UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI 2143

Probability and Statistical Data Analysis

Project 2

NAME: ASWIND SARVANESH VARMAN A/L SARAVANAN

MATRIC NUMBER: A19EC0025

LECTURER'S NAME: Dr. Azurah Abu Samah

Content	Page

Introduction

The case study is about cereals and its nutritional values that it holds inside like sugar, protein, carbohydrates and so on. It also has been segregated into whether it can be served hot or cold. Apart from its nutritional values, there is also a shelf option where we can where see which level it is being stored at. Also, there are also weight and cups variables, which is the weight in ounces and number of cups in one serving. Lastly, all the cereals are rated and given a fair review.

The data were collected by students of Paul Velleman at Cornell University. More specifically, it was done at a local Wegmans supermarket. The datasets then have been gathered and cleaned up by Petra Isenberg, Pierre Dragicevic and Yvonne Jansen.

I carried out my inferential statistics using the dataset provided. For hypothesis 1-sample, I took the ratings variable and tested my hypothesis. For correlation and regression model, however, two variables that is sugar and calories were chosen.

After that, I did my Chi Square test of independence on fat and weight variables. I wanted to test the popular claim that states that cereal's weight depend on fat only. So, testing my hypothesis of these two variables being independent or not, helped me to identify the truth behind this myth.

Hypothesis 1 sample

In the dataset provided, the sample mean rating of 77 cereals are 42.67 %. The population standard deviation is 14.05. I used a 0.05 significance level to test the claim that the mean rating of the 77 cereals is less than 50 %.

H0: $\mu = 50$ %
H1: $\mu < 50$ %
$\alpha = 0.05$
c.v = $z_{0.05} = -1.645$
$z = (\chi - \mu) / (\sigma/ \sqrt{n}) = (42.67 - 50)/ (14.05/ \sqrt{78}) = -4.611$
P-value = P (z < -4.611) = 0.0001

Conclusion: For P-value method, since 0.0001 < 0.05, and z-value falls within the rejection region, we reject H0. There is sufficient evidence to conclude that the mean rating of 77 cereals is lower than 50 %.

Below are the calculations done in R programming.

Console | Terminal | Jobs

C:/Users/Aswind S Varman/Desktop/Semester 2/PSDA/Project 2/Project 2 R coding/

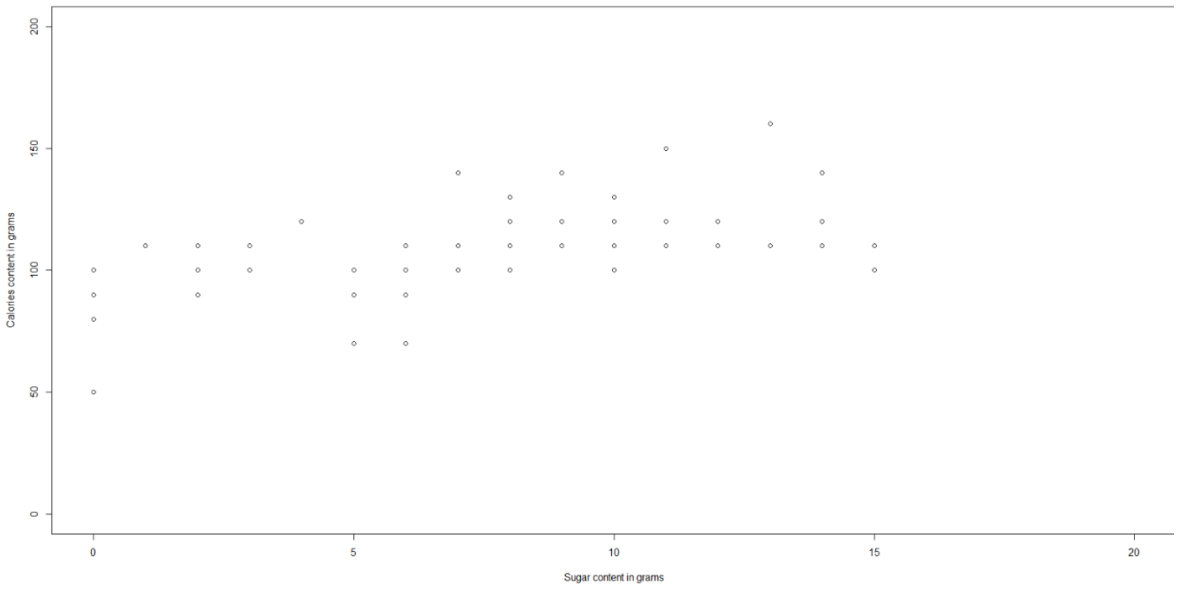
```
> ## Hypothesis 1 sample
> n = 77
> sd = sd(rating)
> xbar = mean(rating)
> mu = 50
> alpha = 0.05
>
> ##Calculate z test statistics
> z = (xbar - mu)/(sd/sqrt(n))
> z.alpha = qnorm(1-alpha)
> z
[1] -4.581537
> z.alpha
[1] 1.644854
>
> ## P-value for calculated test statistics
> pval = pnorm(z, lower.tail = TRUE)
> pval
[1] 2.307851e-06
>
```

