2.2: How to Read a Data Set

Variables are often measured many times in an attempt to better understand them. The sample of cases is gathered and then data are gathered about variables from those cases. The word **case** refers to an individual or on instance of a thing which is being studied. Variables are measured within cases. A **score** is a number yielded for a case. All cases may be looked at together when trying to understand the variable. However, it can be overwhelming to try to understand the variable by looking at all the cases one by one; therefore, statisticians and mathematicians use a variety of techniques to summarize score from all of the cases in a data set.

Before we can begin to use data, we must understand how to read a data set. Data can be collected and organized into a data set a few ways but the most common way to construct a data set is to put the names of variables as column headers with the data for each variable entered in rows below their appropriate headers, as shown in Data Set 2.1 below.

Table 1 Raw Data Set 2.1

ID Number	Major	Job	Years of Experience	Annual Salary	Morale
1	Statistics	UX Researcher	4	184,600	8
2	Statistics	Science Journalist	6	74,490	5
3	Statistics	Sr. Research Fellow	7	190,000	7
4	Statistics	Research Analyst	4	138,000	6
5	Statistics	Biostatistician	5	167,000	4
6	Statistics	Psychometrician	6	120,000	3
7	Statistics	Data Scientist	5	110,240	6
8	Statistics	Research Assistant	5	57,750	7
9	Statistics	Biostatistician	3	102,000	6
10	Statistics	Research Assistant	3	48,200	3
11	Statistics	Data Scientist	4	65,500	5
12	Statistics	Research Analyst	2	74,500	9
13	Statistics	Data Scientist	7	181,000	8
14	Statistics	UX Researcher	6	123,700	10
15	Statistics	Research Analyst	5	65,000	4
16	Statistics	UX Researcher	1	88,950	5
17	Statistics	Research Assistant	1	34,000	7
18	Statistics	Research Analyst	2	52,680	2

In Data Set 2.1, there are 5 columns of data. The data set shown includes raw scores. These can also be referred to as *raw data.* The term **raw scores** refers to data as they were collected and before they have been transformed, summarized, or analyzed using any formulas. Let's take a look at each column (i.e. the vertical sections of the table) of raw data in dataset 2.1 to understand it.

The first column includes ID Numbers. This is a nominal variable used to organize the cases. Each case in Data Set 2.1 was given an ID number in place of a name. For Data Set 2.1, each case has been named with sequential case numbers starting at 1. However, random numbers or letters could also be used to name cases. This can be done to help organize data without using identifying information such as a person's real name, IP address, or Social Security Number. When data are specifically presented this way to protect the anonymity of participants, the data are referred to as **de-identified data.** It is important to note that though ID Number

is a variable, it is not a test variable. This means it is simply used to organize things but that it would not be analyzed or used to test hypotheses.

Each row of the data set (i.e. the horizontal sections of the table) contains the data for an individual case. Each case is a member of the sample. There are 18 rows so the sample size is 18. The symbol for sample size is *n*. Thus, we can summarize the sample size by writing $n = 18$. Sample size is always shown as a whole number because we cannot have half a case or half a person and, therefore, specifying to the hundredths decimal place does not add any information we wouldn't already know.

The second column of data is titled "Major." Major refers to the focus of someone's college degree. This was measured qualitatively on the nominal scale of measurement. However, if we read through the data for Major we will notice that, despite the fact that there are many possible majors, everyone in the sample majored in statistics. Therefore, Major is a constant in this data set and not a variable. It is best used to describe the sample and to whom limited generalizations should be made. Specifically, the fact that major did not vary means the sample of individuals is all persons who studied statistics and, therefore, any other findings or summaries from this sample would best be used to understand the population of persons who studied statistics as well rather than students of all majors.

The last four columns include data for variables which can be summarized or used to test hypotheses. The third column of data is titled "Job." Job refers to each individual's job title. This was measured qualitatively on the nominal scale of measurement. Job is a variable in this data set because not all persons had the same job title. The fourth column of data is titled "Years of Experience." This refers to the number of years of job-relevant experience someone has, rounded to the year. Years of Experience was measured quantitatively on the ratio scale of measurement. The fifth column of data is titled "Annual Salary" and reports each person's income for the year rounded to the dollar. Annual Salary, therefore, was also measured quantitatively on the ratio scale of measurement. Finally, the sixth column of data reports each person's self-reported level of "Morale" on an 11-point scale from 0 to 10 where higher values indicate greater morale. When psychological and emotional variables like morale are measured in this way, they are generally treated as interval scales. This is because the intervals between values are treated as even but the 0 is not known to represent a complete absence of the variable. It is worth noting that whether this kind of measure is truly interval or not has been debated. When certain conditions are met, such as when an 11-point scale is used, it can be appropriate to view and treat variables such as Morale as interval. Though this debate is beyond the scope of this book, those interested in learning more can start by reading Wu and Leung (2017). For this chapter, we will presume the conditions are met and treat Morale as a quantitative variable measured on the interval scale.

It is important to look through a data set to identify which columns include variables and to then assess whether each variable is quantitative or qualitative before identifying which scale of measurement was used for each variable. This is because there are different ways statisticians can summarize and present data and results that are best suited for, or can only be used with, certain kinds of data. Some ways of summarizing data can only be used for quantitative data, some for qualitative data, and some require that both quantitative and qualitative data are used together. For this chapter, we will focus on data univariately. Therefore, in this chapter we will learn how to create and read tables and graphs, some of which are used to summarize one quantitative variable at a time and others which can be used to summarize one qualitative variable at a time. Some of the graph can be extended to include more than one variable. When that is possible, it will be noted.
