Chapter 5 Introduction to Data Analysis

About This Chapter

This chapter discusses data analysis. The expression “data analysis” may sound a little too specialized and technical to you. However, collecting and analyzing data is a common technique widely used in society as well as in almost all academic fields, regardless of whether the field is in the social, human sciences or physical sciences, or is engineering.

Data analysis is supported by advanced mathematics, such as probability theory based on Lebesgue integration. Advanced mathematics in this chapter is used as a tool for analysis. We described data analysis as a “technique” above, and you are going to learn the methods of data analysis which every college-educated student should know. Therefore, our goal is to learn how to select an analysis method and interpret its results. We use computers to avoid going into complicated statistical calculations, but interested students should read the reference literature and study further.

Data gathering and data handling (the processing of data) are important part of actual data analysis. Moreover, in order to use data as basis of your argument, you need to interpret the results of analysis and express the finding in easy-to-understand forms like graphs or tables.

In line with the above ideas, this chapter describes the basics of computation, a process from which computers got their name.

5.1 Types of Data

There are many different types of data which we may like to analyze. In order to analyze any data as numerical values, we need to use a "measuring stick." This measuring stick is called a scale. For example, the data may be as follows:

- Scores for an English quiz
- Favorite days of the week
- Postal codes
- Opinions on the consolidation of City A and City B
- Lengths
- Weights
- Temperatures
- Prices (in yen, dollars, etc.)

Each of these groups of data expresses some sort of characteristics. These characteristics can be roughly divided into two groups: **qualitative data** which does not have a size or a rank order, such as days of the week and opinions; and **quantitative data** which has sizes and/or rank orders. Note that some qualitative data consists of numbers, such as postal codes. In such cases, we call each data item a "numerical character string" instead of a numerical value.

In addition, some data can be sorted without being changed (for example, scores, days of the week, agree/disagree, or price) and for other data you need to specify certain units in order to be able to sort it (length, weight or temperature).

The former types can be sorted according to their observed values, but they can also be expressed in terms of numerical values (called "coding"), such as indicating women with 1 and indicating men with 2.

Note that the observed values of the latter type of data change depending on the precision of the "measuring stick" you use. For example, when measuring length, the measured results change depending on whether you use a ruler graduated in millimeters or tenths of a millimeter. This holds true as well when you use a digital measure or a thermometer. In this case, the measured values change according to the precision of the sensor. Which is to say, when you measure the latter type of data, whichever method of measurement you use, you cannot get the real values, but only an approximation instead.

As has been just described, we can also classify data by whether we can only obtain approximated values or can get exact values. Length, weight and temperature are continuous quantities and a score is a discrete quantity.

Let us look at these types of data from the viewpoint of scales. Let us define a scale as "a criterion to correspond values to data" for now. There are different kinds of scales: nominal scales, ordinal, interval and a ratio scale. Different scales represent different data characteristics. Note that graphic methods (how graphs are created) and computable statistics vary from scale to scale.

5.1.1 Nominal Scale

A nominal scale is a data classification standard in which the rank order of values or the difference between two values does not have any meaning. For example, you cannot obtain any meaningful result from subtracting one postal code from another. Data may be expressed in letters (for example, gender). Even if data is expressed by numbers, these numbers are more like numerical character strings than numerical values. Such a scale is called a nominal scale.

5.1.2 Ordinal Scale

In an ordinal scale, differences between values is meaningless but the rank order of values has meaning. An example of an ordinal scale is choices in a questionnaire, such as (1) Bad, (2) Somewhat bad, (3) Neutral, (4) Somewhat good and (5) Good.

We cannot assume such data values are evenly spaced apart. For example, when you use a five-level rating system to classify likes and dislikes, it does not make any difference if you use -2, -1, 0, 1 and 2 or 1, 5, 15, 25 and 55. But usually we assign consecutive values starting from 1 (1, 2, 3, 4, 5)[1].

In fields like pedagogy, psychology and marketing, the intervals between such values are often assumed to be constant for the sake of analysis. In other words, the scale type in such a case is assumed to be an interval scale, as explained in 5.1.3 “Interval Scale” below.

You cannot calculate the mean in an ordinal scale, but the median (when all data values are laid out back to back from lowest value to highest value, the value in the middle is the median) does have meaning.

5.1.3 Interval Scale

In an interval scale, like temperature, intervals between data values have meaning. In other words, differences between data values have meaning. For example, this means that 100 degrees and 80 degrees have a difference of 20 degrees. Note that, however, that in this case it does not mean that 80 degrees is four fifths of 100 degrees.

