Statistical Data Analysis: Uni-variate Data Analysis Project

Emal Ismail

The American University of Afghanistan
Division of Science, Technology, and Mathematics
Instructor: Asadullah Jawid

November 23, 2019

Contents

1 Summary

In conclusion, based on the mean and standard deviation uni-variate analysis of our continues and discrete variables through sub-sampling we can conclude that in first province which has lower average temperature and comparatively low average rainfall rate as well. Both factors has affected the average livestock income of farmers to be less than the average of the second and third provinces. Furthermore, in sub-sample analysis in-dept comparison cross analysis of different variables has been done which illustrates numerical variables differences through critical measures of dispersion analysis.

2 Introduction

This paper summarizes findings from farmers of 3 different provinces of Afghanistan. The survey sample included 1503 farmers owners/operators. Questionnaires were asked as an structured interview with the aforementioned farmers. Moreover, The survey project was motivated by a number of questions. And all questions asked respondents to comment on farm characteristics, practices, tools, facilities and personal conditions that latter on encoded as number of categorical and numerical variables in accumulated data-set. Hence, categorical variables of our data-set are as follow: Province, Solar Panel, Access to Internet, Generator Main Occupation Farming and Literacy. Furthermore, the numerical variables of our data-set is of eight variables that are Livestock Net Income, Temperature, Rainfall, Mobile, TV , Motor-Bike, Car and No Hrs with Electricity.

3 Methods

In this paper for analysis of categorical and numerical variables we are going to use different numerical measures such as central tendency, variation, distribution and also we have used simple frequency table for categorical data. Since, frequency tables are useful for analyzing categorical data and for screening data for data entry errors we measured our categorical variables with them. Subsequently, in order to better communicate our results we have used visualization tools such as pie-charts for categorical variables as well as histograms for our continues and discrete numerical variables. Furthermore in sub-sample analysis we have used measures of dispersion (mean/standard deviation) in order for assessing our numerical data spread and dispersion in order for better comparison.

4 Results and Discussions

Table 1: Summary Statistics of Categorical Variables

Variable	Categories	Counts	Percent
Province	[1],[2],[3]	[555],[495],[452]	[36.9],[32.9],[30.2]
Solar Panel	Yes[1], No[0]	101[1401]	6.73[93.27]
Access to Internet	Yes[1],No[0]	238[1264]	15.84[84.14]
Generator	Yes[1],No[0]	148[1354]	9.86[90.14]
Main Occupation Farming	Yes[1],No[0]	1194[308]	79.49[20.51]
Literate	Yes[1],No[0]	880[622]	58.58[41.43]

Table 1 has produced frequency counts and percentages for categorical variables of our population. This procedure serves as a summary reporting tool and is often used to analyze survey data. The categorical variables is consist of Province, Solar Panel, Access to Internet, Generator Main Occupation Farming and Literacy.

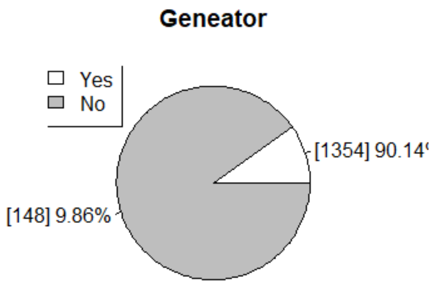

Figure 1: Farmers with Generator.

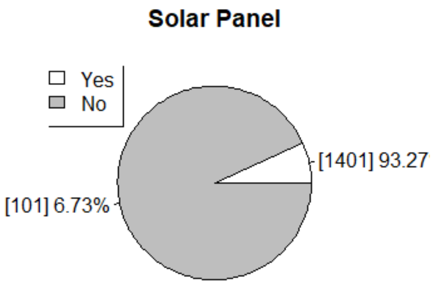

Figure 2: Farmers with Solar Panel.

In Figure 1 we can see that from 1503 framers only 148 of them has generator as an source of alternative electricity generation which makes only 9.86 percent of the whole population. Furthermore, as shown in Figure 2 among 1503 farmers only 101 which makes 6.73 percent of the whole population uses solar panel as an alternative source of electricity generation.

