Dr. Florence Nansubuga

Outline

- Checking errors
- Dealing with missing values
- Normalising data by removing outliers

Checking for errors

Purpose

- Primarily to check whether the values are entered correctly e.g. if sex has two values 1 and 2 any other value that could have been entered erroneously can distort the statistical results if not corrected.

Procedure

1. Inspect the data file and remove cases with a whole range of variable items that are not filled
2. Go to analyse
3. Click descriptive analysis then frequency
4. Select and enter a set of variables with the same values in the dialogue box
5. Click statistics, tick minimum and maximum
6. Click continue and then ok
7. Check the output generated to find the values that are out of range
8. Click search forward to scan for the incorrect value and change it

Missing data

Why do missing values occur?

- Missing values are either random or in a systematic pattern.
- Random missing values may occur because the respondents inadvertently did not answer some questions.
  - For example, the study may be overly complex and/or long, or
  - the subject may be tired and/or not paying attention, and miss the question.
- Random missing values may also occur through data entry mistakes

Missing data

- Patterned missing values may occur because a particular segment of respondents purposefully did not answer some questions.
  - For example, single women my refuse to answer a particular question because of social desirability concerns.
  - Also, the question may not provide appropriate answer choices, such as “no opinion” or “not applicable”, so the subject chooses not to answer the question.

Problematic missing data

Rates of

- Less than 1% missing data are generally considered trivial,
- 1-5% manageable.
- 5-15% require sophisticated methods to handle, and
- More than 15% severely impact any kind of interpretation (can result in type 1 and type 2 errors).
- If a variable or case has more that 5% missing data, consider removing that variable or case before analysis (exclude listwise option).
- According to Figueredo (1999), listwise deletion is less hazardous if it involves minimal loss of sample size and where there is no patterned missing data

Examining Missing data

- The best way to control missing data is to ensure reliability and validity of the measures thru a pilot study.
- In addition quality assurance should be key during the data collection process and the sample size should be adequate.
- However in situations where missing data is inevitable; use the following steps to inspect the nature missing values; *(lets try this using supervision on our trial file)*
  1. Go to analysis
  2. Click missing value analysis
  3. Enter variables in quantitative variables dialogue box
  4. Check EM (Expectation-Maximization). This checks if the subjects with missing values are different from the subjects without missing values.

Examining Missing data

Interpreting analysis results (output)

1. Inspect the MVA (missing value analysis) table to establish the percentage of missing values
2. From our trial data results supper12 and supper16 had the highest missing values of 1.3% and this is manageable.
3. Inspect the ME mean and check the values under MCAR (Missing Completely at Random)
   - *If p<.05, then the subjects with missing values are different from the subjects without missing values., which indicates the missing values are patterned.*
   - *If p>.05 then the missing values are random.*
4. In our trial data for supper, p = .102 > .05, meaning that the missing values are random.

How to deal with missing values

There are three ways of dealing with missing values during analysis

1. Exclude cases list wise: This will include only cases with full data. However this method can significantly affect the sample size. List wise deletion has been shown to produce more biased estimates than alternative methods esp. in situations where the sample size is inadequate or where there is patterned missing data (Little &Rubin, 1987).
2. Exclude cases pairwise: This will exclude cases only if they are missing the data required for a specific analysis. This may affect the sample size to a minimal extent. This option is checked as a default in most of the SPSS tests.
3. Replace missing values with the mean: All missing values will be replaced by the mean value of the variable. However if the percentage of missing values is above 5%, you should not use this option as it will distort the results of your analysis.

Replacing missing values (imputation)

- If only a few percent (<5%) are missing, the data can be replaced using
  - the mean (if the data is normal),
  - median (if the data is skewed) or
  - mode (if the data is categorical).
- Where the goal is to compare several groups (e.g. gender or treatment conditions), it is often desirable to do this replacement within each group.
- As the percentage of missing data approaches or exceeds 5% a new problem arises.

