Disarray to Dataset

2

2.1 Data Collection

Let us begin with the first step of the intelligence cycle: data collection. Many businesses gather crucial information – on expenditures and sales, say – but few enter it into a central database for systematic evaluation. The first task of the statistician is to mine this valuable information. Often, this requires skills of persuasion: employees may be hesitant to give up data for the purpose of systematic analysis, for this may reveal past failures.

But even when a firm has decided to systematically collect data, preparation may be required prior to analysis. Who should be authorized to evaluate the data? Who possesses the skills to do so? And who has the time? Businesses face questions like these on a daily basis, and they are no laughing matter. Consider the following example: when tracking customer purchases with loyalty cards, companies obtain extraordinarily large datasets. Administrative tasks alone can occupy an entire department, and this is before systematic evaluation can even begin.

In addition to the data they collect themselves, firms can also find information in *public databases*. Sometimes these databases are assembled by private marketing research firms such as ACNielsen or the GfK Group, which usually charge a data access fee. The databases of research institutes, federal and local statistics offices, and many international organizations (Eurostat, the OECD, the World Bank, etc.) may be used for free. Either way, public databases often contain valuable information for business decisions. The following Table 2.1 provides a list of links to some interesting sources of data:

Let's take a closer look at how public data can aid business decisions. Imagine a procurement department of a company that manufacturers intermediate goods for machine construction. In order to lower costs, optimize stock levels, and fine-tune

Chapter 2 Translated from the German original, Cleff, T. (2011). 2 Vom Zahlenwust zum Datensatz. In Deskriptive Statistik und moderne Datenanalyse (pp. 15–29) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

DOI 10.1007/978-3-319-01517-0_2, © Springer International Publishing Switzerland 2013

Table 2.1 External data sources at international institutions

German federal statistical office	destatis.de	Offers links to diverse international data bases
Eurostat	epp.eurostat.ec.europa.eu	Various databases
OECD	oecd.org	Various databases
Worldbank	worldbank.org	World & country-specific development indicators
UN	un.org	Diverse databases
ILO	ilo.org	Labour statistics and databases
IMF	imf.org	Global economic indicators, financial statistics, information on direct investment, etc.

order times, the department is tasked with forecasting stochastic demand for materials and operational supplies. They could of course ask the sales department about future orders, and plan production and material needs accordingly. But experience shows that sales departments vastly overestimate projections to ensure delivery capacity. So the procurement (or inventory) department decides to consult the most recent Ifo Business Climate Index.[1] Using this information, the department staff can create a valid forecast of the end-user industry for the next 6 months. If the end-user industry sees business as trending downward, the sales of our manufacturing company are also likely to decline, and vice versa. In this way, the procurement department can make informed order decisions using public data instead of conducting its own surveys.[2]

Public data may come in various states of aggregation. Such data may be based on a category of company or group of people, but only rarely one single firm or individual. For example, the Centre for European Economic Research (ZEW) conducts recurring surveys on industry innovation. These surveys never contain data on a single firm, but rather data on a group of firms – say, the R&D expenditures of chemical companies with between 20 and 49 employees. This information can then be used by individual companies to benchmark their own indices. Another example is the GfK household panel, which contains data on the purchase activity of households, but not of individuals. Loyalty card data also provides, in effect, aggregate information, since purchases cannot be traced back reliably to particular cardholders (as a husband, for example, may have used his wife's card to make a purchase). Objectively speaking, loyalty card data reflects only a household, but not its members.

[1] The Ifo Business Climate Index is released each month by Germany's Ifo Institute. It is based on a monthly survey that queries some 7,000 companies in the manufacturing, construction, wholesaling, and retailing industries about a variety of subjects: the current business climate, domestic production, product inventory, demand, domestic prices, order change over the previous month, foreign orders, exports, employment trends, three-month price outlook, and six-month business outlook.

[2] For more, see the method described in Chap. 5.

To collect information about individual persons or firms, one must conduct a *survey*. Typically, this is most expense form of data collection. But it allows companies to specify their own questions. Depending on the subject, the survey can be oral or written. The traditional form of survey is the questionnaire, though telephone and Internet surveys are also becoming increasingly popular.

2.2 Level of Measurement

It would go beyond the scope of this textbook to present all of the rules for the proper construction of questionnaires. For more on questionnaire design, the reader is encouraged to consult other sources (see, for instance, Malhotra 2010). Consequently, we focus below on the criteria for choosing a specific quantitative assessment method.

