ASSESSMENT 7

# QUESTION 1

Using the merged credit data sets, created in session 6, generate 5-number summaries (minimum, 25th percentile, median, 75th percentile and maximum) for each numeric variable.

For age, amount of loan, duration of loan and instalment rate as a percentage of income, decide using the 5-number summary whether the variable is likely to be normally distributed and explain your reasoning. For each of these variables - what is the dispersion and what is the central tendency as measured by the 5-number summary?

# ANSWER 1

**proc** **means** data =LOAN\_RISK min p25 median p75 max;

var age

amount

duration

instalment;

**run**;

|  |  |  |  |
| --- | --- | --- | --- |
| VARIABLE | DISPERSION | CENTRAL TENDENCY | VARIABLE DISTRIBUTION |
| Age | Minimum – 19years  25th percentile - 26.5  75th percentile - 42.0  Maximum – 74years | Median is 33.000 years | The min, 25th percentile values are closer to the median than the 75th percentile and maximum values. Majority of the data is distributed to the left. Similarly, the average of the min and max (55) and the average of the 25th and 75th percentile (15.5) are not close to the median. This suggests that the shape of the distribution is skewed to the right (positive skew). |
| Amount of loan | Minimum – 276.00  Lower percentile 25% - 1389.50  Upper percentile 75% - 4231.00  Maximum – 15945.00 | Median - 2400 | The min, 25th percentile values are closer to the median than the 75th percentile and maximum values. Majority of the data is distributed to the left. Similarly, the average of the min and max (15,669) and the average of the 25th and 75th percentile (2841.5) are not close to the median. This suggests that the shape of the distribution is skewed to the right (positive skew). |
| Duration of loan | Minimum – 4  Lower percentile 25% - 12  Upper percentile 75% - 24  Maximum – 60 | Median - 18 | The min, 25th percentile values are closer to the median than the 75th percentile and maximum values. Majority of the data is distributed to the left. Similarly, the average of the min and max (54) and the average of the 25th and 75th percentile (12) are not close to the median. This suggests that the shape of the distribution is skewed to the right (positive skew). |
| Instalment rate as a percentage of income | Minimum – 1  Lower percentile 25% - 2  Upper percentile 75% - 4  Maximum – 4 | Median - 3 | Due to the reduced variation in the values of the data, it is difficult to conclude the distribution. The average of the min and max (3) suggests a normal distribution, but the average of the 25th and 75th percentile (2) indicates the data is not normally distributed.  Further analysis will show if the data is skew. |

# Question 2

Univariate analysis: proc univariate

When compared with proc means, proc univariate provides many additional output tables. With the correct choice of options, proc means can also calculate these values but is best used when comparing sub-groups defined by categorical data.

Self-Assessment Question

Using proc univariate on age, amount of loan, duration of loan and instalment rate as a percentage of income, decide whether each variable is normally distributed and justify your answer. Compare the results obtained using proc means with that using proc univariate.

Hint: the purpose of this question is to compare the basic output of proc means with the basic output of proc univariate. Please do not use other features until asked to do so.

# ANSWER 2

The basic output for proc means returns an output of the specified summary statistics defined in the code, in this example, we specified the 5-number summary which returned the min 25 percentile, median, 75th percentile and max. The output can be used for boxplot, which will give a good understanding of the behaviour of the centre, spread, tail and skew of the distribution.

The basic output for proc univariate returns five tables: moments, basic statistical measures, tests of location, quantiles and extreme observations. These tables show descriptive statistics and summary statistics of the variable using the var statement as well as dispersion and central tendency of the distribution.

