

# MATH1324 Assignment 2

## Supermarket Price Wars

### Group Details

- Ashley Mallia (s3773716)
- Daniel Evans (s3766654)

### Executive Statement

**Aim:** The aim of this investigation was to determine if there was a statistically significant difference in the mean price between products sold at coles and woolworths. Or to put it more simply, which major supermarket is cheaper.

**Sample:** To obtain price information for identical products between the two companies, we wrote a javascript program to download the entire website <http://www.grocerycop.com.au> (<http://www.grocerycop.com.au>). This form of convenience sampling allowed us to gather sample data from the population in a fast and efficient manner.

The variables in the data were page number: item number, category, title, company, price and price difference.

We obtained 5552 paired samples across 10 different categories: Fridge, Bakery, Fruit-vegetables, Pantry, Freezer, Baby-health-beauty, Meat-seafood, Entertainment-international-food, Clothing-household-pet, Tobacco-drinks

**Procedure:** After some small manipulations and tidying of the dataset, we performed a paired samples t-test to determine if there was a statistically significant difference between the mean price of products. Our null hypothesis assumes there is no difference in means, while the alternative hypothesis states the opposite.

#### Main Findings:

When taken as a whole, a paired-samples t-test **did not** find evidence of a statistically significant difference between the prices of identical products sold at Coles and Woolworths.  $t(df=5551) = 1.947$ ,  $p > 0.01$ , 99% CI [-0.02 0.142].

When divided by category, a paired-samples t-test **did** find evidence of a statistically significant difference in 4 categories: drinks-tobacco, entertainment-international-food, baby-health-beauty, pantry

#### Conclusion:

The prices between the Coles and Woolworths are kept close together due to competition. Our statistical analysis did find a statistically significance difference in the prices of products in some categories. Since supermarkets change their prices daily, this analysis does not provide practical significance and we could not use our findings to predict which store would be cheaper on any given day.

### Load Packages and Data

```
library(dplyr)
library(tidyr)
library(readr)
library(purrr)
library(ggplot2)
library(broom)
library(car)
library(stats)
library(latex2exp)
grocerycop <- read_csv("grocerycop.csv", col_types = cols(category = col_factor()))
# Removed all rows that contain NA, leaving only paired data
pairs = na.omit(grocerycop)
```

## Summary Statistics

[Hide](#)

