10 Quantitative Data Analysis Approaches

Babak Taheri, Catherine Porter, Christian König and Nikolaos Valantasis-Kanellos

In order to understand data and present findings in an accurate way, researchers and managers need to develop an awareness of statistical analysis techniques. The previous chapter concentrated on quantitative data collection, this chapter delves into the statistical tools used to analyse the data once collected. It focuses on two sets of the most widely used statistical tools – exploring relationships and comparing groups – as shown in the 'Deductive' section in the Data Analysis area of the Methods Map (see Chapter 4). Finally, we briefly explain the nature of Big Data.

Data preparation

Real-life data generally cannot be used directly for data analysis – they are unorganised and filled with different types of problems and errors. We discuss three pre-processing steps that prepare data for further analysis: data entry, data cleaning and data formatting.

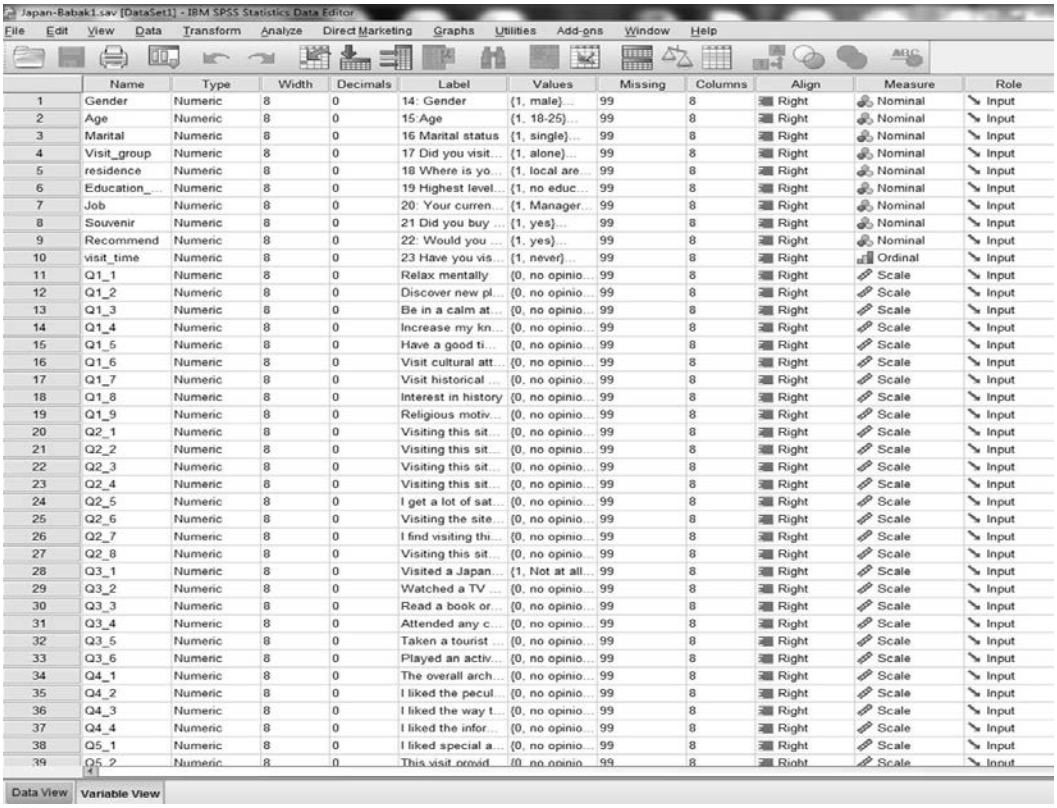
■ Data entry

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.