/\*QUESTION 2\*/

**proc** **univariate** data = LOAN\_RISK;

var age

amount

duration

instalment;

**run**;

|  |  |  |  |
| --- | --- | --- | --- |
| VARIABLE | PROC MEANS | PROC UNIVARIATE | VARIABLE DISTRIBUTION |
| Age | Minimum – 19years  Lower percentile 25% - 26.5  Median is 33.000 years  Upper percentile 75% - 42.0  Maximum – 74years | 0% Min – 19years  25% Q1 - 26.5  50% Median - 33.000 years  75% Q3 - 42.0  100% Max – 74years | Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. It suggests that the shape of the distribution is skewed to the right (positive skew). |
| Amount of loan | Minimum – 276.00  Lower percentile 25% - 1389.50  Median - 2400  Upper percentile 75% - 4231.00  Maximum – 15945.00 | 0% Min – 276.00  25% Q1 – 1389.5  50% Median – 2400  75% Q3 - 4231  100% Max – 15945 | Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. It suggests that the shape of the distribution is skewed to the right (positive skew). |
| Duration of loan | Minimum – 4  Lower percentile 25% - 12  Upper percentile 75% - 24  Maximum – 60 | Median - 18 | Majority of the data is distributed to the left.  The 25th percentile and 75th percentile are close to the median equally. It suggests most of the distribution is close to the centre. The extremes are spread unevenly; the maximum value is farther from the median than the min value. It suggests that the presence of outliers to the right making the distribution positively skewed. |
| Instalment rate as a percentage of income | Minimum – 1  Lower percentile 25% - 2  Upper percentile 75% - 4  Maximum – 4 | Median - 3 | The distribution is probably normal. |

# QUESTION 3

Univariate analysis: histograms

If a variable is normally distributed, then the average of the minimum and maximum, as well as the 25th and 75th percentile, should equal the median. Taking age as an example, these values are 46.5 and 34.25 respectively, where the median is 33, suggesting that age is skewed toward larger values.

###### Self-Assessment Question

The extent to which the distribution deviates from normal can be assessed visually using proc univariate's [histogram](https://documentation.sas.com/?docsetId=procstat&docsetTarget=procstat_univariate_syntax09.htm&docsetVersion=9.4&locale=en) command:

proc univariate data=LOAN\_RISK;  
 histogram age;  
run;

Use the histogram command to examine the distributions of age, amount of loan, duration of loan and instalment rate as a percentage of income.

Review the paper by Park (2008). Manually (do not use SAS) create a table showing how the values of skewness and Kurtosis relate to the histograms you have drawn and the 5-number summary. (At the end of this session, you may find it an interesting challenge to create a sequence of histograms which display the 5-number summary, the skewness and the Kurtosis is an inset table but this is not obligatory.)

Hint: the histogram command allows you to overlay the normal distribution:

proc univariate data= LOAN\_RISK;  
 histogram age / normal(mu= est sigma= est);  
run;

# ANSWER 3

/\*question 3\*/

**proc** **univariate** data= LOAN\_RISK;

histogram age

amount

duration

instalment / normal(mu= est sigma= est);

inset min median skewness kurtosis;

**run**;

|  |  |  |
| --- | --- | --- |
| VARIABLE | SKEWNESS | KURTOSIS |
| Age | From the diagram of the histogram, most of the data is to the left of the distribution. When compared to a normal distribution and the value of skewness 1.062524, this distribution is not normal because the skewness value is >0, which means it is positively skewed. | The value of Kurtosis is 0.663045 (SAS uses Kurtosis-3) the value is more than 0, which suggests the peak is higher than a standard normal distribution. |
| Amount of Loan | From the diagram of the histogram, most of the data is to the left of the distribution. When compared to a normal distribution and the value of skewness 1.831357, this distribution is not normal because the skewness value is >0, which means it is positively skewed. | The value of Kurtosis is 3.747165 (SAS uses Kurtosis-3) the value is more than 0, which suggests the peak is higher than a standard normal distribution with thin tails. |
| Duration of Loan | From the diagram of the histogram, most of the data is to the left of the distribution. When compared to a normal distribution and the value of skewness 1.116218, this distribution is not normal because the skewness value is >0, which means it is positively skewed. | The value of Kurtosis is 0.902889 (SAS uses Kurtosis-3) the value is more than 0, which suggests the peak is higher than a standard normal distribution. |
| Instalment | From the diagram of the histogram, most of the data is to the right of the distribution. When compared to a normal distribution and the value of skewness -0.45983, this distribution is not normal because the skewness value is <0, which means it is negatively skewed. | The value of Kurtosis is -1.24942 (SAS uses Kurtosis-3) the value is less than 0, which suggests the peak is lower than a standard normal distribution. |

