FACULTY OF COMPUTING

SEMESTER 2 2023/2024

SECI1143 – PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 3

PROJECT 2

LECTURER: DR. NOR HAIZAN BT MOHAMED RADZI

GROUP MEMBERS

STUDENT NAME	MATRIC NO
JOANNE CHING YIN XUAN	A23CS0227
EVELYN GOH YUAN QI	A23CS0222
CHUA JIA LIN	A23CS0069

PRESENTATION LINK: https://youtu.be/zZNlqLV88Mk

TABLE OF CONTENT

1.0 INTRODUCTION

In Project 2, we have chosen a dataset from Kaggle website, which is the "Customer_ Spending_Dataset" collected by Aditya Goyal who is a Data Science intern at Celebal Technologies. This dataset provides information of 1000 samples of the customer's name, age, gender, income, education, country, purchase frequency, and spendings.

In this project, we aim to investigate whether there is a relationship between the variables age, income, gender, and spendings. There are a few ways that we decided to help our group to analyze and interpret the dataset. In the data analysis process, we decided to use the hypothesis 1-sample test, correlation test, regression test, and Chi-square test of independence to test on the dataset and provide the results to support our expectations.

Therefore, we expect to see whether the population mean of the customers' age is greater than 50 years old. Besides, we also expect how the customer's income will affect their spendings because we believe that customers with higher incomes will have greater spendings. Other than that, we are interested in the relationship between the customer's age and their spendings. Moreover, we also expect to know the independence between gender and the educational level.

2.0 DATASET

Our data set used is collected from online sources which is Kaggle. This secondary data set consists of 8 different variables, but we only choose 5 out of 8, which are age, gender, income, spending and education level. The table below shows the description of each variable regarding their type of data and level of measurement.

No	Variable/Parameter	Type of Data	Level of Measurement
1	Age	Ratio	Quantitative
2	Gender	Nominal	Qualitative
3	Income	Ratio	Quantitative
4	Spending	Ratio	Quantitative
5	Educational Level	Nominal	Qualitative

The table below shows the description of the statistical test analysis.

Selected Variables	Objectives	Test Analysis and Expected Outcome
Age	To determine whether the mean of customer's age is greater than 50. This hypothesis will help us understand the age distribution of the customers.	**Analysis:** 1 Sample Hypothesis Testing (Test on Mean, Variance Unknown) **Expected Outcome:** By conducting 1 sample t-test, we can determine whether the mean of customers' age is greater than 50. If the p-value is less and equal to the 0.05 level of significance, it would indicate that the mean of customer's age is greater than 50.
Income, Spending	To examine the relationship between customers' incomes and their spendings.	**Analysis:** Correlation analysis **Expected Outcome:** By calculating the correlation coefficient between customers' income and their spendings, we can determine the strength and direction of the relationship. If the correlation coefficient is positive and statistically significant, it would

		indicate that as the customers' income increases, their spending tends to increase.
Age, Spending	To assess if the customer's age is a predictor of their spendings.	**Analysis:** Simple linear regression **Expected Outcome:** By conducting this simple linear regression analysis, we can determine the relationship between the customer's age and their spendings. If the regression coefficient is statistically significant (typically $p < 0.05$), it would suggest that at the customer's age we can quantify the extent of this relationship through this regression equation.
Gender, Education Level	To assess the independence between gender and educational level	**Analysis:** Chi-square test of independence **Expected Outcome:** By conducting a chi-square test of independence, we can examine if there is a statistically significant association between gender and educational level. If the p-value is less than or equal to the 0.05 level of significance, it would indicate that gender and educational level are not independent and there is a relationship between them.

3.0 Data Analysis

3.1 Hypothesis 1 Sample Test

In this analysis, we will use variable age, where we will test whether the mean of the customer's age is greater than 50 years old at 0.05 significance level. The sample size of the dataset is 1000. So, the hypothesis test on mean for 1 sample is carried out and the population variances are unknown.

