Report on Analysis of Empirical Work using SAS EG

STAT2110 STATISTICAL DATA PROCESSING SAS EG

Name: Onyebuchi Dalbert Zimuzochukwu
Student Number: Z109554
Email Address: Z109554@student.uwasa.fi

| Course Title : STAT2110|

Date: 11.12.2016

INTRODUCTION

The aim of this project is to study the statistics of the presented data. Here a

For this project, I have chosen the following variables:

a. Age (Quantitative)
b. Education (Categorical: Ordinal)
c. Employment (Quantitative)
d. Income (Quantitative)
e. Debt to Income Ratio (Quantitative : Ratio level)
f. Credit Card Debit (Quantitative)
g. Default on Loans. (Categorical: Nominal)
h. Residence (Categorical: Nominal)
i. Other debts (Quantitative)

My reason for chosen all these is because I want to evaluate the empirical data holistically (As a Whole) to get a complete picture about the population.

In this work, the following hypotheses will be examined:

Hypothesis: Younger people have more defaults on loans due to lifestyle preferences.

Hypothesis: College students have more loan debts (default on loans) due to college expenses.

Hypothesis: More Income leads to less debt, as income earners have the means to pay back.
Hypothesis: There is a strong relationship between Debit to income ratio and Credit Card Debt and a positive relationship exist between the two variables.

Hypothesis: More income leads to more Credit Card Debt as high-income earners take out more credit card loan from the banks.

Hypothesis: More income leads to less default on loans as income earners have the means to pay back.

Hypothesis: More income leads to less other debts.

2.0. STATISTICAL DESCRIPTION OF DATA

a. AGE (QUANTITATIVE DATA)

From SECTION 1 A (See Appendix 1), the age population appears to be a young population as the median is 34 years old.

In addition, people of 29 years appears to be the highest occurrence in the population (MODE).

From SECTION 1 B (See Appendix 1), here we notice that the distribution of age is multimodal.

The distribution is positively skewed and platykurtic. Skewness = 0.322, Kurtosis = -0.73

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR is quite small (13) compared with range (36), so the dispersion of the variable values is quite weak.

From SECTION 1 C (See Appendix 1)

A greater number of the population lie between 29 and 42 years old. There are no outliers.

b. EDUCATION (CATEGORICAL DATA: Ordinal)

From Section 2 B (SEE APPENDIX 2) and Section 2 C ((SEE APPENDIX 2), the percentage of people who completed a high school degree is the highest, followed by those who did not complete a high school degree.
Those who went ahead to complete further higher education degrees appears to be smaller, With those with a Master of PHD degree the lowest.
It can be deduced that people represented by this statistic may not be motivated to pursue higher degrees as their basic high school degree can get them a form of comfortable employment. (As is noticed when compared to the employment distribution).
The distribution is unimodal, and the distribution is left-skewed and is platykurtic.
Skewness = 0.87, Kurtosis = -0.36

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

c. EMPLOYMENT (Quantitative Data)

From Section 3 A (SEE APPENDIX 3), the Mean = 8.17. Median = 7, Standard Deviation = 6.
Skewness = 0.87, Kurtosis = -0.313

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

From Section 3 B (SEE APPENDIX 3), the distribution is unimodal, mesokurtic and positively skewed.
Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR is quite small (10) compared with range (31), so the dispersion of the variable values is quite weak.
From Section 3 C (SEE APPENDIX 3), we notice that the mean number of years spent with employer is 8 and also a few outliers. This could also indicate that job mobility is high as only few stay longer at a particular work place.

d. INCOME (Quantitative Data)
From Section 4 A (SEE APPENDIX 4), Mean = 48.13, Median = 36 (Which shows the economy is fine when compared to modern day). Standard Deviation: 40.32 (very Large)

From Section 4 B (SEE APPENDIX 4), the distribution is unimodal, leptokurtic and positively skewed.
Skewness = 3.84, Kurtosis = 24.29

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR is quite small (32) compared with range (432), so the dispersion of the variable values is quite weak.

From Section 4 C (SEE APPENDIX 4), we have a few outliers, People with very high incomes.

e. DEBT TO INCOME RATIO (Quantitative : Ratio Level)
From Section 5 A (SEE APPENDIX 5), Mean = 10.35, Median = 8.5, Standard Deviation: 6.89

From Section 5 B (SEE APPENDIX 5), the distribution is unimodal, Leptokurtic and positively skewed.
Skewness = 1.09, Kurtosis = 1.36

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR is quite small (9.5) compared with range (40.7), so the dispersion of the variable values is quite weak.