# QUESTION 4

##### Univariate analysis: theory-driven plots

Park (2008) describes the P-P plot and the Q-Q plot as theory-driven visualisations of fit with the normal distribution. These can be obtained using the following code:

proc univariate data= LOAN\_RISK;  
 var age;   
 ppplot age;  
run;

or

proc univariate data= LOAN\_RISK;  
 var age;  
 qqplot age / normal(mu= est sigma= est);  
run;

###### Self-Assessment Question

The univariate procedure produces a table of goodness-of-fit tests for a variable by adding normaltest to the procedure options after data=. Manually (do not use SAS) create a table showing how the results of these tests relate to the P-P plot and the Q-Q plot for each of age, amount of loan, duration of loan and instalment rate as a percentage of income.

# ANSWER 4

Including the normaltest statement returns a summary output table called ‘Test for Normality’ which includes statistical numeric methods:

Shapiro-Wilk (W)

Kolmogorov-Smirnov (D)

Cramer-Von Mises (W-Sq)

Anderson-Darling (A-Sq)

When the p-Value(α) is less than the chosen significant level 0.05, the null hypothesis is rejected, and the distribution is not normally distributed.

It is important to note that it would not be enough to use just the numeric method. A graphic method by including **p-p plots (probability plot) and q-q plot (quantile plot) to** the procedure will compare the deviation of our distribution to a standard p-p plot and q-q plot.

/\*QUESTION 4\*/

**proc** **univariate** data= LOAN\_RISK normaltest;

var age

amount

duration

instalment;

ppplot age

amount

duration

instalment;

**run**;

**proc** **univariate** data= LOAN\_RISK normal;

var age

amount

duration

instalment;

qqplot age

amount

duration

instalment / normal(mu= est sigma= est);

**run**;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Shapiro-Wilk (W)** | **Kolmogorov-Smirnov (D)** | **Cramer-Von Mises (W-Sq)** | **Anderson-Darling (A-Sq)** | **P-P Plot and Q-Q plot** |
| Age | p-Value = <0.0001  The null hypothesis that suggests the distribution is normally distributed is rejected because the p-value is less than 0.05. | p-Value = <0.0100  The null hypothesis that suggests the distribution is normally distributed is rejected because the p-value is less than 0.05. | p-Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p-Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | The P-P plot derived from the distribution show the residuals deviate from the line of fit with most points at the top of the line.  The Q-Q plot is curved around the line with most values away from the line.  This confirms the data is not normally distributed. |
| Amount | p Value = <0.0001  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0100  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | The P-P plot derived from the distribution show the residuals deviate from the best line of fit with most points at the top of the line.  The Q-Q plot is upwards of the line with most values away from the line.  This confirms the data is not normally distributed. |
| Duration | p Value = <0.0001  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0100  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | The P-P plot derived from the distribution show the residuals deviate from the best line of fit with most points at the top of the line. The Q-Q plot have step like residual with most values away from the line.  This confirms the data is not normally distributed. |
| Instalment | p Value = <0.0001  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0100  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | p Value = <0.0050  The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. | The P-P plot derived from the distribution show the residuals deviate from the best line of fit with most points scattered above the line.  The Q-Q plot have individual plots around the line with most values away from the line.  This confirms the data is not normally distributed. |

# QUESTION 5

##### Univariate analysis: extremes of the distribution

The 5 most extreme observations are listed in the usual output of proc univariate. It is sometimes useful to examine more observations than 5 and to be able to select the actual from a data set containing a primary key variable that has a unique value for each data subject.

###### Self-Assessment Question

Part a: write code to list the frequency of the most extreme values from the distributions of age, amount of loan, duration of loan and instalment rate as a percentage of income, using proc univariate.   
  
Hints:

* surrounding proc univariate with ods select extreme values (before) and ods select all (after) causes the procedure to print only the table of extreme values.
* Proc univariate's nextrval option controls the number of values output.

Part b: write code to print in separate tables the 10 most extreme observations from the distributions of age, amount of loan, duration of loan and instalment rate as a percentage of income.   
  