Hypothesis Statement:

Null Hypothesis, $H_0 : \mu = 50$

Alternative Hypothesis, $H_1 : \mu > 50$

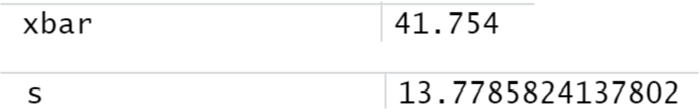
where μ represent the population mean of the customer's age.

Test Statistic:

Given $\alpha = 0.05$ and $n = 1000$, the test statistic can be calculated using the formula below:

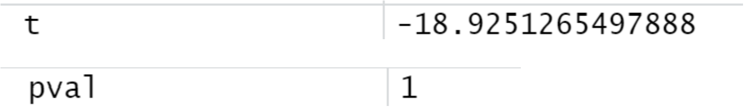
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

By using RStudio, the sample mean of customer's age, $\bar{x}$ and the sample standard deviation of customer's age, s are calculated. Hence, $\bar{X} = 41.754$ and $s = 13.77858$.



xbar	41.754
s	13.7785824137802

From the RStudio, we obtain the test statistic value, $t = -18.925$. The probability value that we obtain is P-value = 1.0



t	-18.9251265497888
pval	1

Conclusion:

Since P-value > α (1.0 > 0.05), we fail to reject the null hypothesis, H_0. There is no sufficient evidence at 0.05 significance level to support the claim that the mean of the customer's age is greater than 50 years old.

3.2 Correlation Test

In this analysis, we will use the variables income and spending to determine whether there is a linear relationship between the customer's income and spending at a 95% significance level in a sample size of 1000. Since both of our data are ratio type, Pearson's technique was selected to calculate the sample correlation coefficient, r.

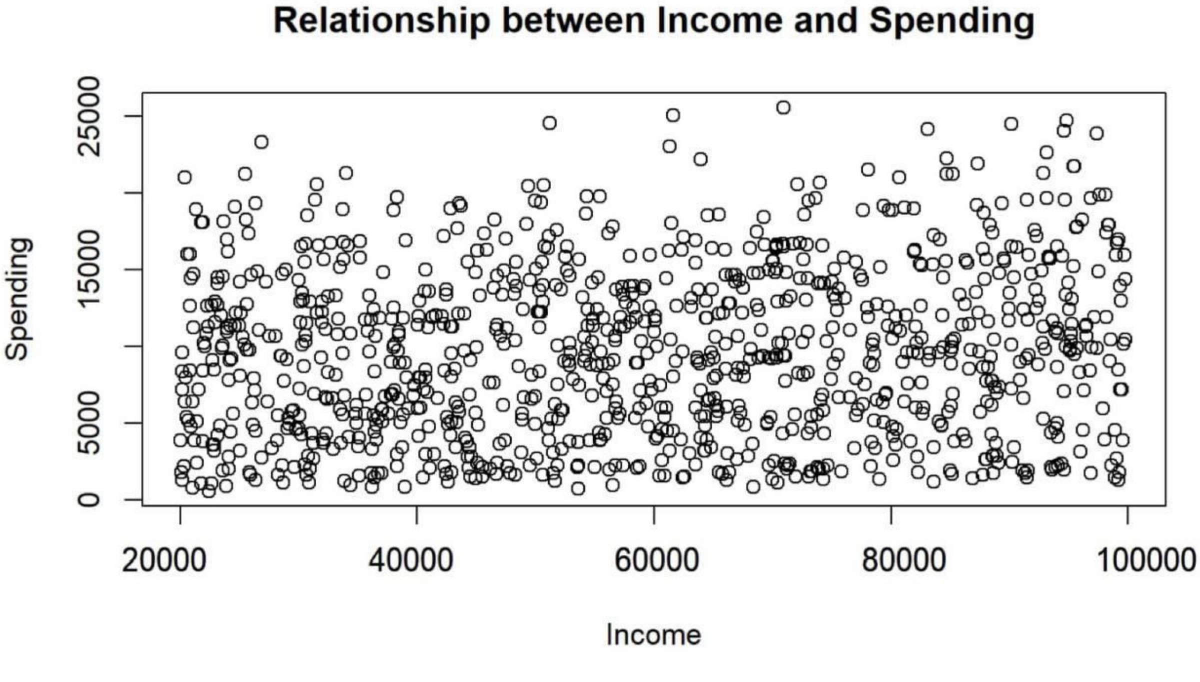


Figure 3.1: Scatter plot relationship between income and spending

From the scatter plot in Figure 3.1, we can see that there is a weak positive correlation relationship between the customer's income and spending.