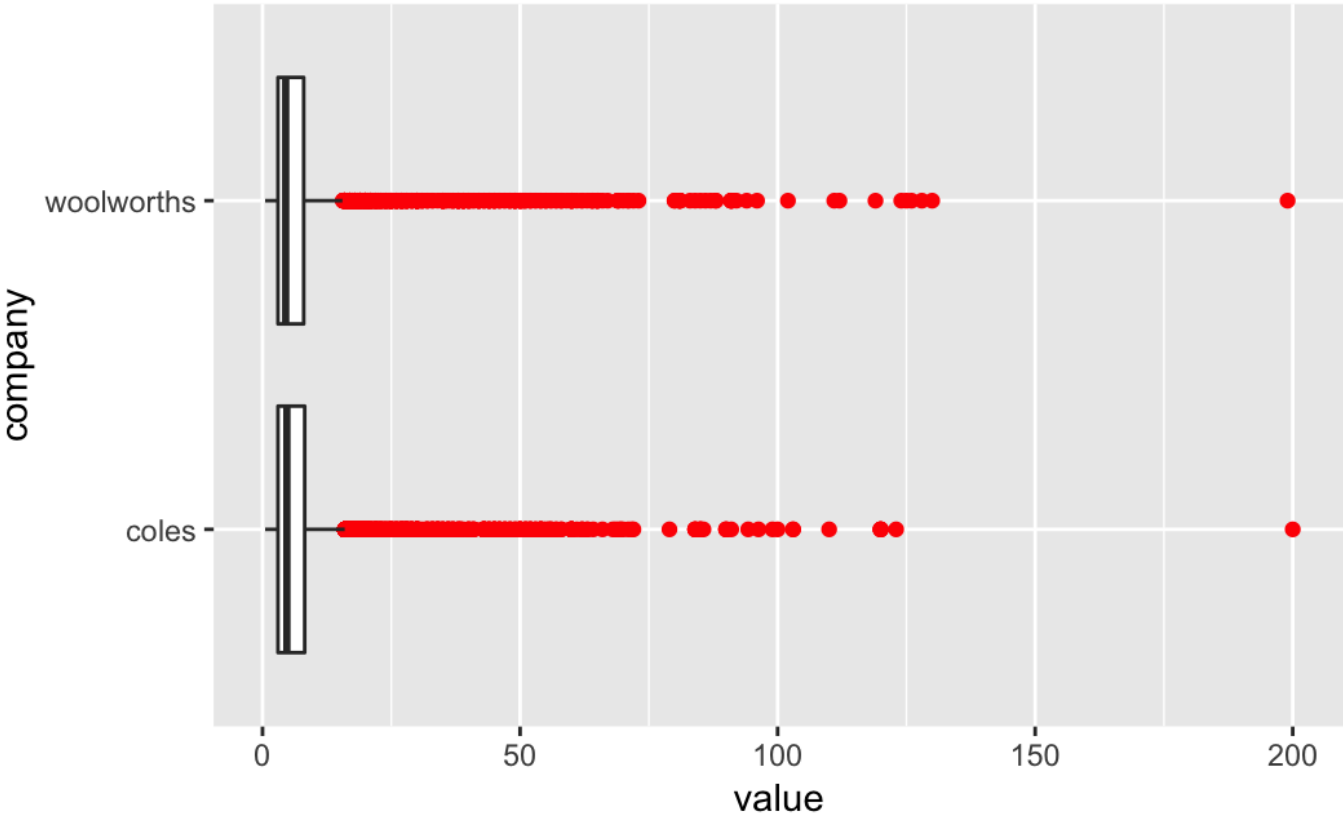
```
g <- pairs %>% gather(company, value, coles:woolworths, factor_key = T)
g %>% group_by(company) %>% summarise(
  N = n(),
  MEAN = mean(value, na.rm = T),
  SD = sd(value, na.rm = T),
  MIN = min(value, na.rm = T),
  Q1 = quantile(value, .25, na.rm = T),
  MEDIAN = quantile(value, .5, na.rm = T),
  Q3 = quantile(value, .75, na.rm = T),
  MAX = max(value, na.rm = T),
  IQR = Q3-Q1,
  LF = Q1 - 1.5*IQR,
  UF = Q3 + 1.5*IQR,
  LOUT = sum(value < LF, na.rm = T),
  UOUT = sum(value > UF, na.rm = T)
) %>% arrange(company) %>% modify_if(is.numeric, round, 2)
```

company	N	M...	SD	...	Q1	MEDI...	Q3	...	IQR	LF	UF	L...	U...
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
coles	5552	7.92	10.96	0.5	3	4.73	8.19	200	5.19	-4.78	15.97	0	565
woolworths	5552	7.86	11.21	0.5	3	4.50	8.00	199	5.00	-4.50	15.50	0	562

2 rows

[Hide](#)

```
ggplot(g, aes(x = company, y = value)) + geom_boxplot(outlier.colour = "red") + coord_
_flip()
```



From the above side-by-side boxplot, it appears that the spread of the data is very similar between coles and woolworths. They both appear to have similar means, interquartile ranges and number of outliers. The summary statistics table above confirms quantitatively that the data for coles and woolworths follow similar trends.

Hide

```
# add the difference column
pairs = pairs %>% mutate(diffcw = coles - woolworths)
summary <- pairs %>% summarise(
  N = n(),
  MEAN = mean(diffcw, na.rm = T),
  SD = sd(diffcw, na.rm = T),
  MIN = min(diffcw, na.rm = T),
  Q1 = quantile(diffcw, .25, na.rm = T),
  MEDIAN = quantile(diffcw, .5, na.rm = T),
  Q3 = quantile(diffcw, .75, na.rm = T),
  MAX = max(diffcw, na.rm = T),
  IQR = Q3-Q1
)
summary
```

N	MEAN	SD	MIN	Q1	MEDIAN	Q3	M...	IQR
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
5552	0.06129503	2.346334	-21.11	-0.05	0	0.39	50	0.44
1 row								

The statistical summary above shows that the total difference in means is  $\Delta = \mu_c - \mu_w = 0.061$  with  $sd = 2.346$ .