Table 2: Descriptive Statistics of Numerical Variables

Variable	Mean	Std.Dev	Min	Max	Skew	Kurt
Livestock Net Income	32871.58	26886.325	-22400	108000	0.5272	-0.4759
Temperature	10.95	2.764	7.027	17.392	0.2456	-1.0627
Rainfall	34.5422	18.658	10.809	80.396	0.6872	-0.5856
Mobile	3.161	2.237	0	20	2.062	7.74
TV	0.7483	0.7194	0	12	5.339	78.692
Motor-Bike	0.7523	0.7742	0	10	3.2033	29.764
Car	0.2396	0.4743	0	3	1.9868	4.3821
No Hrs with Electricity	4.083	4.029	0	20	1.2566	1.4968

As shown in Table-2 we have produced means, standard deviations, Minimums, Maximums, skewness and Kurtosis for various numerical variables of our data-set. It is consist of eight variables that are Livestock Net Income, Temperature, Rainfall, Mobile, TV , Motor-Bike, Car and No Hrs with Electricity. we can use this table to summarize data containing a combination of continuous and categorical variables for our sub-sample analysis.

Hence, in skewness and Kurtosis columns we can see that four of the variables seems to be out of range for acceptable normal distribution of data. Based on the skewness of Motor-Bike variable shown in Tale 2, which is equal to 3.20336 we can conclude that the distribution seems to be highly skewed. Additionally, kurtosis of Motor-Bike variable which is equal to: 29.76484 is greater or less than -2 and +2. Therefore it's not considered to be acceptable in order to prove normal uni-variate distribution.

subsequently, for the three other variables TV, Mobile and car as shown in the table two their skewness is not in range of -1 and 1 and also their kurtosis is greater than 2. Therefore, these variables distribution is highly skewed as well as there kurtosis is not considered to be acceptable in order to prove normal uni-variate distribution.

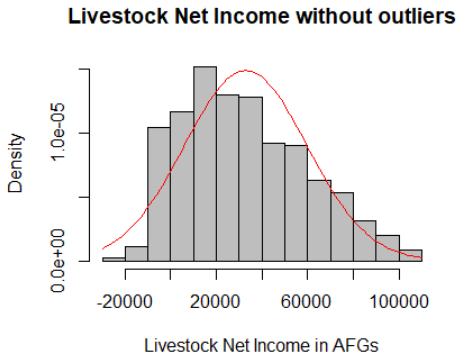

Figure 3: Livestock Net Income without Outliers.

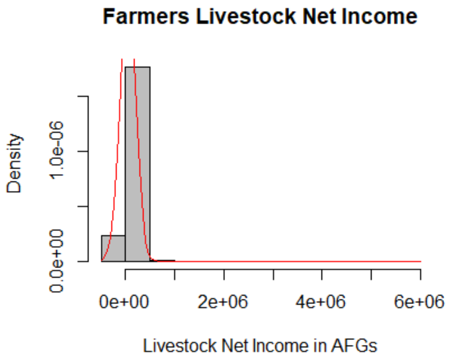

Figure 4: Livestock Net Income with Outliers.

As shown in Figure 2 Livestock Net Income variable data normal range is between -181600 and 5754000 which is not a good representation of our sample Because of huge outliers in Q1 and Q3. subsequently, Based on the skewness of Figure 1. which is equal to: 31.48262 we can conclude that the distribution seems to be highly skewed. Additionally, kurtosis of Figure 2 which is equal to: 1115.594 is greater or less than -2 and +2. Therefore, it's not considered to be acceptable in order to prove normal uni-variate distribution.

As shown in Figure 1. as well as Table 2. for better data analysis we have removed the outliers of Livestock Net Income numerical continues variables through limiting the data between -25000 and 110000. The data greater or smaller the aforementioned range considered outliers or in simple words those data was not able to illustrate the average required data for the livestock net income of farmers. subsequently, based on the skewness of Livestock net income which is equal to: 0.527215 and also shown in Table 2 we can conclude that the distribution is moderately Skewed. Where it has long right tail therefore it is Right-skewed distributions and also called positive-skew distribution. Additionally kurtosis of Livestock net income which is equal to: -0.4759578 is between -2 and +2 which is considered acceptable in order to prove normal uni-variate distribution.

