UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SEMESTER 2

SESSION 2020/2021

SCSI2143 – PROBABILITY AND STATISTICAL DATA ANALYSIS

PROJECT 2: NUTRITION DATA ON 77 CEREAL PRODUCTS

PREPARED BY :
MUHAMMAD SYAHIR BIN SULAIMAN(A20EC0101)
IRSYAD ROS BIN HISYAM ROS(A20EC0052)

LECTURER'S NAME: Dr. Nor Azizah Ali
SECTION : 04
SUBMITTED ON : 3/7/2021

1.0 INTRODUCTION

This project purposely to identify the nutrition data on 77 cereal products. This data set was taken from website https://www.kaggle.com/crawford/80-cereals which is a website that provides lots of datasets that we can use especially for students to do research and their statistical case study. These datasets have been gathered and cleaned up by Petra Isenberg, Pierre Dragicevic and Yvonne Jansen. The original source can be found from http://lib.stat.cmu.edu/datasets/1993.expo/. This dataset contains 16 variables as following:

- Name: Name of cereal
- mfr: Manufacturer of cereal
    - A = American Home Food Products;
    - G = General Mills
    - K = Kelloggs
    - N = Nabisco
    - P = Post
    - Q = Quaker Oats
    - R = Ralston Purina
- type:
    - cold
    - hot
- calories: calories per serving
- protein: grams of protein
- fat: grams of fat
- sodium: milligrams of sodium
- fiber: grams of dietary fiber
- carbo: grams of complex carbohydrates
- sugars: grams of sugars
- potass: milligrams of potassium
- vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
- shelf: display shelf (1, 2, or 3, counting from the floor)
- weight: weight in ounces of one serving
- cups: number of cups in one serving
- rating: a rating of the cereals

2.0 ANALYSIS AND RESULT

2.1 HYPOTHESIS TESTING 1 SAMPLE

For testing, we used the rating variables from our dataset to determine whether a product is satisfactory or not. To determine if a product is satisfactory, it must have a rating more than 50%. The null hypothesis is that 20% of the products are satisfactory. The alternative hypothesis is that 20% of the products are not satisfactory.

$$H_0: p = 0.20$$
$$H_1: p > 0.20$$

Based on the dataset, only 21 of the 77 products have a rating higher than 50%.

N = 77

$\hat{p} = \frac{21}{77} = 0.2727$

$p = 0.20$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{N}}}$$
$$= \frac{0.2727 - 0.20}{\sqrt{\frac{(0.20)(0.80)}{77}}}$$
$$= 1.59$$

P(z > 1.59) = 0.0559

By using a 95% confidence level, we will compare the P-value with alpha value (significance level).

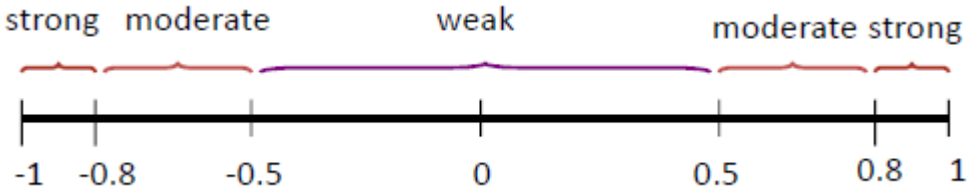
$$P\text{-}value \mid \alpha\text{-}value$$
$$0.0559 > 0.05$$
$$H_0 \text{ is failed to be rejected}$$

Since that the P-value is more than the alpha value, H_0 is failed to be rejected. There is sufficient evidence to support the claim that only 20% of the products are satisfactory. From here, we can conclude that our null hypothesis is true. This might be because that the minimum satisfactory rating of 50% is too high and thus leading to only 20% of products to meet the conditions. Alternatively, it might be because that our confidence level when comparing P-value with alpha value is too high which caused the null hypothesis to be rejected compared to if the confidence level was 90% instead.

2.1 CORRELATION TEST

Correlation is a statistic that measures the degree to which two variables move in relation to each other. Correlation shows the strength of a relationship between two variables and is expressed numerically by the correlation coefficient. The correlation coefficient's value range between -1.0 and 1.0.

This is the guideline to identify the strength of correlation:



CORRELATION BETWEEN POTASSIUM AND PROTEIN

In this study, we will measure the strength of the association or linear relationship between potassium and protein in 80 cereals. If the correlation is positive, so potassium increases when total spend increases.

Hypothesis statement:

H0: p = 0 (no linear correlation)

H1: p ≠ 0 (linear correlation exist)

Find r:

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{\left[\left(\sum x^2\right) - \left(\sum x\right)^2/n\right]\left[\left(\sum y^2\right) - \left(\sum y\right)^2/n\right]}}$$

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Using this Pearson's product-moment formula, the correlation coefficient,r is 0.54941. Thus, the relationship between Potassium and Protein is moderately positive correlation. The figure below shows scatter graph of potassium and protein.