Let us begin with an example. Imagine you own a little grocery store in a small town. Several customers have requested that you expand your selection of butter and margarine. Because you have limited space for display and storage, you want to know whether this request is *representative* of the preferences of all your customers. You thus hire a group of students to conduct a survey using the short questionnaire in Fig. 2.1.

Within a week the students have collected questionnaires from 850 customers. Each individual survey is a *statistical unit* with certain relevant *traits*. In this questionnaire the relevant traits are *sex*, *age*, *body weight*, *preferred bread spread*, and *selection rating*. One customer – we'll call him Mr. Smith – has the *trait values* of *male*, *67 years old*, *74 kg*, *margarine*, and *fair*. Every survey requires that the designer first define the statistical unit (who to question?), the relevant traits or variables (what to question?), and the trait values (what answers can be given?).

Variables can be classified as either discrete or continuous variables. *Discrete variables* can only take on certain given numbers – normally whole numbers – as possible values. There are usually gaps between two consecutive outcomes. The *size of a family*(1, 2, 3, …) is an example of a discrete variable. *Continuous variables* can take on any value within an interval of numbers. All numbers within this interval are possible. Examples are variables such as *weight* or *height*.

Generally speaking, the statistical units are the subjects (or objects) of the survey. They differ in terms of their values for specific traits. The traits *gender*, *selection rating*, and *age* shown in Fig. 2.2 represent the three levels of measurement in quantitative analysis: the nominal scale, the ordinal scale, and the cardinal scale, respectively.

The lowest level of measurement is the *nominal scale*. With this level of measurement, a number is assigned to each possible trait (e.g. $x_i = 1$ for *male* or $x_i = 2$ for *female*). A *nominal variable* is sometimes also referred to as *qualitative variable,* or *attribute*. The values serve to assign each statistical unit to a specific group (e.g. the group of *male* respondents) in order to differentiate it from another group (e.g. the *female* respondents). Every statistical unit can only be assigned to one group and all statistical units with the same trait status receive the same number. Since the numbers merely indicate a group, they do not express qualities

Sex: □ male □ female

Age: _____

Body weight: _____ kg

Which spread do you prefer? *(Choose one answer)*
□ butter □ margarine □ other

On a scale of 1 (poor) to 5 (excellent) how do rate the selection of your preferred spread at our store?

□(1)	□(2)	□(3)	□(4)	□(5)
poor	fair	average	good	excellent

Fig. 2.1 Retail questionnaire

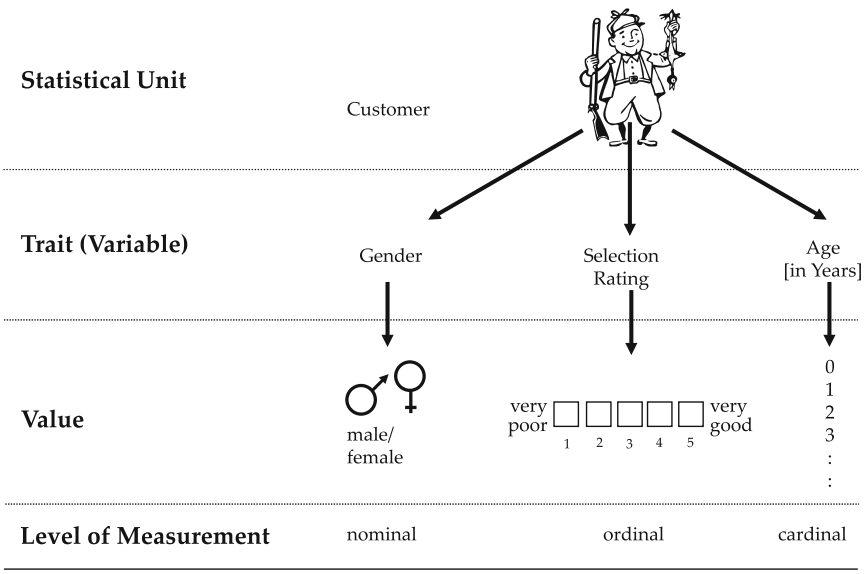

Fig. 2.2 Statistical units/Traits/Trait values/Level of measurement

such as larger/smaller, less/more, or better/worse. They only designate membership or non-membership in a group ($x_i=x_j$ versus $x_i\neq x_j$). In the case of the trait *sex*, a *1* for *male* is no better or worse than a *2* for *female*; the data are merely segmented in

terms of male and female respondents. Neither does rank play a role in other nominal traits, including profession(e.g. 1=*butcher*; 2=*baker*; 3=*chimney sweep*), nationality, class year, etc.

