MATH1324 Assignment 3

Code

Supermarket Price Wars

Group/Individual Details

- Anshit Malik (s3631281)
- Mohammad (s3650497)
- Namita Chhibba (s3631442)

Executive Statement

Problem Statement

This report is commissioned to examine the variability between the prices of two renowned super markets in Australia viz. Wesfarmers subsidiary “Coles” and “Woolworths Ltd”.

Procedure

- Raw data was created to list the count of products present in the products’ hierarchies on the Coles website. We considered Coles website rather than Woolworths’ because it demonstrated a better understanding of the products listing.
- Random sampling was done which allowed all the products in the population to have equal probability of being chosen.
- Paired T-test was applied.
- Results from test were interpreted and conclusion was drawn.

Sample

Products are segmented into a hierarchy elaborated into primary, secondary and tertiary categories with a definite count of total products sold in both supermarkets. We took a sample size of 99 products and collected the prices of each sampled product from the Coles and Woolworths’ website. Furthermore, the quantity represented in our data is taken at a minimum which represents the accurate amount without any biasness of the bulk factor. Process of sampling as given below:

- Raw Data was created manually on Microsoft Excel and was transferred to R.
- A dummy population was created
- The product ids were assigned to the products’ population
- Randomly product ids were chosen from the population without replacement
- set.seed () function was used when running simulation to ensure the result is reproducible.
- The data was distributed amongst the participants to record the prices for the randomly generated products.

Assumption: Occurrence of a special/home brand product is discarded and the next product on the same page was considered, the same process was followed when the same product was not found on the Woolworth’s webpage.

Variables creation

The following variables were created:

- Webpage number: To locate the webpage where sampled product is present.
- Product number: To locate the location of product on the webpage.
- Coles price
- Woolworths price

Conclusion

It was concluded that one has to pay less for most of the products at Woolworth's as compared to Coles.

Load Packages & Data

Hide

```
library(openxlsx)
library(psych)
library(dplyr)
library(ggplot2)
library(plotly)
ass3_data_1<-read.xlsx("C:\\Users\\anshi\\Desktop\\Semester 1\\Intro To Stats\\Assignment 3
\\Data Excel\\File4R.xlsx")
ass3_data_1
```

	Primary.Product <chr>	Secondary.Product <chr>	Tertiary.Product <chr>
1	Bread & Bakery	From Our Bakery	Bread
2	Bread & Bakery	From Our Bakery	Bakery Cakes
3	Bread & Bakery	From Our Bakery	Cupcakes & Muffins
4	Bread & Bakery	From Our Bakery	Donuts & Cookies
5	Bread & Bakery	From Our Bakery	Pastries & Desserts
6	Bread & Bakery	From Our Bakery	Crumpets, Muffins and Bagels
7	Bread & Bakery	From Our Bakery	Wraps, Pita & Flat Bread
8	Bread & Bakery	Packaged Bread & Bakery	Packaged Bread
9	Bread & Bakery	Packaged Bread & Bakery	Cakes
10	Bread & Bakery	Packaged Bread & Bakery	Cookies & Biscuits

1-10 of 661 rows Previous **1** 2 3 4 5 6 ... 67 Next

Dummy Population Creation

Hide

```
CumulativeFreq<-cumsum(ass3_data_1$Product.Count)
Calc_1<-c(0,(CumulativeFreq[1:(length(CumulativeFreq)-1)]+1))
Calc_2<-c(0,(CumulativeFreq[1:(length(CumulativeFreq)-1)]))
ass3_data_1<-data.frame(cbind(ass3_data_1,CumulativeFreq,Calc_1,Calc_2))
Population.df<-ass3_data_1[rep(row.names(ass3_data_1), ass3_data_1$Product.Count),1:ncol(ass3
_data_1)]
row.names(Population.df)<-1:nrow(Population.df)
ProductID<-1:nrow(Population.df)
Data_4_analysis<-data.frame(cbind(ProductID,Population.df))
Actual_web_page<-(Data_4_analysis$ProductID-Data_4_analysis$Calc_1-1)%/%24
Actual_web_page[Actual_web_page==-1]<-0
Web_Page_Num<-Actual_web_page+1
(Data_4_analysis$ProductID - Data_4_analysis$Calc_2)->x1
x2<-(x1)%%24
x2[x2==0]<-24
Product_Num<-x2
Final_Data<-data.frame(cbind(Data_4_analysis,Web_Page_Num,Product_Num))
Final_Data_4_Sampling<-Final_Data[,c(1:4,9:10)]
Final_Data_4_Sampling
```

