Chapter 2: Investigating your data

For an overview of key terms, refer to page 32-34

Graphical examination of your data
The more complex an analysis becomes, the need and level of understanding increases. The starting point for understanding the nature of any variable is to characterize the shape of its distribution. Many times the researcher can get a perspective of the variable by creating a **Histogram** = A graphical representation of a single variable that represents the frequency of occurrences within categories. >> If the histogram is bell shaped, it is a *normal distribution*. The histogram can be used to examine every kind of metric variable.

If the researcher is interested in examining the relations between two, creating a **scatterplot** may be useful. The points in the graph represent joint values of the variables for any case. The patterns of these points predict a relationship between the variables. If the points are close to each other and showing a straight line, there is a *linear relationship or correlation*. A curved pattern may show a nonlinear relationship, and random points may show no relationship. The researcher also has to understand the extent and character of differences. Assessing group differences is done through univariate analysis such as t-tests and analysis of variance and through multivariate techniques of discriminant analysis and multivariate analysis of variance.

Another important aspect is identifying outliers. This can be done by making a **boxplot**. The upper and lower quartiles of the data distribution form the upper and lower boundaries of the box (25^{th}-75^{th} percentile). The box presents 50 percent of the data values. The larger the box, the greater the spread. The median is depicted by a solid line in the box. Outliers (1-1.5 quartiles away from the box) and extreme values (more than 1.5 quartiles away from the box) are displayed outside of the **whiskers**.

Sometimes a researcher needs to compare observations characterized on a multivariate profile, in that case a number of **multivariate displays** center around one of three types of graphs.

1. Direct portrayal of data values either by (a) **glyphs or metroglyphs**, which are some form of circle with radii that correspond to a data value, or (b) multivariate profiles, which portray barlike profile for each observation.
2. Mathematical transformation of the original data into a mathematical relationship, which can be portrayed graphically. Andrew's Fourier transformation is the most common.
3. Graphical displays with iconic representativeness, the most popular being a face.

Missing data
Missing data, where valid values on one or more variables are not available for analysis, are a fact of life in multivariate analysis. To identify patterns in the missing data that would characterize the missing data process, the researcher asks questions like: Are the missing data randomly throughout the observations or are distinct patterns identifiable? How prevalent are the missing data? Both substantive and practical considerations necessitate an examination of the missing data process:

- The practical impact of missing data is the reduction of the sample size available for analysis.
- From the substantive perspective any statistical result based on data with a nonrandom

missing data process could be biased.

The concern for missing data processes is similar to the need to understand the causes of non-response in the data collection process.

A four step process for identifying missing data and applying remedies:

1. **Determine the type of missing data**
   Are the missing data part of the research design or are the causes truly unknown?

- Ignorable missing data: remedies are not needed. The missing data process is operation at random. There are three instances in which a researcher most often encounters ignorable missing data:

1. Resulting from taking a sample from the population instead of using the entire population
2. Due to a specific design of the data collection process, for example skipping questions that are not applicable.
3. When data are censored. Censored data are observations not complete because of their stage in the missing data process. A typical example is causes of death, people who are still living cannot provide all information about this.

- Non-ignorable missing data: these fall in two categories based on their source: known versus unknown processes:

1. Many missing data processes are known to the researcher in that they can be identified due to procedural factors, such as errors in data entry that create invalid codes, disclosure restrictions, failure to complete the entire questionnaire, or morbidity of the respondent. Researcher has little control over these factors.
2. Unknown missing data processes are less easily identified and accommodated. These instances are often related directly to the respondent

2. **Determine the extent of missing data**
   The researcher must examine the pattern of missing data and determine the extent of missing data for individual variables, individual cases and even overall. The primary issue is to explore if the missing data affect the outcomes. If it has a small effect, any of the approaches can be chosen. If the effect is larger, we first have to determine the randomness of the missing data process before selecting a remedy step.
   The most direct means of assessing the extent of missing data is by tabulating the percentage of missing variables for each case and the number of cases with missing data for each variable. The researcher should look for any nonrandomness in the data. Finally the researcher should determine the number of cases with no missing data on any of the variables which will provide the sample size for analysis if remedies are not applied. If it is determined that the extent is acceptably low and no specific randomness patterns appear, the researcher can apply any of the imputation techniques. Before proceeding to the formalized methods, the researcher should consider the simple remedy of deleting offending cases with excessive levels of missing data . This should always be based on both empirical and theoretical considerations (Rules of thumb 2-2).