This leads us to the next highest level of measurement, the *ordinal scale*. With this level of measurement, numbers are also assigned to individual value traits, but here they express a rank. The typical examples are answers based on scales from 1 to *x*, as with the trait *selection rating* in the sample survey. This level of measurement allows researchers to determine the intensity of a trait value for a statistical unit compared to that of other statistical units. If Ms. Peters and Ms. Miller both check the third box under *selection rating*, we can assume that both have the same perception of the store's selection. As with the nominal scale, statistical units with the same values receive the same number. If Mr. Martin checks the fourth box, this means both that his perception is different from that of Ms. Peters and Ms. Miller, and that he thinks the selection is *better* than they do. With an ordinal scale, traits can be ordered, leading to qualities such as larger/smaller, less/more, and better/worse ($x_i=x_j$; $x_i>x_j$; $x_i<x_j$).

What we cannot say is how large the distance is between the third and fourth boxes. We cannot even assume that the distance between the first and second boxes is as large as that between other neighbouring boxes. Consider an everyday example of an ordinal scale: standings at athletic competitions. The difference between each place does not necessary indicate a proportional difference in performance. In a swimming competition the time separating first and second place may be one one-thousandth of a second, with third place coming in two seconds later, yet only one place separates each.

The highest level of measurement is the *metric* or *cardinal scale*. It contains not only the information of the ordinal scales – larger/smaller, less/more, better/worse ($x_i=x_j$; $x_i>x_j$; $x_i<x_j$) – but also the distance between value traits held by two statistical units. Age is one example. A 20 year old is not only older than an 18 year old; a 20 year old is exactly 2 years older than an 18 year old. Moreover, the distance between a 20 year old and a 30 year old is just as large as the distance between an 80 year old and a 90 year old. The graduations on a cardinal scale are always equidistant. In addition to age, typical examples for cardinal scales are currency, weight, length, and speed.

Cardinal scales are frequently differentiated into absolute scales,[3] ratio scales,[4] and interval scales.[5] These distinctions tend to be academic and seldom play much role in deciding which statistical method to apply. This cannot be said of the distinction between cardinal and ordinal scale variables, however. On account of the much greater variety of analysis methods for cardinal scales in relation to ordinal methods, researchers often tend to see ordinal variables as cardinal in nature. For example,

[3] A metric scale with a natural zero point and a natural unit (e.g. age).

[4] A metric scale with a natural zero point but without a natural unit (e.g. surface).

[5] A metric scale without a natural zero point and without a natural unit (e.g. geographical longitude).

researchers might assume that the gradations on the five-point scale used for rating selection in our survey example are identical. We frequently find such assumptions in empirical studies. More serious researchers note in passing that *equidistance* has been assumed or offer justification for such *equidistance*. Schmidt and Opp (1976, p. 35) have proposed a rule of thumb according to which ordinal scaled variables can be treated as cardinal scaled variables: the ordinal scale must have more than four possible outcomes and the survey must have more than 100 observations. Still, interpreting a difference of 0.5 between two ordinal scale averages is difficult, and is a source of many headaches among empirical researchers.

As this section makes clear, a variable's scale is crucial because it determines which statistical method to apply. For a nominal variable like *profession* it is impossible to determine the mean value of three backers, five butchers, and two chimney sweeps. Later in the book I will discuss which statistical method goes with which level of measurement or combination of measurements.

Before data analysis can begin, the collected data must be transferred from paper to a form that can be read and processed by a computer. We will continue to use the 850 questionnaires collected by the students as an example.

2.3 Scaling and Coding

To emphasize again, the first step in conducting a survey is to define the level of measurement for each trait. In most cases, it is impossible to raise the level of measurement after a survey has been implemented (i.e. from nominal to ordinal, or from ordinal to cardinal). If a survey asks respondents to indicate their age not by years but by age group, this variable must remain on the ordinal scale. This can be a great source of frustration: among other things, it makes it impossible to determine the average age of respondents in retrospect. It is therefore always advisable to set a variable's level of measurement as high as possible beforehand (e.g. age in years, or expenditures for a consumer good).

The group or person who commissions a survey may stipulate that questions remain on a lower level of measurement in order to ensure anonymity. When a company's works council is involved in implementing a survey, for example, one may encounter such a request. Researchers are normally obligated to accommodate such wishes.