Correlation

For correlation, I wanted to test the relationship between the sugar variable and calories variable of the 77 cereals. Since both the data type is ratio, I used the Pearson's technique.

From R programming, the value of r is 0.5623, which is a relatively strong positive linear correlation association between sugar and calories variable. A scatter plot is plotted to show the relationship between the two variables.

```
Console  Terminal  Jobs
C:/Users/Aswind S Varman/Desktop/Semester 2/PSDA/Project 2/Project 2 R coding/
> ## Calculating corr. coefficient
> x <- sugars
> y <- calories
> cor(x,y)
[1] 0.5623403
> plot(x,y, xlim = c(0,20), ylim = c(0,200), xlab = "Sugar content in gram
s", ylab = "Calories content in grams")
>
```



Scatter plot of sugar content in grams against the calories content in grams

Hypotheses:

H0: $p = 0$ (No linear correlation)

H1: $p \neq 0$ (Linear correlation exists)

Test statistic:

$t = r / (\sqrt{1-r^2/n-2}) = = 0.5623 / (\sqrt{1-0.5623^2/77-2}) = 5.889$

$\alpha = 0.05$

t-value (from table): +/- 1.9921

Since test statistics value > t-value from table, we reject H0.

There is enough evidence to conclude that there is a linear relationship between sugar content and calories content at the 5% level of significance.

Regression

For linear regression model, I took the same scatter plot of the sugar content in grams against the calories content in grams. I wished to test to see the relationship between the sugar variable and calories content using regression.

```
> plot(x,y, xlim = c(0,20), ylim = c(0,200), xlab = "Sugar content in grams", ylab = "Calories content in grams")
>
> ## Linear Regression
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     89.820        2.465

> abline(model)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-39.82  -9.40   0.46  12.64  38.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.8201     3.4366  26.137  < 2e-16 ***
x             2.4650     0.4185   5.889 1.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.22 on 75 degrees of freedom
Multiple R-squared:  0.3162,    Adjusted R-squared:  0.3071
F-statistic: 34.69 on 1 and 75 DF,  p-value: 1.025e-07

> |
```

From the calculation done in R programming, the population linear regression is 89.82 + 2.465 x, where the intercept is 89.82 and the slope of the line is 2.465. The R^2 value is seen as 0.3162. The estimate of the standard error of the least squares slope is 0.4185.

I did a t-test for the population slope to test whether there is a linear relationship between x and y.