3. **Diagnose the randomness of the missing data process**
   Diagnosing the randomness is necessary to determine the appropriate remedies. Of the two levels of randomness when assessing missing data, one requires special methods to accommodate a nonrandom component (Missing at Random, MAR). A second level (Missing Completely at Random, or MCAR) is sufficiently random to accommodate any type of missing data remedy. Only MCAR allows for the use of any remedy desired. The

distinction between these two levels is in the generalizability to the population, as described here:

- MAR: if the missing values of Y depend on X but not on Y. (the observed Y values represent a random sample of the actual Y samples for each value of X, but not necessarily truly random for all Y values)
- MCAR: the observed values of Y are a truly random sample of all Y values, with no underlying process that causes bias to the observed data.

  As sample size and the number of variables increases, so does the need for empirical diagnostic tests. This can be done in SPSS (Missing value analysis). These tests generally include one or both diagnostic tests:
- The first assesses the missing data process of a single variable Y by forming two groups: observation with missing data for Y and observations with valid data for Y. Statistical tests are then performed to test if there are significant differences. Significant differences indicate the possibility of a nonrandom missing data process. If the variable being compared is metric, t-tests are performed.
- The second is an overall test of randomness that determines whether the missing data can be classified as MCAR. The pattern of missing data on all variables is analyzed and compared with the pattern expected for a missing data process. If no significant difference is found, the data can be classified as MCAR.

  As a result of these tests, the missing data process is classified as either MAR or MCAR.

4. **Select the imputation method**

   Imputation is the proves of estimating the missing value based on valid values of other variables and/or cases in the sample. The objective is to employ known relationships that can be identifies in the valid values of the sample to assist in estimating the missing values. This should be done carefully because of the potential impact on the outcomes. All of the imputation methods are used for metric variables, nonmetric missings are usually left missing unless a specific modelling approach is employed. Imputation of the MAR missing data process. The researcher should apply only one remedy- the specifically designed modeling approach. This set of procedures explicitly incorporates the missing data into the analysis, either through a process specifically designed for missing data estimation or as an integral portion of the standard multivariate analysis. The first approach involves maximum likelihood estimation techniques that attempts to model the process underlying the missing data and to make the most accurate and reasonable estimates possible. One example is the EM approach. Stage E makes the possible best estimates and stage M then makes estimates of the parameters assuming the missing data were replaced. Comparable procedures employ structural equation modeling to estimate the missing data. The second approach involves the inclusion of missing data directly into the analysis, defining observations with missing data as a select subset of the sample. This is most appropriate for missing data in the independent variables of a dependent relationship.

   When the missing data occur on a nonmetric variable, the researcher can define those observations as a separate group and then include them in any analysis. When the missing data are present on a metric independent variable in a dependent relationship, the observations are incorporated into the analysis while maintaining the relationship among the valid values. The first step is to code all the observations that have missing values with a dummy variable. Then the missing values are imputes by mean substitution method. Finally the relationship is estimated by normal means. The dummy variable represents the

difference for the dependent variable between those observations with missing data and those observations with valid data. The coefficient of the original variable represents the relationship for all cases with non-missing data.

Imputation method of a MCAR missing data process; there are two basic approaches.

The first approach is imputation using only valid data.