In our above sample survey the following levels of measurement were used:

- Nominal: gender; preferred spread
- Ordinal: selection rating
- Cardinal: age; body weight

Now, how can we *communicate* this information to the computer? Every statistics application contains an Excel-like spreadsheet in which data can be entered directly (see, for instance, Fig. 3.1, p. 24). While columns in Excel spreadsheets are typically named A, B, C, etc., the columns in more professional spreadsheets are labelled with the *variable name*. Typically, variable names may be no longer than eight characters. So, for instance, the variable *selection rating* is given as "*selectio*".

```
-------------------------------------------------------------------------
value label: selectio
-------------------------------------------------------------------------
  definition
           1   poor
           2   fair
           3   average
           4   good
           5   excellent
   variables:  selection
-------------------------------------------------------------------------
value label: brd sprd
-------------------------------------------------------------------------
  definition
           0   butter
           1   margarine
           2   other
   variables:  bread spread
-------------------------------------------------------------------------
value label: sex
-------------------------------------------------------------------------
  definition
           0   male
           1   female
   variables:  gender
```

Fig. 2.3 Label book

For clarity's sake, a variable name can be linked to a longer *variable label* or to an entire survey question. The software commands use the variable names – e.g. "Compute graphic for the variable selectio" – while the printout of the results displays the complete label.

The next step is to enter the survey results into the spreadsheet. The answers from questionnaire #1 go in the first row, those from questionnaire #2 go in the second row, and so on. A computer can only "understand" numbers. For cardinal scale variables this is no problem, since all of the values are numbers anyway. Suppose person #1 is 31 years old and weighs 63 kg. Simply enter the numbers 31 and 63 in the appropriate row for respondent #1. Nominal and ordinal variables are more difficult and require that all contents be coded with a number. In the sample dataset, for instance, the nominal scale traits *male* and *female* are assigned the numbers "0" and "1", respectively. The number assignments are recorded in a label book, as shown in Fig. 2.3. Using this system, you can now enter the remaining results.

2.4 Missing Values

A problem that becomes immediately apparent when evaluating survey data is the omission of answers and frequent lack of opinion (i.e. responses like *I don't know*). The reasons can be various: deliberate refusal, missing information, respondent inability, indecision, etc.

Faulkenberry and Mason (1978, p. 533) distinguish between two main types of answer omissions:

(a) *No opinion*: respondents are indecisive about an answer (due to an ambiguous question, say).
(b) *Non-opinion*: respondents have no opinion about a topic.

The authors find that respondents who tend to give the first type of omission (no opinion) are more reflective and better educated than respondents who tend to give the second type of omission (non-opinion). They also note that the gender, age, and ethnic background of the respondents (among other variables) can influence the likelihood of an answer omission.

This observation brings us to the problem of *systematic bias* caused by answer omission. Some studies show that lack of opinion can be up to 30% higher when respondents are given the option of *I don't know* (Schumann & Presser 1981, p. 117). But simply eliminating this option as a strategy for its avoidance can lead to biased results. This is because the respondents who tend to choose *I don't know* often do not feel obliged to give truthful answers when the *I don't know* option is not available. Such respondents typically react by giving a random answer or no answer at all. This creates the danger that an identifiable, systematic error attributable to frequent *I don't know* responses will be transformed into an undiscovered, systematic error at the level of actual findings. From this perspective, it is hard to understand those who recommend the elimination of the *I don't know* option. More important is the question of how to approach answer omissions during data analysis.

In principle, the omissions of answers should not lead to values that are interpreted during analysis, which is why some analysis methods do not permit the use of missing values. The presence of missing values can even necessitate that other data be excluded. In regression or factor analysis, for example, when a respondent has missing values, the remaining values for that respondent must be omitted as well. Since answer omissions often occur and no one wants large losses of information, the best alternative is to use some form of substitution. There are five general approaches:

(a) The best and most time-consuming way to eliminate missing values is to fill them in yourself, provided it is possible to obtain accurate information through further research. In many cases, missing information in questionnaires on revenue, R&D expenditures, etc. can be discovered through a careful study of financial reports and other published materials.
(b) If the variables in question are qualitative (nominally scaled), missing values can be avoided by creating a new class. Consider a survey in which some respondents check the box *previous customer*, some the box *not a previous customer*, and others check *neither*. In this case, the respondents who provided no answer can be assigned to a new class; let's call it *customer status unknown*. In the frequency tables this class then appears in a separate line titled *missing values*. Even with complex techniques such as regression analysis, it is usually possible to interpret missing values to some extent. We'll address this issue again in later chapters.