From Section 5 C (SEE APPENDIX 5), there are outliers, people who are in serious debt.

f. CREDIT CARD DEBT (Quantitative Data)
From Section 6 A (SEE APPENDIX 6), Mean = 1.60, Median = 0.90, Mode = 0.086, Standard Deviation: 2.10.

From Section 6 B (SEE APPENDIX 6), the distribution is unimodal, Leptokurtic and positively skewed.
Skewness = 3.41, Kurtosis 15.8

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR is quite small (16.01) compared with range (1.59), so the dispersion of the variable values is quite weak.

From Section 6 C (SEE APPENDIX 6), we have many outliers, which shows that many people in the sample are in debt.

g. <u>DEFAULT ON LOANS (Categorical Data: Nominal)</u>
From Section 7 A (SEE APPENDIX 7), Mean = 0.29, Median = 0.00, Mode = 0.00, Standard Deviation: 0.45.

From Section 7 B (SEE APPENDIX 7), the distribution is unimodal, platykurtic and positively skewed.

Skewness = 0.922, Kurtosis = -1.15

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR (1) is equal with range (1), so the dispersion of the variable values is quite Strong. Of course, it should be as the variables are nominal.

The frequency of people who defaulted on loans were smaller than those who did not.

About 29% of the population defaulted on loans.

From Section 7 C (SEE APPENDIX 7), there are no outliers.

h. <u>RESIDENCE (Categorical: Nominal)</u>

From Section 8 A (SEE APPENDIX 8), Mean = 1.96, Median = 2.00, Mode = 2.00, Standard Deviation: 0.71.

From Section 8 B (SEE APPENDIX 8), the distribution is unimodal, platykurtic and symmetric.
Skewness = 0.059, Kurtosis = -0.99

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability. The IQR (2) is equal with range (1), so the dispersion of the variable values is quite Strong. Of course, it should be as the variables are nominal. The frequency of people who live in downtown is more than that of other places.

From Section 8 C (SEE APPENDIX 8), there are no outliers.

i. <u>OTHER DEBTS (Quantitative)</u>

From Section 9 A (SEE APPENDIX 9), Mean = 3.24, Median = 2.10, Mode = 7.82, Standard Deviation: 3.485.

From Section 8 B (SEE APPENDIX 9), the distribution is the distribution is unimodal, Leptokurtic and positively skewed.

Skewness = 2.72, Kurtosis = 10.093

By the way, the number of cases is about 498, so the standard error for skewness is about 0.109 and the standard error for kurtosis is about 0.2195.)

Because of the skewness, I compared interquartile range (IQR) and range to figure out the situation of variability.

The IQR (3.07) is equal with range (26.998), so the dispersion of the variable values is quite weak.

From Section 9 C (SEE APPENDIX 9), there are so many outliers.

3.0. ANALYZES OF STATISTICAL DEPENDENCY

a. Age in Years (Quantitative) Versus Residence (Nominal)

From APPENDIX 10, Likelihood Ratio Chi-square probability = 0.3490, Phi Coefficient = 0.379, Contingency Coefficient = 0.3542 and Cramer's V = 0.2678.
All the measures show that there is a moderate relationship between the variables. The percentage distribution of Age is different for different places of residence.
There is also a negative correlation (-0.024) and the P-Value = 0.593. There is a very weak negative linear relationship between the variables: Age and Residence.
The Correlation is not SIGNIFICANT.

b. Age in Years (Quantitative) Versus Years in Current Employer (Quantitative)

From APPENDIX 11, Likelihood Ratio Chi-square probability = 0.99, Phi Coefficient = 1.73 Contingency Coefficient = 0.87 and Cramer's V = 0.32.

All the measures show that there is a strong relationship between the variables.

H0: The population distribution of the variable is independent.
H1: The population distribution is NOT independent.

We can accept the alternate Hypothesis H1 that in the population, the variables are NOT independent.