Sample correlation coefficient:

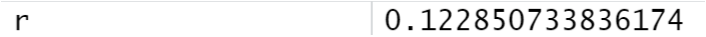
We can calculate the sample correlation coefficient using Pearson’s method by:

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

where:

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

By using RStudio, we can see that the sample correlation coefficient, r is 0.12285, which indicates that there is a relatively weak positive linear correlation between the income and spending of the customers.



```
r                          0.122850733836174
```

Hypothesis Statement:

Null Hypothesis, $H_0 : \rho = 0$ (no linear correlation)

Alternative Hypothesis, $H_1 : \rho \neq 0$ (linear correlation exists)

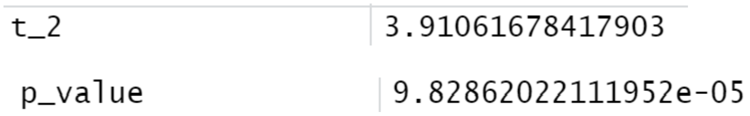
where ρ represent the population correlation coefficient.

Test Statistic:

Given α = 0.05 and n = 1000, the test statistic can be calculated using the formula below:

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

By using RStudio, the test statistic, t is 3.9106. The probability value that we obtain is P-value = 9.829×10^{-5}



```
t_2          3.91061678417903
p_value      9.82862022111952e-05
```

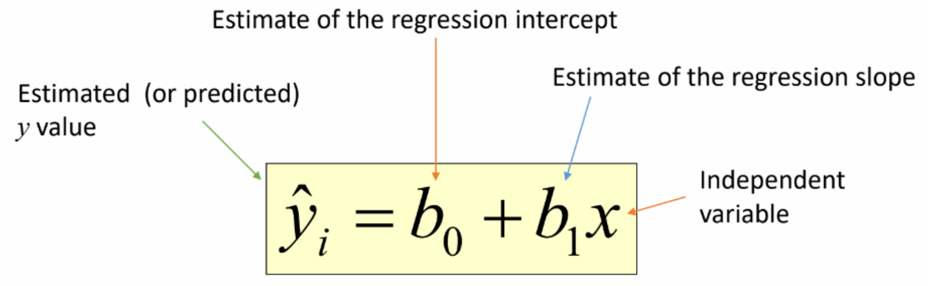
Conclusion:

Since P-value < α ($9.829 \times 10^{-5} < 0.05$), we reject the null hypothesis, H_0. There is sufficient evidence at 0.05 significance level to support the claims that there is a linear relationship between the customer’s income and spending. The sample correlation coefficient, r is 0.12285, indicating a weak positive linear correlation between the income and spending of the customers.

3.3 Regression Test

In the regression test, we measure the relationship between the customer's age and their spendings for each customer with a sample size, n = 1000. The independent variable, x is the age of the customer and the dependent variable, y is the customer's spendings. The sample regression line provides an estimate of the population regression line.

