x2
  Paired t-test
  data:  pricesDifference$Coles_price and
  pricesDifference$Woolworths_price
  t = 1.191, df = 49, p-value = 0.2394
  alternative hypothesis: true difference in means is not
  equal to 0
  95 percent confidence interval:
   -0.3025513  1.1829513
  sample estimates:
  mean of the differences
                   0.4402
  ```

-
  ```
  Paired t-test
  #the mean of x2-x1
  data:  pricesDifference$Woolworths_price and
  pricesDifference$Coles_price
  t = -1.191, df = 49, p-value = 0.2394
  alternative hypothesis: true difference in means is not
  equal to 0
  95 percent confidence interval:
   -1.1829513  0.3025513
  sample estimates:
  mean of the differences
                  -0.4402
  # we can also use one-sample t-test of the mean difference
   data:  pricesDifference$d
  ```

```
t = 1.191, df = 49, p-value = 0.2394
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.3025513  1.1829513
sample estimates:
mean of x  0.4402
```

- **T-test result summary:**

  We assumed normality n > 30 (n = the size of both Coles and Woolworths).
  p-value = 0.2394: p > 0.05.
  Estimated difference between means: $\bar{x}1 - \bar{x}2$= 0.4402, $\bar{x}2 - \bar{x}1$= -0.4402.
  95% CI of difference between means [-0.3025513 1.1829513].

- **Decision:**

  Fail to reject H_0.

- **What do we conclude?**

  The decision fail to reject H_0, where the mean of the difference = 0.4402 around 0 as the p > 0.05. The confidence interval of the population [-0.3025513 1.1829513] captures H_0. The result of paired-sample t-test is not statistically significant, which means there is no statistically significant difference between Coles and Woolworths prices.

Discussion:

The investigation about finding which supermarket, Coles or Woolworths, has lower prices. The null hypothesis is that the two supermarkets have the same prices, which is represented by $\mu\Delta = 0$, and that means the difference between them equal to zero. If the t-test rejects, H_0, then the alternative hypothesis will be accepted ($\mu\Delta \neq 0$).
The t-test failed to reject this hypothesis, because the probability of the observed sample result is higher than 0.05(significance level). From the result, we concluded that there is no significant difference. May there is a very slight difference when we consider the mean value of the difference between them : : $\bar{x}1 - \bar{x}2$= 0.4402, $\bar{x}2 - \bar{x}1$ = -0.4402, which means that $x1$ (the mean of Coles) is bigger than $\bar{x}2$ (the mean of Woolworths) that results in Woolworths is cheaper. However, it is a tiny difference, and they are near to have the same prices. According to the t-test, they have the same prices.
The strengths of this investigation: the products are randomly selected with sufficient size for normally distributed data. The null hypothesis helps this investigation, because if it is rejected, then there should be a difference.
The limitations of this investigation: the data does not have a variety. The values very close together in a specific range, which makes the visualizations of the data more likely to be not normally distributed. What improvements could be made? Increase the size of the data and chose the data randomly with respect to different prices and different categories.