1. **Complete Case approach** = Include only observations with complete data (listwise method in SPSS). The approach has two disadvantages:
   1. It is most affected by any nonrandom missing data processes, because the cases with any missing data are deleted from the analysis.
   2. This approach results in the greatest reduction in sample size, because missing data on any variable eliminates the entire case.
   This approach is best suited for instances in which the extent of missing data is small, the sample is sufficiently large to allow for deletion and the relationships are strong.
2. **Using all-available data** = This method imputes the distribution characteristics or relationships from every valid value (PAIRWISE method SPSS). This method is primarily used to estimate correlations and maximize pairwise information available in the sample. The distinguishing characteristic is that the characteristic of a variable is based on a potentially unique set of observations. Missing data are not replaced, but instead the obtained correlations are used on just the valid cases as representative for the entire sample. Several problems can still arise:
   1. Correlations may be calculated that are out of range and inconsistent with other correlations. Any correlation between X and Y us constrained by their correlation to a third variable Z: range of r*xy* = r*xz*r*yz* +/-
   The correlation between X and Y can very between -1 and 1 if X and Y have zero correlation with all other variables in the correlation matrix.
   As the correlations with other variables increase, the range of the correlations between X and Y decreases. , which increases the potential for the correlation in a unique set of cases to be inconsistent with correlations derived from other sets of cases. An associated problem is that the eigenvalues in the correlation matrix can become negative , thus altering the variance properties of the correlation matrix

Imputation using replacement values

1. **Using known replacement values** = The common characteristic in these methods is to identify a known value, most often from a single observation, that is used to replace the missing data.

- Hot or cold deck imputation: in this approach the researcher substitutes a value from another source for the missing values. In the hot deck method the value comes from another observation in the sample that is deemed similar. Cold deck imputation derives the replacement value from an external source. Here the researcher must be sure that the replacement value from an external source is more valid than an internally generated value.
- Case substitution: in this method, entire observations with missing data are replaced by choosing another non sampled observation.

2. **Calculating replacement values** = The second basic approach involves calculating a replacement value from a set of observations with valid data in the sample.

- Mean substitution: one of the most widely used methods, mean substitution replaces missing values with the mean value for that variable. This approach has several disadvantages. It understates variance estimates by using the mean for all missing data.

Second, the distribution of data is distorted. Third, this method depresses the observed correlation because all missing data will have a single constant value. However, this method is easily implemented.

- Regression imputation: in this method regression analysis is use to predict the missing values based on its relationship with other variables. First, a predictive equation is formed for each variable with missing data. Then replacement values for each missing value are calculated from that observations calculated in the predictive equation this method also has several disadvantages.

First, it reinforces the relationships already in the data. Second, unless stochastic terms are added to the estimated data, the variance is understated. Third, this method assumes that the variable with missing data has substantial correlation with the other variables. Fourth, the sample size must be large enough to allow for a sufficient number of observations to be used in each prediction. Finally the regression equation is not constrained in the estimates it makes. The imputation methods are summarized on page 53.

A recap of the missing value analysis: we can summarize 4 conclusions.

1. The missing data process is MCAR. Such a finding provides 2 advantages to the researcher. First, it should not involve any hidden impact on the results that need to be considered when interpreting the results. Second, any of the imputation methods can be applied as remedies for the missing data.
2. Imputation is the most logical course of action. Some form of imputation is needed in order to keep a sufficient sample size for any multivariate analysis.
3. Imputed correlations differ across techniques. When estimating correlations among the variables in the presence of missing data , the researcher can choose between four different techniques: the complete case method, the all-available information method, the mean substitution method and the EM method. The researcher will however obtain different results by using different methods.
4. Multiple methods for replacing the missing data are available and appropriate. The presence of several acceptable methods enables the researcher to combine estimates into a single composite, hopefully mitigating any effects strictly due to one of the methods

Outliers

Outliers = Observations with a unique combination of characteristics identifiable as distinctly different from the other observations. In assessing the impact of outliers, we must consider practical an substantive considerations: From a practical standpoint, outliers can have a marked effect on any type of empirical analysis. In substantive terms, the outlier must be viewed in light of how representative it is of the population

Why do outliers occur? Outliers can be classified into 1 of 4 categories based on the source of their uniqueness:

1. The first class arises from procedural error, such as a mistake in data entry.
2. The second class is the result of an extraordinary event, which accounts for the uniqueness of the observation.
3. The third class of outliers contains extraordinary observations for which the researcher has no explanation.
4. The last class contains observations that are unique in their combination of values across the variables.

