Getting started with SAS Studio

Prior to performing statistical analysis, the type of variables present in the analysis must be determined to ensure that the appropriate methods of analysis are used.

For the purpose of this course these can be broadly defined into **two** categories:

Continuous data – this represents *quantitative* data having a continuous range of values, for example, the height (in cm) of trees in the Amazon rain forest.

Categorical data – by contrast, this represents *qualitative* data and are discrete, meaning that they can assume only certain fixed numeric or non-numeric (text) values; in this type of data the order the data appears does not always matter.

For example, we may use and 1 and 2 to represent male and female, clearly female is not twice male; thus these numbers are simply labels.

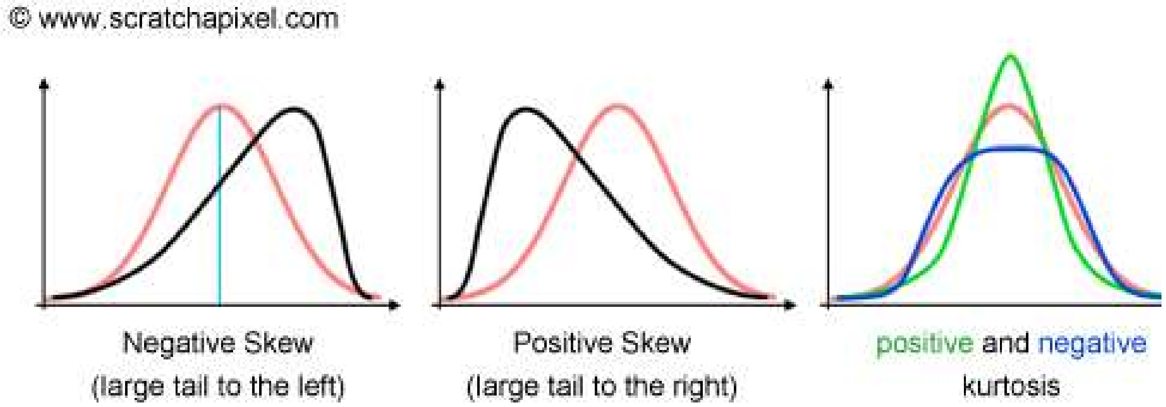
On the other hand, in surveys people are commonly asked to express their opinion by giving a number (e.g. 0 to 100), meaning that the order is important. Summary statistics are not always appropriate with this type of data.

Summary Statistics

The main use of summary statistics is to obtain a succinct description of the distributional behaviour of the data being analysed. Several summary statistics are available in SAS Studio to describe the distribution of continuous data used in the analysis.

Sample size	This is the number of observations in the data.
Mean	The mean (arithmetic) is the sum of the response divided by the number of observations (this is commonly referred to as the average).
Median	This is the middle value in the data when ranked in ascending order.
Mode	The mode is the most common (frequent value), and can be seen as a peak on a histogram.
Interquartile range	This is the difference between the upper and lower quartiles. If we divide the ascending data into two separate groups (high and low) and calculate the median of both of these groups, we find that Median (high) – Median (low) = Interquartile Range.
Variance	This provides a measure of the dispersion of the data around the mean. It is the average of the sum of the squared distances from the mean.
Standard Deviation	This is the square root of the variance.

Coefficient of variation	This is generally expressed as a percentage, it is equal to: $\frac{\text{Standard Deviation}}{\text{Mean}}$ (x 100 for percentage). The use of the coefficient of variation lies partly in the fact that the mean and standard deviation tend to change together in many experiments. A knowledge of relative variation is valuable in evaluating experiments. The smaller this quantity is, the less variation there is in the data.
Correlation	The correlation coefficient (*r)* measures the strength of association between two variables. A correlation coefficient near to 1 implies a strong positive relationship, near 0 implies the variables are independent and near –1 implies a negative relationship. See Figures (a) – (d) below for examples.
Skewness	This is the degree of asymmetry of a distribution. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is heavier than the right tail. Similarly, skewed right means that the right tail is heavier than the left tail.
Kurtosis	This represents the degree of peakdness of the data. In SAS a normal distribution has a kurtosis of 0. A distribution with a high peak (value > 0) is called *leptokurtic*, a flat-topped curve is called (value < 0) *platykurtic* and the normal distribution (value = 0) *mesokurtik*.



Exercise 1

The nationality, sex and key IQ characteristics of 40 delegates attending a Psychology seminar in USW were recorded.

