UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING

SESSION 2019/2020 SEMESTER 2

PROBABILITY AND STATISTICAL DATA ANALYSIS (SECI2143-02)

PROJECT 2: STARBUCK'S BEVERAGE ANALYSIS

NAME: AHMAD ZULHAKIM BIN ZAINAL (A19EC0179)

LECTURER: CHAN WENG HOWE

Contents

Introduction

Handcrafted drinks are famous these days. It can be proved by the blooming beverages' industry such as Tealive, Chatime, and Starbucks branching their premises in various place in the nation. It manages to attract customers with their variety of flavoursome drinks even with their slightly higher price compared to local store drinks. Since these drinks have become a part of people's life, it is important for them to know what they are consumed in term of nutrition because it is commonly known that these type of drinks have high amount of sugar in every serving etc. Given that sugar is one of the sources of calories, addiction to these types of drinks can affect their health and brings diseases. Besides, people should also take note of the other nutrition it contained. So, this project's objective is to investigate these beverages' calories content and the ingredients that may contribute to the factor. The sample will be taken from the Starbucks' Drink Menu to be analysed.

Starbucks' Beverage Nutritional Facts

Toward analysing the nutritional facts of handcrafted drinks, the dataset of Starbucks' Beverages will be used. This data was collected and published by Starbucks on its public domain. The purpose of the dataset is to illustrates the nutritional information for Starbucks' drink menu items. All nutritional information for drinks is for the 12oz serving size, and consisting drinks from 5 types of beverages, each with different preps. The total sample size is 224.

Table 1 illustrates the variables in the dataset

Name of Variables	Type of Data
Beverage	Nominal
Beverage Prep	Nominal
Calories	Ratio
Total Fat	Ratio
Trans Fat	Ratio
Saturated Fat	Ratio
Sodium	Ratio
Total Carbohydrate	Ratio
Cholesterol	Ratio
Dietary Fibre	Ratio
Sugars	Ratio
Protein	Ratio
Vitamin A	Ratio
Vitamin C	Ratio
Calcium	Ratio
Iron	Ratio
Caffeine	Ratio

In this analysis, several variables will be used for specific test which are **Beverage, Sugars, Calories, Saturated Fat,** and **Caffeine.** Four different statistical tests will be done between these variables.

Table 2 shows the statistical test and its objectives

Statistical Test	Objectives
Hypothesis Testing on One Sample	To test whether the population mean for Calories' content of targeted population is equal to 150 Calories with confidence level of 95%
Correlation	To see the relationship between Calories and Caffeine (mg)
Regression	Regression of Calories and Sugar (g)
Chi Square Test of Independence	To test the dependency between Beverage and Saturated Fat (g)

Analysis of The Beverages

The Population Mean for Calories

The method that was used to test the claimed that the population mean for Calories content of the Starbucks' drinks is equal to 150 is hypothesis testing on one sample with significance level of 0.05. A z-test statistic will used since we want to test the population mean.

Hypothesis Statement:

H_0: $\mu_0 = 150$ H_1: $\mu_1 \neq 150$

The analysis is done by using R as below:

```
> cal=starbucks_menu$Calories
> s=sd(cal)
> mu=150
> n=242
> alpha=0.05
> xbar=mean(cal)
> z=(xbar-mu)/(s/sqrt(n))
> alpha=0.05
> z.alpha=qnorm(1-(alpha/2))
> c(-z,z)
```

After calculating the test statistic and the p-value in R, the results are:

Z=6.34889 P-value= (-1.9510, 1.9510)

Since the z=6.34889 exceed the p-value, we reject the null hypothesis. There is sufficient evidence to support that population mean for Calories content of Starbucks' drinks is not equal to 150.