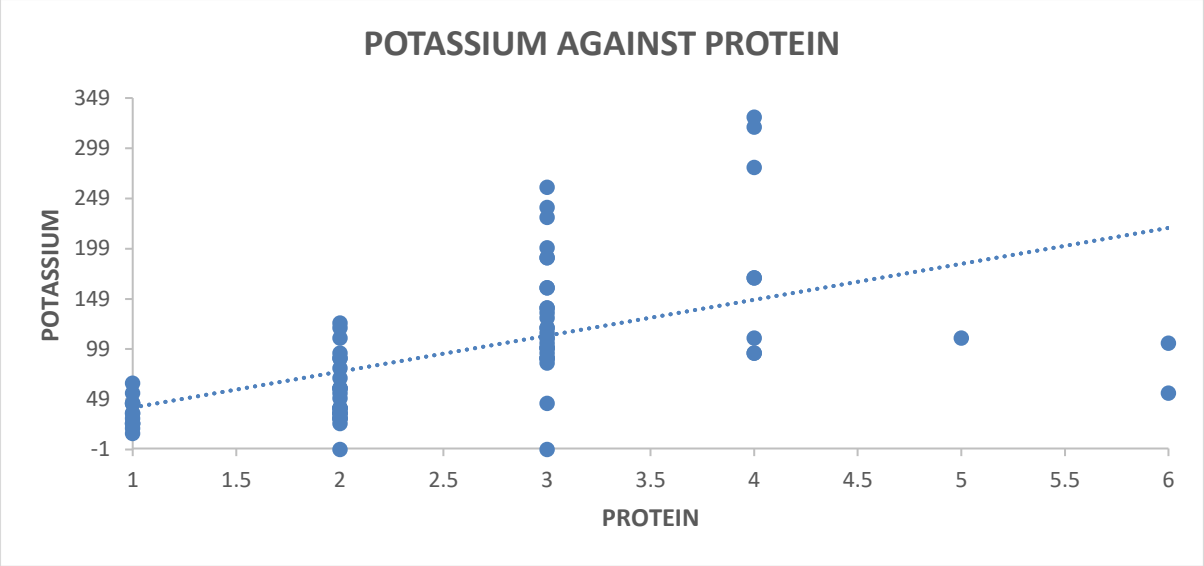


Figure 1-Scatter graph of 'Potassium' and 'Protein'

Next, do the test statistic using t value formula:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Using this formula, the value of t is 5.69447.
The critical value for tα/2=0.025, df=75 is 1.992
Therefore, since the t=5.69447>1.992, we reject the null hypothesis. There is sufficient evidence of a linear relationship between Potassium and Protein at the 5% significance level.

2.4 CHI-SQUARE GOODNESS OF FIT TEST

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

Situation 1:

A study was conducted on a person consumed 10 different cereals containing calories on 10 days

Days	1	2	3	4	5	6	7	8	9	10
Calories intake	Bran (70)	Natural Bran (120)	All-Bran (70)	All-Bran extra fiber (50)	Almond Delight (110)	Apple Cinnamon Cheerios (110)	Apple Jacks (110)	Basic 4 (130)	Bran Chex (90)	Bran Flakes (90)

Solution:

Hypothesis statement:

H0: p1 = p2 = p3 = p4 = p5

H1: At least 1 of the 10 proportions is different from others.

Expected frequencies:

E=$\frac{n}{k}$ = $\frac{950}{10}$ = 95

Days	1	2	3	4	5	6	7	8	9	10
Calories intake (observed)	(70)	(120)	(70)	(50)	(110)	(110)	(110)	(130)	(90)	(90)
expected	95	95	95	95	95	95	95	95	95	95
$\frac{(O-E)^2}{E}$	6.5789	6.5789	6.5789	21.3157	2.3684	2.3684	2.3684	12.8947	0.2632	0.2632

Test statistics:

$\chi 2 = \sum \frac{(O-E)^2}{E}$ =

6.5789+6.5789+6.5789+21.3157+2.3684+2.3684+12.8947+0.2632+0.2632 =61.5787

Critical value:

$X_{9,0.05}$ = 16.919

Conclusion:

Since $\chi 2$= 61.5787 > $X_{9,0.05}$ =16.919, we reject the hypothesis null. That is, we reject claim that the calories intake with equal proportions on the 10 days.

CONCLUSION

In conclusion, there are several benefits for me when working on the project, from selecting data to completing the project report. This project our improves our statistical analysis skill and allowing us to learn new things. The method that we used to do this research is really a fantastic method for people who wish to conduct data analysis and also make their job easier. All the tests we ran on this dataset were really helpful in identifying all the components in a more scientific manner.

REFFERENCE

1. https://www.kaggle.com/crawford/80-cereals
2. https://perso.telecom-paristech.fr/eagan/class/igr204/datasets
3. http://lib.stat.cmu.edu/datasets/1993.expo/