In order to determine if this difference is statistically significant, we conduct a paired samples t-test.

# Hypothesis Test

For our paired samples, we want to determine if the difference between the population means is statistically significant. Our null hypothesis assumes there is no statistically significant difference in means, while the alternative hypothesis states the opposite.

$$H_0 : \Delta = 0$$

$$H_A : \Delta \neq 0$$

## Significance Level

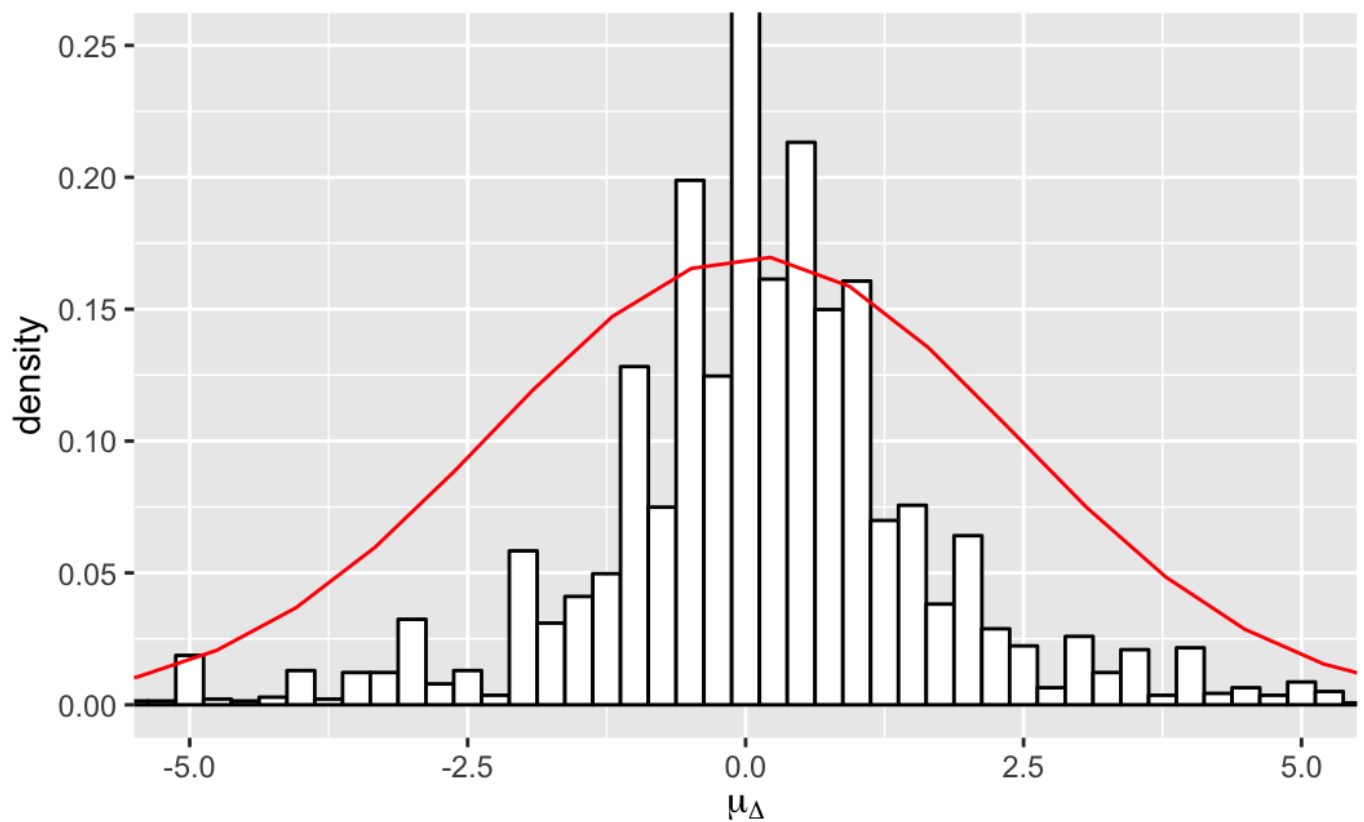
- Due to our large sample size we choose to use a significance level of 0.01