The data can be found in IQ Data, and the variables used in the analysis were as follows:

Sex: Male or Female;

Nationality: British, French, German or Spanish;

FSIQ: Full Scale IQ scores based on the four Wechsler (1981) subtests;

VIQ: Verbal IQ scores based on the four Wechsler (1981) subtests;

PIQ: Performance IQ scores based on the four Wechsler (1981) subtests;

MRI_Count: total pixel Count from 18 MRI scans.

To obtain Summary statistics for this data:

- Select **Tasks** then **Statistics** then **Summary Statistics**.
- Under the **DATA** tab select the **IQ** data set.
- Add **FSIQ**, **VIQ** and **PIQ** to the Analysis variable list.
- Add **Nationality** to the Classification variable list.
- Select **edit** to add to the statistics.
- Select the **Options** tab to view the Statistics that can be selected. By default this includes **Mean**, **Standard deviation, Min, Max** and **Number of observations**. There are a number of other measures that can be selected under the Additional Statistics drop down. Select **Median** and **Coefficient of variation** under the Additional tab.

- To run the code select the SAS running image. 🏃

Your output will be displayed in the right hand pane.

The following descriptive statistics are obtained for the IQ scores, classified by the nationality of the individuals:

Nationality	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	N
British	17	FSIQ	112.4705882	24.6148270	81.0000000	144.0000000	17
		VIQ	112.7058824	22.8439399	83.0000000	150.0000000	17
		PIQ	109.0000000	23.6034955	74.0000000	150.0000000	17
French	7	FSIQ	105.5714286	24.3437681	85.0000000	141.0000000	7
		VIQ	105.5714286	23.4652732	86.0000000	150.0000000	7
		PIQ	104.5714286	24.1098676	84.0000000	147.0000000	7
German	9	FSIQ	123.0000000	18.7616630	96.0000000	140.0000000	9
		VIQ	119.8888889	21.3157480	90.0000000	150.0000000	9
		PIQ	120.3333333	17.1755640	90.0000000	147.0000000	9
Spanish	7	FSIQ	111.4285714	29.5852280	77.0000000	141.0000000	7
		VIQ	108.5714286	30.3526887	71.0000000	145.0000000	7
		PIQ	110.4285714	25.1253998	72.0000000	132.0000000	7

If we view the code produced for this analysis within the **CODE** window which is automatically generated, we can identify the **Means Procedure** below.

```
proc means data=MS4S08_1.IQ_DATA chartype mean std
min max n vardef=df;
    var FSIQ VIQ PIQ;
    class Nationality;
run;
```

We are able to edit this code by selecting edit along the toolbar. To add the variables Skewness and Kurtosis to the summary table we add,

```
skew kurt
```

to the list of statistics. We then select **Run** again**.**

Once the output has been generated it is important to analyse the results. It can be useful to add some plots to interpret this.

What does the data tell us about the different IQ scores for different Nationalities?

A - Germany are the smartest country, followed by Britain, then Spain and lastly French?

Exercise 2

A government research group have been asked to investigate the fuel consumption habits of the UK. They collected data from 468 households in the UK. The data can be found in the SAS data sets **HH_Car_Survey**, **HH_Surveyor** and **Manufacturer_Data**. Descriptions of the variables used in the study are outlined below.

HH_Car_Survey
HH_ID - Household Id
Make - Make of the household's primary car
Model - Model of the household's primary car
Fuel - Type of fuel (petrol or diesel)
Engine_size_l - Engine size in litres
Annual_Mileage - Estimated annual household mileage
Annual_Cost - Annual car running cost (£s)
Gender - Gender of primary driver (1=male, 0=female)
Age - Age of primary driver
Region - Region of residence
Years_licence - No. of years of holding a driving licence

HH_Surveyor
HH_ID - Household Id
Surveyor1- Surveyor 1 score on driving efficiency of driver (1=inefficient & 10 = efficient)
Surveyor2 - Surveyor 2 score on driving efficiency of driver (1=inefficient & 10 = efficient)

Manufacturer_Data
Make - Make of car
Model - Model
Fuel - Type of fuel (petrol or diesel)
Engine_size_l – Engine size in litres
Engine_size_cc – Engine size in cc
C02_emission - CO2 emissions (g/km)
MPG - Fuel consumption of car in miles per gallon (mpg)
Zero_to_60 - Time taken in seconds to go from 0 to 60 mph

You are required to **explore** the variables to see if there are any interesting results, that is: were there observable differences in **Age, Gender, Make, Annual_Mileage, Annual_Cost** and **Region**. Identification of rogue values (outliers) is also important so that they can be omitted prior to the analysis.

The structure for reporting on this task could be:

- Outline the purpose of the analysis
- Describe the methods used to look at the data and assumptions made
- Summarise the key results obtained from the analysis
- Make overall conclusions from the data based on the output obtained