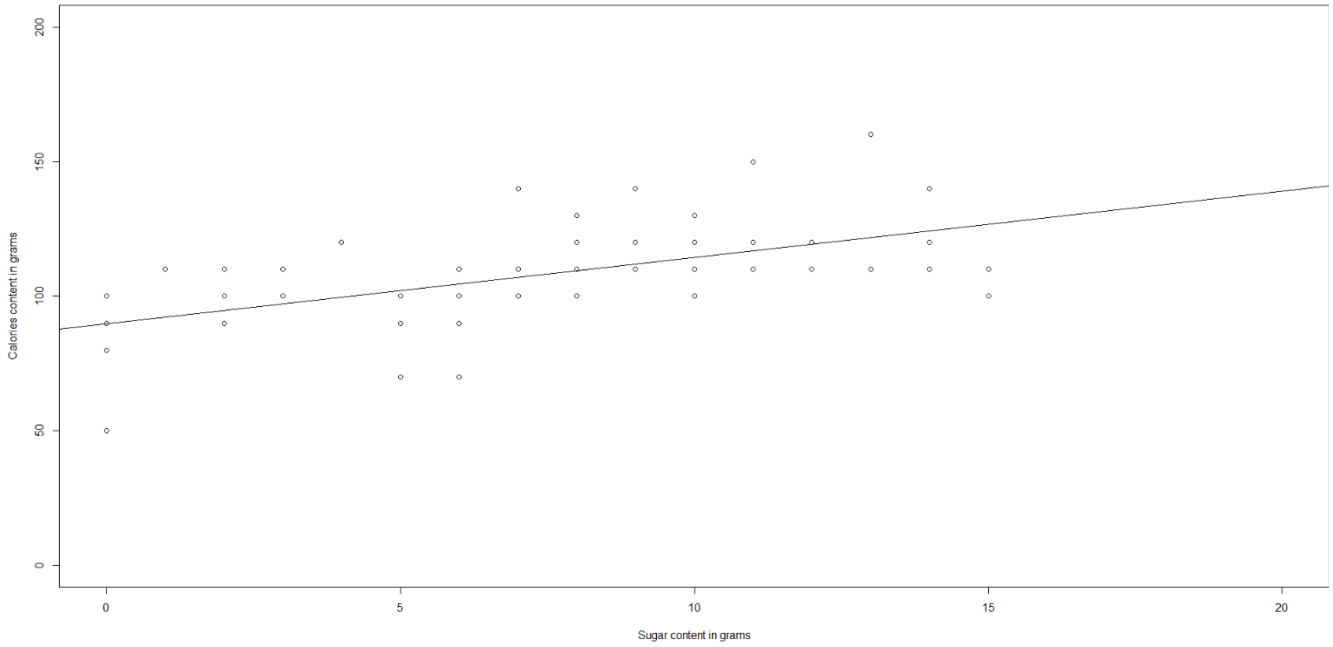
H0: $\beta = 0$ (No linear relationship)

H1: $\beta \neq 0$ (Linear relationship does exist)

Test statistic formula is $t = (b_1-\beta_1)/s_{b1} = (2.465 - 0)/0.4815 = 5.119$.

Hence, the test statistic value is 5.12. From the table, the t value is 1.9921.

Thus, we reject H0 because the test statistics value does fall in the rejection zone. As conclusion, there is sufficient evidence that sugar content affects the calories content in cereals.