Detecting and handling outliers

Methods of detecting outliers:

- **Univariate detection** = Examines the distribution of observations for each variable in the analysis and selects the outliers as those cases falling at the outer ranges of the distribution.
- **Bivariate detection** = Pairs of variables can be assessed jointly through a scatterplot. Cases that fall markedly outside the range of the other observations will be seen as isolated points in the scatterplot. To assist in determining this two-dimensional portrayal, an ellipse representing a bivariate normal distribution's confidence interval is superimposed over the scatterplot. This ellipse provides a graphical portrayal of the confidence limits and facilitates identification of the outliers. A variant of the scatterplot is the influence plot, with each point varying in size in relation to its influence on the relationship. A drawback of the bivariate method is the potentially large number of scatterplots that can arise.
- **Multivariate detection** = When more than 2 variables are considered, the researcher needs a means to objectively measure the multidimensional position of each observation relative to some common point. This issue is addressed by Mahalanobis D2 measure. Higher D values represent observations located farther away from the general distribution in the multidimensional space. However, this method is only providing an overall view.

D2/df is approximately distributes as a t value. So if the t values are larger than 2.5 in small samples, or exceeding 3 or 4 for large samples these are outliers.

- Outlier designation. The researcher must select only observations that demonstrate real uniqueness in comparison with the remainder of the population across as many perspectives as possible. The researcher must refrain from designating too many observations as outliers.
- Outlier description and profiling. Once the outliers are identifies, the researcher should generate profiles of each outlier observation and identify the variables responsible for its being an outlier. Discriminant analysis and multiple regression analysis can be used. If possible the researcher should categorize the outlier into one of the 4 categories described before.
- Retention or deletion of the outlier. After these steps, the researcher should decide on the retention or deletion of every outlier. They should be retained unless there is proof that they are truly aberrant and not representative of any of the observations in the population.

Testing the assumptions of the multivariate analysis

Some techniques are less affected by violating certain assumptions, which is termed robustness, but in all cases meeting some of the assumptions will be critical. The need to check assumptions is more important in multivariate analysis because of two characteristics of multivariate analysis. First, the complexity of the relationships, makes potential distortions and biases more potent. Second, the complexity of analyses and results may mask the indicators of assumption violations apparent in the simpler univariate analyses.

Assessing individual variables versus the variate.

Multivariate analysis requires that the assumptions underlying the statistical techniques be tested twice: for the separate variables and for the model variate.

Four important statistical assumptions

Four of the assumptions potentially affect every univariate and multivariate statistical technique.

1. **Normality**. The most fundamental assumption is normality, referring to the shape of data distribution for an individual metric variable and its correspondence to the normal distribution. If there is no normality, all statistical tests are invalid. Multivariate normality means that all individual variables are normally distributes and that the combinations are also normal. So if a variable is multivariate normal, it is also univariate normal. The reverse is not true.

- Assessing the impact of violating the normality distribution. The severity of non-normality is based on two dimensions; the shape of the offending distribution and the sample size. Impacts due to the shape of the distribution. There are two dimensions : Kurtosis, which refers to the peakedness or flatness of the distribution. If there are a lot of peaks this is called **leptokurtic**, when the distribution is flat this is called **patykurtic**. The second dimension is Skewness. This is concerning the balance of the distribution. Is it shifted to one side or centered and symmetrical? Positively or negatively skewed? Impacts due to the sample size large samples sizes reduce the detrimental effects of non-normality.
- Graphical analysis of normality. The simplest check for normality is creating a histogram that compares the observed values with a distribution approximating the normal distribution., but this method is problematic for small sample sizes. A more reliable approach is the normal probability plot, which compares the cumulative distribution of actual data values with the cumulative distribution of a normal distribution. In figure 2.6, different normal probability plots are shown.
- Statistical tests of normality. An easy test is a rule of thumb based on the skewness and kurtosis values. The z value for the skewness is calculated as:

  $z_{skewness} = \frac{skewness}{\sqrt{\frac{6}{N}}}$, where N is the sample size

  $z_{kurtosis} = \frac{kurtosis}{\sqrt{\frac{24}{N}}}$

  If either calculated z values exceeds the specified critical value, then the distribution is non-normal in terms of that characteristic. Specific statistical tests for normality are also available in SPSS. These are the Shapiro-Wilks test and a modification of the Kolmogorov-Smirnov test

2. **Homoscedacity** = The assumption that dependent variables exhibit equal levels of variance across the range of predictor variables. The variance od the dependent variable values must be relatively equal. If this is not the case, the relationship is heteroscedastic. The dependent variables should be metric, but the independent variables can either be metric or nonmetric. The two most common sources of heteroscedacity are:
   1. Variable type. Many variables have a natural tendency toward differences in dispersion.
   2. Skewed distribution of one or both variables

- Graphical tests of equal variance dispersion. The test for homoscedacity is best examined graphically. (figure 2.7). the most common application is multiple regression. Boxplots also work well to show the degree of variation between groups formed by a categorical value.
- Statistical tests for homoscedacity. The most common, the levene test, is used to assess whether the variances of a single metric variable are equal across any number of groups. If more than one variable is being tested, a Box M test can be used.
- Remedies for heteroscedacity. Heteroscedastic variables can be remedies through data

transformations similar to those used to achieve normality.

3. **Linearity**.
   Identifying nonlinear relationships; The most common way is to examine scatterplots of the variables and to identify any other nonlinear patterns in the data. An alternative approach is to run a simple regression and to examine the residuals. A third approach is to explicitly model a nonlinear relationship by the testing of alternative model specifications. Remedies for nonlinearity; The most direct approach is to transform one or two variables to achieve linearity. An alternative is the creation of new variables.
4. **Absence of correlated errors**
   Identifying correlated errors; Similar factors that affect one group may not affect another. If groups are analyzed separately, the effects are constant within each group. But if observations from both groups are combined, this can lead to biased results because an unspecified cause is affecting the estimation of the relationship. Another common source of correlated error is time series data. To identify correlated errors, the researcher must first identify potential causes.

Values for a variable should be grouped and ordered on the suspected variable and then examines for any patterns.
Remedies for correlated errors;. Correlated errors must be corrected by including the omitted causal factor into the multivariate analysis. The most common remedy is the addition of a variable that represents the omitting factor.

Overview of testing for statistical assumptions
Data transformations
Data transformations provide a means of modifying variables for one of two reasons:

1. To correct violations of the statistical assumptions underlying the multivariate techniques
2. To improve the relationship between variables

Transformations to achieve normality and homoscedacity
For non-normal distributions, the most common patterns are flat distributions and skewed distributions. For the flat distribution, the most common transformation is the inverse (1/X). Skewed distributions can be transformed by taking the squared or cubed transformation for negative skewness, and logarithm and root for positive skewness. For heteroscedacity: if the cone opens to the right, taking the inverse is the best transformation. If the cone opens to the left: take the square root.

Transformation to achieve linearity
Numerous procedures are available for achieving linearity between two variables, but most simple nonlinear relationships can be placed in one of four categories (Figure 2.8).

Incorporating nonmetric data with dummy variables
In many instances metric data must be used as independent variables. A researcher has available a method for using dichotomous variables, known as dummy variables, which act as replacement variables for the nonmetric variable. Any nonmetric value with k categories can be represented as k-1 dummy variables. In constructing dummy variables, two approaches can be used to represent the categories, and more importantly, the category that is omitted, known as the reference category or comparison group.

1. The first approach is known as indicator coding. An important consideration is the reference category, the category that received all zeros for the dummy variables. The deviations represent the differences between the dependent variable mean score and the comparison group. This form is most appropriate in a logical comparison group.
2. An alternative method is effects coding. It is the same as indicator coding except that the comparison group is given -1 score instead of 0 for the dummy variables.