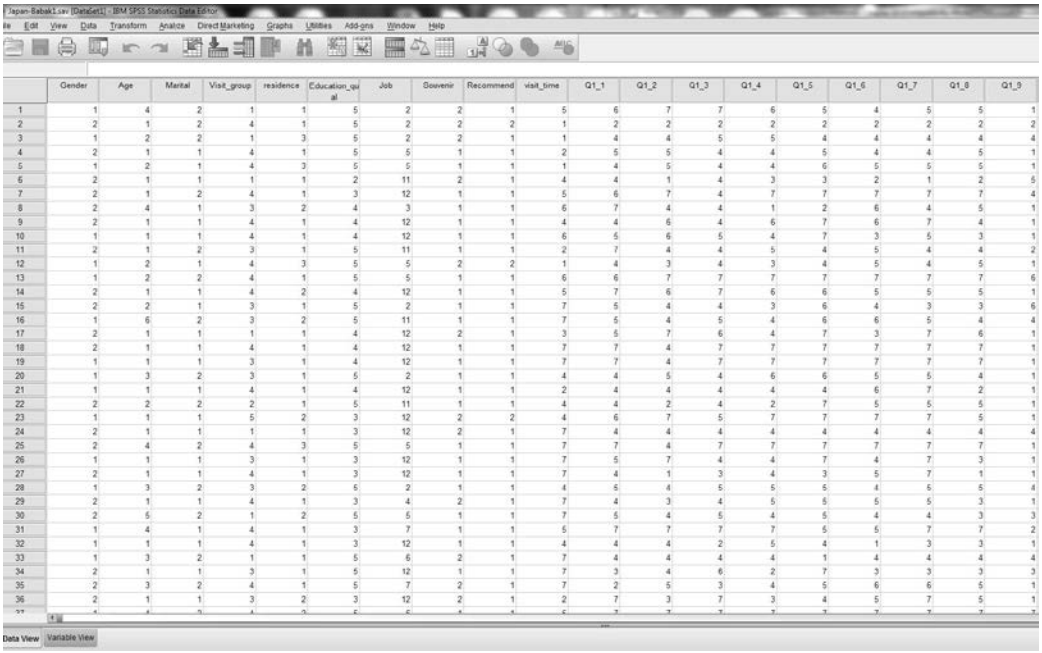
No matter how careful one is, it is difficult to avoid making mistakes when entering data. To maintain a certain level of precision, one could use *double entry*. Its idea is very simple – let two individuals enter the same content and compare their inputs. When discrepancies are found, one shall verify and maintain the correct copy. By doubling 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 gender information such as ‘male’ or ‘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. 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 (Statistical package for the social sciences). SPSS can be used to analyse questionnaire-based and other data organised as cases with particular variables. Figure 10.1 illustrates a snapshot of variable view (information on variables is entered in the SPSS) and data value (data entered directly or can be imported from a spreadsheet file) on SPSS. Table 10.1 explains the information required for each variable in the questionnaire.

Table 10.1: Information required for each variable in the questionnaire in variable view in SPSS

Variable Label	Short Description
Name	Up to 8 characters (no spaces), starting with a letter Not allowed: ALL, AND, BY, EQ, GT, LE, LT, NE, NOT, WITH, OR, TO Can be: short version of item description e.g., var01, Q1a
Width	Max. no. of characters
Decimal places	Decimal places for numbers
Label	Longer version of name
Values	Values for coded variables
Missing	Blanks, no answer, etc
Columns	No. of columns in data view screen
Alignment	Left, right, centre
Types of measure	Nominal, ordinal, scales