Table 3: Sub-Sample Analysis

Variable	Mean/Std.Dev Province[1]	Mean/Std.Dev Province[2]	Mean/Std.Dev Province[3]
Livestock Net Income	24990(26138)	28654(28201)	30517(27651)
Rainfall	31.2 (10.8)	50.4(18.6)	21.2(12.7)
Temperature	8.29(1.18)	11.6(2.3)	13.4(1.4)

As shown in Table 3 we have analyzed variables across various sub-samples. In the above table, we investigate the livestock net income , temperature and rainfall of our variables across three provinces of the accumulated data-set. The comparison between each variable has been done through measures of spread or dispersion. Standard deviation can be illustrated as the square root of the average squared deviation of the data from the mean which explain the spread of dataset and can be used for comparison between two numerical variables.

As show in Figure 5 as well as table 2 the Livestock Net Income Mean/Std.Dev across all three provinces seems to have approximately the same with standard deviations, ranging between 26000 and 27600, but with slightly different means. Comparing the three dataset standard deviations shows that the data in the first and second province is slightly less spread out than the data in the third province dataset. Therefore, we can conclude that the average livestock net income of the third province seems to be higher compared to the first and second provinces.

Subsequently, based on table 2 and Figure 6 rainfall variable's Mean/Std.Dev across all three provinces seems to have different mean and as well as very different standard deviations. Comparing the three dataset standard deviations shows that the data in the first, second and third province dataset has differently spread out as shown in the Table 2 as well. Therefore, we can conclude that the average rainfall rate of the second province seems to be higher compared to the first and third provinces and also the rate of rainfall for second province is slightly higher than third province.

Finally, according to data of Table 2 and Figure 7 temperature variable's Mean/Std.Dev across all three provinces seems to have slightly different mean as well as approximately same standard deviations. Comparing the three dataset standard deviations shows that the data in the first, second and third province dataset is almost spread out the same as shown in the Table 2. Therefore, we can conclude that the average Temperature rate of the all three provinces are almost the same range between 8 and 14 but the third province temperature seems to be higher compared to first and second. Additionally, the second province temperature is higher compared to the first province.

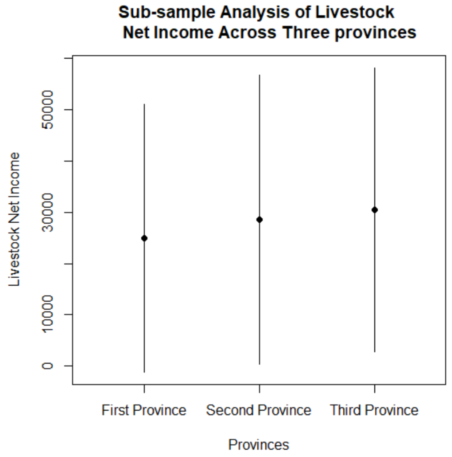

Figure 5: Livestock Net Income Mean/Std.Dev across three provinces.

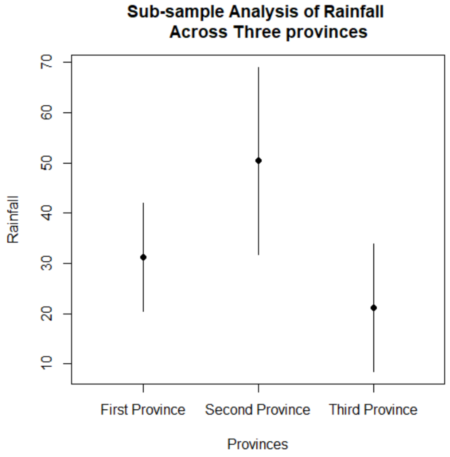

Figure 6: Rainfall Mean/Std.Dev across three provinces.

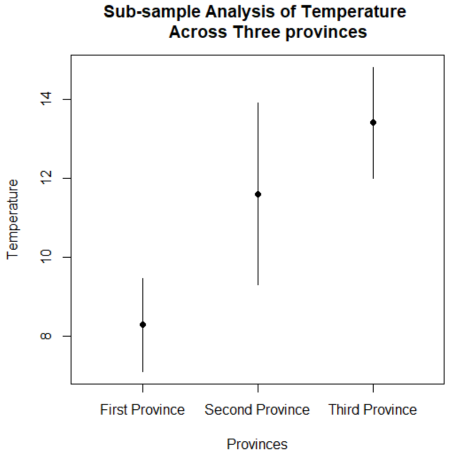

Figure 7: Temperature Mean/Std.Dev across three provinces.