(c) If it is not possible to address missing values conducting additional research or creating a new category, missing variables can be substituted with the total arithmetic mean of existing values, provided they are on a cardinal scale.
(d) Missing cardinal values can also be substituted with the arithmetic mean of a group. For instance, in a survey gathering statistics on students at a given university, missing information is better replaced by the arithmetic mean of students in the respective course of study rather than by the arithmetic mean of the entire student body.
(e) We must remember to verify that the omitted answers are indeed non-systematic; otherwise, attempts to compensate for missing values will produce grave distortions. When answers are omitted in non-systematic fashion, missing values can be estimated with relative accuracy. Nevertheless, care must be taken not to understate value distribution and, by extension, misrepresent the results. "In particular", note Roderick et al. "variances from filled-in data are clearly understated by imputing means, and associations between variables are distorted. Thus, the method yields an inconsistent estimate of the covariance matrix" (1995, p. 45). The use of complicated estimation techniques becomes necessary when the number of missing values is large enough that the insertion of mean values significantly changes the statistical indices. These techniques mostly rely on regression analysis, which estimates missing values using existing dependent variables in the dataset. Say a company provides incomplete information about their R&D expenditures. If you know that R&D expenditures depend on company sector, company size, and company location (West Germany or East Germany, for instance), you can use available data to roughly extrapolate the missing data. Regression analysis is discussed in more detail in Chap. 5.

Generally, you should take care when subsequently filling in missing values. Whenever possible, the reasons for the missing values should remain clear. In a telephone interview, for instance, you can distinguish between:

- Respondents who do not provide a response because they do not know the answer;
- Respondents who have an answer but do not want to communicate it; and
- Respondents who do not provide a response because the question is directed to a different age group than theirs.

In the last case, an answer is frequently just omitted (missing value due to study design). In the first two cases, however, values may be assigned but are later defined as *missing values* by the analysis software.

2.5 Outliers and Obviously Incorrect Values

A problem similar to missing values is that of obviously incorrect values. Standardized customer surveys often contain both. Sometimes a respondent checks the box marked *unemployed* when asked about job status but enters some outlandish figure like €1,000,000,000 when asked about income. If this response were included in a survey of 500 people, the average income would increase by €2,000,000.

This is why obviously incorrect answers must be eliminated from the dataset. Here, the intentionally wrong income figure could be marked as a missing value or given an estimated value using one of the techniques described in Sect. 2.4.

Obviously incorrect values are not always deliberate. They can also be the result of error. Business surveys, for instance, often ask for revenue figures in thousands of euros, but some respondents invariably provide absolute values, thus indicating revenues one-thousand times higher than they actually are. If discovered, mistakes like these must be corrected before data analysis.

A more difficult case is when the data are unintentionally false but cannot be easily corrected. For example, when you ask businesses to provide a breakdown of their expenditures by category and per cent, you frequently receive total values amounting to more than 100%. Similar errors also occur with private individuals.

Another tricky case is when the value is correct but an outlier. Suppose a company wants to calculate future employee pensions. To find the average retirement age, they average the ages at which workers retired in recent years. Now suppose that of one of the recent retirees, the company's founder, left the business just shy of 80. Though this information is correct – and though the founder is part of the target group of retired employees – the inclusion of this value would distort the average retirement age, since it is very unlikely that other employees will also retire so late in the game. Under certain circumstances it thus makes sense to exclude outliers from the analysis – provided, of course, that the context warrants it. One general solution is to *trim* the dataset values, eliminating the highest and lowest five per cent. I will return to this topic once more in Sect. 3.2.2.

2.6 Chapter Exercises

Exercise 1:

For each of the following statistical units, provide traits and trait values:

(a) Patient cause of death
(b) Length of university study
(c) Alcohol content of a drink

Exercise 2:

For each of the following traits, indicate the appropriate level of measurement:

(a) Student part-time jobs
(b) Market share of a product between 0% and 100%
(c) Students' chosen programme of study
(d) Time of day
(e) Blood alcohol level
(f) Vehicle fuel economy
(g) IQ
(h) Star rating for a restaurant

Exercise 3:

Use Stata, SPSS, or Excel for the questionnaire in Fig. 2.1 (p. 16) and enter the data from Fig. 3.1 (p. 24). Allow for missing values in the dataset.

http://www.springer.com/978-3-319-01516-3

Exploratory Data Analysis in Business and Economics
An Introduction Using SPSS, Stata, and Excel
Cleff, Th.
2014, XXII, 215 p. 130 illus., 11 illus. in color., Softcover
ISBN: 978-3-319-01516-3