Estimated Regression Model:



Estimate of the regression intercept

Estimated (or predicted) y value

Estimate of the regression slope

$$\boxed{\hat{y}_i = b_0 + b_1 x}$$

Independent variable

Test statistic formula:

$$\boxed{t = \frac{b_1 - \beta_1}{s_{b_1}}}$$

$$\text{d.f.} = n - 2$$

where,
b_1 = Sample regression slope coefficient
β_1 = Hypothesized slope
s_{b1} = Estimator of the standard error of the slope

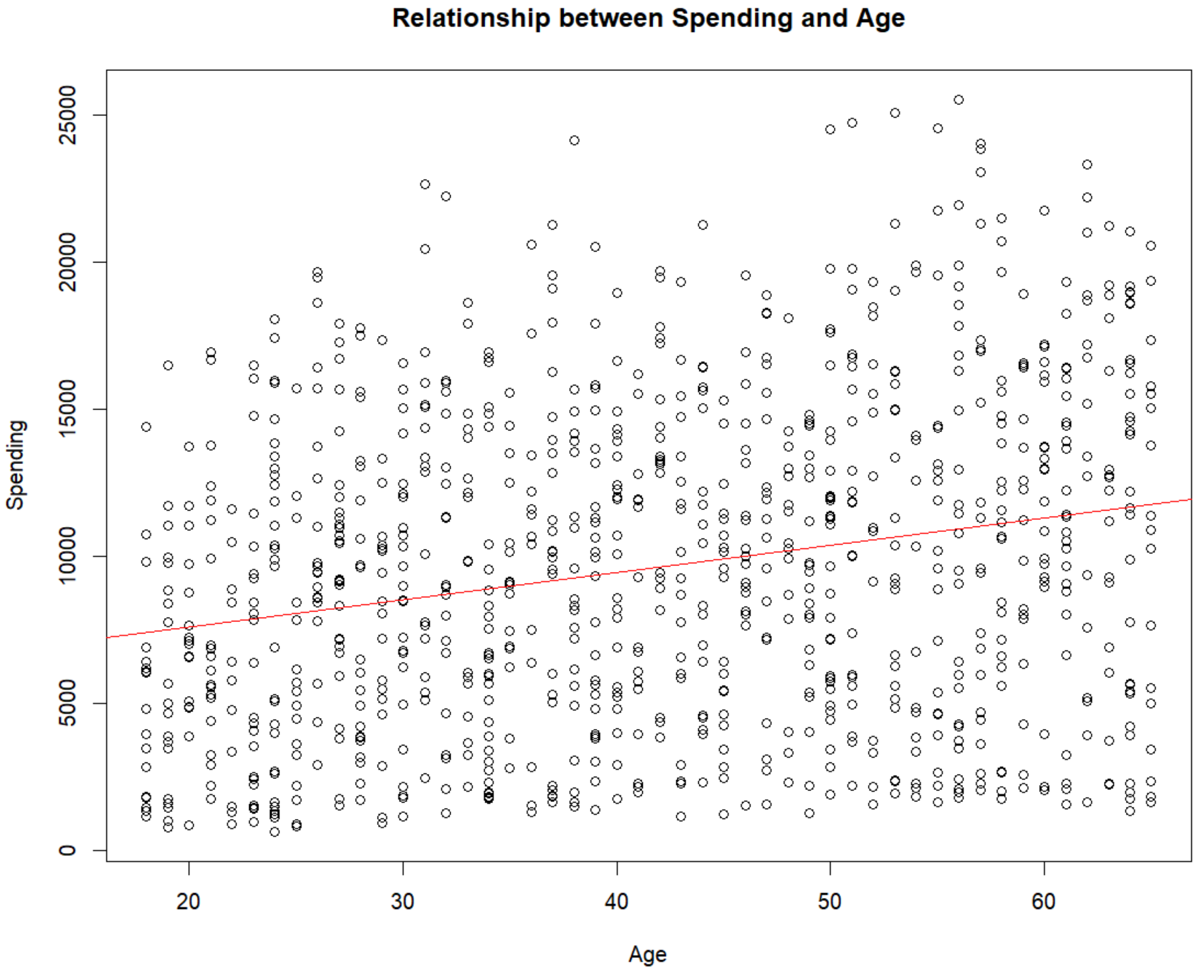
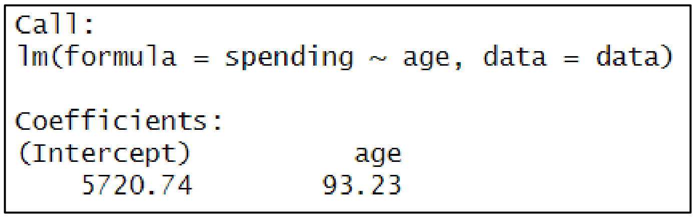