## Assumptions

- This data is continuous
- TRUE
- The sample is representative of a randomly selected portion of the total population
- We expect our sample is representative
- the sample size is large enough for CLT to be effective
- Our sample is large with  $n=5552$
- The difference in the population means is approximately normally distributed
- This histogram below shows that price differences is approximately normally distributed

[Hide](#)

```
ggplot(pairs, aes(x = diffcw)) +
  geom_histogram(aes(y=..density..), binwidth = .25, colour="black", fill="white") +
  coord_cartesian(xlim=c(-5,5), ylim=c(0, .25)) +
  stat_function(fun = function(x) dnorm(x, summary$MEAN, summary$SD), color="red") +
  xlab(TeX("$\\mu_\\Delta$"))
```


[Hide](#)

```
lt.p.value <- leveneTest(value ~ company, data = g, center=mean) %>% tidy %>%
{ .[1,]$p.value } %>% round(3)
```

- The data has homogeneity of variance
- A Levene's test was conducted to determine the assumption of equal variance in our datasets with the following statistical hypothesis:  $H_{0lt} : \sigma_1^2 = \sigma_2^2$
- The p value for Levene's test  $p = 0.898 > 0.05$  therefore we **Fail to Reject**  $H_{0lt}$  meaning we can assume our data has homogeneity of variance.

## Decision Rules

- **Reject**  $H_0$  if
- $p\text{-value} < 0.01$  (the significance level)
- 99% CI of the mean difference does not capture  $H_0 : \Delta = 0$
- Otherwise **Fail to Reject**  $H_0$

## Paired T-Test

[Hide](#)

```
alpha = 0.01
result = pairs %>% nest() %>%
  mutate(
    out = map(data, ~ t.test(.x$coles, .x$woolworths, paired = T, conf.level = 1-alpha) %>%
      tidy %>% rename(mean = estimate))
  ) %>%
  unnest(out, .drop = T) %>%
  mutate(alpha = alpha,
    conf.level = (1-alpha)*100,
    ptest = ifelse(p.value < alpha, paste0("p < ",alpha), paste0("p > ",alpha)),
    htest = ifelse(p.value < alpha, "Reject H0", "Fail to Reject H0")) %>%
  arrange(p.value) %>%
  modify_if(is.numeric, round, 3)
result %>% select(mean, p.value, ptest, htest, conf.level, conf.low, conf.high)
```

mean	p.value	ptest	htest	conf.level	conf.low	conf.high
<dbl>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
0.061	0.052	p > 0.01	Fail to Reject H0	99	-0.02	0.142

1 row

## Interpretation

A paired-samples t-test was used to test for a statistically significant difference between the prices of identical products sold at Coles and Woolworths.

The mean difference between Coles and Woolworths prices was found to be 0.061

All assumptions for using a t test were tested. A histogram was used to test normality and Levene's test was used to test homogeneity of variance.

The paired-samples t-test results are:

- $t(df=5551) = 1.947$ ,  $p > 0.01$ , 99% CI [-0.02 0.142].
- Given these findings we **Fail to Reject H0**