Correlation between Calories and Caffeine

In this part, we want to see the correlation between Caffeine and the Total Fat (g) of the beverages. Let x=Calories and Y=Caffeine, and calculate the correlation coefficient and plot the data as Scatter Diagram in R. The analysis is as below:

After calculating the correlation coefficient using R, the result as below:

```
> alpha=0.05
> x<-starbucks_menu$Calories
> y<-starbucks_menu$`Caffeine (mg)`
> plot(x,y,xlim=c(0,600), ylim=c(0,500),xlab="Calories",ylab="Caffeine (mg)")
> r=cor(x,y)
> t=r/(sqrt((1-(r*r))/(n-2)))
> t.alpha<-qt((alpha/2),df=n-2,lower.tail = FALSE)
> c(-t.alpha,t.alpha)
```

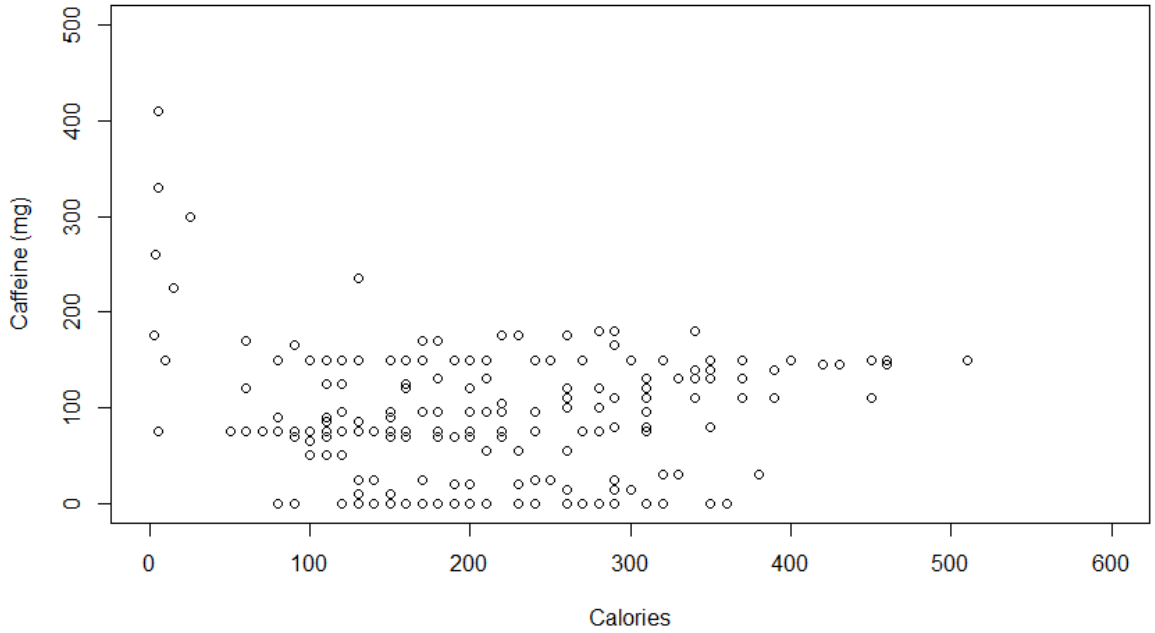


Figure 1 shows the correlation between Calories and Caffeine

Result:

Correlation coefficient=0.0503

The correlation coefficient is close to 0, meaning that it has weaker linear relationship.

Due to the low correlation coefficient, a significance test for correlation is done to check whether these variables have linear relationship at 0.05 significance level.

Hypothesis Statement:

H_0: ρ= 0 (no linear correlation) H_1: ρ≠ 0 (linear correlation exist)

```
> alpha=0.05
> t=r/(sqrt((1-(r*r))/(n-2)))
> t.alpha<-qt((alpha/2),df=n-2,lower.tail = FALSE)
> c(-t.alpha,t.alpha)
```

Results:

T-statistic=0.7807

p-value=(-1.9699, 1.9699)

Since the T-statistic=0.7807 does not fall in the critical region, p-value= (-1.9699, 1.9699), we failed to reject the null hypothesis. There is sufficient evidence to support that the Calories and Caffeine are not correlated to each other.

Regression between Calories and Sugar

Since sugar is one of main source of calories, we want to prove that the number of calories is affected by the sugar amount. Let x=Sugars, and y=Calories. Then, we want to decide if these variables have significant relationship.