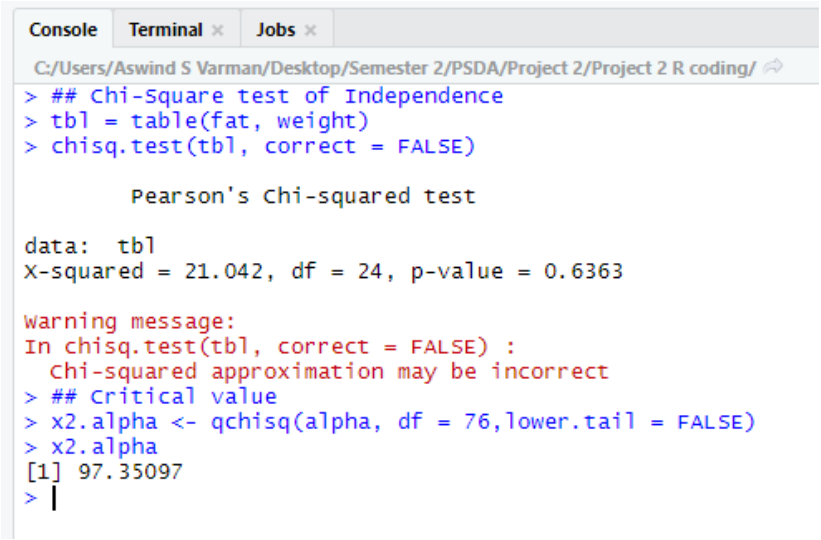


Here is the graphical presentation of the scatter plot and regression line done in R programming. Let's interpret the intersection coefficient, b_0. It is the estimated average value of calories content when the value of sugar content is zero. Looking at the graph, when the sugar content is zero, the calories content is 89.82 grams. This means that there is always the presence of calories in a cereal even though there is no presence of sugar in its content.

Also, let's take a look at interpretation of the slope coefficient, b_1. It measures the estimated change in the average value of calories content as a result of a one-unit change in sugar content. According to the graph, $b_1 = 2.465$ tells us that the average value of calories content increases by 2.465 grams, on average, for each additional one gram of sugar is added.

Chi-Square test of Independence

For this Chi-Square test of Independence, I decided to choose two variables that is fat variable and weight variable. I tested whether the fat variable and weight variable are dependent on each other using a significance level of 0.05.



I used to R programming to carry out the Chi-Square test of Independence.

H0: No relationship between fat and weight variables.

H1: Variables has relationship

$\alpha = 0.05$

Test statistic: $\chi^2 = 21.042$

Critical value: $\chi^2_{k=76,\ \alpha = 0.05} = 97.35097$

Since test statistic value < critical value, thus do not reject H0 at $\alpha = 0.05$.

As conclusion, there is sufficient evidence to claim that fat and weight variables are independent.

Discussion

1. For hypothesis 1 sample, the variable rating was selected to be tested. I tested the claim that the mean rating of all the cereals is lower than 50 %. After testing, it was proven that the mean rating is lower than 50 %. What this analysis indicates is that all the cereals have a relatively low ratings by customers and need to improve their product so that their collective rating could see an increase.

2. Sugar and calories variables were chosen for both correlation and regression model. After plotting the scatter plot, it can be said that there is a moderate relationship between the two variables. After doing significance test for the correlation, we could conclude that there was a linear relationship between the two variables.

3. For linear regression model, it was further explained in my data analysis.

4. I chose two variables that is weight, in ounces per serving, and fat variables, in grams, to test the claim that they are independent of each other using Chi Square test of independence. From my analysis, it could be proven that both variables are independent and one does not affect another in any way. It means that the fat content from cereals does not influence the weight content of those cereals. So, no matter how high or low the fat content is in a cereal, the weight is not that affected by it because it takes in a lot of other nutrients as well like sugar, carbohydrates, protein and so on.

Conclusion

As conclusion, there is a common saying that states breakfast is the most important meal of the day as it helps to kickstart the day in a positive mood and keeps us going until the end of that day. Thus, eating cereals during breakfast should not be taken lightly as it can have drastic effects on us mentally and also physically.

As stated by my statistics, there are many nutrients that play their own each but very important role in every type of cereal. For someone who is very active in the daily life, their carbohydrate and sugar intake must be high in their cereals so that it may provide the body with the necessary glucose it needs. But for someone with diabetes or high blood pressure, they may need to consume a cereal that has a low percentage of sugar and cholesterol content in it.

The analysis that I performed helps to see the relationship between certain content in cereals and how it affects one another. This data is important for those who want to take care of their personal health as eating the wrong type of cereal might be damaging for their body system in the long term. I hope to provide awareness among the public of the importance of eating cereals during breakfast and also the importance of correctly identifying the correct cereal for oneself.

References

1. https://www.kaggle.com/crawford/80-cereals
2. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
3. http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r
4. https://www.betterhealth.vic.gov.au/health/healthyliving/cereals-and-wholegrain-foods