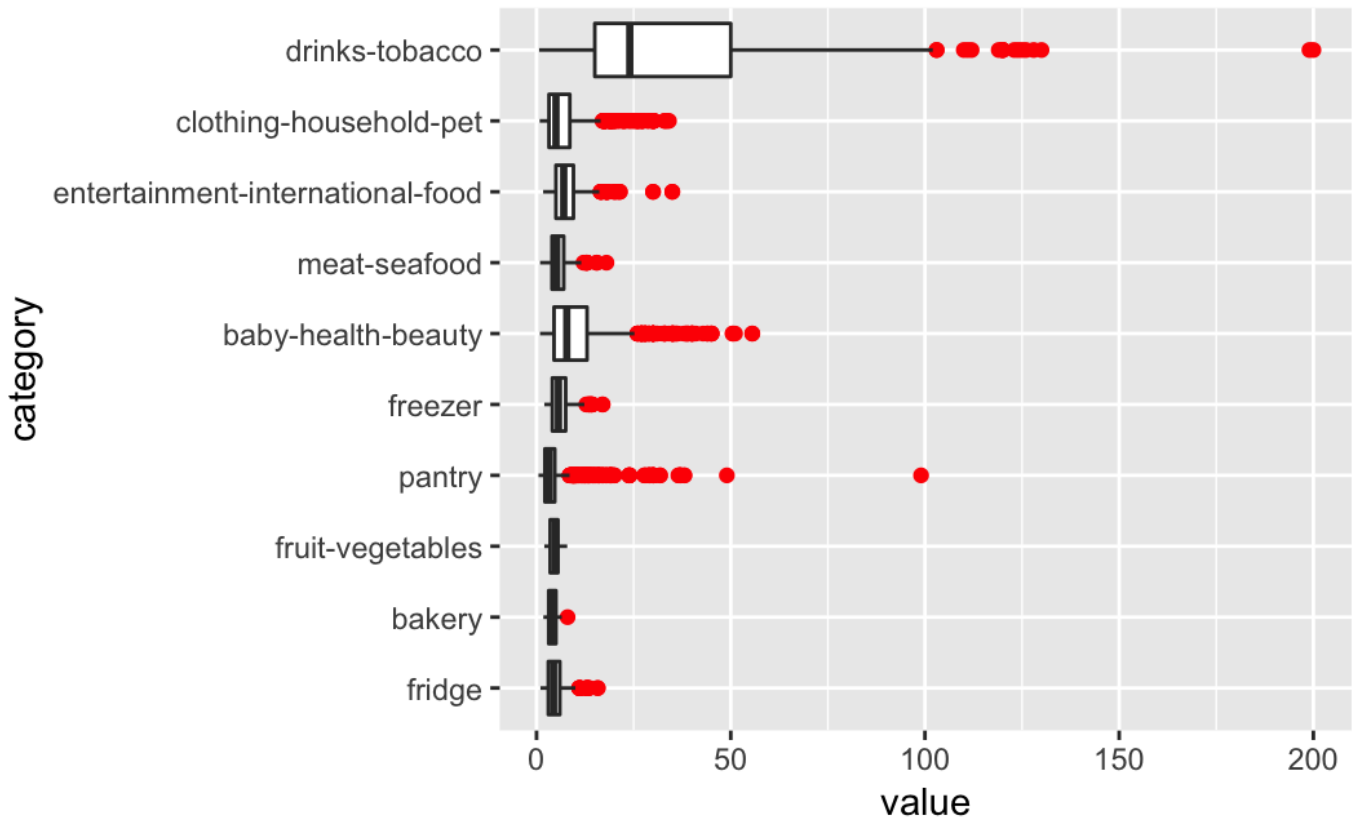
When our whole sample was analysed, the paired-samples t-test **did not** find evidence of a statistically significant difference between the prices of identical products sold at Coles and Woolworths.

## Further Interpretation (by category)

When the data was taken as a collective, we did not find a statistically significant difference. We decided to perform this analysis again, this time grouping the data by Category.

Hide

```
ggplot(g, aes(x = category, y = value)) + geom_boxplot(outlier.colour = "red") + coord_flip()
```


[Hide](#)

```
result2 = pairs %>% group_by(category) %>% nest() %>%
  mutate(
    out = map(data, ~ t.test(.x$coles, .x$woolworths, paired = T, conf.level = 1-alpha
a) %>%
      tidy %>% rename(mean = estimate, df = parameter))
  ) %>%
  unnest(out, .drop = T) %>%
  mutate(alpha = alpha,
    conf.level = (1-alpha)*100,
    ptest = ifelse(p.value < alpha, paste0("p < ",alpha), paste0("p > ",alpha)),
    htest = ifelse(p.value < alpha, "Reject H0", "Fail to Reject H0")) %>%
  arrange(ptest, desc(abs(mean))) %>%
  modify_if(is.numeric, round, 3)
result2 %>% select(category, df, mean, p.value, ptest, htest, conf.level, conf.low, c
onf.high)
```