	ProductID <int>	Primary.Product <chr>	Secondary.Product <chr>	Tertiary.Product <chr>	Web_Page_... <dbl>	I
1	1	Bread & Bakery	From Our Bakery	Bread	1	
2	2	Bread & Bakery	From Our Bakery	Bread	1	
3	3	Bread & Bakery	From Our Bakery	Bread	1	
4	4	Bread & Bakery	From Our Bakery	Bread	1	
5	5	Bread & Bakery	From Our Bakery	Bread	1	
6	6	Bread & Bakery	From Our Bakery	Bread	1	
7	7	Bread & Bakery	From Our Bakery	Bread	1	
8	8	Bread & Bakery	From Our Bakery	Bread	1	
9	9	Bread & Bakery	From Our Bakery	Bread	1	
10	10	Bread & Bakery	From Our Bakery	Bread	1	

1-10 of 26,306 rows Previous **1** 2 3 4 5 6 ... 100 Next

Random Sampling

Hide

```
set.seed(12345)
random_productID<-sample(1:26306, 99, replace=F)
Random_Sample_data<-data.frame(random_productID)
names(Random_Sample_data)<-"ProductID"
Final_Sample_data<-merge(Random_Sample_data,Final_Data_4_Sampling,by="ProductID")
Final_Sample_data
```

ProductID <int>	Primary.Product <chr>	Secondary.Product <chr>	Tertiary.Product <chr>

30	Bread & Bakery	From Our Bakery	Bread
158	Bread & Bakery	From Our Bakery	Crumpets, Muffins and I
227	Bread & Bakery	From Our Bakery	Wraps, Pita & Flat Bread
504	Bread & Bakery	Packaged Bread & Bakery	Rolls & Bagels
909	Fruit & Vegetables	Vegetables	Broccoli & Cauliflower
1142	Fruit & Vegetables	Fresh Flowers	Flowers
1266	Meat, Seafood & Deli	Meat	Chicken
1446	Meat, Seafood & Deli	Meat	Turkey
1581	Meat, Seafood & Deli	Seafood	Fish
1590	Meat, Seafood & Deli	Seafood	Fish

1-10 of 99 rows | 1-4 of 6 columns Previous **1** 2 3 4 5 6 ... 10 Next

Data after recording Prices for the sample products

Hide

```
Data_Mart<-read.xlsx("C:\\Users\\anshi\\Desktop\\Semester 1\\Intro To Stats\\Assignment 3\\Da
ta excel\\R Data.xlsx")
Data_Mart
```

	ProductID <dbl>	Primary.Category <chr>	Secondary.Category <chr>	Tertiary.Category <chr>
1	30	Bread & Bakery	From Our Bakery	Crumpets, Muffins ar
2	158	Bread & Bakery	Packaged Bread & Bakery	Cakes
3	227	Bread & Bakery	Packaged Bread & Bakery	Kosher Bakery
4	504	Fruit & Vegetables	Fruit	Organic Fruit
5	909	Fruit & Vegetables	Fruit	Plums & Apricots
6	1142	Meat, Seafood & Deli	Deli Meats	Sliced Meats
7	1266	Meat, Seafood & Deli	Deli Specialty	Deli Gourmet Cheese
8	1446	Dairy, Eggs & Meals	Ready to Eat Meals	Italian Meals & Pasta
9	1581	Dairy, Eggs & Meals	Ready to Eat Meals	Italian Meals & Pasta
10	1590	Health & Beauty	Dental	Toothpaste