Temperature, for example 100 degrees Celsius, is simply a relatively measured value based on artificial criteria. In other words, the reference point 0 degrees Celsius has no absolute meaning. When you think about how many times hotter 100 degrees is compared to 0 degrees, you can easily grasp this concept. As we have seen, strictly speaking, the ratio of two values has no meaning in an interval scale.

In an interval scale, it is possible to calculate most major statistics, such as the mean and the standard deviation.

5.1.4 Ratio Scale

A ratio scale deals with data where 0 (zero) has meaning (in other words, the ratio of two values has meaning), such as length and weight. For example, it can be said that 10 kg is twice as heavy as 5 kg or that the difference of 10 kg and 5 kg is 5 kg.

In a ratio scale, it is also possible to calculate most major statistics.

5.2 Basics of Calc and Its Components

This section explains how to use spreadsheet software, which is frequently used in the compilation and analysis of numerical data. By using spreadsheet software, you can enter numerical values and characters and visualize the data by summarizing it in a table or a graph. Moreover, since spreadsheet software allows you to perform calculations using functions and provides convenient features for the carrying out of the same calculations repeatedly, it is widely used, especially in businesses.

[1] By doing so, we can apply such nonparametric statistical techniques as rank correlation and rank sum tests.

5.2.1 Rows, Columns and Cells

In this section, while learning the basic operations of OpenOffice.org Calc (hereinafter called "Calc"), you will actual enter the various types of data you learned about in the previous section and create tables and graphs. First, start Calc.

Spreadsheet software commonly has rows, columns and cells.

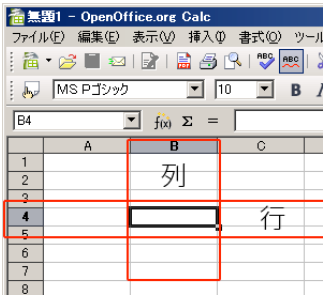

Figure 5.1: Row, Column and Cell

The foundation of a spreadsheet is the cell. A spreadsheet has rows and columns and a cell is the intersection of a row and a column. The current cell is the cell which you can edit at present.

Now, in order to distinguish one cell from the many other cells, each cell is assigned with something like an address. This is called the "cell address" or simply the "address." A cell address is expressed as a combination of a column letter and a row number. For example, the address of the leftmost and top cell is "**A1**" because it is located on column "**A**" and row "**1**." Similarly there are cells such as **A2**, **A3**, **B1** and **B2**. In addition, Calc can handle up to 256 columns and 65,536 rows. Columns **A-Z** are followed by columns **AA-AZ**, **BA-BZ** and so forth and the last column is **IV**[2]. The address of the current cell is displayed on the upper-left portion of the spreadsheet, like "**B4**" in Figure 5.1.

Cells are the foundations of a spreadsheet. All data is entered in cells.

5.2.2 Moving Current Cell and Entering and Editing Data

In order to enter data, first you need to move the current cell to the location of your choice. There are several ways to accomplish this, but the easiest is to use the Up, Down, Left and Right arrow keys (↑, ↓, ← and →). Pressing an arrow key moves the current cell.

After moving the current cell to the location you want, type any character. You can enter data this way. The data can be letters or numbers. After typing, press the Enter key. Pressing the Enter key confirms your data entry.

When the entered data consists of alphabetic characters, it should be displayed left-justified. When the data consists of numerical characters, it should be displayed right-justified. Data in certain format may be handled differently. For example, if you enter "**10/1**," the spreadsheet interprets it as October 1st (without asking you) and handles it as such.

There are three or so methods of confirming data entry and the current cell moves differently afterwards depending on the methods you use. Firstly, if you press the Enter key on the keyboard, the current cell moves down by one cell.

[2] This limit varies from software to software. For example, Microsoft Office 2007 can display more rows and columns. Nonetheless, spreadsheet software is not suitable for processing large amounts of data items.

Secondly, if you press the Tab key, the current cell moves to the right by one cell. Finally, if you press the Enter key on the numeric keypad (usually located at the lower right corner of the keyboard), the current cell does not move.

When you use a spreadsheet, you often enter items of data one after another in one direction (to the right or down, for example), so knowing the above will help.