Replacing missing values (imputation)

There are various estimation methods for replacing missing values namely;

1. Series mean. Replaces missing values with the mean for the entire series.
2. Mean of nearby points (moving average). Replaces missing values with the mean of valid surrounding values. The span of nearby points is the number of valid values above and below the missing value used to compute the mean.
3. Median of nearby points. Replaces missing values with the median of valid surrounding values. The span of nearby points is the number of valid values above and below the missing value used to compute the median.
4. Linear interpolation. Replaces missing values using a linear interpolation. The last valid value before the missing value and the first valid value after the missing value are used for the interpolation. If the first or last case in the series has a missing value, the missing value is not replaced.
5. Linear trend at point. Replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1 to n. Missing values are replaced with their predicted values.

Replacing missing values (imputation)

Procedure

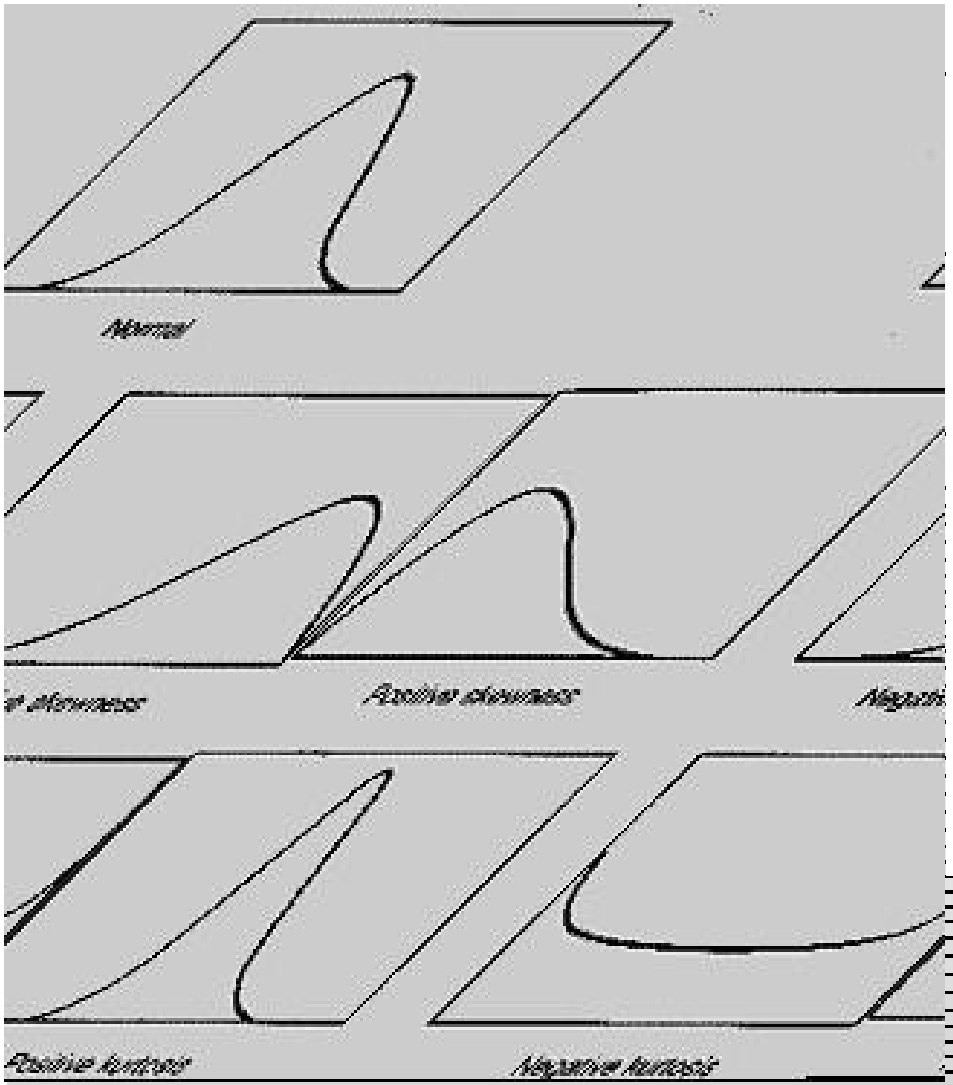
- Click transform then click replace missing values
- Select the estimation method of your choice
- Enter a set of variables with missing values in the new variables box (a maximum of 100)
- Click Ok