Figure 3.2 Scatter plot relationship between spending and age

From the scatter plot in Figure 3.2, we can see that the best fit line shows an upward inclining indicating there is a positive linear relationship between the customer’s age and their spendings.



```
Call:
lm(formula = spending ~ age, data = data)

Coefficients:
(Intercept)          age
    5720.74        93.23
```

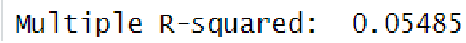
By using the RStudio, we can get the formula for estimated regression model which is

$$Y = 5720.74 + 93.23X$$

From the formula, the value of $b_0 = 5720.74$ is the estimated value of y when the value of x is equal to 0. While the value of $b_1 = 93.23$ measures the estimated change in the average value of customer’s spendings, y because of one-unit change in customer’s age, x.

The coefficient of determination is the portion of the total variation in the dependent variable explained by variation in the independent variable.

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$



```
Multiple R-squared:  0.05485
```

From the RStudio, we obtain the value of coefficient of determination, $R^2 = 0.\ 05485$.

Since $0 < R^2 < 1$, the result shows a weaker linear relationship between the customer’s age, x and customer’s spendings, y. Some but not all the variation in customer’s spendings, y is explained by the variation in customer's age, x.

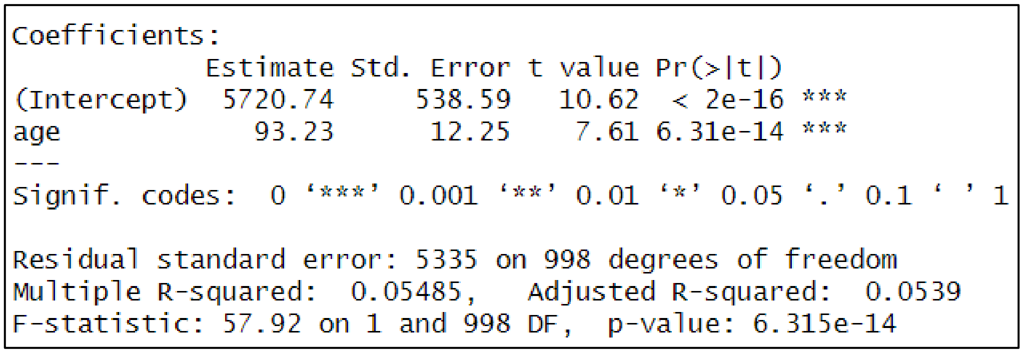
Hypothesis statement:

Null hypothesis, H_0: $\beta_1 = 0$ (no linear relationship)

Alternative hypothesis, H_1: $\beta_1 \neq 0$ (linear relationship does exists)

where β represents the population slope coefficient (hypothesized slope).

