10 Analysing Quantitative Data

Babak Taheri, Liang Lu and Christian König

In this chapter ...

Researchers and managers need to develop an awareness of statistical analysis techniques if they want to understand the data and present the findings in an accurate way. The previous chapter concentrated on data collection, this chapter delves into the statistical tools used to analyse data. Our focus in the chapter is on two sets of the most widely used statistical tools for exploring relationships and comparing groups.

Data preparation

Real-life data generally cannot be used directly for data analysis – it is unorganised and filled with different types of problems. We begin by discussing three pre-processing steps that prepare data for further analysis: data entry, data cleaning and data formatting.

■ Data entry

A conventional way to organise data is to use tables, with *records* as rows and *attributes* as columns. A record is an identifiable piece of information which contains a set of values of attributes to the record. For example, one may organise the information collected from questionnaires in the following way: each record corresponds to all the answers from a respondent, with each attribute associated with the answer to one question.

No matter how careful one is, it is difficult to refrain from making mistakes when entering data. To maintain a certain level of precision, one could use *double entry*. Its idea is very simple – let two individuals enter the same content and compare their inputs. When discrepancies are found, a quick check verifies and maintains the correct copy. Though doubling the efforts, double entry is very efficient in preventing entry mistakes. Another method is to use encoding to avoid entering text data directly. For example, when entering information such as gender, 'male' and 'female', in text forms, some may introduce typos such as 'mael' and 'femeal', and some may capitalize the first letters as 'Female' and 'Male', which could be interpreted as different words with 'female' and 'male'. Alternatively, one can encode 'male' as '0' and 'female' as '1', so that one could enter 0s and 1s instead. The encoding function is explicitly provided in many data analysis software such as SPSS.

■ Data cleaning

Even if there are no errors introduced during entry phase, real-life data needs to be cleaned because it is often *incomplete*, *noisy* and *inconsistent* (Han, Kamber, & Pei, 2011). Incompleteness arises when for some records the values for some attributes are missing. There are mainly two ways to deal with this issue. First, delete the cases with missing data; this could be viable when the number of cases with missing data is relatively small compared to the whole dataset. Second, fill the missing values; one can use the expected value on the corresponding attribute or regression on other attributes to predict the missing value. Noises refer to random factors that can only be quantified in a probabilistic way. Noises confound observations and cause *outliers* that are far away from normal observations. A primary task of data cleaning is to identify and 'smooth' out these outliers. Inconsistencies often arise when one combines information from different sources. For example, combining datasets with both American and British date information may cause confusion (i.e. the 3rd of April 1990 could be displayed as both 4/3/90 and 3/4/90).

■ Data formatting

The final step is to format the data in a way that is consistent with the requirement of the statistical test used to analyse the data. This often involves transforming one type of measurement into another. Let us consider a dataset

summarising the output of all producers in a specific industry. Instead of getting to know their exact outputs, say, in tons, one may be more interested in categorizing them in terms of scales – small, medium or large producers. An arithmetic operation (e.g. log) may be needed to realise this transformation from output to scale.

Preliminary analysis

■ Describing data

To present a sample in an illustrative way, one can either use descriptive statistics (numbers) or graphs, or both; it is just a matter of opinion - some prefer descriptive statistics because they are quantifiable while others prefer graphs because they are more intuitive. Therefore, when deciding in which form to present data, it is important to know who your target audience are.

If the sample is of a non-metric type, frequency and ratio are two commonly used statistics. Frequency counts the number of occurrences of a specific category, and ratio calculates the corresponding percentage of frequency in the entire sample. Non-metric data can be visualised through pie chart or bar chart.

We give an example on the cut quality of diamonds based on a dataset with 53,940 records (Source: http://vincentarelbundock.github.io/Rdatasets/datasets.html). The cut quality of diamonds is a non-metric measurement and has five categories: fair, good, very good, premium and ideal. Table 10.1 summarises the frequencies and ratios of all five categories. Figure 10.1 plots both with a bar graph and a pie chart.

Table 10.1: Cut quality of diamonds: frequencies and ratios

Cut quality	Frequencies	Ratios
Fair	1610	2.98%
Good	4906	9.10%
Very good	12082	22.40%
Premium	13791	25.57%
Ideal	21551	39.95%

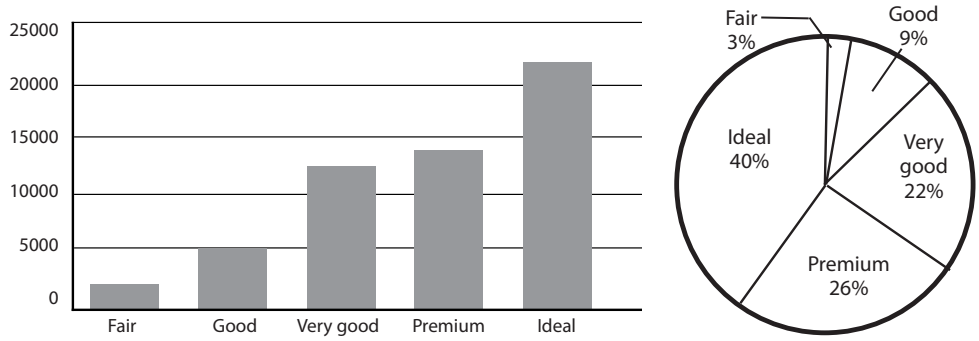

Figure 10.1: Cut quality of diamonds: bar graph and pie chart

If the sample is of a metric type, it makes sense to calculate all sorts of statistics measuring the basic characteristics of the sample such as *centre* and *dispersion*. The centre denotes a typical value that represents the entire sample and can be measured by *mean* (average) and *median* (the value of the middle case in a series). Dispersion accesses the variation across the sample and can be measured by *variance* (the average of the squared differences from the mean), *standard deviation* (a measure of how spread-out numbers are), *coefficient of variation* (a measure of the dispersion of data points in a data series around the mean), *range* (the difference between the lowest and highest values), etc. We summarise these statistics in Table 10.2.

Table 10.2: Commonly used statistics for metric sample.

Suppose the sample is represented by a set of values $x_1, x_2, \ldots, x_n$ and are sequenced as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ in either decreasing or increasing order.

Sample characteristics	Statistics
Centre	Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ Median $\tilde{x} = \begin{cases} x_{(\frac{n}{2})} \text{ if } n \text{ is even} \\ x_{(\frac{n+1}{2})} \text{ if } n \text{ is odd} \end{cases}$
Dispersion	Standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ Coefficient of Variation $c_v = \frac{s}{\bar{x}}$ Range $L = x_{(n)} - x_{(1)}$