There is also a positive correlation (0.5515) and the P-Value < 0.0001. The correlation is significant at 0.01 % significance level.
There seems to be moderate positive linear relationship between the variables: Age and Employment.

c. Age in Years (Quantitative) Versus Education (Ordinal)

From APPENDIX 12, Likelihood Ratio Chi-square probability = 0.0018, Phi Coefficient = 0.59, Contingency Coefficient = 0.5 and Cramer's V = 0.297.

All the measures show that there is a moderate relationship between the variables.

There is also a positive correlation (0.00335) and the P-Value = 0.94. The Correlation is not SIGNIFICANT.

There is a very weak linear relationship between the variables: Age and Education.

d. Age in Years (Quantitative) Versus Income (Quantitative)

From APPENDIX 13, Likelihood Ratio Chi-square probability = 1, Phi Coefficient = 3.19, Contingency Coefficient = 0.95 and Cramer's V = 0.53.

All the measures show that there is a strong relationship between the variables.

Ho: The population distribution of the variable is independent.

H1: The population distribution is NOT independent.

We can accept the alternate Hypothesis H1 that in the population, the variables are NOT independent.

There is also a positive correlation (0.45) and the P-Value < 0.001. The correlation is significant at 0.01 % significance level.

There is a moderate positive linear relationship between the variables: Age and Income.

e. Age in Years (Quantitative) Versus default on Loans (Nominal)

From APPENDIX 14, Likelihood Ratio Chi-square probability = 0.0051, Phi Coefficient = 0.33, Contingency Coefficient = 0.32 and Cramer's V = 0.33.

All the measures show that there is a moderate relationship between the variables.

There is also a negative correlation (-0.16) and the P-Value < 0.0002. The correlation is significant at 0.02 % significance level.

Ho: Younger people do NOT have more defaults on loans.

H1: Younger people have more defaults on loans.

We can accept the alternate Hypothesis H1 that Younger people do have more defaults on loans. There are prevailing factors peculiar with the young: fashion, compulsive Disorder e.g.: shopaholics, little or no stable employment, loan seeking for education, misplaced or misguided priorities, and social factors.

f. Education (Ordinal) Versus Residence (Nominal)

From APPENDIX 15, Likelihood Ratio Chi-square probability = 0.22, Phi Coefficient = 0.15, Contingency Coefficient = 0.15 and Cramer's V = 0.1.

All the measures show that there is weak relationship between the variables.

There is also a negative correlation (-0.06) and the P-Value 0.1825.

g. Education (Ordinal) Versus Default on Loans (Nominal)
H_1 = College students have more loan debts (default on loans)
H_o = College students have less loan debts (default on loans)
From APPENDIX 16, Likelihood Ratio Chi-square probability = 0.17, Phi Coefficient = 0.12, Contingency Coefficient = 0.12 and Cramer's V = 0.12.
All the measures show that there is weak relationship between the variables.
There is also a positive correlation (0.0883) and the P-Value = 0.0489. Since the P value is small, the result is significant. The correlation is significant at 5 % significance level. The null Hypothesis H_o is rejected and the alternative Hypothesis is accepted.

We also noticed that the number of people who defaulted on loans were more among those with a college degree (22 out of 30: 73%).

Moreover, this is expected, as there is currently a rise in loan debts among college students.

The Cocharan-Armitage Trend test is as follows: Statistics (Z) = -1.9702, One-sided Pr < Z = 0.0244, Two-sided Pr < |Z| = 0.0488.

h. Education (Ordinal) Versus Debt to income Ratio (Quantitative : Ratio level)

Ho: The population distribution of the variable is independent.
H1: The population distribution is NOT independent.

From APPENDIX 17, Likelihood Ratio Chi-square probability = 0.98, Phi Coefficient = 1.26, Contingency Coefficient = 0.78 and Cramer's V = 0.63.
All the measures show that there is very strong relationship between the variables.

There is also a positive correlation (0.0217) and the P-Value = 0.6284.

An unusually high p value indicates the data match the null model suspiciously well, suggesting perhaps… the data were fabricated.

We can accept the alternate Hypothesis H1 that in the population, the variables are NOT independent.

i. Debt to Income Ratio (Quantitative : Ratio Level) Vs Income (Quantitative)

Ho: More Income leads to more debt
H1: More Income leads to less debt

From APPENDIX 18, Likelihood Ratio Chi-square probability = 1, Phi Coefficient = 6.43, Contingency Coefficient = 0.99 and Cramer's V = 0.62.