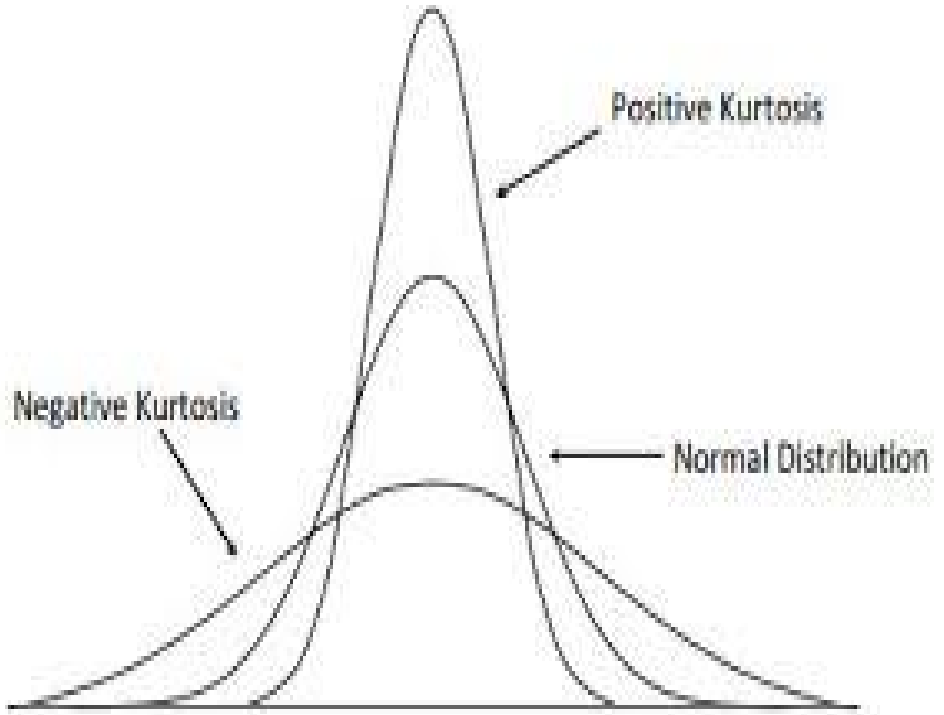
Assessing normality

Terminologies

- Normal data (bell shaped) -
- Skeweness - symmetry of distribution
- Positive skeweness– scores clustered to the left hand side of the graph at the low values (y>0)
- Negative skeweness – scores clustered to the right side of the graph at the high values (y<0)
- Kurtosis – peakedness of distribution
- Positive kurtosis – distribution is rather peaked (clustered in the centre) with long thin tails (k>0)
- Negative kurtosis – relatively flat distribution with to many cases in the extremes (k<0)
- Outliers – extreme values (activities in the tail) which affects normality

Normal curve, Skewness and Kurtosis





Assessing normality

Procedure

- Click analyse, descriptive statistics and then explore
- Click a variable of interest (e.g. Performance) and move it into the dependent list box
- Click any independent or grouping variables that you wish to split your sample (e.g. education) and move it into the factor list box
- Click on plots, under descriptive uncheck stem and leaf and check histogram
- Click continue and Ok

Assessing normality

Interpreting results

- A list of common statistics is provided e.g. mean, median, SD, range etc.
- One statistics you may not know is 5% trimmed mean. This is obtain by removing the top and bottom 5% of the cases and SPSS recalculates the new mean values. The assumption is that the top and bottom 5% of the cases in extremes (that their scores are in the tails) and they are influencing the original mean.
- Our major interest is on the test of normality. The Kolmogorov-Smirnov statistic gives the result of normality. A non significant result (p>.05) indicates normality
- In our trial data it is among the PhD group (p=.20>.05) the performance scores had a normal distributed. The rest had activities in the tails (outliers)

More about outliers

- You can now examine all items measuring performance to explore outliers item by item
- Click analyse, descriptive statistics and then explore
- Click all items measuring performance and move them into the dependent list box
- Click on statistics and then outliers, click continue
- Click on plots, under descriptive uncheck stem and leaf and check histogram
- Click continue and Ok
- Inspect the output especially the poly boxes
- Remove outliers identifies for each item