Test statistic:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5720.74     538.59   10.62  < 2e-16 ***
age            93.23      12.25    7.61 6.31e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5335 on 998 degrees of freedom
Multiple R-squared:  0.05485,	Adjusted R-squared:  0.0539 
F-statistic: 57.92 on 1 and 998 DF,  p-value: 6.315e-14
```

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that there exists a linear relationship between the customer's age and the customer's spendings for each customer.

From the RStudio, we obtain the sample regression slope coefficient, $b_1 = 93.23$, estimator of the standard error of the slope, $s_{b1} = 12.25$, test statistic value, $t = 7.61$ and the degree of freedom is $df = 998$. The probability value that we obtained is $P - \text{value} = 6.315 \times 10^{-14}$.

Conclusion:

Since $P - \text{value} < \alpha$ or ($6.315 \times 10^{-14} < 0.050$), thus reject null hypothesis, H_0.

There is sufficient evidence at 0.05 significance level to support the claims that there exists a linear relationship between the customer's age and the customer's spendings for each customer. We can see the residual and residual standard error which is the difference between y and predicted y is larger, r square is low, so there is no strong linear relationship. But we can't draw the conclusion that there is no relationship between the customer's age and their spendings, because we just prove that there is a linear relationship.

3.4 Chi-square Test of Independence

In this analysis, we will use the variables Gender and Education Level to assess the independence of these variables using a Two-Way Contingency Table at a 95% confidence level. As a result, we employ the Chi-Square Test of Independence in conjunction with a two-way contingency table.

Education	Female	Male
Bachelor	141	130
High School	112	133
Master	121	115
PhD	125	123
Total	499	501

Observed Frequencies for variables Gender and Education Level

1.State the test hypothesis:

H_0: Gender and Education Level are independent.
H_1: Gender and Education Level are dependent.

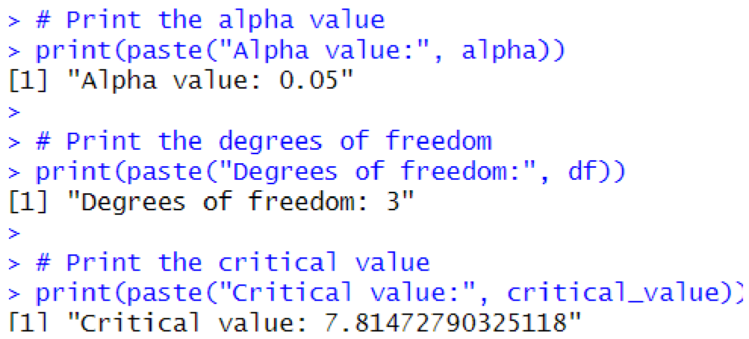
2.Find the critical value:

Table :

Critical value,χ^2 =7.815 (with df=(2−1)(4−1)=3, α=0.05)

By RStudio:

Critical value,χ^2 = 7.8147279



```
> # Print the alpha value
> print(paste("Alpha value:", alpha))
[1] "Alpha value: 0.05"
>
> # Print the degrees of freedom
> print(paste("Degrees of freedom:", df))
[1] "Degrees of freedom: 3"
>
> # Print the critical value
> print(paste("Critical value:", critical_value))
[1] "Critical value: 7.81472790325118"
```

3.Calculate the expected counts:

Gender	Educational Level				Total
	Bachelor	High School	Master	PhD	

	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
Male	130	135.77	133	122.75	115	118.24	123	124.25	501
Female	141	135.23	112	122.26	121	117.76	125	123.75	499
Total	271	271	245	245	236	236	248	248	1000

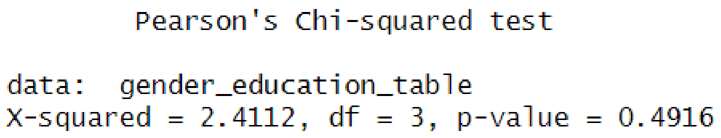
4.Calculate the test statistic value:

-*Calculate manually:*

Cell,ij	Observed Count,o_{ij}	Expected Count,e_{ij}	$\frac{(o_{ij-}e_{ij})^2}{e_{ij}}$
1,1	130	135.77	0.25
1,2	133	122.75	0.86
1,3	115	118.24	0.09
1,4	123	124.25	0.01
2,1	141	135.23	0.25
2,2	112	122.26	0.86
2,3	121	117.76	0.09
2,4	125	123.75	0.01
		χ^2	2.42

When we calculate test statistic manually, we get test statistic, $\chi^2 = 2.42$

-Using RStudio



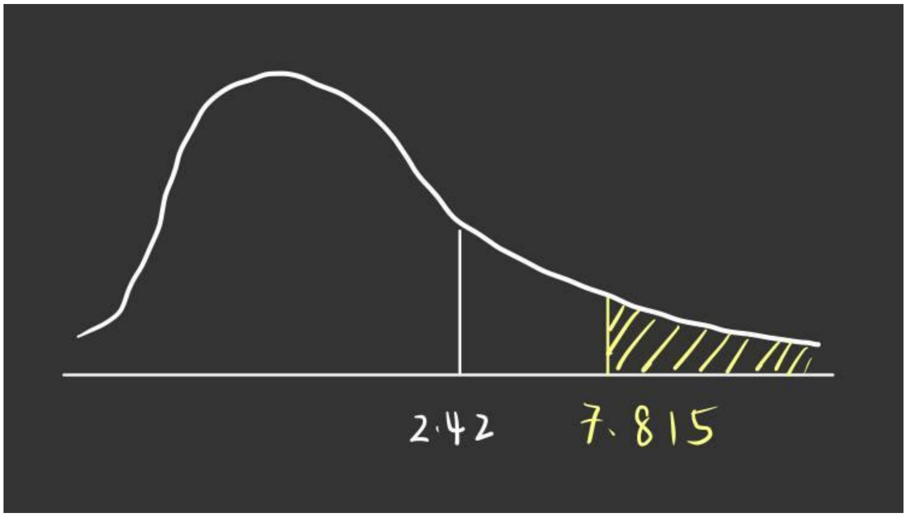
```
        Pearson's Chi-squared test

