Code ▾

# MATH1324 Assignment 3

### *Supermarket Price Wars*

# Group/Individual Details

- Abhay Rao (s3649215)
- Dilip Chandra (s3574580)
- Saurabh Mallik (s3623575)

# Executive Statement

Supermarkets today, are raising the bar of price competitiveness. Each store is committing to offer lower prices on products across various categories. To gain some insight into the price war between Coles and Woolworths, we have collected random sample data (90 matched products across three segments ??? Beverages, Health & Beauty and Snacks & Confectionaries) over a period of one day (May 7th, 2017).

To carry out this investigation, we browsed Coles??? & Woolworths??? websites and collected prices of 30 matched products from each of the three categories by using simple random sample technique.

We used Google???s random number generator between 1 to 50. For every random number generated, we check the product available in Woolworths??? website in that category & then matched the same item on the Coles website. After numerous attempts, we were then able to get a total of 90 matched products & created a dataset.

The dataset created from this investigation consisted of four variables ??? Category, Store, Product and Price (In Australian dollars).

We used summary statistics with R Studio functions and boxplot to understand the variability of prices within the categories. Further, we used the help of QQ Plots and T-Test to get insights as to be able to understand whether a statistical significance in difference of means existed.

From the statistical summaries and analyses it was evident that there was no statistically significant difference in the means (as $p$ value was greater than 0.05) and the null hypothesis of the t-test failed to get rejected.

In conclusion, the t-test conducted, fails to reject the hypothesis, hence we cannot determine which supermarket is cheaper.

# Load Packages and Data

1. Loading packages to initiate data loading and cleaning.

Hide

```
library(readr)
library(dplyr)
library(car)
```

2. The dataset created on excel (csv format) is to be loaded in order to undertake the hypothesis testing. Next, the dataset is viewed to check for correct input values.

Hide

```
PWD <- read_csv("~/Desktop/Assignment 3/PWD.csv")
View(PWD)
```

3. Since, three seperate hypothesis testing needs to be undertaken for the three different categories (“Health and Beauty”, “Beverages”" and “Snacks & Confectioneries”), we subset the dataset into three separate dataframes.

Hide

```
PW_Health <- PWD %>% filter(Category == "Health and Beauty")
PW_Bev <- PWD %>% filter(Category == "Beverages")
PW_Snk <- PWD %>% filter(Category == "Snacks & Confectioneries")
```

Since the sample data was collected and entered manually, there were no null values in the dataset, nor any outliers, hence the dataframes are tidy and ready to be analysed.

# Summary Statistics

In order to proceed to hypothesis testing, we first need to summarise the subsetted dataframes based on stores for each category that needs to be tested.

1. Firstly, the subsetted data for Beverages needs to be summarised in order to figure out the minimum, maximum, mean etc, along with an appropriate plot.

Hide

```
PW_Bev %>% group_by(Store) %>% summarise(Min = min(Price),
                                         Q1 = quantile(Price, probs = .25),
                                         Median = median(Price),
                                         Q3 = quantile(Price, probs = .75),
                                         Max = max(Price),
                                         Mean = mean(Price),
                                         SD = sd(Price),
                                         Count = n())

boxplot(Price ~ Store, data = PW_Bev)
```

2. Next, the subsetted data for “Health and Beauty”" needs to be summarised in order to figure out the minimum, maximum, mean etc, along with an appropriate plot.

Hide

```
PW_Health %>% group_by(Store) %>% summarise(Min = min(Price),
                                         Q1 = quantile(Price, probs = .25),
                                         Median = median(Price),
                                         Q3 = quantile(Price, probs = .75),
                                         Max = max(Price),
                                         Mean = mean(Price),
                                         SD = sd(Price),
                                         Count = n())

boxplot(Price ~ Store, data = PW_Health)
```

3. Finally, the subsetted data for “Snacks & Confectioneries” needs to be summarised in order to figure out the minimum, maximum, mean etc, along with an appropriate plot.