category <fctr>	df <dbl>	mean <dbl>	p.value <dbl>	ptest <chr>	htest <chr>
drinks-tobacco	343	-1.131	0.000	p < 0.01	Reject H0
entertainment-international-food	222	0.487	0.006	p < 0.01	Reject H0
baby-health-beauty	1331	0.236	0.001	p < 0.01	Reject H0
pantry	2368	0.090	0.007	p < 0.01	Reject H0
fruit-vegetables	8	-0.526	0.142	p > 0.01	Fail to Reject H0
meat-seafood	41	-0.154	0.447	p > 0.01	Fail to Reject H0
freezer	259	0.145	0.089	p > 0.01	Fail to Reject H0

category <fctr>	df <dbl>	mean <dbl>	p.value <dbl>	ptest <chr>	htest <chr>
bakery	40	-0.126	0.339	p > 0.01	Fail to Reject H0
clothing-household-pet	572	0.080	0.254	p > 0.01	Fail to Reject H0
fridge	358	0.074	0.073	p > 0.01	Fail to Reject H0

1-10 of 10 rows | 1-7 of 9 columns

The results of the table above will be discussed in the next section

## Discussion

### What did we conclude?

Using a paired-samples t-test we determined:

- when our sample was analysed as a whole, we **did not** find evidence for statistically significant price differences

When our sample was analysed by category:

- in 8 categories we **did not** find a statistically significant difference between the prices
- in 4 categories we **did** find a statistically significant difference between the prices

[Hide](#)

```
result2 %>%
  filter(p.value < alpha) %>%
  mutate(conc = ifelse(mean < 0, "coles was on average cheaper than woolworths", "woolworths was on average cheaper than coles"),
         conclusions = paste0(conc, " by $", abs(mean %>% round(2)))) %>%
  arrange(desc(abs(mean))) %>%
  select(category, conclusions)
```

category <fctr>	conclusions <chr>
drinks-tobacco	coles was on average cheaper than woolworths by \$1.13
entertainment-international-food	woolworths was on average cheaper than coles by \$0.49
baby-health-beauty	woolworths was on average cheaper than coles by \$0.24
pantry	woolworths was on average cheaper than coles by \$0.09

4 rows

### What were the strengths and limitations of your investigation?

A strength of our investigation was the sample size. As we were able to write a java script to scrape the entire data from the grocerycop website, we were able to analyse a very large sample size. This lead to a more accurate representation of the population and allowed us to assume the data was normally distributed when conducting the paired samples t-test.

We also performed an analysis by category using very efficient code.

### What improvements could be made or what would you do differently next time?



In our analysis we set the value of  $\alpha$  (the probability of Type I error) to a low value. Since  $\alpha = \Pr(\text{Reject } H_0 \mid H_0 \text{ is true})$  is the probability that we detect an effect (Reject  $H_0$ ) when an effect does not exist ( $H_0$  is true).

There is another type of error we did not consider,  $\beta$  (the probability of Type II error). Since  $\beta = \Pr(\text{Fail to Reject } H_0 \mid H_0 \text{ is false})$  is the probability that we detect no effect (Fail to Reject  $H_0$ ) when an effect does exist.

Statistical power is defined as  $\text{power} = 1 - \beta$ . We want high power so that our hypothesis test is good at detecting a false null hypothesis.

Below I have run a power test to determine what the power level for our t.test is:

[Hide](#)

```
ptt <- power.t.test(power = NULL, n = nrow(pairs), delta = summary$MEAN, sd = summary$SD, sig.level = alpha, type = "paired", alternative = "two.sided")
ptt2 <- power.t.test(power = 0.8, n = NULL, delta = summary$MEAN, sd = summary$SD, sig.level = alpha, type = "paired", alternative = "two.sided")
```

For our sample size of  $n=5552$  (pairs) the  $\text{power}=0.264$  meaning that our test was **underpowered**. Therefore, the t.test would not have been able to reliably detect false null hypothesis, given the very small number of cents of difference we were trying to detect.

In order to achieve  $\text{power}=0.8$ , which is typically used in studies, our sample size would need to be increased to  $n=17117$  (pairs) in order for the t.test to be able to reliably detect false null hypothesis.