Japan-Babak1.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	0	14: Gender	{1, male}...	99	8	Right	Nominal	Input
2	Age	Numeric	8	0	15:Age	{1, 18-25}...	99	8	Right	Nominal	Input
3	Marital	Numeric	8	0	16 Marital status	{1, single}...	99	8	Right	Nominal	Input
4	Visit_group	Numeric	8	0	17 Did you visit...	{1, alone}...	99	8	Right	Nominal	Input
5	residence	Numeric	8	0	18 Where is yo...	{1, local are...	99	8	Right	Nominal	Input
6	Education_...	Numeric	8	0	19 Highest level...	{1, no educ...	99	8	Right	Nominal	Input
7	Job	Numeric	8	0	20: Your curren...	{1, Manager...	99	8	Right	Nominal	Input
8	Souvenir	Numeric	8	0	21 Did you buy ...	{1, yes}...	99	8	Right	Nominal	Input
9	Recommend	Numeric	8	0	22: Would you ...	{1, yes}...	99	8	Right	Nominal	Input
10	visit_time	Numeric	8	0	23 Have you vis...	{1, never}...	99	8	Right	Ordinal	Input
11	Q1_1	Numeric	8	0	Relax mentally	{0, no opinio...	99	8	Right	Scale	Input
12	Q1_2	Numeric	8	0	Discover new pl...	{0, no opinio...	99	8	Right	Scale	Input
13	Q1_3	Numeric	8	0	Be in a calm at...	{0, no opinio...	99	8	Right	Scale	Input
14	Q1_4	Numeric	8	0	Increase my kn...	{0, no opinio...	99	8	Right	Scale	Input
15	Q1_5	Numeric	8	0	Have a good ti...	{0, no opinio...	99	8	Right	Scale	Input
16	Q1_6	Numeric	8	0	Visit cultural att...	{0, no opinio...	99	8	Right	Scale	Input
17	Q1_7	Numeric	8	0	Visit historical ...	{0, no opinio...	99	8	Right	Scale	Input
18	Q1_8	Numeric	8	0	Interest in history	{0, no opinio...	99	8	Right	Scale	Input
19	Q1_9	Numeric	8	0	Religious motiv...	{0, no opinio...	99	8	Right	Scale	Input
20	Q2_1	Numeric	8	0	Visiting this sit...	{0, no opinio...	99	8	Right	Scale	Input
21	Q2_2	Numeric	8	0	Visiting this sit...	{0, no opinio...	99	8	Right	Scale	Input
22	Q2_3	Numeric	8	0	Visiting this sit...	{0, no opinio...	99	8	Right	Scale	Input
23	Q2_4	Numeric	8	0	Visiting this sit...	{0, no opinio...	99	8	Right	Scale	Input
24	Q2_5	Numeric	8	0	I get a lot of sat...	{0, no opinio...	99	8	Right	Scale	Input
25	Q2_6	Numeric	8	0	Visiting the site...	{0, no opinio...	99	8	Right	Scale	Input
26	Q2_7	Numeric	8	0	I find visiting thi...	{0, no opinio...	99	8	Right	Scale	Input
27	Q2_8	Numeric	8	0	Visiting this sit...	{0, no opinio...	99	8	Right	Scale	Input
28	Q3_1	Numeric	8	0	Visited a Japan...	{1, Not at all...	99	8	Right	Scale	Input
29	Q3_2	Numeric	8	0	Watched a TV ...	{0, no opinio...	99	8	Right	Scale	Input
30	Q3_3	Numeric	8	0	Read a book or...	{0, no opinio...	99	8	Right	Scale	Input
31	Q3_4	Numeric	8	0	Attended any c...	{0, no opinio...	99	8	Right	Scale	Input
32	Q3_5	Numeric	8	0	Taken a tourist ...	{0, no opinio...	99	8	Right	Scale	Input
33	Q3_6	Numeric	8	0	Played an activ...	{0, no opinio...	99	8	Right	Scale	Input
34	Q4_1	Numeric	8	0	The overall arch...	{0, no opinio...	99	8	Right	Scale	Input
35	Q4_2	Numeric	8	0	I liked the pecul...	{0, no opinio...	99	8	Right	Scale	Input
36	Q4_3	Numeric	8	0	I liked the way t...	{0, no opinio...	99	8	Right	Scale	Input
37	Q4_4	Numeric	8	0	I liked the infor...	{0, no opinio...	99	8	Right	Scale	Input
38	Q5_1	Numeric	8	0	I liked special a...	{0, no opinio...	99	8	Right	Scale	Input
39	Q5_2	Numeric	8	0	This visit provid...	{0, no opinio...	99	8	Right	Scale	Input


Japan-Babak1.sav [DataSet1] - IBM SPSS Statistics Data Editor
ile Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View