You can also move the current cell by using the mouse. Simply click on a cell to make it the current cell. When you drag on a certain range, the color of the cells in the range is highlighted. If you enter data in this state, the current cell will only move within that range.

Figure 5.2: When Multiple Cells Are Selected

In Figure 5.2, cells from A2 to B7 are selected. The current cell is **B7**. If you press the Enter key in this situation, the next current cell will be A2.

You can select the range of cells with the keyboard in addition to by using the mouse. To select the range, press an arrow key while holding down the Shift key. If you want to select an exact range, use the keyboard for better maneuverability.

To correct data which is already in a cell, select the cell and press the F2 key.

5.2.3 Formulas

A spreadsheet is not a spreadsheet without formulas and functions. Although you can use a spreadsheet just as a table, you need to use formulas in order to truly realize the utility of a computer.

To enter a formula, first type "**=**" in the current cell. Then type an expression using numbers, symbols, etc. Note that we use so-called "single-byte" characters here.

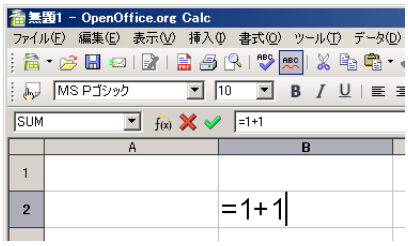

Figure 5.3: Entering a Formula

In Figure 5.3, "**=1+1**" has been entered in a cell. After finishing typing, press the Enter key. The result should be that "**2**" is displayed.

Addition and subtraction are expressed using "+" and "-," respectively. Multiplication and division are "*" and "/," respectively.

Formulas can be as simple, as above, or more complex formulas can be used. As we learned during arithmetic, multiplication and division are performed before addition and subtraction in a formula. When parentheses are used, any expression in the parentheses is calculated first. For example, let us consider the following formula.

```
=20*(6+4)/2-9*(4+8)/2-5
```

In this formula, the expressions in the parentheses are calculated first. The formula is converted as follows.

```
=20*10/2-9*12/2-5
```

After the parentheses are removed, calculate the formula from the left, carrying out multiplication and division first.

```
=200/2-108/2-5
=100-54-5
=41
```

5.2.4 Functions

Functions are useful, too. A function is a formula which produces an output value when one or more input values are given as the so-called parameter(s) of the formula. Parameters are sometimes called arguments. For example, let us consider the square root of 2. This square root can be calculated only after the number 2 is given. In other words, when the value 2 is given, the value of this function is uniquely determined. This "2" is called a parameter or an argument. To calculate the square root of 2, type as follows.

```
=sqrt(2)
```

You can enter an arbitrary value in an argument. Try many different values. When you enter a value in an argument, the function will give you a "return value."

Some functions take only one argument, like **`sqrt`**, while others take multiple arguments. Some functions have a fixed number of arguments and others do not. For example, the SUM function "**`sum`**" takes up to 30 arguments (the minimum number of arguments is 1). Note that 30 is a limit set by Calc.

```
=sum(1;2;3;4;5;6;7;8;9;10)
```

For the **`SUM`** function, use semicolon (;) to separate multiple arguments when specifying them. These arguments are the numbers to be totaled.

There are many kinds of functions. General classification and descriptions are shown in Table 5.1. There are several hundred functions available and you can even create functions of your own.

Some functions involve complicated calculations. When you use a function, however, do not think of it as a "black box." That is, you should be aware of how the calculations are carried out by the function. No matter which spreadsheet software you use, you are the one who is responsible for the entered data and the resulting data.

Anyway, since it is almost impossible to memorize all the functions, you should not even try. Except for the basic ones, a correct approach is to memorize the functions you need when you need them. What you need to know is the way of finding out how to use a particular function.

For any function, you can find how to use it and what it can do by clicking **OpenOffice.org Help** on the Calc **Help** menu. To search for how to use Calc, including its functions, type the word you want to know in the **Search term** text field. However, note that Calc will tell you that you can calculate the "variance" using the **`VAR`** function, for example, but it will not tell you what a "variance" is.

Date & Time	Dates and time calculation
Financial	Calculation of interest rates, depreciation, etc.
Information	Information about the state of cells
Logical	Logical calculation
Mathematical	Absolute values, logarithms, trigonometric functions, etc.
Array	Matrix operations, regression analysis, etc.
Statistics	Averages, variances, probability density distributions, tests, etc.
Spreadsheet	String searching in documents, etc.
Text	Extraction and conversion of strings, etc.
Database	Calculation of information about data in a specific range

Table 5.1: Classification of Calc Functions

5.2.5 Use of Cell Address in Formulas and Functions

Up until now, we directly specified numbers in a formula or a function. In this section, we will specify a cell instead of a number.