Hide

```
PW_Snk %>% group_by(Store) %>% summarise(Min = min(Price),
                                         Q1 = quantile(Price, probs = .25),
                                         Median = median(Price),
                                         Q3 = quantile(Price, probs = .75),
                                         Max = max(Price),
                                         Mean = mean(Price),
                                         SD = sd(Price),
                                         Count = n())

boxplot(Price ~ Store, data = PW_Snk)
```

4. Next after filtering the subsets based on stores, we test for normality for each category.

Hide

```
PW_Bev_Coles <- PW_Bev %>% filter(Store == "Coles")
PW_Bev_Coles$Price %>% qqPlot(dist = "norm", main = "Coles Beverages", ylab = "Price")

PW_Bev_Wool <- PW_Bev %>% filter(Store == "Woolworths")
PW_Bev_Wool$Price %>% qqPlot(dist = "norm", main = "Woolworths Beverages", ylab = "Price")

PW_Health_Coles <- PW_Health %>% filter(Store == "Coles")
PW_Health_Coles$Price %>% qqPlot(dist = "norm", main = "Coles Health and Beauty", ylab = "Price")

PW_Health_Wool <- PW_Health %>% filter(Store == "Woolworths")
PW_Health_Wool$Price %>% qqPlot(dist = "norm", main = "Woolworths Health and Beauty", ylab = "Price")

PW_Snk_Coles <- PW_Snk %>% filter(Store == "Coles")
PW_Snk_Coles$Price %>% qqPlot(dist = "norm", main = "Coles Snacks & Confectioneries", ylab = "Price")

PW_Snk_Wool <- PW_Snk %>% filter(Store == "Woolworths")
PW_Snk_Wool$Price %>% qqPlot(dist = "norm", main = "Woolworths Snacks & Confectioneries", ylab = "Price")
```

From the qqplots we see can infer that “Beverages” and “Health and Beauty” segments are not normally distributed, whereas “Snacks & Confectioneries” is normally distributed with a right skew.

The next step would be to undertake hypothesis testing for the three categories.

# Hypothesis Test

The hypothesis test that will be used to determine which supermarket is cheaper will be “Two-Sample T-Test”. The significance level ($\alpha$) of 0.05 has been used to reject/not reject the null hypothesis, assuming, there is no variation in mean price in the supermarket over one day (As data collected is for one day i.e. 7th May, 2017).

i. $H_0$ : $\mu$Coles - $\mu$Woolworths = 0
ii. $H_A$ : $\mu$Coles - $\mu$Woolworths $\neq$ 0

A. T-Test for Beverages category

Hide

```
t.test(Price ~ Store, data = PW_Bev)
```

B. T-Test for Health and Beauty category

Hide

```
t.test(Price ~ Store, data = PW_Health)
```

C. T-Test for Snacks & Confectioneries category

Hide

```
t.test(Price ~ Store, data = PW_Snk)
```

# Interpretation

Based on the hypothesis tests, the following are the interpretations:

a. For Beverage category, estimated difference between means (8.689 - 7.543) is .627 (i.e. $p$ value), hence $p$ value is greater than 0.05 and hence we fail to reject $H_0$.

b. For Health and Beauty category, estimated difference between means (9.980 - 9.067) is .451 (i.e. $p$ value), hence $p$ value is greater than 0.05 and hence we fail to reject $H_0$.

c. For Snacks and Confectioneries category, estimated difference between means (4.234 - 3.812) is .417 (i.e. $p$ value), hence $p$ value is greater than 0.05 and hence we fail to reject $H_0$.

# Discussion

From the above summary statistics, analysis and interpretation we find that over the three categories - “Beverages”, “Health and Beauty” and “Snacks and Confectioneries”- the results of the study, a statistically significant mean difference between Coles and Woolworths prices was not found.

a. For beverages: $t(df = 53.935)$. $p = .627$, 95% CI [-3.557, 5.849]. Woolworths although cheaper slightly based on plot comparisons, doesn’t show a statically siginificant difference.

b. For health and beauty: $t(df = 57.643)$. $p = .451$, 95% CI [-1.494, 3.320]. Woolworths although cheaper slightly based on plot comparisons, doesn’t show a statically siginificant difference.

c. For snacks & confectioneries: $t(df = 57.883)$. $p = .417$, 95% CI [-0.612, 1.455]. Woolworths although cheaper slightly based on plot comparisons, doesn’t show a statically siginificant difference.

Since the data was collected over only one day and it consisted of only 30 randomly selected matched products over three categories, the study was extremely limited in scope.

With supermarket product prices varying everyday and with special discounts and offers, our dataset was limited in respect to cohesive values.

In order to get a better idea of which supermarket is cheaper, a detailed investigation across all categories will need to be conducted over numerous days. This will help get a mean price of every product on offer in the different supermarkets and provide better numbers to work with.