1-10 of 99 rows | 1-5 of 8 columns Previous **1** 2 3 4 5 6 ... 10 Next

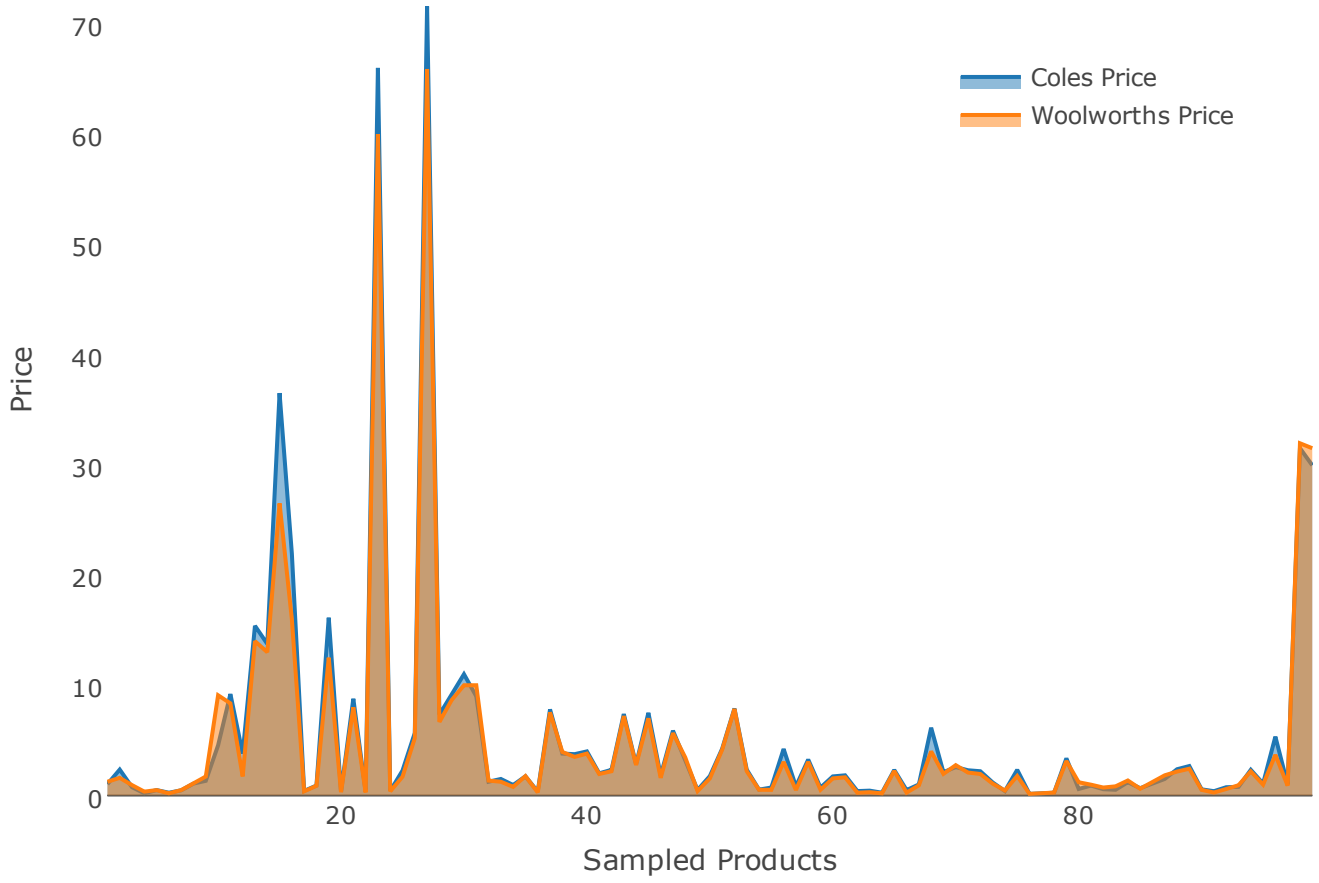
Summary Statistics

Visualisation Code for trend

Insight: Our visualization shows the comparison of the prices of all the 99 products in the sample from the two supermarkets. And it is observed that major price difference exists for the products with prices greater than $10 and in majority of these cases Coles is expensive than Woolworths.