All the measures show that there is very strong relationship between the variables.

There is also a negative correlation (-0.053) and the P-Value = 0.2078.

This means that as Income increases, the debit to income ratio decreases. It therefore means that people with larger incomes have lower debts. So we accept the alternate hypothesis H1.

j. Credit Card Debt (Quantitative) vs Debt to Income Ratio (Quantitative : Ratio Level)
Ho: More Debit to income ratio leads to more Credit Card Debt
H1: More Debit to income ratio leads to less Credit Card Debt

From APPENDIX 19, Likelihood Ratio Chi-square probability = 1, Phi Coefficient = 10.78, Contingency Coefficient = 0.99 and Cramer's V = 0.76.
All the measures show that there is very strong relationship between the variables.
There is also a positive correlation (0.485) and the P-Value = < 0.0001. The correlation is significant at 0.01%.

This means that, as debit to income ratio increases, there is a tendency for credit card debt to rise too. Therefore, we accept the null hypothesis H0.

k. Income (Quantitative) vs Credit Card Debt (Quantitative)

H0: More income leads to less Credit Card Debt
H1: More income leads to more Credit Card Debt

From APPENDIX 20, Likelihood Ratio Chi-square probability = 1, Phi Coefficient = 8.29, Contingency Coefficient = 0.99 and Cramer's V = 0.8.
All the measures show that there is strong relationship between the variables.
There is also a positive correlation (0.586) and the P-Value = < 0.0001. The correlation is significant at 0.01%.
This means that, income increases, there is a tendency for credit card debt to rise too. Therefore, we reject the null hypothesis H0, and accept the alternate hypothesis H1.
People with more money tend to spend more with credit cards, as the bank places a higher limit on the amount of money they can spend on their credit cards given the fact that their higher income is a guarantee they will pay back.

l. Income (Quantitative) vs default on loans (Nominal)
H0: More income leads to less default on loans
H1: More income leads to more default on loans
From APPENDIX 21, Likelihood Ratio Chi-square probability = 0.02, Phi Coefficient = 0.48, Contingency Coefficient = 0.43 and Cramer's V = 0.48.
All the measures show that there is moderate relationship between the variables.
There is also a negative correlation (-0.085) and the P-Value = 0.059. The correlation is significant at 5%.
This means that, income increases, there is a decrease in the default on loans. Therefore, we accept the null hypothesis H0.

m. Income (Quantitative) vs other debts (Quantitative)
H0: More income leads to less other debts
H1: More income leads to more other debts
From APPENDIX 22, Likelihood Ratio Chi-square probability = 1, Phi Coefficient = 9.04, Contingency Coefficient = 0.99 and Cramer's V = 0.87.
All the measures show that there is strong relationship between the variables.
There is also a positive correlation (0.6045) and the P-Value = < 0.0001. The correlation is significant at 0.01%.
This means that, as income increases, there is an increase in other debts. Therefore, we accept the alternate hypothesis H1. More income earners spend more and incur more other debts.

4.0. SUMMARY

From the analysis above, the following were observed:

1. The population is not ageing as a lot of the population is between 29 and 42.
2. Only few persons pursued higher education degrees beyond the high school degree.
3. Very few members of the population earned higher than the rest of the population.
4. Job mobility is high as only few stay longer at a particular work place.
5. Many people in the population are in debt.
6. About 29% of the population defaulted on loans.
7. Younger people had more defaults on loans. There are prevailing factors peculiar with the young: fashion, compulsive Disorder e.g. : shopaholics, little or no stable employment, loan seeking for education, misplaced or misguided priorities, and social factors.
8. We also noticed that the number of people who defaulted on loans were more among those with a college degree (22 out of 30: 73%). Moreover, this was expected, as there is currently a rise in college loan debts among college students.
9. People with larger incomes had lower debts (negative correlation between income and debt to income ratio).
10. As debit to income ratio increased, there was a tendency for credit card debt to rise too.
11. As income increased, there is a tendency for credit card debt to rise too. People with more money rend to spend more with credit cards, as the bank places a higher limit on the amount of money they can spend on their credit cards given the fact that their higher income is a guarantee they will pay back.
12. As income increased, there is a decrease in the default on loans.
13. As income increased, there is an increase in other debts. More income earners spend more and incur more debts.