data:  gender_education_table
X-squared = 2.4112, df = 3, p-value = 0.4916
```

When we calculate test statistic using RStudio, we also get test statistic , $\chi^2 = 2.4112$, with

p-value= 0.4916

5.The decision:



Since the statistic value($\chi^2 = 2.42$) <critical value ($\chi^2_{k=3, \alpha=0.05} = 7.815$), it does not fall within the critical region. Thus, we **fail to reject** the null hypothesis, H_0. Therefore, there is insufficient evidence to conclude that there is a relationship between Gender and Education Level, at α=0.05 . Gender and Education Level are independent.

4.0 Conclusion

For Hypothesis 1 Sample Test for age, we aimed to test whether the mean age of customers is greater than 50 years at a 0.05 significance level. With a sample size of 1000, the test resulted in failing to reject the null hypothesis (H0). There is insufficient evidence to support the claim that the mean age of customers is greater than 50 years. This result suggests that, within the sample, the average age of customers does not significantly exceed 50 years.

Next is the correlation between income and spending. The correlation analysis between customer income and spending, using Pearson’s correlation coefficient, indicated a weak positive correlation with a P-value less than 0.05. Therefore, we reject the null hypothesis (H0) and conclude that there is sufficient evidence to support a linear relationship between customer income and spending. Despite the weak correlation, the relationship is statistically significant, indicating that as income increases, spending tends to increase slightly as well.

For the regression analysis, the regression analysis aimed to assess the relationship between customer age and spending. The results indicated a positive linear relationship with a significant regression slope coefficient and a very low P-value. Therefore, we reject the null hypothesis (H0) and conclude that there is sufficient evidence to support a linear relationship between age and spending. However, the low R-squared value and large residual standard error suggest that this linear relationship is not strong. This implies that while age and spending are related, other factors may also significantly influence customer spending behavior.

Lastly, the Chi-Square Test of Independence assessed the relationship between gender and education level. The test resulted in failing to reject the null hypothesis (H_0), indicating insufficient evidence to support a dependency between gender and education level. This suggests that, within the sample, gender and education level are independent of each other, meaning that the distribution of education levels is similar across different genders.

In conclusion, we found that the mean age of customers is not significantly greater than 50 years, customer income has a weak but significant positive impact on spending, age positively influences spending, and gender and education level are independent of each other. At the end of this Project 2, we will be skilled in using R-Studio to conduct test analyses, including hypothesis testing, correlation analysis, regression analysis. and chi-square test of independence using R-Studio. Our ability to analyze data is being enhanced by this project, which is beneficial for our future. Also, special thanks to our lecturer, Dr. Nor Haizan for or her support and guidance throughout this project.

5.0 Appendix

Customer_Spending_Dataset. (2023, May 30). Kaggle. https://www.kaggle.com/datasets/goyaladi/customer-spending-dataset