Hints:

* surrounding proc univariate with ods select extreme jobs (before) and ods select all (after) prints only the table of extreme observations.
* Proc univariate's nextrobs option controls the number of values output.
* Use of by followed by the primary key variable within proc univariate sets the key value for the list of extreme observations
* A possible strategy for this task is:
  1. Create the correct output using proc univariate
  2. Surround the statement with ods csvall file= '/path/to/the/CSV/file' (before) and ods csvall close (after) to output the results to a CSV file
  3. Import the CSV file into a suitable SAS data set
  4. Use proc SQL to create a table containing the appropriate observations
  5. Use proc print to print a report of the observations

# Answer 5

/\*Question 5a\*/

title 'Extreme Observations 5';

ods select ExtremeObs;

**proc** **univariate** data= LOAN\_RISK nextrval= **5**;

ods select all;

var age

amount

duration

instalment;

**run**;

/\*Questiion 5b\*/

/\*create CSV file\*/

title 'Age ExtremeObs';

ods select extremeObs;

ods csvall file='C:\Users\Folashikemi\OneDrive - De Montfort University\P2586104\Documents\IMAT5168 Analytical Programming Mark\Lab 7\age\_extremeobs.csv';

**proc** **univariate** data= LOAN\_RISK nextrobs= **10**;

var age;

id customer;

**run**;

ods select all;

ods csvall close;

title 'Amount ExtremeObs';

ods select extremeObs;

ods csvall file='C:\Users\Folashikemi\OneDrive - De Montfort University\P2586104\Documents\IMAT5168 Analytical Programming Mark\Lab 7\amount\_extremeobs.csv';

**proc** **univariate** data= LOAN\_RISK nextrobs= **10**;

var amount;

id customer;

**run**;

ods select all;

ods csvall close;

title 'Duration ExtremeObs';

ods select extremeObs;

ods csvall file='C:\Users\Folashikemi\OneDrive - De Montfort University\P2586104\Documents\IMAT5168 Analytical Programming Mark\Lab 7\duration\_extremeobs.csv';

**proc** **univariate** data= LOAN\_RISK nextrobs= **10**;

var duration;

id customer;

**run**;

ods select all;

ods csvall close;

title 'Instalment ExtremeObs';

ods select extremeObs;

ods csvall file='C:\Users\Folashikemi\OneDrive - De Montfort University\P2586104\Documents\IMAT5168 Analytical Programming Mark\Lab 7\instalment\_extremeobs.csv';

**proc** **univariate** data= LOAN\_RISK nextrobs= **10**;

var instalment;

id customer;

**run**;

ods select all;

ods csvall close;

# QUESTION 6

##### Univariate analysis: transforming data

When a measurement variable does not fit a normal distribution and you are using parametric tests (that depend on this being the case) then it is reasonable to try a transformation.

###### Self-Assessment Question

For each of age, amount of loan, duration of loan and instalment rate as a percentage of income that is not normally distributed, find a suitable transformation that improves the fit with a normal distribution.   
  
Hint: Search for suitable transformations using the search terms: "statistics transforming data".

# ANSWER 6

/\*Question 6\*/

**data** LOANRISK;

set LOAN\_RISK;

/\* log each variable to base to shrink the extreme values\*/

LOGage=log10(age);

LOGamount=log10(amount);

LOGduration=log(duration);

LOGinstalment=log(instalment);

**run**;

title 'Transformed Histogram';

**proc** **univariate** data = LOANRISK nextrval=**5**;

var LOGage

LOGamount

LOGduration

LOGinstalment;

histogram / normal(color=red mu=**10** sigma=**0.5**);

inset mean std normal / position= ne;

**run**;

### Question 7

##### Univariate analysis: frequencies

A fundamental requirement of relational database tables is that the primary key is unique and not null (or missing).

###### Self assessment question

How would you check this requirement for customer id number of the credit data set and its ancillary files (the ratings and the dates of assessment) before the files are joined together?

# ANSWER 7

/\*Question 7\*/

**proc** **freq** data = customer;

tables customer;

**run**;

**proc** **freq** data = assessment;

tables customer;

**run**;

**proc** **freq** data = rating;

tables customer;

**run**;