Hide

```
Coles_Price <- Data_Mart$Coles.Price
count<-1:99
Woolies_Price <- Data_Mart$Woolsworth.Price
trend_plot <-plot_ly(y = ~Coles_Price, x = ~1:99, type = 'scatter', mode = 'lines', name = 'C
oles Price', fill = 'tozeroy',fillcolor = 'rgba(168, 216, 234)',line = list(width = 2)) %>%
  add_trace(y = ~Woolies_Price, x = ~1:99, name = 'Woolworths Price', fill = 'tozeroy', fillc
olor = 'rgba(255, 212, 96)') %>% layout(xaxis = list(title = 'Sampled Products',showgrid = FA
LSE),yaxis = list(title = 'Price',showgrid = FALSE),autosize = T,legend = list(x = 0.7, y =
0.9))
trend_plot
```



Audit Report of Data

Hide

```
Audit_Data<-Data_Mart
apply(Audit_Data,2,function(x) length(unique(x)))->unique_data
apply(Audit_Data,2,function(x) sum(is.na(x)))->missing_data
duplicate_data<-nrow(Audit_Data)-unique_data
non_missing<-nrow(Audit_Data)-missing_data
Total_rows<-nrow(Audit_Data)
Audit_Data<-cbind(Total_rows,missing_data,non_missing,duplicate_data,unique_data) %>% as.dat
a.frame()
names(Audit_Data)<-c("Total Obs","Missing Obs","Non Missing Obs","Duplicate Obs","Unique
Obs")
Audit_Data
```

	Total Obs <int>	Missing Obs <int>	Non Missing Obs <int>	Duplicate Obs <int>	Unique Obs <int>
ProductID	99	0	99	0	99
Primary.Category	99	0	99	83	16
Secondary.Category	99	0	99	57	42
Tertiary.Category	99	0	99	27	72
Product.Name	99	0	99	0	99
Quantity	99	0	99	94	5
Coles.Price	99	0	99	13	86
Woolsworth.Price	99	0	99	13	86

8 rows

Summary Statistics of Data

Hide

```
Summary_Statistics<-Data_Mart[7:8] %>% describeBy() %>% as.data.frame()
```

```
no grouping variable requested
```

Hide

```
Quantile_values_Coles<-Data_Mart %>%
summarise(Q1=quantile(Coles.Price,prob=0.25),Q3=quantile(Coles.Price,prob=0.75))
row.names(Quantile_values_Coles)<-"Coles.Price"
Quantile_values_Woolsworth<-Data_Mart %>% summarise(Q1=quantile(Woolsworth.Price,prob=0.25),Q
3=quantile(Woolsworth.Price,prob=0.75))
row.names(Quantile_values_Woolsworth)<-"Woolsworth.Price"
Quantiles<-rbind.data.frame(Quantile_values_Coles,Quantile_values_Woolsworth)
Summary_Statistics <-
cbind.data.frame(Summary_Statistics[8],Quantiles[1],Summary_Statistics[5],Quantiles[2],Summar
y_Statistics[9],Summary_Statistics[c(2:4,13)])
names(Summary_Statistics)<- c("Min","Q1","Median","Q3","Max","N","Mean","Standard
Deviation","Standard Error")
Summary_Statistics
```

	Min <dbl>	Q1 <dbl>	Med… <dbl>	Q3 <dbl>	Max <dbl>	N <dbl>	Mean <dbl>	Standard Deviation <dbl>	Sta
Coles.Price	0.143	0.6850	1.82	4.230	71.58	99	5.302899	11.17091	
Woolsworth.Price	0.148	0.6015	1.70	3.875	65.90	99	4.868626	10.17632	

2 rows

Hypothesis Test

Paired t-test is used for analysis because: - Coles and Woolworth's product's prices are continuous. - Price of each product is independent of other products prices. - The prices are normally distributed as sample size is >30 i.e. 99 (as per Central Limit Theorem). - Product prices of Coles and Woolworth's are recorded for the same number of products and hence we had equal number of observations/pairs.