Above, we directly specified 2 in order to calculate the square root of 2, for example. Here, we enter **`2`** in cell **`A1`** and instruct cell **`B1`** to calculate the square root of what is in cell **`A1`**.

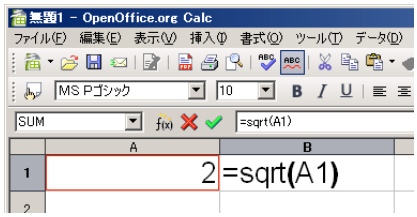

Figure 5.4: Entering a Function with a Cell Address Argument

On pressing the Enter key, the same result should be displayed as if you directly specified the number. Here, “A1” on the screen indicates the cell address “A1.”

Next, replace the “2” in cell A1 with another number. It can be 3, 4, 100 or 100,000. As the numerical value in **`A1`** changes, the square root in A2 also changes. As we have seen, by making a spreadsheet refer to a cell address, we can perform calculations which satisfy various conditions.

The first spreadsheet program for PC was called VisiCalc, which was developed by a student of Harvard Business School. This student was himself looking for such functionality. In the past, special paper for tallying or graphing data was sold at stationery stores to aid manual calculations. But accounting and financial simulations require an enormous amount of calculations.

For example, repeated calculations are required to find out how much a 1% change in sales affects the bottom line or how gradual changes in advertising expenses impacts sales and profits. That means, what is needed is a way of evaluating a computational expression using different parameters. Spreadsheet software is perfect for this type of task.

5.2.6 Relative and Absolute References

Next, let us see another convenient aspect of using cell addresses in formulas. Firstly, enter numbers 1 to 5 in cells A1-A5 as shown in Figure 5.5. Now, suppose we want to see at a glance what will happen if we take the square roots of these numbers.

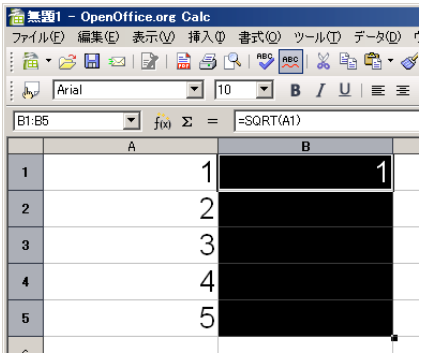

Figure 5.5: Preparing for Copying

Of course, you can enter the formulas five times, such as "**=sqrt(1)**," "**=sqrt(2)**,"... but you can do better than this. After entering the numbers in **A1-A5** as in the above figure, enter "**=sqrt(a1)**" in **B1** and press the Enter key. Next, highlight cells B1-B5. You can select the range of cells with the mouse or with the Shift key and the arrow keys.

Next, on the **Edit** menu, click **Fill → Down**. You will see the respective square roots calculated in **B2-B5**.

Move the current cell to **B2** or **B3** and check the formula bar in each case. You see the formula "**=SQRT(A2)**" in **B2** (Figure 5.6) though you have not typed it yourself. Check **B3**, **B4** and **B5**, and you can see that the arguments of **SQRT** are entered in sequence (A3, A4 and A5)

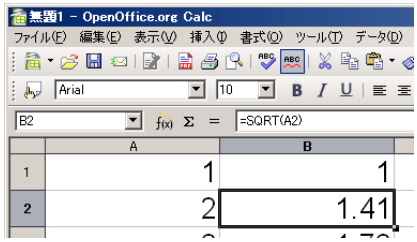

Figure 5.6: Copied Formulas

This feature is certainly convenient and is exactly what we want. The logic behind the feature is as follows. The formula "**=SQRT(A1)**" refers to the cell using a system called relative reference, so the arguments of the formula shift accordingly (relatively) when the formula is copied.

Relative reference refers to the relative positional relationship of cells. This is how it works. Cell B1 refers to cell **A1**. However, **B1** is not directly referring to **A1**, but B1 recognizes A1 merely as being "to its immediate left." Therefore, when you copy a formula and move it downward, the copied formula down below also refers "to its immediate left." If we illustrate it basically, what is entered and copied within Calc is something like the figure below.