Hypothesis Statement:

H_0: $\beta_0 = 0$ (no linear correlation) H_1: $\beta_1 \neq 0$ (linear correlation exist)

Analysis using R:

```
> x2<-starbucks_menu$`Sugars (g)`
> y2<-starbucks_menu$Calories
> model<-lm(y2~x2)
> model
> plot(x2,y2,xlim=c(0,80), ylim=c(0,500),xlab="Sugars(g)",ylab="Calories")
> abline(model)
> summary(model)
```

Output when summary(model) is called:

```
Call:
lm(formula = y2 ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-79.930 -33.424  -8.424  24.033 121.506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.5429     5.3665   6.996 2.61e-11 ***
x2            4.7426     0.1398  33.932  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.81 on 240 degrees of freedom
Multiple R-squared:  0.8275,      Adjusted R-squared:  0.8268
F-statistic:  1151 on 1 and 240 DF,  p-value: < 2.2e-16
```

Results:

y-intercept=37.5429

m=4.7426

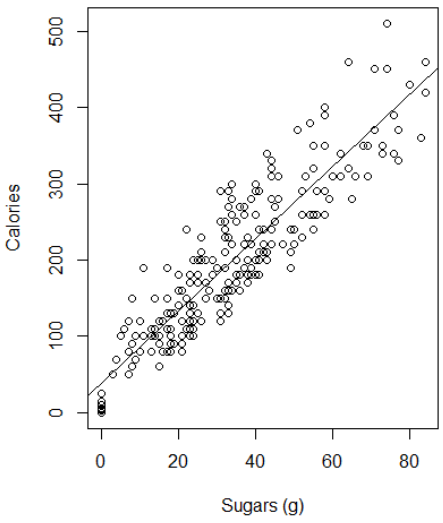
R^2=0.8275

p-value=2.2x10^{-16}

Since p-value is smaller than 0.05 significance level, thus we rejec t the null hypothesis. So, there is sufficient evidence to support th at there is linear correlation between sugar and calories.

But based on the results, it seems that R^2 =0.8275 which is in the range of 0 - 1. This portrays that it has a weaker linear relationshi p and some but not all the variation in Calories are explained by v ariation in Sugars.

Figure 2 shows the relationship between Sugars and Calories



Dependency of Beverage and Saturated Fat

Now, we want to test whether saturated fat is independent towards beverage at 0.05 significance level.

Hypothesis Statement:

H_0: Saturated fat is independent to Beverage H_1: Saturated Fat is related to Beverage

Analysis using R:

```
> library(plyr)
> sed<-subset(starbucks_menu, `Beverage_category`=='Signature Espresso
Drinks',select=c(Beverage:`Caffeine (mg)`))
> beverage<-count(sed$Beverage)
> names(beverage)[1]='beverage'
> saturated_fat<-count(sed$`Saturated Fat (g)`)
> names(saturated_fat)[1]='saturated_fat'
> d<-data.frame(beverage, saturated_fat)
> d<-data.frame(beverage$freq, saturated_fat$freq)
> chisq.test(d, correct=FALSE)
> alpha2=0.05
> x2.alpha<-qchisq(alpha2, df=3, lower.tail = FALSE)
```

Results:

X^2 = 32.19

critical value = 7.8147

Since X^2 = 32.19 is larger than p-value=7.8147, we reject the null hypothesis. There is sufficient evidence to support that saturated fat is related to beverage.

Conclusion

Post-analysis process, some of our assumption are proved to be right while the rest are rejected. It is shown that the population mean for Calories content of Starbucks' drinks is not equal to 150. Besides, there is sufficient evidence to support that Calories and Caffeine have a weak linear relationship. The result is further proved through T-statistic test which shown that the Calories and Caffeine are not correlated to each other. Although Calories does not relate to Caffeine, it does have a significant relationship with Sugar. From the regression analysis, there is evidence to support that there is linear correlation between Sugar and Calories although the linear relationship is not strong enough. Finally, saturated fat has been proved to relate with beverage after conducting the chi-square test of dependency.

Appendix

Source of dataset:
https://www.kaggle.com/starbucks/starbucks-menu?select=starbucks_drinkMenu_expanded.csv