$$H_0 : \mu_\Delta(Coles.Price - Woolworths.Price) = 0$$

$$H_A : \mu_\Delta(Coles.Price - Woolworths.Price) \neq 0$$

$$\alpha\ (significance.level) = 0.05$$

Hide

```
Price_Test<-t.test(Data_Mart$Coles.Price,Data_Mart$Woolsworth.Price, paired=T)
Price_Test
```

```
    Paired t-test

data:  Data_Mart$Coles.Price and Data_Mart$Woolsworth.Price
t = 2.7334, df = 98, p-value = 0.007437
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1189927 0.7495528
sample estimates:
mean of the differences
              0.4342727
```

Summary Statistics of Paired t-test

Hide

```
Tabulated_t_statistic<-qt(0.975,df=Price_Test$parameter)
Test_t_statistic<-round(Price_Test$statistic,3)
Test_p_value<-round(Price_Test$p.value,3)
Lower_bound_interval<-round(Price_Test$conf.int[1],3)
Upper_bound_interval<-round(Price_Test$conf.int[2],3)
Test_Summary<-data.frame(Test_p_value,Test_t_statistic,Tabulated_t_statistic,Lower_bound_inte
rval,Upper_bound_interval)
row.names(Test_Summary)<-"=>"
names(Test_Summary)<-c("p Value","Test t-statistic","Tabulated t-statistic","Lower Bound
CI","Upper Bound CI")
Test_Summary
```

	p Value <dbl>	Test t-statistic <dbl>	Tabulated t-statistic <dbl>	Lower Bound CI <dbl>	Upper Bound CI <dbl>
=>	0.007	2.733	1.984467	0.119	0.75

1 row

Interpretation

The research question is that “Which out of the two supermarkets viz. Coles and Woolworths is cheaper?”And to answer this question we used Hypothesis Testing. Our null hypothesis was “The product prices in Coles and Woolworths are equivalent” and the alternative hypothesis stated that “The product prices in Coles and Woolworths are not equivalent”. We decided on to use Paired T-test as we found it to be the best fit to draw the correct interpretation from our data.

Following the procedure of paired t-test we came up with the following observations:

- Firstly, p-value came out to be 0.007 which is <0.05 stating that if we conclude that population prices are different then we only have a 0.7% chance of making an error.
- Secondly, the mean of price difference (=0) does not lie in the 95% confidence interval (0.119-0.749).
- Thirdly, our test’s t-statistics value (2.73) is greater than the tabulated t-statistics value (1.98). Hence, lies in the rejection region.

Based on these three outcomes we have got statistically significant evidence in support of rejecting the null hypothesis and considering the alternative hypothesis i.e. the product prices in Coles and Woolworths are not equivalent(mean price difference = 0.434)

Discussion

Findings & Conclusion

Product prices in Coles and Woolworths are not equivalent and as the mean of product prices at Woolworths is lower than the mean of product prices at Coles, consequently it is concluded that Woolworths is cheaper than the Coles.

Strengths

- For creating the simple random sample we were successful in targeting the whole population of products by assigning each of them a unique product ID and then randomly choosing 99 items from the whole list without replacement, representing our sample.
- We used paired t-test which helped us detect differences that the (unpaired) t-test might have missed. Also statistically the paired t-test has greater power than the normal t-test and hence, our final interpretation holds lower probability of error.

Limitations Though we had whole population at our disposal and we attempted to choose all product Ids at random but we had to skip the product Ids associated with the products which were either in special category or home brand products or which were available in Coles but not in Woolworths. And to make up for it we selected the next product in line.

Scope for future investigation

We can segregate the products according to their prices, brands and units. As in, we can hold separate study for the low, medium and high valued products or we can do it for the products of different brands, or the study could even focus on the comparison of products measured in volume, weight or piecewise. This would give us an eagle eye view of the pricing structure of both supermarkets, providing not just overall difference in prices but the details regarding which categories and products hold that difference.