The formula "**=sqrt(A1)**" entered in **B1** actually means "**=sqrt**(immediate left)" in Calc. Therefore, whenever this formula is copied, the copied formula always refers to its immediate left.

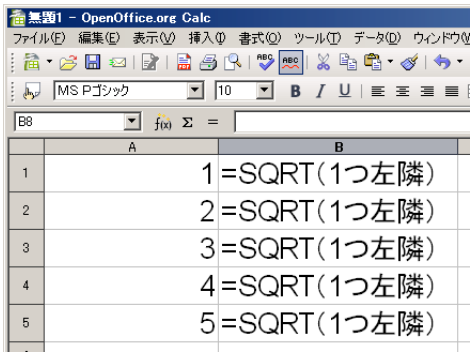

Figure 5.7: Concept of Relative Reference Copy

This is how **B2** can refer to **A2** and **B3** refer to **A3** just by the copying of the formula. In addition to horizontally, you can use relative referencing vertically too, or using a combination of them. Relative references save us from entering the same formula over and over again.

That being said, we will have problems if we only use relative references. That is why we also have a feature called absolute reference. We can make a reference absolute by attaching a "**$**" sign before each column letter and each row number: for example, **a1**. Moreover, we can make only columns absolute and keep rows relative, or vice versa, as in **$a1** and **a$1**.

5.2.7 Specification of Range in References

We talked about the **SUM** function earlier and said that this function could take only up to 30 arguments. But this does not mean you can total only up to 30 cells. There is a more efficient way to specify the arguments.

For example, suppose you want to calculate:

```
=sum(A1;A2;A3;A4;A5)
```

Instead of specifying **A1**, **A2**, **A3**, **A4** and **A5** individually, you can specify the same range using a colon (**:**) as follows.

```
=sum(A1:A5)
```

This way of specifying a range is not limited to a single row or column, but can be used more freely. You can specify a range which covers multiple rows and columns. The range is a rectangle with its vertices being the top-left cell and the bottom-right cell.

5.2.8 Do Not Use a Spreadsheet to Calculate Data You Will Use in Your Paper

Microsoft Excel and Calc can handle calculations of the levels so far described, but they are not dependable enough to be used for academic reports and papers[3]. That is because Excel and Calc are not designed to calculate fractional parts of numbers accurately. See Figure 5.8. In this figure, **Number** is selected as the cell format and calculations are performed to display 20 decimal places. Numerical values of 3.2 and 3.3 are entered in cells A1 and A2, respectively and the formula "=A2-A1" is entered in cell A3.

[3] This is true for other uses, including business.

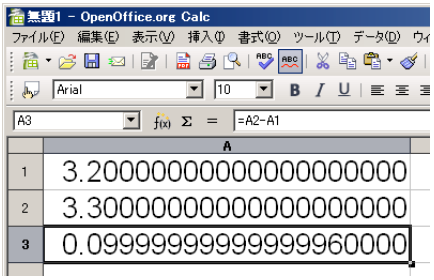

Figure 5.8: Calculation after Decimal Point by Calc

Calc is used here, but Excel will generate exactly the same results. We shall omit a detailed explanation of why this happens, but spreadsheet software always involves errors in calculation after decimal point and the displayed calculation is the result of rounding the error. Depending on the calculation method, calculation result and display method, you may or may not actually see the error. Although each error may be very small, it is true that in many cases users unknowingly carry out inaccurate calculations. Therefore it is even risky to use spreadsheet software for calculation where a slight discrepancy in calculation results will cause a big problem.

Excel has advanced statistic calculation functions, such as regression analysis and analysis of variance, but you should not use these either. Many of these functions are written as Visual Basic for Applications (VBA) programs in Excel, but many problems have been reported in the past about the calculation results.

Since programs are written by humans, it is impossible to eliminate the possibility of error. This holds true with any programs. However, there are reliable applications software in whose development many experts have participated and that have been verified by many users. There are techniques for carrying out error-free calculation using spreadsheet software; however, instead you should use software with guaranteed calculation accuracy.

You can use a large number of applications software free of charge, such as “R” which is good for statistic calculation, numerical calculation and graphing, Maxima, which is formula processing software and Octave whose strong suit is numerical calculation, modeling and simulation.

Waseda University makes commercial software, like SAS and SPSS, available to all Waseda students, but you may not find yourself in such a privileged environment after graduation. Those who wish to introduce a certain amount of statistical processing or numerical calculation in their papers should take advantage of the free software mentioned above.