	Gender	Age	Marital	Visit_group	residence	Education_qual	Job	Souvenir	Recommend	visit_time	Q1_1	Q1_2	Q1_3	Q1_4	Q1_5	Q1_6	Q1_7	Q1_8	Q1_9
1	1	4	2	1	1	5	2	2	1	5	6	7	7	6	5	4	5	5	1
2	2	1	2	4	1	5	2	2	2	1	2	2	2	2	2	2	2	2	2
3	1	2	2	1	3	5	2	2	1	1	4	4	5	5	4	4	4	4	4
4	2	1	1	4	1	5	5	1	1	2	5	5	4	4	5	4	4	5	1
5	1	2	1	4	3	5	5	1	1	1	4	5	4	4	6	5	5	5	1
6	2	1	1	1	1	2	11	2	1	4	4	1	4	3	3	2	1	2	5
7	2	1	2	4	1	3	12	1	1	5	6	7	4	7	7	7	7	7	4
8	2	4	1	3	2	4	3	1	1	6	7	4	4	1	2	6	4	5	1
9	2	1	1	4	1	4	12	1	1	4	4	6	4	6	7	6	7	4	1
10	1	1	1	4	1	4	12	1	1	6	5	6	5	4	7	3	5	3	1
11	2	1	2	3	1	5	11	1	1	2	7	4	4	5	4	5	4	4	2
12	1	2	1	4	3	5	5	2	2	1	4	3	4	3	4	5	4	5	1
13	1	2	2	4	1	5	5	1	1	6	6	7	7	7	7	7	7	7	6
14	2	1	1	4	2	4	12	1	1	5	7	6	7	6	6	5	5	5	1
15	2	2	1	3	1	5	2	1	1	7	5	4	4	3	6	4	3	3	6
16	1	6	2	3	2	5	11	1	1	7	5	4	5	4	6	6	5	4	4
17	2	1	1	1	1	4	12	2	1	3	5	7	6	4	7	3	7	6	1
18	2	1	1	4	1	4	12	1	1	7	7	4	7	7	7	7	7	7	1
19	1	1	1	3	1	4	12	1	1	7	7	4	7	7	7	7	7	7	1
20	1	3	2	3	1	5	2	1	1	4	4	5	4	6	6	5	5	4	1
21	1	1	1	4	1	4	12	1	1	2	4	4	4	4	4	6	7	2	1
22	2	2	2	2	1	5	11	1	1	4	4	2	4	2	7	5	5	5	1
23	1	1	1	5	2	3	12	2	2	4	6	7	5	7	7	7	7	5	1
24	2	1	1	1	1	3	12	2	1	7	4	4	4	4	4	4	4	4	4
25	2	4	2	4	3	5	5	1	1	7	7	4	7	7	7	7	7	7	1
26	1	1	1	3	1	3	12	1	1	7	5	7	4	4	7	4	7	3	1
27	2	1	1	4	1	3	12	1	1	7	4	1	3	4	3	5	7	1	1
28	1	3	2	3	2	5	2	1	1	4	5	4	5	5	6	4	5	5	4
29	2	1	1	4	1	3	4	2	1	7	4	3	4	5	5	5	5	3	1
30	2	5	2	1	2	5	5	1	1	7	5	4	5	4	5	4	4	3	3
31	1	4	1	4	1	3	7	1	1	5	7	7	7	7	5	5	7	7	2
32	1	1	1	4	1	3	12	1	1	4	4	4	2	5	4	1	3	3	1
33	1	3	2	1	1	5	6	2	1	7	4	4	4	4	1	4	4	4	4
34	2	1	1	3	1	5	12	1	1	7	3	4	6	2	7	3	3	3	3
35	2	3	2	4	1	5	7	2	1	7	2	5	3	4	5	6	6	5	1
36	2	1	1	3	2	3	12	2	1	2	7	3	7	3	4	5	7	5	1

Figure 10.1: Example of (top) variable view and (bottom) data view in SPSS software

Data cleaning

Even if there are no errors introduced during entry phase, real-life data need to be cleaned because they are 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 whole record that misses data; this could be viable when the number of records 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).

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 a matter of personal preference – some prefer descriptive statistics because they are quantifiable while others prefer graphs because they are more intuitive. Therefore, when deciding which form to present data, it is important to know who your target audience is.

If the sample is of a nonmetric type (for example an ordinal scale as described in Chapter 9), *frequency* and *ratio* are two commonly used descriptive statistics. Frequency counts the number of occurrences of a specific category, and ratio calculates the corresponding percentage of frequency in the entire sample. Nonmetric data can be visualised through pie charts or bar charts. We give an example on the cut quality of diamonds based on a dataset with 53940 records (Source: http://vincentarelbundock.github.io/Rdatasets/datasets.html). The cut quality of diamonds is a nonmetric measurement and has five categories: fair, good, very good, premium and